Ph.D. DISSERTATION

3D Depth Reconstruction from Focal Stack and Depth Refinement

초점 스택에서 3D 깊이 재구성 및 깊이 개선

February 2021

DEPARTMENT OF ELECTRICAL ENGINEERING &
COMPUTER SCIENCE
THE GRADUATE SCHOOL
SEOUL NATIONAL UNIVERSITY

ZHIQIANG MA

3D Depth Reconstruction from Focal Stack and Depth Refinement

초점 스택에서 3D 깊이 재구성 및 깊이 개선

지도교수 신 영 길

이 논문을 공학박사 학위논문으로 제출함

2020 년 12 월

서울대학교 대학원

전기.컴퓨터공학부

마지강

마지강의 공학박사 학위논문을 인준함

2020 년 12 월

위 원 장 ______ 김 명 수 ______ (인)

부위원장 ______ 신 영 길 ______ (인)

위 원 ______ 서 진 욱 ______ (인)

위 원 ______ 김 보 형 ______ (인)

위 원 ______ 이 정 진 ______ (인)

Abstract

ZHIQIANG MA
Department of Electrical Engineering & Computer Science
The Graduate School
Seoul National University

Three-dimensional (3D) depth recovery from two-dimensional images is a fundamental and challenging objective in computer vision, and is one of the most important prerequisites for many applications such as 3D measurement, robot location and navigation, self-driving, and so on. Depth-from-focus (DFF) is one of the important methods to reconstruct a 3D depth in the use of focus information. Reconstructing a 3D depth from texture-less regions is a typical issue associated with the conventional DFF. Further more, it is difficult for the conventional DFF reconstruction techniques to preserve depth edges and fine details while maintaining spatial consistency. In this dissertation, we address these problems and propose an DFF depth recovery framework which is robust over texture-less regions, and can reconstruct a depth image with clear edges and fine details.

The depth recovery framework proposed in this dissertation is composed of two processes: depth reconstruction and depth refinement. To recovery an accurate 3D depth, We first formulate the depth reconstruction as a maximum a posterior (MAP) estimation problem with the inclusion of matting Laplacian prior. The nonlocal principle is adopted during the construction stage of the matting Laplacian matrix to preserve depth edges and fine details. Additionally, a depth variance based confidence measure with the combination of the reliability measure of focus measure is proposed to maintain the spatial smoothness, such that the smooth depth regions in initial depth could have high confidence value and the reconstructed depth could be more derived from the initial depth. As the nonlocal principle breaks the spatial consistency, the reconstructed depth image is spatially inconsistent. Meanwhile, it suffers from texture-copy artifacts. To smooth the noise and suppress the texture-copy artifacts introduced in the reconstructed depth image, we propose a closed-form edge-preserving depth refinement algorithm that formulates the depth refinement as a MAP estimation problem using Markov random fields (MRFs). With the incorporation of pre-estimated depth edges and mutual structure information into our energy function and the specially designed smoothness weight, the proposed refinement method can effectively suppress noise and texture-copy artifacts while preserving depth edges. Additionally, with the construction of undirected weighted graph representing the energy function, a closed-form solution is obtained by using the Laplacian matrix corresponding to the graph.

The proposed framework presents a novel method of 3D depth recovery from a focal stack. The proposed algorithm shows the superiority in depth recovery over texture-less regions owing to the effective variance based confidence level computation and the matting Laplacian prior. Additionally, this proposed reconstruction method can obtain a depth image with clear edges and fine details due to the adoption of nonlocal principle in the construction of matting Laplacian matrix. The proposed closed-form depth refinement approach shows that the ability in noise removal while preserving object structure with the usage of common edges. Additionally, it is able to effectively suppress texture-copy artifacts by utilizing mutual structure information. The proposed depth refinement provides a general idea for edge-preserving image smoothing, especially for depth related refinement such as stereo vision.

Both quantitative and qualitative experimental results show the supremacy of the proposed method in terms of robustness in texture-less regions, accuracy, and ability to preserve object structure while maintaining spatial smoothness.

Keywords: Depth-from-Focus, Depth Recovery, Matting Laplacian, Depth Refinement, Image Denoising, Markov Random Field

Student Number: 2012-31287

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

With the development of semiconductor industry, digital camera imaging devices are rapidly becoming small, cheap, and portable. As three-dimensional (3D) vision provides much richer information than 2D, 3D depth reconstruction from two-dimensional images becomes more and more important in computer vision. 3D recovery from 2D image sequences can be boardly categorized into active and passive methods. In active methods, laser scanning, sonar and many more like them are included. Normally, devices used in active methods are expensive and sometimes impractical to use. On the contrary, passive methods including stereo vision, shape-from-motion, shape-from-shading and shape-from-texture are more popular for being cheap and easy to implement. Shape-from-stereo can extract the depth information by measuring the dispar-

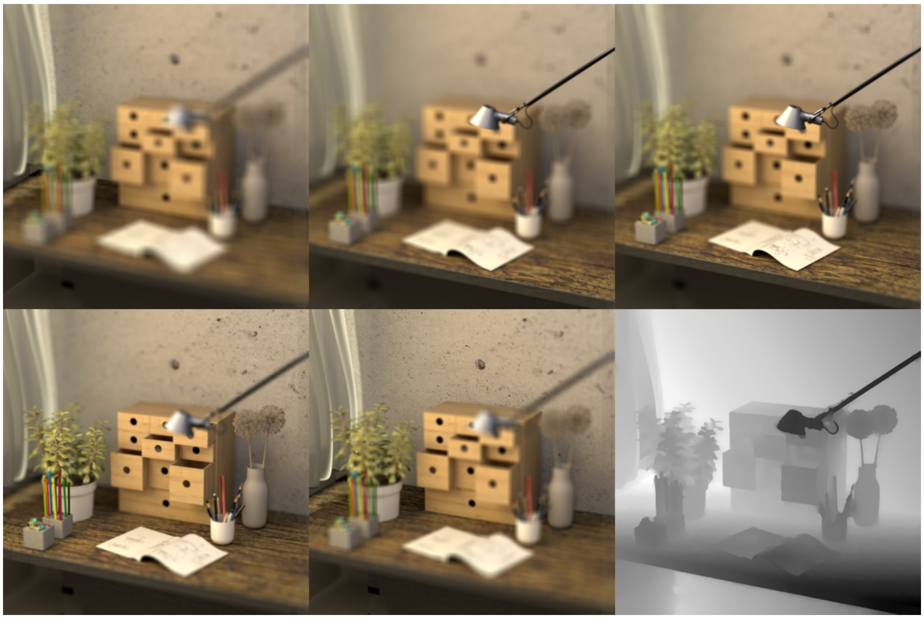

Figure 1.1: Example of shape recovery from a focal stack.

ities between a pair of images taken from different viewpoints. Structure-from-motion computes the correspondences between images to obtain the 2D motion field, which in turn used to recover 3D motion and the depth. Depth-from-focus (DFF) and depth-from-defocus (DFD) are two representative passive methods utilizing focus information to recovery a 3D depth image from a focal stack. In contrast to the multi-camera system, DFF and DFD approaches require only a monocular camera whose extrinsic or intrinsic setting can be modified, thus preventing matching ambiguities typically found in stereo devices.

DFD attempts to recover depth based on the direct relationships among the depth, amount of blur, and camera parameters. The disadvantage of DFD is the requirement of accurate camera calibration. Unlike DFD, DFF estimates the depth value for each pixel from a sequence of images acquired with different focal settings. This technique assumes that a one-to-one correspondence exists between the depth of one point in the scene and the focal setting. Subsequently, the depth per pixel can be roughly inferred by choosing the focal setting at which the pixel is most focused or sharpest. The algorithm used to measure the sharpness level for each pixel is typically referred to as a focus measure operator. As DFF uses a larger number of observations, the performance of DFF is generally better compared with that of DFD [1].

1.2 Motivation

Many researches regrading to DFF have been conducted and can be roughly divided into four directions: 1) development of various kind of focus measure operators; 2) improvement of focus measure accuracy [1–8]; 3) formulating the DFF into a reconstruction process [3, 9–13] ; 4) deep learning based methods [14].

The inherent problem in the first direction is the sensitivity of the window size. Focus measure operators using small window size can derive a depth image with clear edges and details but at a cost of introducing noise. On the contrary, a large window size can make it roust to noise but at a cost of blurring edges and losing object details. Another problem for the traditional focus measure operators is that they would yield spurious responses in texture-less regions owing to the weak variation in the gray level. The researches aiming at improving focus measure accuracy can only basically improve the errors caused by the noise and they have limitations to accurately recover the depth from texture-less regions. Typically, there two main issues existing in current DFF reconstruction approaches: depth recovery from texture-less regions and object structure or edge preservation while maintain spatial consistency. Most of current researches focus either on robustness in depth recovery over texture-less regions or on the object structure preservation. It is still difficult to solve above issues at the same time. Deep learning based methods have a strong ability to extract meaningful image features and correlate pixel information via convolutions. However, as a lot of ground truth training data are required for the deep learning based methods and it is difficult to acquire ground truth depth images, till now, there are not many deep learning based DFF method have been proposed. Additionally, intrinsic parameters are necessary when recovery a 3D depth image using DFF method. It is impractical to train a general model to process a focal stack acquired with different devices. Thus, the current deep learning based methods are basically used for focus measures. Even being used for focus measures, there are still some limitations when applying deep learning techniques. Firstly, once one model is trained, it is impossible to modify image resolution for new input image sequences but reshape the input data to adapt the trained model. Another limitation when applying deep learning techniques is the size of focal stack. Similar to image resolution, once the model is trained, the size of focal stack for the input data should be consistent with the trained data, which has a great influence on the accuracy of depth recovery.

Depth refinement is a common operation after depth reconstruction to make depth spatially smooth while preserving object structures. The commonly used methods are edge-preserving image filters such as anisotropic fusion filter, bilateral filter and nonlocal means filter. Even though the traditional edge-preserving filters have good property in noise smoothing while preserving object structures. However, there is one common but challenging issue named texture-copy artifacts that they are hard to handle it. Texture-copy artifacts are the fake edges appear in recovered depth image due to the structure inconsistency between guided color image and initial depth image. Those fake edges might be preserved or even enhanced by using traditional edge-preserving filter if the texture-copy artifacts have large edge gradient. In this dissertation, texture-copy artifacts reduction is one important target that needs to be taken into account.

Consequently, it is meaningful and necessary to develop a DFF method that should be able to recover depth from texture-less regions and preserve objective structures and details while maintain spatial consistency.

1.3 Contribution

In this dissertation, a framework of depth recovery from focal stack is presented. The proposed framework aims to improve the recovery accuracy and robustness in texture-less regions, and maintain spatial smoothness while preserving object structure. In the proposed framework, two processes, depth reconstruction and depth refinement, are included.

Dealing with texture-less regions and preserving object structure are two main issues in the process of depth reconstruction. This work thus presents a depth reconstruction method using matting Laplacian prior. In the depth reconstruction process, the matting Laplacian is employed to improve the robustness in texture-less regions. Besides, the nonlocal principle is adopted in the construction of matting Laplacian to preserve object structure and fine details. Meanwhile, a variance-based confidence measure is proposed to help maintain depth spatial smoothness. The proposed method with the inclusion of nonlocal Laplacian prior can effectively recover a depth image in texture-less regions, and can preserve with clear edges and rich details.

While the nonlocal principle can help preserve object structure and fine details, it breaks the depth spatial consistency and thus noise and texture-copy artifacts can be introduced. To smooth noise and suppress texture-copy artifacts, an closed-form edge-preserving depth refinement method is presented. The proposed method treats depth refinement as a MAP estimation problem based on Gaussian MRFs Model. As Gaussian MRFs tends to over-smooth depth image and blur real depth edges, specially designed smoothness weight and mutual structure information are incorporated into the proposed method, and therefore can better suppress noise and texture-copy artifacts while preserving depth edges. Additionally, the proposed method can obtain an global optimum by utilizing the Laplacian matrix based on the undirected weighted graph representing the energy function.

The architecture of proposed framework is illustrated in Figure 1.2. Given an image with different focus settings, the focus measure is first computed to derive depth image and a Gaussian interpolation around the peak of the focus measure profile is subsequently performed to generate a relatively smooth and reliable initial depth image. With the focus measures, a probability-based scheme is proposed to generate an all-in-focus image. Additionally, an effective

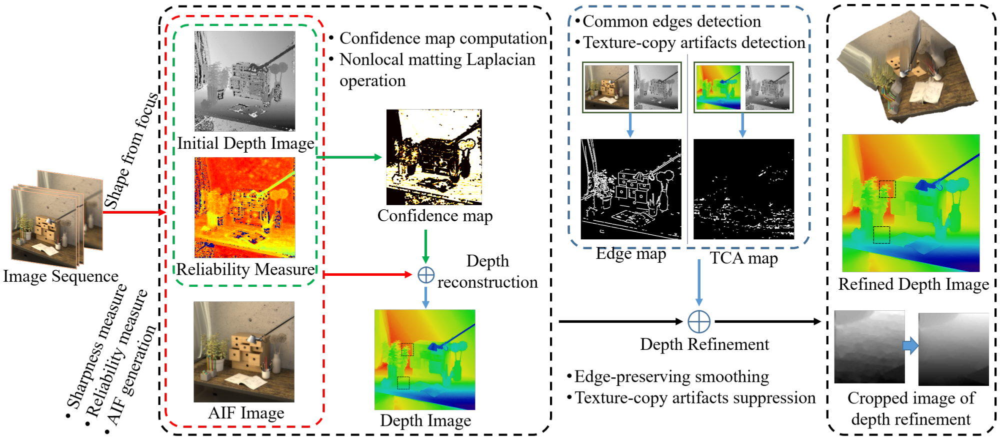

Figure 1.2: The architecture of proposed two-stage framework.

variance based confidence measure scheme is proposed to compute a confidence map for the initial depth image. By combining all-in-focus image, initial depth image and corresponding confidence map, a depth reconstruction algorithm using the MAP framework is proposed, in which the likelihood model is built based on the initial depth image and the prior model is derived using the affinity matrix embedded in nonlocal matting Laplacian matrix. After the process of depth reconstruction, a closed-form MAP-MRF based depth refinement algorithm is proposed, in which the pre-estimated depth edges and mutual structure information are incorporated into the proposed energy function to effectively smooth the noise and suppress the texture-copy artifacts introduced in the reconstructed depth image. Additionally, a closed-form solution can be obtained with the construction of undirected weighted graph representing the energy function by using the Laplacian matrix corresponding to the graph.

1.4 Organization

The following chapters of this dissertation are organized as follows. The background and related works is presented in chapter 2. In chapter 2, the basic knowledge of DFF and several focus measure operators are first described. Subsequently, a literature review of depth reconstruction from a focal stack described is presented. Finally, a literature review of edge-preserving image smoothing algorithms is introduced. A depth reconstruction method using matting Laplacian prior is introduced in chapter 3. A closed-form MAP-MRF based edge-preserving depth refinement is presented in chapter 4. Experimental comparison between the proposed method and the state-of-the-art algorithms are reported in chapter 5. The conclusion and future works are presented in chapter 6.

Chapter 2

Related Works

2.1 Overview

In this chapter, an introduction of related knowledge and works regarding DFF is presented. First, the thin lens model and principle of DFF are described separately. Several representative focus measure operators used for initial depth estimation are then introduced. The following parts of this chapter are the literature review regarding DFF reconstruction methods and edge-preserving image smoothing approaches. Additionally, several representative methods of DFF reconstruction and image denoising are simply introduced, respectively.

2.2 Principle of depth-from-focus

The DFF approach is a method to estimate 3D depth from focal stack acquired with varying focus settings. In order to derive the depth image of scene from

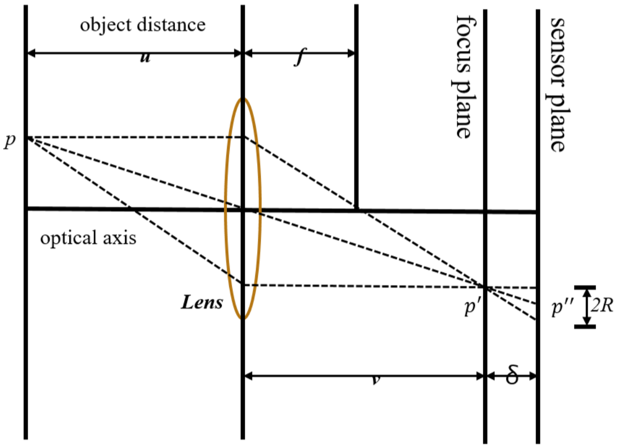

Figure 2.1: Illustration of focused and defocused images using thin lens model.

a focal stack, it is necessary to estimate the psychical distance of each point in the scene by measuring its relative degree of focus in the images where that point appears. This technique attempts to recover depth based on the direct relationships among the depth, amount of blur, and camera settings. Figure 2.1 gives an illustration of the effect of defocus in an image using thin lens model. If the sensor plane is located at a distance of δ from focus plane, point p would be projected as a circle of radius R at imaging sensor. That means p is defocused. The image distance v is determined by the focal length f of the lens and the object distance u. The geometry relationship between these three variables can be derived by the Gaussian lens formula:

$$1/f = 1/u + 1/v. \tag{2.1}$$

By assuming that there is a one-to-one correspondence between image dis-

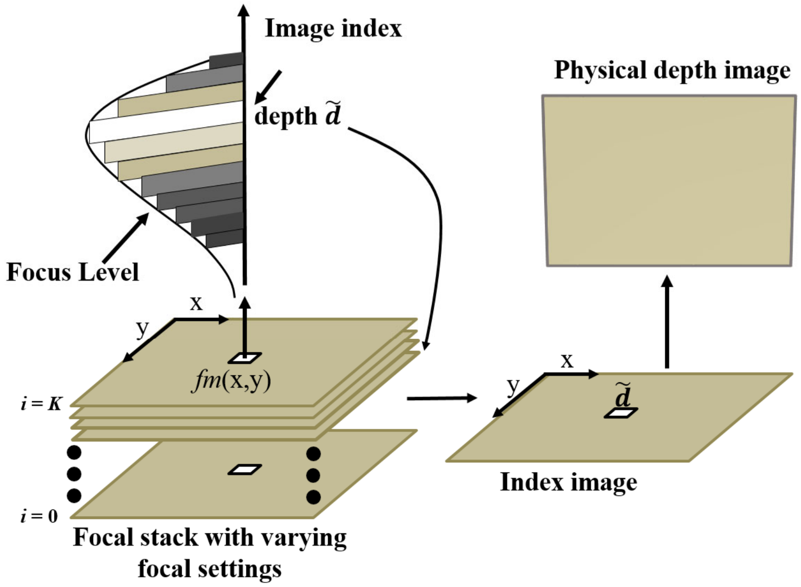

Figure 2.2: Illustration of the process of depth recovery from a focal stack.

tance v and object distance u, the most focused situation can only be achieved at a certain object distance. Thus, for a given focal stack with varying focal settings, the depth per pixel can be roughly inferred by choosing the focal setting at which the pixel is most focused or sharpest. Figure 2.2 illustrates the process of depth recovery from a focal stack. As shown in Figure 2.2, degree of focus at each position (x, y) for all images is first measured and then find the image index where the pixel is most focused. Based on the index image and Gaussian lens formula, the depth image can be finally derived.

Therefore, the measuring the focus level becomes critical in DFF application.

2.2.1 Focus measure operators

The algorithm used to measure the sharpness level for each pixel is typically referred to as a focus measure operator. Comparative studies targeted at the focus measure operator have been conducted, such as a second-derivative-based focus measure operator called summed modified Laplacian (SMLAP) in [2], the gradient-based operator in [15], wavelet-based operator in [16], statistics-based operator in [17], discrete cosine transform (DCT) based operator [18], and miscellaneous operators [19]. Several representative focus measure operators are presented in this dissertation.

Tenengrad

A popular focus measure based on the magnitude of image gradient is defined as [20, 21]

$$fm(x,y) = \sum_{(i,j)\in\Omega(x,y)} \Big(G_x(i,j)^2 + G_y(i,j)^2\Big), \tag{2.2}$$

where G_x and G_y are image gradients computed by convolving the give image using Sobel operators in X and Y direction respectively.

Diagonal Laplacian

Thelen et al. in [6] proposed a diagonal Laplacian based focus measure operator that takes both the horizontal and vertical variations of the image into consideration. This diagonal Laplacian operator can be defined as follows:

$$fm\,(x,y) = |I * \mathcal{L}_x| + |I * \mathcal{L}_y| + |I * \mathcal{L}_{x1}| + |I * \mathcal{L}_{y2}| \tag{2.3}$$

where $\mathcal{L}_x$, $\mathcal{L}_y$, $\mathcal{L}_{x2}$ and $\mathcal{L}_{y2}$ are convolution masks to compute the diagonal Laplacian, and they are defined as

$$\mathcal{L}_x = \begin{bmatrix} -1 & 2 & 1 \end{bmatrix} \quad and \;\; \mathcal{L}_y = {\mathcal{L}_x}^T. \tag{2.4}$$

and

$$L_{x2} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \;\; L_{y2} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.5}$$

Gray-level variance

The variance of image gray-levels is also one of the most used methods to estimate degree of focus. It has been applied in many applications such as auto-focusing, DFF, image segmentation, image restoration, and so on. This focus measure operator is defined as

$$fm(x,y) = \sum_{(i,j)\in\Omega(x,y)} (I(i,j) - \mu)^2, \tag{2.6}$$

where μ is the mean grey-value of pixels with $\Omega_{(x,y)}$.

Summed of wavelet coefficients

Wavelet-based focus measure operators are mostly based on the statistical properties of the discrete wavelet transform (DWT) coefficients. In the first level DWT, the image is decomposed in to four sub-images, where W_{LH1},W_{HL1},W_{HH1}

and W_{LL1} denote the three detail sub-bands and the coarse approximation sub-band,respectively. For a higher level DWT, the coarse approximation is successively decomposed into detail and coarse sub-bands. The information of the detail and coarse sub-bands is then used to compute the focus measure. Yang and Nelson [22] presented a focus measure operator for the use of auto-focusing, which is calculated from sub-bands

$$fm = \sum_{(i,j)\in\Omega_D} |W_{LH1}(i,j)| + |W_{HL1}(i,j)| + |W_{HH1}(i,j)|\,, \tag{2.7}$$

where Ω_D is the corresponding window of Ω in the DWT sub-bands. In this work, the focus measure of all the wavelet-based operators has been computed using the coefficients of the over-complete wavelet transform, thus avoiding the need for computing the corresponding neighborhood within each sub-band.

In general, reliable depth estimation can be obtained if focused regions contain sharp edges or high frequency patterns. However, most of those traditional operators would yield spurious responses to texture-less regions owing to the weak variation in the gray level. Another problem of those algorithms is the sensitivity to the window size. A small window size can preserve depth discontinuities but increase the sensitivity to noise. Meanwhile, a large window size performs better for noisy images but at the cost of blurring sharp edges.

2.3 Depth-from-focus reconstruction

As aforementioned, there are several issues to recover a depth only using focus measure operator. First of all, those focus measure operators are sensitive to window size. A small window size can preserve object structures but increase the sensitivity to noise. A large window size makes it robust to noise but at a cost of blurring edges. Another issue for these operators is the poor performance in yielding spurious responses to texture-less image regions. Many studies addressing these problems have been performed. Those efforts can be roughly categorized into three types: 1) methods aiming at the improvement of focus measure accuracy, 2) methods using depth reconstruction process, and 3) deep learning based methods.

Methods aiming at the improvement of focus measure accuracy

Many researches regarding the improvement of focus measure accuracy have been conducted. In [2], the authors presented a scheme to perform a Gaussian interpolation around the peak detected in the focus measure profile. Instead of using Gaussian interpolation, the authors in [1, 3] used a polynomial to interpolate the focus measure profile. In microscopy, Muhammad and Choi [4] proposed a Lorentzian-Cauchy fit for the focus measure profile. Aydin and Akgul [5] suggested an adaptive focus measure operator using adaptively shaped and weighted support windows that are determined from the image characteristics of an all-in-focus image of a scene. Thelen et al. [6] discussed the importance of window size and proposed an adaptive method to select the effective window size among several neighborhood sizes for the local operator based on a confidence criterion. More recently, Surh et al. [7] presented a new ring difference filter focus measure operator. The structure of this focus measure operator is a combination of ring and disk, which utilizes both local and nonlocal characteristics and thus makes the focus measure more robust against noise. Sakurikar and Narayanan in [8] proposed a composite scheme using various different focus measure operators to derive more accurate depth image.

In this dissertation, a representative method proposed by Aydin and Akgul [5] is described. They proposed a new adaptive focus measure operator, which is achieved by assigning weights to each pixel in the support window. The weights are computed according to the similarity and proximity levels between the pixels enclosed by the window and the pixel for which window is computed. The adaptive focus measure operator is defined as

$$AFM(x_0, y_0) = \sum_{(x,y)\in\Omega_{x_0,y_0}} \omega_{x_0,y_0} fm(x, y), \tag{2.8}$$

where $fm(x, y)$ is the focus measure operator, ω_{x_0,y_o} is the weight of support window Ω_{x_0,y_o} centered at the pixel (x_0, y_0), which can be calculated using all-in-focus image I_f according to the following equation

$$\omega_{x_0,y_o}(x, y) = e^{-(\Delta d/\gamma_1 + \Delta I_f/\gamma_2)}. \tag{2.9}$$

In (2.9), ΔI_f is the euclidean distance in color space, Δd is the euclidean distance in spatial domain. They are defined as follows:

$$\Delta d = \sqrt{(x - x_0)^2 + (y - y_0)^2}, \quad (x, y) \in \Omega_{x_0,y_o}, \tag{2.10}$$

and

$$\Delta I_f = \sqrt{\begin{array}{l} \Big(I_f^r(x,y) - I_f^r(x_0,y_0)\Big)^2 + \Big(I_f^g(x,y) - I_f^g(x_0,y_0)\Big)^2 \\ \\ +\Big(I_f^b(x,y) - I_f^b(x_0,y_0)\Big)^2 \end{array}}, \qquad (2.11)$$

where $I_f^r(x,y)$, $I_f^g(x,y)$, and $I_f^b(x,y)$ are the intensity values of R, G, and B color channels, respectively. γ_1 and γ_2 are parameters supervise relative weights. As we can see, the weight of adaptive focus measure operator consists of two Gaussian kernels on space distance Δd and range distance ΔI. It resembles the kernels used in Bilateral filter [23], which are designed for edge preserving image smoothing. By employing adaptively shapped and weighted windows, this method can partially solve the problems due to the depth discontinuities and edge bleeding, which are difficult for traditional focus measure operators. However, to accurately recover depth from texture-less regions, large enough support window is needed. Computing weights in the support window for each pixel in each frame has much higher computational complexity compared to traditional focus measure operators. Additionally, a very large support window could also introduce visual artifacts.

Methods using depth reconstruction process

In addition to the efforts on improving the focus measure accuracy, some researchers have attempted to formulate the DFF problem as depth reconstruction to derive more reliable and accurate depth image from a noisy depth image. Because depth reconstruction with a limited number of sequence images is an ill-posed problem, prior knowledge is required to regularize the solution. Gaganov and Ignateko [9] proposed an MRF-based framework to derive an energy function consisting of two truncated quadratic functions, and yields an optimal depth estimation with enforced smoothness constraints. The energy optimization algorithm used in that framework is α-expansion based on graph cut. Even though reasonable results can be obtained, it is prone to obtaining the local minima energy owing to its nonconvex property. Another MRF-based approach was proposed in [10] to extract smooth and texture-less objects using iterative conditional modes. Their approach is robust against texture-less regions but surfers from the high computation cost, which is the inherent problem in MRF-based algorithms. The authors in [3] proposed a variational approach and solved it using an efficient nonconvex minimization scheme. The primary problem of this algorithm is that the depth image is over-smoothed, and thus the edges and fine details cannot be preserved. Tseng and Wang in [11] presented a depth reconstruction algorithm with spatial coherency prior based on matting Laplacian matrix constructed from the all-in-focus image. A local learning scheme to derive a spatial coherency prior directly from a multi-focus image sequence was proposed in [12]. Depth reconstruction methods using matting Laplacian prior assume that a typical depth image can be approximated by a set of piece-wise of affine transformations of image features within the corresponding windows. It is robust over low-contrast regions and can reduce edge bleeding artifacts. However, the assumption of this local spatial coherency prior does not hold in highly textured regions, and thus the texture-copy artifacts could be introduced. Javidnia and Corcoran in [13] proposed a depth reconstruction algorithm based on Preconditioned Alternating Direction Method of Multipliers (PADMM) to refine the depth discontinuities and derive a noise-free depth image.

In this dissertation, a reconstruction method proposed by Moeller et al. [3] is described. In [3], the authors try to formulate the DFF problem as a variational problem. The proposed objective function includes a smooth but nonconvex data fidelity term and a convex nonsmooth regularization, which makes the the method robust to noise and leads to more realistic depth map. It is defined as

$$\hat{d} = \arg\ \min_{d} D(d) + \alpha R(d), \tag{2.12}$$

where $D(d)$ is the data fidelity term that takes the dependence on the measured data into account, $R(d)$ is the regularization term. α is a parameter controlling the balance between fidelity and regularity. As reconstructing depth map by an energy minimization is designed, negative contrast measure at each pixel is chosen as the data fidelity term that can be computed as

$$D(d) = -\sum_{i}\sum_{j} c_{i,j}(d_{i,j}), \tag{2.13}$$

where $c_{i,j}$ is the continuous contrast function that maps a depth to its corresponding contrast value. The regularization term R acts as smoothness term to impose spatial smoothness on reconstructed depth image. The regularization term normally depend on the prior knowledge about the depth that need to be recovered. To to this, the authors used discrete isotropic total variation (TV),

$R(d) = \|Kd\|_{2,1}$, where K is the linear operator such that Kd becomes a matrix with the x-derivative in the first column and y-derivative in the second column, then the total variation can be defined as

$$\|g\|_{2,1} := \sum_i \sqrt{\sum_j (g_{i,j})^2}. \tag{2.14}$$

To minimize such nonconvex energy function, many approaches such as forward-backward splittings (FBS) [24, 25], and methods based on the difference of convex functions [26]. To decrease computational complexity, the authors proposed to apply the alternating directions of multipliers (ADMM) [27] as if the energy was convex. To do this, a new variable g under the constrain $g = Kd$ is introduced, then the energy function can be rewritten as

$$\left(\hat{d}, \hat{g}\right) = \arg \min_{d,g} D(d) + \alpha\|g\|_{2,1} \quad \textit{such that} \;\; g = Kd. \tag{2.15}$$

The constraint $g = Kd$ is enforced iteratively by using augmented Lagrangian method. The minimization for d and g can be down in an alternating way by using applying ADMM and FBS, which can yield

$$\begin{aligned} d^{k+1} &= \arg \min_d \tfrac{\lambda}{2} \left\|Kd - g^k + b^k\right\|_2^2 + \tfrac{1}{2} \left\|d - d^k + \tau\nabla D(d^k)\right\|^2, \\ g^{k+1} &= \arg \min_g \tfrac{\lambda}{2} \left\|g - Kd^{k+1} + b^k\right\|_2^2 + \alpha\|g\|_{2,1}, \\ b^{k+1} &= b^k + \left(Kd^{k+1} - d^{k+1}\right) \end{aligned} \tag{2.16}$$

Due to the excellent performance in edge preservation and smoothing of flat regions, total variation being as a regularization has often been used in many tasks such as image denosing, restoration, reconstruction and so on. However, the performance of Moeller's method relies on the initial depth derived using focus measure operator. To make it robust in texture-less regions, a large kernel for focus measure operator is used, such that object structure and details would be lost in initial depth image. Additional, as it only relies on the initial depth and no additional information such as color information utilized in the optimization framework, numerous fine details are lost and depth edges are blurred.

Deep learning based methods

Recently, deep learning has drawn considerable attention in both academia and industry. As for the problem of DFF, Hazirbas et al. [14] proposed a deep learning method to depth disparity via an auto-encoder-style convolutional neural network named Deep Depth From Focus Network (DDFFNet). The network proposed in this paper is the first end-to-end learning approach to DFF problem. To train such a convolutional neural network, the authors created a dataset with a large number of light-field images and co-registered ground truth depth images recorded with an RGB-D camera. The network takes a focal stack S of refocused images $I \in R^{H \times W \times C}$, and the corresponding target disparity map $D \in R^{H \times W}$ as the input. Then the loss function between the estimated disparity $f(S)$ and the target D can be defines as

$$Loss = \sum_{p}^{HW} M(p) \left\| f_W(S,p) - D(p) \right\|_2^2 + \lambda \left\| \Theta \right\|_2^2 . \tag{2.17}$$

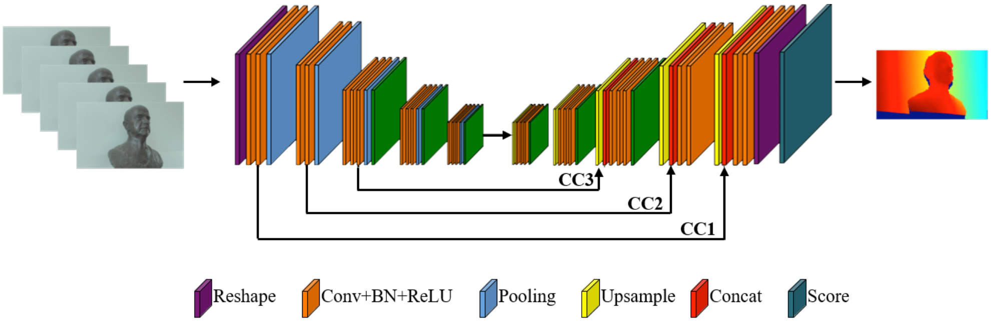

Figure 2.3: DDFFNet proposed in [14]. This network takes a focal stack as a input, and the output is a disparity map.

The loss function is the summation over all valid pixels p where $D(p) > 0$ indicated by the mask $M(p)$. $f : R^{S\times H\times W\times C} \rightarrow R^{H\times W}$ is a convolutional neural network with weights Θ penalized in L_2 norm. Figure 2.3 gives the architecture of the DDFFNet. In DDFFNet, the VGG-16 net[28] was utilized as a baseline for the encoder network that consists of 13 convolutional layers, 5 poolings and 3 fully-connected layers. In order to reconstruct the input size, the authors removed the fully-connected layers and reconstructed the decoder part of the network by mirroring the encoder layers. They inverted the 2×2 pooling operation with 4×4 upconvolution (deconvolution) [29] with a stride of 2 and initialized the weights of the upconvolution layers with bilinear interpolation, depicted as upsample in Figure 2.3. Similar to the encoder part, they utilized convolutions after upconvolution layers to further sharpen the activation results. To accelerate convergence, they added batch normalization [30] after each convolution and learned the scale and shift parameters during training. Batch normalization layers were followed by rectified linear unit (ReLU) activation. Moreover, after the 3rd, 4th and 5th poolings and before the corresponding upconvolutions, they applied dropout with 0.5 probability during training similar to [31]. In order to preserve the sharp object boundaries, the authors concatenated the feature maps of early convolutions conv1_2, conv2_2, conv3_3 with the decoder feature maps: output of the convolutions were concatenated with the output of corresponding upconvolutions.

The DDFFNet basically is a deep learning based sharpness measure method, which takes the focal stack and its corresponding disparity map as the input. Its design allows the network to learn the sharpness for each pixel, and from the sharpness level, the regression layer denoted as *Score* in Figure 2.3 regresses the depth from sharpness. Even though the experimental results demonstrated the robustness and accuracy, there is a significant need to improve the ability of preserving object structure.

2.4 Edge-preserving image denoising

In this section, a literature review of methods for edge-preserving image denoising methods is presented. Several classical edge-preserving algorithms including anisotropic diffusion filter, bilateral filter, nonlocal means filter and mutual structure for joint filter are presented respectively.

Anisotropic diffusion

Anisotropic diffusion in [32] is inspired by interpreting the Gaussian blur as a heat conduction partial differential equation (PDE) $\frac{\partial I}{\partial t} = -\triangle I$. That is, the intensity I of each pixel is seen as heat and is propagated over time to its 4 neighbors based on the heat spatial variation.

Perona and Malik in [32] introduced and edge-stopping function g that varies the conductance according to the image gradient. The prevents heat flows across edges:

$$\begin{aligned} \tfrac{\partial I}{\partial t} &= div\left(c(x,y,t)\nabla I\right) \\ &= \nabla c \nabla I + c(x,y,t)\Delta I, \end{aligned} \tag{2.18}$$

where c is the edge-stopping function suggested as

$$c\left(\|\nabla I\|\right) = e^{-(\|\nabla I\|/K)^2}, \tag{2.19}$$

or

$$c\left(\|\nabla I\|\right) = \frac{1}{1+\left(\frac{\|\nabla I\|^2}{K}\right)} \tag{2.20}$$

, in which ∇ is the Laplacian operator, Δ is the gradient operator, and $c(x,y,t)$ is the diffusion coefficient controlling what gradient intensity should stop diffusion. K is an edge magnitude parameter in the intensity domain. In diffusion process, this gradient magnitude is adopted to detect image edges or boundaries as a step. The diffusion coefficient $c(x,y) \to 0$ if $\nabla I \gg K$, such that the diffusion is "stepped" across edges. On the contrary, $c(x,y) \to 1$ if $\nabla I \ll K$, such that it becomes isotropic diffusion (Gaussian filtering). The discrete Perona-Malik diffusion equation is given by

$$I_{t+1} = I_t + \lambda\left(cN_{x,y}\nabla_N\left(I_t\right) + cS_{x,y}\nabla_S\left(I_t\right) + cE_{x,y}\nabla_E\left(I_t\right) + cW_{x,y}\nabla_W\left(I_t\right)\right) \tag{2.21}$$

where t describes discrete time steps and λ is a scalar that determines the diffusion rate. ∇_N, ∇_S, ∇_E, and ∇_W are gradient operator in north, south, east and west directions, respectively. cN_x, cS_x, cE_x, and cW_x are diffusion coefficients in corresponding directions, respectively. The calculations of those operators are reported as follows:

$$\begin{aligned}
\nabla_N\left(I_t\right) &= I_{x,y_{-1}} - I_{x,y} \\
\nabla_S\left(I_t\right) &= I_{x,y+1} - I_{x,y} \\
\nabla_E\left(I_t\right) &= I_{x-1,y} - I_{x,y} \\
\nabla_W\left(I_t\right) &= I_{x+1,y} - I_{x,y},
\end{aligned} \tag{2.22}$$

and

$$\begin{aligned}
cN_{x,y} &= e^{(-\|\nabla_N(I)\|^2/K^2)} \\
cS_{x,y} &= e^{(-\|\nabla_S(I)\|^2/K^2)} \\
cE_{x,y} &= e^{(-\|\nabla_E(I)\|^2/K^2)} \\
cW_{x,y} &= e^{(-\|\nabla_W(I)\|^2/K^2)}.
\end{aligned} \tag{2.23}$$

Although anisotropic diffusion is a very powerful filter in edge-preserving image denoising, it is very difficult to find the proper parameter settings especially stopping time to get satisfactory results. Figure 2.4 illustrates the denoising effects on varying parameter settings.

Bilateral filter

Bilateral filter was proposed by Tomasi and Manduchi [23] as an alternative to anisotropic diffusion. It is a non-linear filter where the output is a weighted

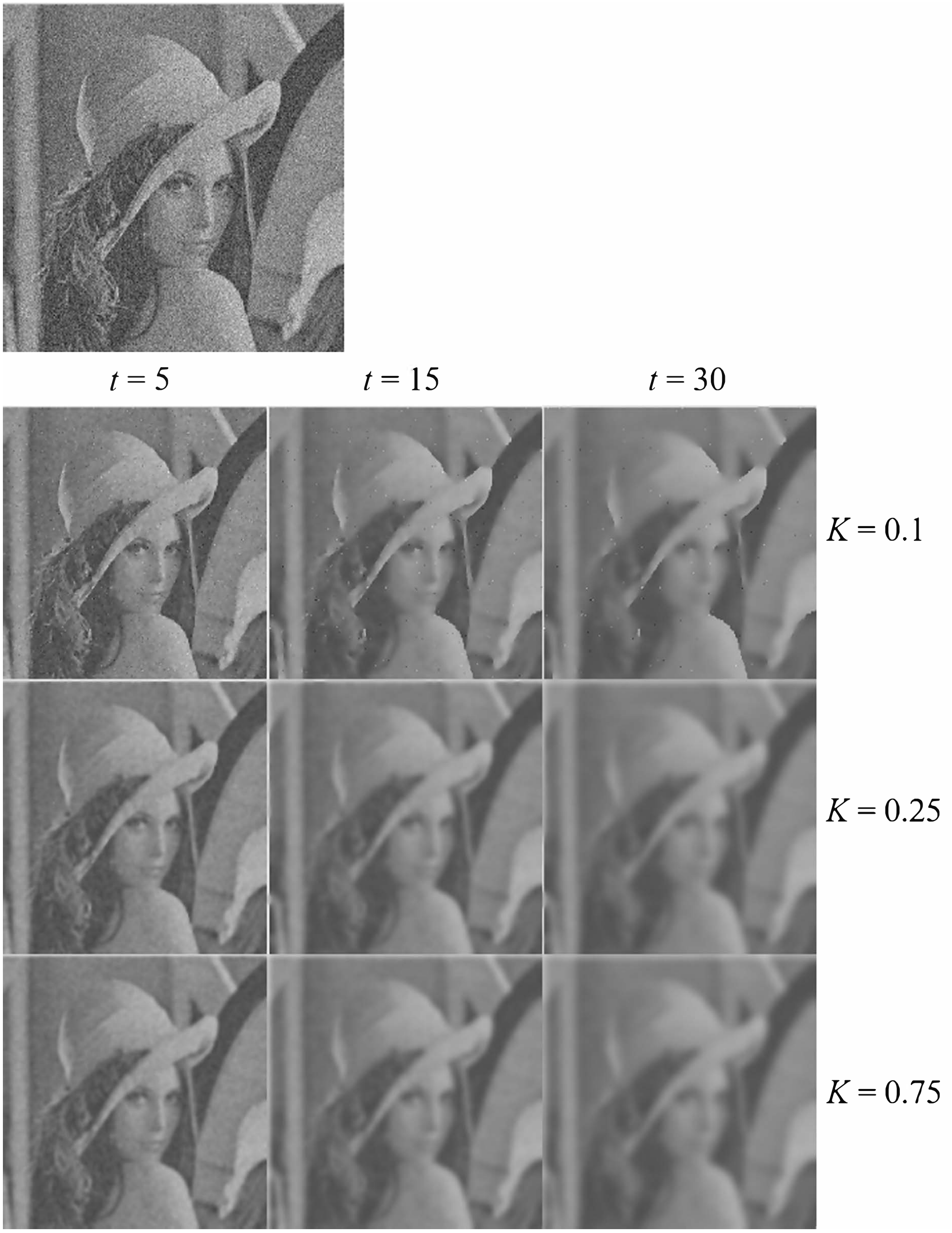

Figure 2.4: Illustration of the effects in anisotropic diffusion using various parameters setting. From top row to bottom row, the edge magnitude parameter K is set to 0.1, 0.25, and 0.75, respectively. From left column to right column, the time steps t is set to 5, 15, and 30, respectively.

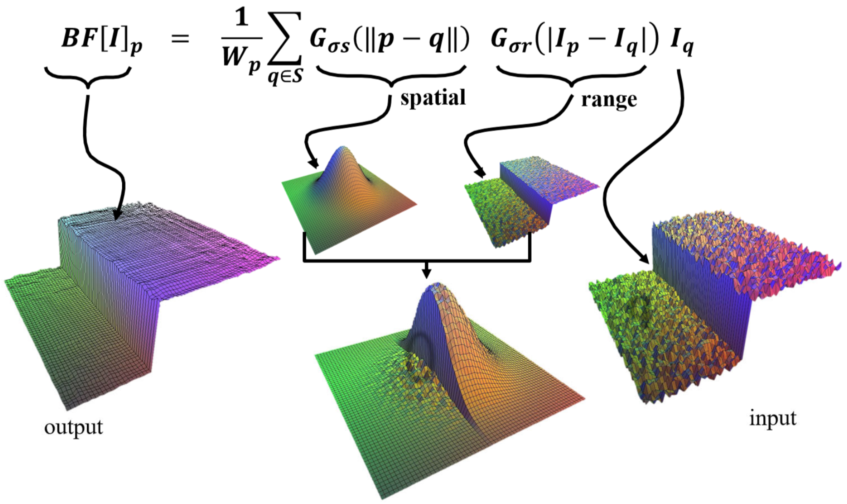
$\underbrace{BF[I]_p} = \frac{1}{W_p}\sum_{q\in S} \underbrace{G_{\sigma s}(\|p-q\|)}_{\text{spatial}} \; \underbrace{G_{\sigma r}(|I_p - I_q|)}_{\text{range}} \; I_q$
output
input

Figure 2.5: Illustration of how the spatial kernel and color kernel combine to preserve edges. This figure is reorganized from images in [33].

average of the input. The bilateral filter is defined as follows:

$$BF[I]_p = \frac{1}{W_p}\sum_{q\in S} G_{\sigma_s}(\|p-q\|)G_{\sigma_r}(|I_p - I_q|) * I_q, \tag{2.24}$$

and Figure 2.5 illustrates the process of bilateral filtering. As shown, the bilateral filter consists of two kernels: spatial kernel G_{σ_s} and range kernel G_{σ_r}, which are defined as follows:

$$G_{\sigma_s} = \exp\left(-\frac{\|p-q\|^2}{2{\sigma_s}^2}\right), \tag{2.25}$$

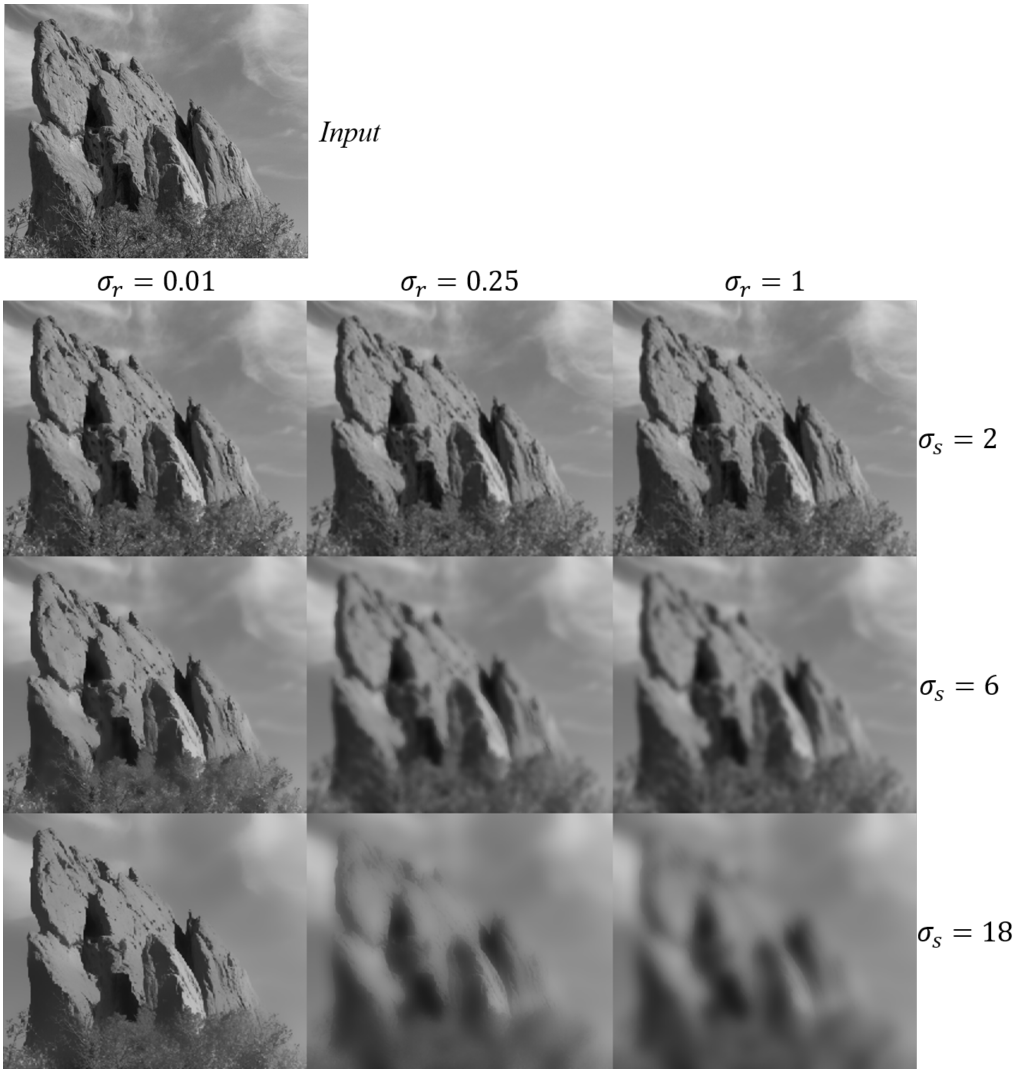

Figure 2.6: Results of bilateral filter with vary spatial and range parameter settings. From top row to bottom row, the spatial parameter is set to 2, 6, and 18, respectively. From left column to right column, the range parameter is set to 0.01, 0.25, and 1, respectively.

and

$$G_{\sigma_r} = \exp\left(-\frac{\|I_p - I_q\|^2}{2{\sigma_r}^2}\right). \tag{2.26}$$

The weight of a pixel not only depends on a spatial function, but also on a function of range function, which is able to decrease the weight of pixels with large intensity differences. Similar with anisotropic diffusion, the range function acts as an edge-stopping function. With Eqs. (2.25) and (2.26), the bilateral filter can be rewritten as

$$\begin{aligned} BF[I]_p &= \tfrac{1}{W_p}\sum_{q\in S} G_{\sigma_s}(\|p-q\|)G_{\sigma_r}(|I_p-I_q|) * I_q \\ &= \tfrac{1}{W_p}\sum_{q\in S} \exp\left(-\tfrac{\|p-q\|^2}{2{\sigma_s}^2}\right)\exp\left(-\tfrac{\|I_p-I_q\|^2}{2{\sigma_r}^2}\right) * I_q \end{aligned}, \tag{2.27}$$

where W_p is a normalization factor:

$$W_p = \sum_{q\in S} \exp\left(-\frac{\|p-q\|^2}{2{\sigma_s}^2}\right)\exp\left(-\frac{\|I_p-I_q\|^2}{2{\sigma_r}^2}\right) \tag{2.28}$$

As an alternative to anisotropic filter, bilateral filter computes the weight of each pixel using a Gaussian in spatial domain multiplied by an influence function (range function) in intensity domain that can decrease the weight of pixels with large intensity differences. Figure 2.6 shows some smoothing results using bilateral filter with varying spatial and range parameter settings. Even though both anisotropic fusion and bilateral filter are able to prevent averaging across edges, bilateral filter has some advantages compared to anisotropic fusion. Bilateral filter does not involve the solution of partial differential equations can be implemented in a single iteration.

Nonlocal means filter

Unlike the other "local mean" filters that take the mean value of of a group of pixels surrounding a target pixel to smooth the image, nonlocal means filter proposed in [34] takes a mean of all pixels in the image, weighted by how similar these pixels are to the target pixel. Given a discrete noisy image $v = \{v(x)|x \in I\}$, the denoised value $NL(x)$, for a pixel x, can be calculated as weighted average of all pixels in the whole image,

$$NL(x) = \sum_{y \in I} w(x, y) * v(y) \tag{2.29}$$

where the family of weights $\{w(x, y)\}_y$ depend on the similarity of the pixels x and y, and satisfy the conditions $0 \leq w(x, y) \geq 1$ and $\sum_y w(x, y) = 1$. The similarity between two pixels x and y depends on the similarity of intensity vectors $v(N_x)$ and $v(N_y)$, where the N_k represents neighborhood pixels in a fixed sized local window centered at pixel k. This similarity is measured as a function of weighted Euclidean distance $\|v(N_x) - v(N_y)\|^2$. The weights then can be computed as

$$w(x, y) = \frac{1}{Z(x)} \exp\left(-\frac{\|v(N_x) - v(N_y)\|^2}{h^2}\right) \tag{2.30}$$

where $Z(x)$ is the normalization factor, which can be defined as follows:

$$Z(x) = \sum_{y} \exp\left(-\frac{\|v(N_x) - v(N_y)\|^2}{h^2}\right), \tag{2.31}$$

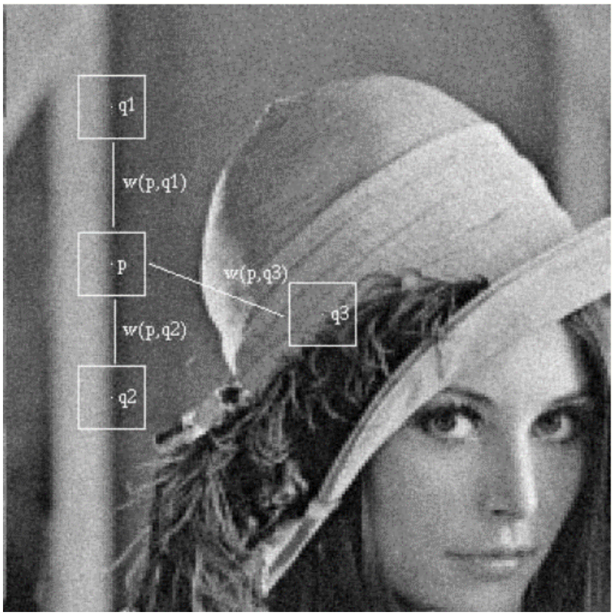

Figure 2.7: Scheme of nonlocal means filter. Similar patches give a large weight, $w(p, q1)$ and $w(p, q2)$, on the contrary, different neighborhoods give a small weight $w(p, q3)$ [35].

where h is a parameter controlling the degree of filtering. Nonlocal means filter not only compares the gray level in a single point but geometrical configuration in a whole neighborhood. This fact allows a more robust comparison than local neighborhood filters. Figure 2.7 gives an illustration of this fact. As shown the pixel $q3$ has the similar intensity value of pixel p, but the neighborhood pixels are much different and thus, the weight of $w(p, q3)$ becomes small. Figure 2.8 gives an example of image smoothing result using nonlocal means filter.

The nonlocal filter has better performance in noise removal and structure preservation by adopting the local and nonlocal geometry of the image compared with other local means filter. However, in terms of depth refinement, there are texture-copy artifacts caused by color inconsistency between depth image

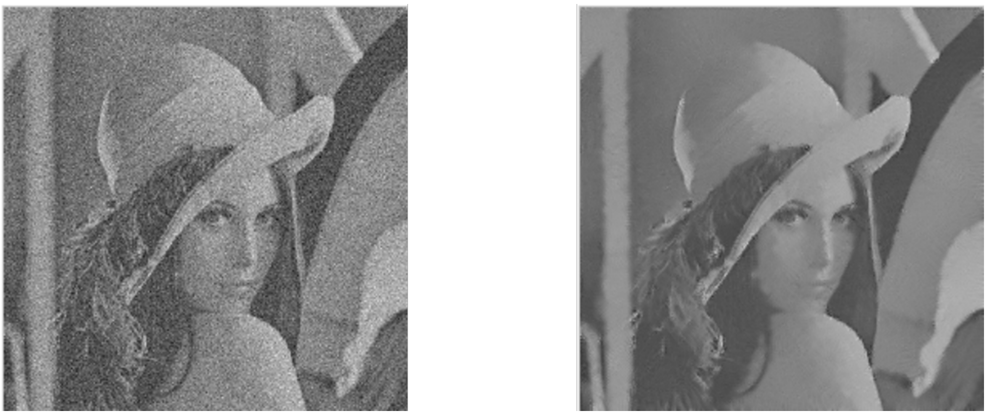

Figure 2.8: An example of image smoothing using nonlocal means filter.

and color image. Some artifacts contain clear edges, which would be treated as geometrical structures and preserved in the smoothed image.

Mutual structure for joint filtering

In image filtering, there is one kind of filtering with a guidance image as a prior and transfer the structure details from guidance image to target image, known as joint or guided filtering. Joint image filtering has been successfully been applied to a variety of computer vision and computer graphics tasks such as depth enhancement [16, 36, 37], joint upsampling [38], and cross-modality noise reduction [39–41]. Generally, there is one assumption for joint filtering that the guided image has perfect structural information. However, there may be completely different edges in guided image and target image. Simply passing all structures into target image could introduce significant errors. Figure 2.9 gives an example of inconsistent edges in target image and guided image. To solve this issue, Shen et al. in [41] propose a joint filtering using mutual structure. The

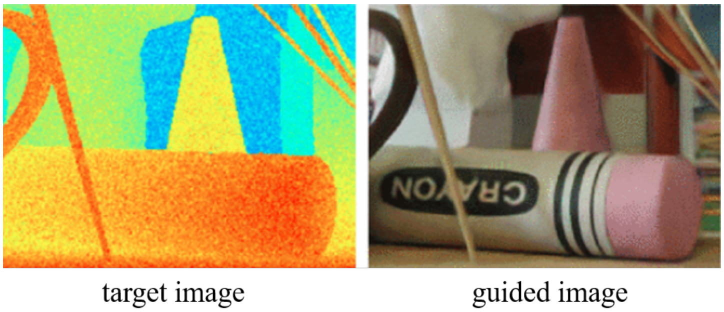

Figure 2.9: Example of inconsistent edges in target image and guidance image [41].

basic idea of this method is trying to utilize both target image and guided image, not guided image only. To address this structure inconsistency problem, they proposed the the concept of mutual-structure, which refers to the structural information both contained in target image and guided image. Thus, target image can be safely enhanced by joint filtering.

Mutual structure formulation

As target image and guided image are hardly with exact same structures, the authors roughly categorize it into three types: mutual structures, inconsistent structures, and smooth regions. Mutual structure can be intuitively understood as common edges that are not necessary with same gradient direction or magnitude. Inconsistent structures are defined when one edge appears only in one image but not in another image. Smooth regions can be easily understood as non-edge regions that are easily influenced by noise. Among these three types, inconsistent structures generally cause significant errors when transferring erroneous structures to target image. To solve this problem, authors try to find the mutual structures between target image and guided image and let it guides joint filtering process. Accordingly, filtering process is not only applied on the target image but the guided image.

The structure similarity measure between corresponding patches in target image I and guided image G is given as

$$\rho(I_p, G_p) = \frac{cov(I_p, G_p)}{\sqrt{\sigma(I_p) + \sigma(G_p)}}, \tag{2.32}$$

where $cov(I_p, G_p)$ is the covariance of patch intensities. $\sigma(I_p)$ and $\sigma(G_p)$ denote the variance.

As $\rho(I_p, G_p)$ is nonlinear operator, it is difficult to use it directly for structure optimization. To make the problem trackable, the relationship between $\rho(I_p, G_p)$ and least-square regression is built as follows:

$$f(I, G, a_p^1, a_p^0) = \sum_{q \in N(p)} \left(a_p^1 I_q + a_p^0 - G_q\right)^2, \tag{2.33}$$

where a_p^1 and a_p^0 are the regression coefficients assumed to be constant in local window $N(p)$ of pixel p. This function linearly represents one patch in G by that in I. To determine the regression coefficients, a error function $e(I_p, G_p)^2$ is defined as

$$e(I_p, G_p)^2 = \min_{a_p^1, a_p^0} \frac{1}{|N|} f(I, G, a_p^1, a_p^0), \tag{2.34}$$

where $|N|$ is number of pixels in $N(p)$. It is claimed that the relationship be-

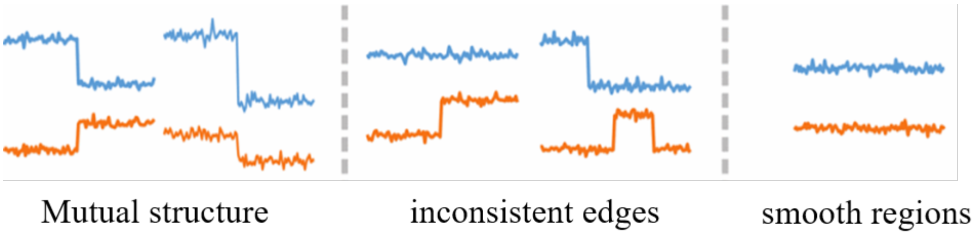

Figure 2.10: 1-D example of three types of different structures [41].

tween the mean square error $e(I_p, G_p)^2$ and $\rho(I_p, G_p)$ is

$$e(I_p, G_p) = \sigma(G_p)(1 - \rho(I_p, G_p)^2), \tag{2.35}$$

where $\sigma(G_p)$ is the variance of patch centered at p in G. As $e(I_p, G_p)$ is not a symmetrical function, the final patch similarity measure is defined as a sum of $e(I_p, G_p)$ and $e(G_p, I_p)$, which is defined as follows:

$$S(I_p, G_p) = \left(\sigma(I_p)^2 + \sigma(G_p)^2\right)\left(1 - \rho(I_p, G_p)^2\right)^2. \tag{2.36}$$

Figure 2.10 gives an example of three different structures. When two patches contain same edges, $|\rho(I_p, G_p)| = 1$. Otherwise, $|\rho(I_p, G_p)|$ is small when patch structures are different. In texture-copy regions where the edges appear in the reconstructed depth image but not in the initial depth image, $\sigma(G_p)$ is large and $\sigma(I_p)$ is small. $S(I_p, G_p)$ therefore outputs a relatively large number. On the other hand, when common edges appear in two patches or when both patches do not contain any significant edges, $S(I_p, G_p)$ would be a small value.

Considering the trivial solution could be produced when the whole images of I and G contain no edges at all, and requirement to smooth the target image by removing noise and visual artifacts, some regularization terms are added into the final objective function that is

$$E(I,G,a,b) = E_S(I,G,a,b) + E_d(I,G) + E_r(a,b), \tag{2.37}$$

where $E_S(I,G,a,b)$ is the essential image structure similarity term, $E_d(I,G)$ is a regularization term used to avoid trivial solution, and $E_r(a,b)$ is another regularization term used to smooth target image. Those three terms are defined as follows:

$$E_S(I,G,a,b) = \sum_p f(I,G,a_p^1,a_p^0) + f(G,I,b_p^1,b_p^0), \tag{2.38}$$

$$E_d(I,G) = \sum_p \lambda \left\| G_p - G_{0,p} \right\| + \beta \left\| I_p - I_{0,p} \right\|, \tag{2.39}$$

where λ and β two parameters.

$$E_r(a,b) = \sum_p \left(\varepsilon_1 {a_p^1}^2 + \varepsilon_2 {b_p^1}^2 \right) \tag{2.40}$$

where ε_1 and ε_2 are parameters controlling smoothness strength on G and I respectively.

The optimization process is able to efficiently get filtered output I and G from I_0 and G_0 using derivatives and Jacobi method [39]. Mutual structure based joint filtering aims to preserve the mutually consistent structures while suppressing that not commonly shared in both images, which enables it have a better better performance compared with bilateral filtering. However, in DFF, it is difficult for this filter to recover initial depth derived using focus measure operator, especially in texture-less regions. Additionally, as the reconstructed depth image using nonlocal Laplacian prior already suffers from inconsistent structure from color image, mutual structure based joint filter could only add to this problem.

Chapter 3

Depth-from-Focus Reconstruction using Nonlocal Matting Laplacian Prior

3.1 Overview

Depth reconstruction from 2D image is a fundamental problem for various kinds of applications such as 3D measurement and 3D object segmentation. Object structure plays an very important role in those applications. In the field of DFF, to preserve object structure and fine details, many researches proposed to adopt color image as guidance, and transfer the object structures to the reconstructed depth image. The advantage of such kind of methods is the ability to preserve object structures. However, as the structure in color image may exist some structures that are not expected to transfer to target depth image, it is inevitable to introduce some visual artifacts called texture-copy artifacts. This

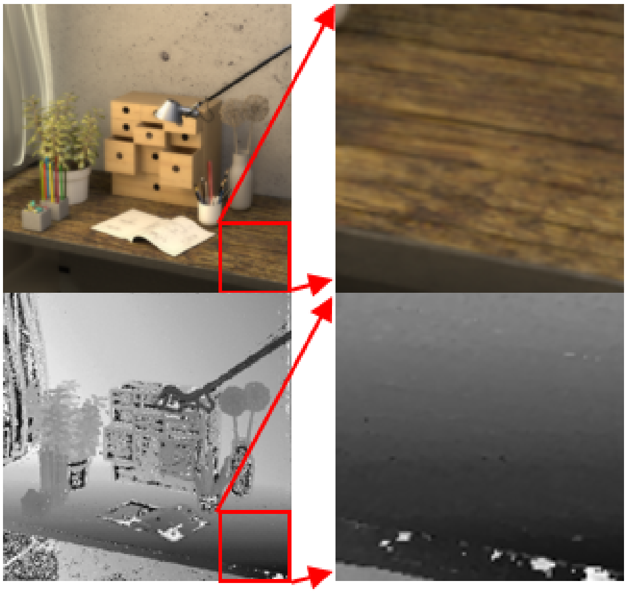

Figure 3.1: An example of inconsistent edges between color image and initial depth image.

is caused by the structure inconsistency between initial depth image and the color image. Figure 3.1 gives and example of structure inconsistency. Now that, it is difficult to avoid the texture-copy artifacts, we can achieve our final goal in an indirect way. That is the two-stage framework for the depth recovery. In the reconstruction stage, preserving the object structure and fine details as much as we can becomes the temporary objective, and the noise removal and texture-copy suppression can be left to the depth refinement stage.

In this chapter, a DFF reconstruction approach with the inclusion of matting Laplacian prior is presented. As the texture-copy artifacts suppression in this stage is not our concern, nonlocal principle is a adopted in the matting Laplacian matrix construction to keep object structure and preserve fine details. Additionally, an effective variance based confidence level measure is also proposed to suppress the texture-copy artifacts. The design of confidence measure does not only consider individual pixels but also its local consistency. Generally, the regions in initial depth image with small depth variances are generally highly textured, thus it can help suppress the texture-copy artifacts to some degree.

3.2 Image matting and matting Laplacian

Image matting is a problem of accurate foreground extraction from an image based on limited user input. Formally, image matting algorithms take an image I as input, which is composed of a foreground image F and a background image B. The image value I_i at pixel i is assumed to be linear combination of a foreground color value F_i and a background color value B_i. That is,

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i, \tag{3.1}$$

where α_i corresponds to the pixel's foreground opacity and usually named the alpha matte value ranging from 0 to 1. In image matting, as all quantities on the right-hand side of Equation (3.1) are unknown, the image matting process is basically an ill-posed problem. Typically, to solve this severely under-constrained problem, user interaction is usually needed to provide a trimap as a starting point. Figure 3.2 gives an example of image matting. The trimap consists of a small set of labelled foreground pixels marked white color, background pixels marked black and unknown pixels marked gray. Such kind of algorithms are

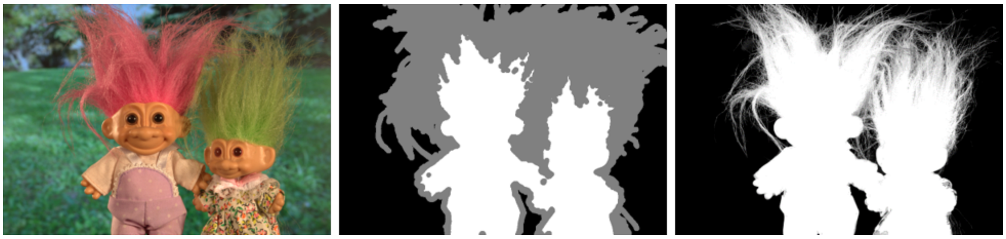

Figure 3.2: Example of image matting. From left to right: Color image, input trimap, and result of matting respectively. The white, black, and gray regions in middle image represent foreground, background and unknown regions, respectively.

called supervised image matting [42–45]. In contrast, other approaches decompose the color image into foreground and background parts automatically called unsupervised matting. [46–49].

Among aforementioned image matting approaches, matting Laplacian based method introduced by Levin et al. in [42] is one of the prominent works. In that work, Levin et al. propose the use of the matting Laplacian, which is basically a graph-based method to compute alpha matte. As solving the alpha in Equation (3.1) is a severely under-constrained problem, Levin et al. assume that the both foreground and background are approximately constant over a small window around each pixel. More specially, the approach is based on a color line model, which assumes that each of F and B in a small window w (3×3 or 5×5) lies roughly on a line in the color space. The α then can be expressed as a linear combination of the color values:

$$\alpha_i = aI_i + b, \forall_i \in \omega_i, \tag{3.2}$$

where $a = \frac{1}{F-B}$ and $b = \frac{-B}{F-B}$. By scanning the window w_j over the image, the matting problem can be formulated as the minimization of the following cost function:

$$J(\alpha, a, b) = \sum_{j \in I} \left(\sum_{i \in \omega_j} (\alpha_i - a_j I_i - b_i)^2 + \varepsilon {a_j}^2 \right), \tag{3.3}$$

where w_j is a local window centered at pixel j, and ε is a regularization parameter. By minimizing the cost function, a and b can be ultimately eliminated from equation 3.3, and a quadratic form respect to α can be derived as

$$J(\alpha) = \alpha^T L \alpha. \tag{3.4}$$

In (3.4), α is a $N \times 1$ vector, where N is the number of image pixels. L is a sparse, symmetric, and positive semi-definite $N \times N$ matrix called the matting Laplacian matrix. The entry $L(i, j)$ of the matting Laplacian can be calculated as

$$\sum_{k|(i,j) \in w_k} \left(\delta_{ij} - \frac{1}{|w_k|} \left(1 + (I_i - \mu_k)^T \left(\Sigma_k + \frac{\varepsilon}{|w_k|} U \right)^{-1} (I_j - \mu_k) \right) \right), \tag{3.5}$$

where I_i and I_j are the color vectors of input image at pixels i and j, respectively, $\delta_{i,j}$ is the Kronecker delta, μ_k is the 3×1 mean vector in window w_k centered at k, Σ_k is the 3×3 covariance matrix of the color intensities in the same window , and $|w_k|$ is the number of pixels in this window.

This matrix is originally proposed to solve the matting problem, and later applied in many fields such as: image dehazing [50], image deblurring [51, 52], depth reconstruction [11, 12, 53], and so on. An example application is presented to show the use of matting Laplacian. Figure 3.3 illustrates an example of haze removal using the method proposed in [50]. As shown the refined transmission

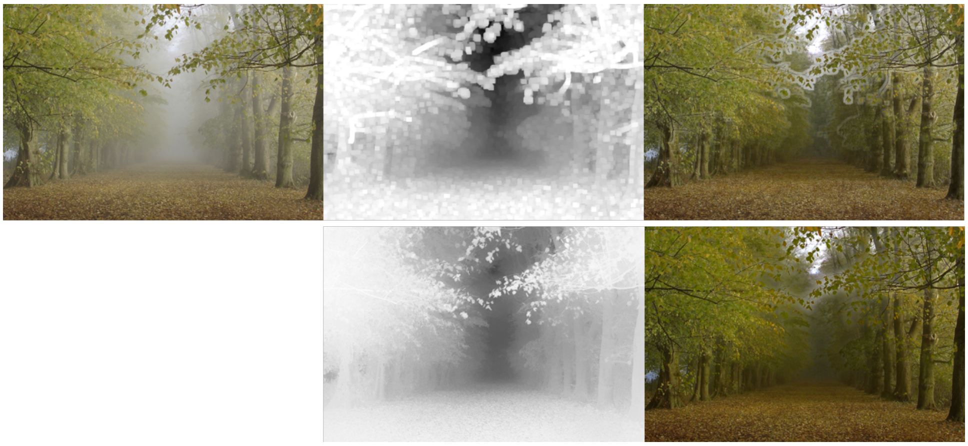

Figure 3.3: Haze removal. The top-left image is the input haze image. The top-middle image is the estimated transmission map before soft matting. The bottom-middle image is the refined transmission map after soft matting. The top-right and bottom right images are the recovered images using unrefined and refined transmission maps, respectively.

map using matting Laplacian can help achieve much better performance than unrefined transmission.

3.3 Depth-from-focus

In this dissertation, the summed-modified Laplacian (SMLAP) operator is employed to measure the degree of focus due to its better performance and less computational complexity compared to other operators [54]. Then, for a given image focal stack $I_{set} = \{I^1, I^2, ...I^K\}$, the focus measure for a pixel i at position (x, y) on k-th image can be calculated as follows:

$$fm_k\left(x, y\right) = \sum_{(i,j)\in\Omega(x,y)} \Delta I_k\left(i, j\right), \tag{3.6}$$

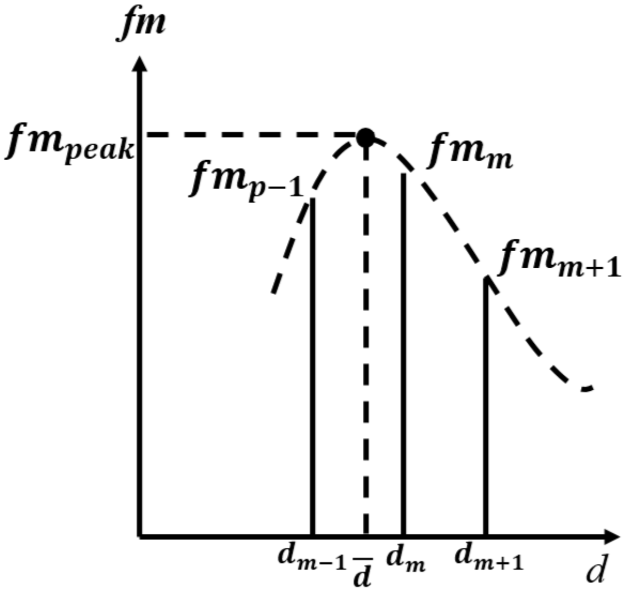

Figure 3.4: Gaussian approximation of focus measure function.

where fm_k is the modified Laplacian of I_k which can be calculated as

$$\Delta I_k = |I_k * \mathcal{L}_x| + |I_k * \mathcal{L}_y|. \tag{3.7}$$

The convolution filter kernels used to calculate the modified Laplacian are

$$\mathcal{L}_x = \begin{bmatrix} -1 & 2 & 1 \end{bmatrix} \quad and \;\; \mathcal{L}_y = {\mathcal{L}_x}^T. \tag{3.8}$$

To improve the accuracy of the initial depth estimation, the authors in [2] proposed to model the focus measure function $fm(x, y)$ of k as a Gaussian function of continuous depth z whose maximum corresponds to the position of

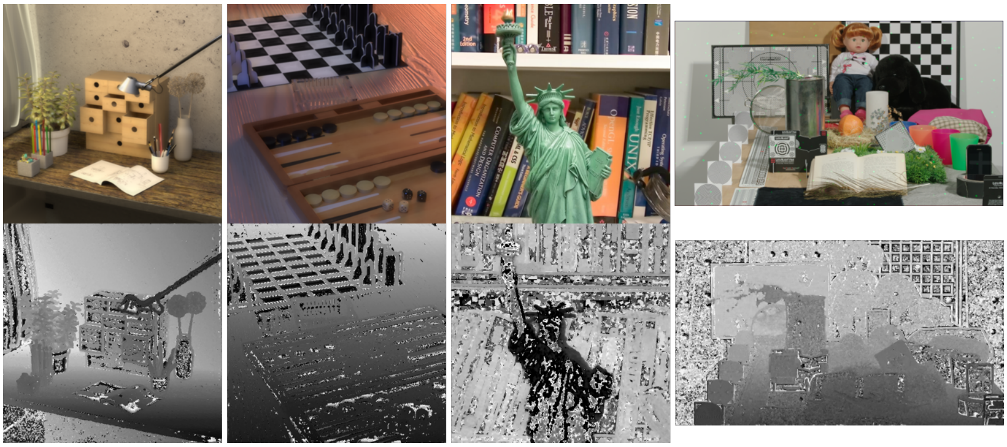

Figure 3.5: Color images in various of scenes and its initial depth images using SMLAP operator after Gaussian interpolation.

best focus. The ideal Gaussian function at position (x, y) is defined as

$$G_{x,y}(z) = A \cdot \exp\left(-\frac{1}{2}\left(\frac{z-\mu}{\sigma_{fm}}\right)^2\right), \tag{3.9}$$

where A is the maximum value of the Gaussian function, μ and σ_{fm} are its mean value and standard deviation, respectively. Following [2], A, μ, and σ_{fm} can be obtained by interpolation as a function of depth z. The estimated initial depth $\tilde{D}(x, y)$ after Gaussian interpolation corresponds to the location that maximizes $G_{x,y}(z)$. Then, we have

$$\tilde{D}(x, y) = \arg\max_{z}\left(G_{x,y}\left(z\right)\right). \tag{3.10}$$

Figure 3.5 presents some example of the initial depth images after Gaussian interpolation from depth images derived using SMLAP. Owing to the limitations of the focus measure operator, some unreliable depths appear in texture-less,

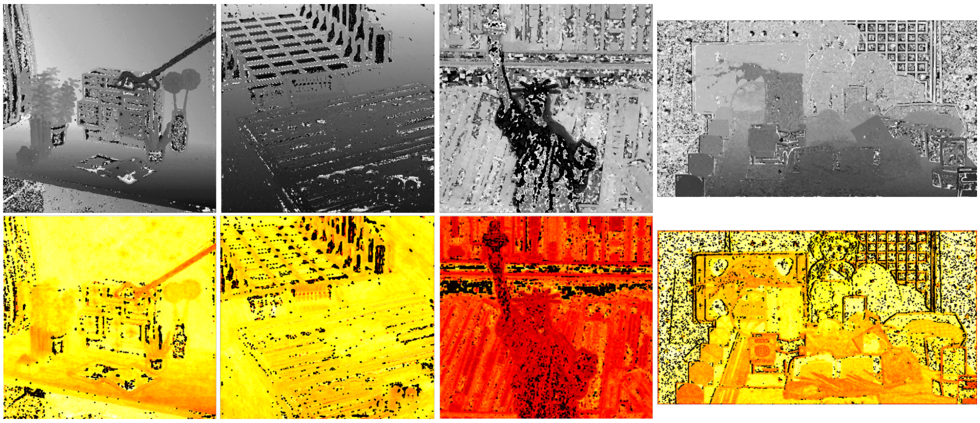

Figure 3.6: Reliability measure for the initial depth images derived using SM-LAP operator after Gaussian interpolation. The reliability measure map indicates that the brighter pixels have higher reliability levels.

underexposed, or overexposed regions. To detect such unreliable depth regions, Pertuz et al. [54] proposed a reliability measure scheme by measuring the fit of focus measure function to the Gaussian model to predict the performance of initial depth estimation. The reliability measure is calculated as

$$RM(x,y) = 20 \cdot \log\left(\frac{fm_{\max}}{\frac{1}{K}\sum\limits_{k=1}^{K}|fm^k(x,y) - G_{x,y}(k)|}\right), \tag{3.11}$$

where $fm_{max} = max\{fm(x,y)\}$ is the normalization factor. Figure 3.6 shows some examples of the reliability measure for the initial depth images after Gaussian interpolation.

By expecting that a larger focus measure value at an image pixel typically indicates that the image pixel is more focused, we assume that the probability p_i^k of pixel i at position (x,y) on the k-th frame is proportional to the corresponding focus measure value fm^k. Here, we define

$$p_i{}^k = \frac{fm^k(x,y)}{\sum\limits_{j=1}^{K} fm^j(x,y)}. \tag{3.12}$$

Based on the assumption, all-in-focused image *AIF* can be calculated as follows:

$$\mathit{AIF}_i = \sum_{k=1}^{K} p_i{}^k \cdot I_i{}^k. \tag{3.13}$$

3.4 Depth reconstruction

3.4.1 Problem statement

In this section, we aim to reconstruct a reliable depth image with clear edges and fine details from a sequence of multi-focus images. Herein, we denote the depth image D and treat the depth reconstruction as a MAP estimation problem in the Bayesian network. For a given image sequence $I^{set} = \{I^1, I^2, ...I^K\}$, we seek a D^* to maximize

$$D^* = \arg\max_D \{p(D|I^{set})\}. \tag{3.14}$$

According to the Bayesian rule, the posterior probability $P(D|I^{set})$ can be decomposed into the product of likelihood function $P(I^{set}|D)$ and prior probability function $P(D)$. Then, we have

$$\begin{aligned} D^* &= \arg\max_D \{p(D|I^{set})\} \\ &= \arg\max_D p(I^{set}|D)p(D). \end{aligned} \tag{3.15}$$

In the following subsections, we will describe the construction of the likelihood and prior models. The likelihood model is constructed based on depth prediction with spatially varying precision, which can properly improve the robustness of depth estimation over texture-less regions. On the other hand, the prior model is derived using the affinity matrix embedded in nonlocal matting Laplacian matrix. The property of this prior achieves the propagation of high-confident depth values to unreliable depth values. With the likelihood and prior models, we can reconstruct a reliable depth image by solving an optimization problem.

3.4.2 Likelihood model

As it is an ill-posed problem to directly model the relation between image set I^{set} and depth image D, we cannot explicitly model the likelihood function $p(I^{set}|D)$. On the other side, the initial depth image $\tilde{D}$ is inferred from image set I^{set} using SMLAP operator. Thus, we employ the indirect relationship between the initial image $\tilde{D}$ and depth image D to formulate $p(I^{set}|D)$ which can be expressed as

$$p(I^{set}|D) \equiv p(\tilde{D}|D). \tag{3.16}$$

Here, we regard $\tilde{D}$ as a random variable governed by the hidden depth D and model it as an identically independent Gaussian distribution with spatially varying precision Λ. Hence, our negative logarithmic likelihood function can be formulated as

$$\begin{aligned} -\ln\left(p(\tilde{D}|D)\right) &\equiv -\ln(N(\tilde{D}|D,\Lambda^{-1})) \\ &\propto \sum_{i=1}^{n} \lambda_i \cdot |\tilde{d}_i - d_i|^2 \\ &= (\tilde{D}-D)^T \Lambda (\tilde{D}-D), \end{aligned} \tag{3.17}$$

where $\tilde{D}$ and D are represented as $n \times 1$ vector and n denotes the total number of pixels of the depth image. d_i and $\tilde{d}_i$ are the values of hidden depth image and observed initial depth image at pixel i. Λ is an $n \times n$ diagonal matrix, in which the element $\Lambda(i,i)$ equals λ_i. Basically, λ_i models the confidence level of initial depth pixel $\tilde{d}_i$. Consequently, a large value of λ_i implies that the reconstructed depth d_i would be more derived from the observed initial depth $\tilde{d}_i$. For the texture-less regions in all-in-focus image, we believe that the confidence levels would decrease. In this dissertation, the design of the confidence level λ is based on the observation that a confident pixel generally exhibits a high reliability measure value and a small depth variance in a local window. On the contrary, a noisy pixel typically has a large depth variance and low reliability measure value. Herein, the depth variance based confidence level λ is defined as follows:

$$\lambda_i = \begin{cases} \mathit{coef}_l \cdot \left(\exp\left(-\sigma_{\tilde{D}}(i)\right)\right) & \sigma_{\tilde{D}}(i) \le t_{var} \; \& \; \mathit{RM}_i > t_{\mathit{RM}}, \\ C_l & \sigma_{\tilde{D}}(i) > t_{var} \; \& \; \mathit{RM}_i > t_{\mathit{RM}}, \\ 0 & \mathit{RM}_i \le t_{\mathit{RM}}, \end{cases} \tag{3.18}$$

where coef_l is a user defined coefficient of confidence level λ. $\sigma_{\tilde{D}}$ is the normalized variance of the initial depth $\tilde{D}$. RM is the reliability measures calculated using Equation 3.11; and t_{RM} is a user specified threshold to separate reliable and unreliable depth regions. t_{var} is the depth variance threshold to partition the smooth depth regions from depth discontinuities, which is derived through Otsu's method [55] on reliable depth regions. To maintain the depth image consistency, the λ value of each reliable pixel is set inversely proportional to the depth variance value. A positive and small constant value C_l is assigned to reliable depth regions with large depth variances to balance the estimation between the observed initial depth and hidden depth. For unreliable regions in the initial depth image, we simply set the confidence levels to zero. Our confidence level design is reasonable and effective as it considers both the independent pixel and its local consistency. As shown in Figure 3.7, the reliable pixels with a small depth variances are assigned with large λ values. As the regions in initial depth image with small depth variances are generally highly textured in its corresponding all-in-focus image, our λ design can thus help maintain the spatial consistency to some degree.

3.4.3 Nonlocal matting Laplacian prior model

As DFF is a classic ill-posed inverse problem, it is necessary to employ image priors to regularize it into a well-posed problem. In [42], Levin et al. introduced a matting Laplacian matrix to compute alpha matte based on a local linear color model. This matrix is originally proposed to solve the matting problem, and later used in dehazing [50] and depth reconstruction [11, 12, 53]. The authors in [11, 53] attempted to reconstruct a depth image by employing a spatial-coherence prior constructed from a graph-based affinity matrix embedded in a

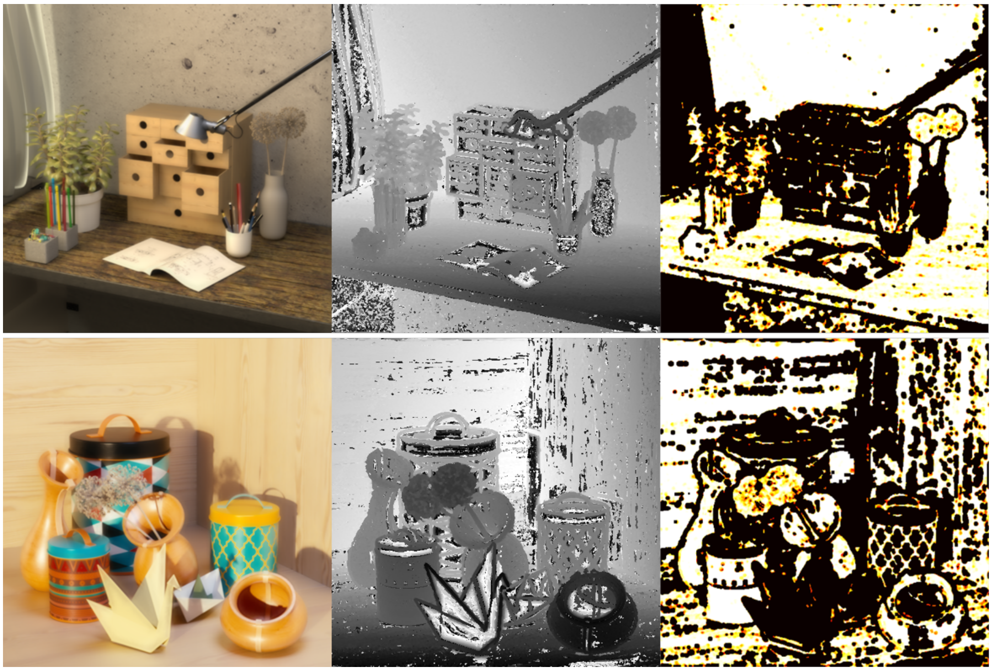

Figure 3.7: Confidence map of selected synthetic data from 4D light field benchmark [56]. From left column to right column are all-in-focus images, the estimated initial depth images; and the corresponding confidence maps, respectively. The confidence map indicates that brighter pixels have higher confidence levels.

matting Laplacian matrix. Tseng and Wang in [12] constructed a local prior model through a local learning scheme under the assumption that the depth value of each pixel can be predicted by an affine transformation of its image features. Good results can be guaranteed if the local linear color model holds. Recently the nonlocal principle has drawn significant attention owing to its excellent edge preservation property in image and movie denoising [34] and image matting [57]. In this dissertation, the nonlocal principle is adopted in the matting Laplacian matrix construction to preserve clear edges and fine details. Rather than using a larger kernel or other nonlocal matting methods performing in spatial domain, our nonlocal principle is implemented by computing the K nearest neighborhood (KNN) in the feature space. For a given pixel i, the feature space is defined as follows:

$$H_i = (x, y, r, g, b)_i. \tag{3.19}$$

The feature vector H_i is constructed using spatial coordinates and RGB color: x, y are the normalized spatial coordinates and r, g, b are the normalized RGB values, respectively. To enforce spatial consistency, r, g, b are scaled by a factor of C_s after normalization. Subsequently, we search for KNN in the feature space using the Euclidean distance $\|H_i - H_j\|$. Figure 3.8 shows a visualization comparison of nine nearest neighbors in spatial domain and defined feature space.

Here, we assume that the KNN of a pixel in the feature space satisfies the color line model. The element of the matting Laplacian matrix $L(i,j)$ can be calculated as follows:

$$\sum_{q|(i,j)\in(Nq)} \left(\delta_{ij} - \frac{1}{K_{NN}}\left(1 + (AIF_i - \mu_q)^T \left(\Sigma_q + \frac{\varepsilon}{K_{NN}} I\right)^{-1} (AIF_j - \mu_q)\right)\right), \tag{3.20}$$

where L is an $n \times n$ matrix; AIF_i and AIF_j are the color vectors of the all-in-focus image at pixels i and j, respectively; δ_{ij} is the Kronecker delta; μ_q and Σ_q are the mean and covariance matrices of the color intensity values in the nonlocal neighbors $N(q)$ of q, respectively; K_{NN} is the element number of $N(q)$; I is a 3×3 identity matrix; and ε is the regularizing parameter. More information

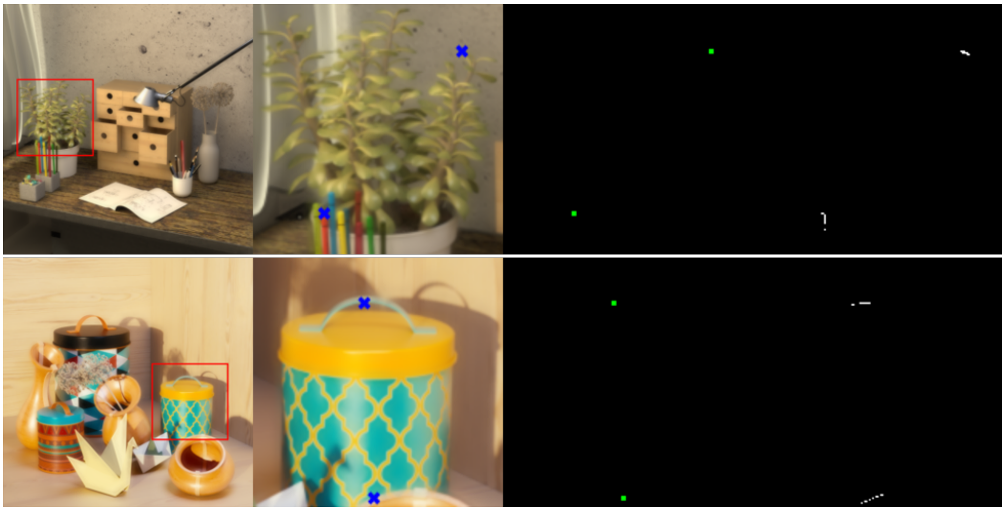

Figure 3.8: Visualization of neighbors used for matting Laplacian matrix construction. The first and second column are the all-in-focus image and its cropped all-in-focus images. The last two columns represent the spatial neighbors marked green and the nonlocal neighbors in our feature space marked white.

can be found in [42]. In the proposed method, unlike the traditional matting Laplacian matrix, the $(i,j)-th$ entry of the modified matting Laplacian matrix is given as

$$\sum_{q|(i,j)\in N(q)}\left(\delta_{ij}-\frac{1}{K_{NN}}\left(1+(X_i-\mu_q)^T\left(\Sigma_q+\frac{\varepsilon}{K_{NN}}I\right)^{-1}(X_j-\mu_q)\right)\right), \tag{3.21}$$

where

$$X_{i\in N(q)}=\begin{cases} AIF_i & RM_i \leq t_{RM}, \\ \mu_q & others. \end{cases} \tag{3.22}$$

For unreliable regions, especially for the unreliable regions that are highly textured in-all-in focus image, we enhance the spatial consistency by reinforcing color smoothness. Our KNN matting Laplacian matrix is constructed based on the nonlocal color smoothness assumption. During the depth reconstruction process, we use this nonlocal color smoothness assumption and employ the matting Laplacian matrix to construct our prior model, which is defined as

$$-\ln\left(P(D)\right) = D^T LD. \tag{3.23}$$

Under this nonlocal color smoothness assumption, there would be a slow depth change between adjacent pixels in our defined feature space if the intensity or colors at these pixels are similar, and a quick depth change if the colors or intensity at these pixels are apparently different. With the likelihood function (3.17) and prior function (3.23), our depth reconstruction is equivalent to minimizing the following energy function:

$$E(D) = (\tilde{D} - D)^T \Lambda(\tilde{D} - D) + D^T LD. \tag{3.24}$$

With the precision matrix Λ and our matting Laplacian matrix L, the closed-form solution of our energy function with respect to D can be obtained by solving the following linear equation

$$\left(L + \Lambda\right) D = \Lambda\widetilde{D} \tag{3.25}$$

3.5 Experimental results

3.5.1 Overview

As the reconstruction process is just one stage of the proposed depth recovery framework, quantitative comparison between the proposed reconstruction method and other related DFF reconstruction methods are not performed in this section. The objective of this set of experiments is to test the performance of depth recovery over texture-less regions and the ability to preserve object structures and fine details in the reconstructed depth image. Several experiments are conducted to validate the proposed method. Experiments for testing the performance of depth recovery over texture-less regions are firstly presented. After that, some experimental comparisons between reconstructed depth images using the proposed nonlocal matting Laplacian prior and local matting Laplacian prior are conducted to evaluate the ability to preserve object structures and fine details. Additionally, a spatial consistency analysis is conduct to compare the difference between reconstruction results using modified nonlocal matting Laplacian and traditional matting Laplacian.

3.5.2 Data configuration

The image sequences used in our experiments are derived from the 4D Light Fields benchmark [56] as it provides several carefully designed synthetic and densely sampled 4D light fields with a highly accurate disparity ground truth. For each dataset, we generate thirty refocused images using the toolbox function LFFiltShiftSum [58] with the range parameters provided in [56].

3.5.3 Reconstruction results

In this experiments, the performance of the depth recovery over texture-less regions are qualitatively evaluated. Figure 3.9 gives some reconstruction results using the proposed reconstruction method. As shown, the result obtained by the SMLAP with Gaussian interpolation is quite noisy and the depth estimation over texture-less regions shows numerous errors. On the other hand, the proposed reconstruction approach with the inclusion of matting Laplacian prior exhibits the good performance in the depth recovery over texture-less regions.

3.5.4 Comparison between reconstruction using local and nonlocal matting Laplacian

The objective of this experiment is to validate the effectiveness of adopting nonlocal principle into the proposed reconstruction method. As the texture-copy artifacts is not taken into account in this reconstruction stage, the purpose of adopting nonlocal principle into the proposed method is trying to preserve object structure and fine details as much as possible. Figure 3.10 shows the comparison results of the reconstructed depth image using traditional local matting and our modified nonlocal matting Laplacian. As shown, the proposed modified nonlocal method can produce depth images with more clear edges and fine details compared with the traditional local method. On the contrary, the local method blurs real depth edges and some details are lost though the rough object structures can be preserved. This is because our nonlocal principle can achieve better clustering performance especially when consistent edges exist in both depth image and the associated all-in-focus image. Additionally, with our

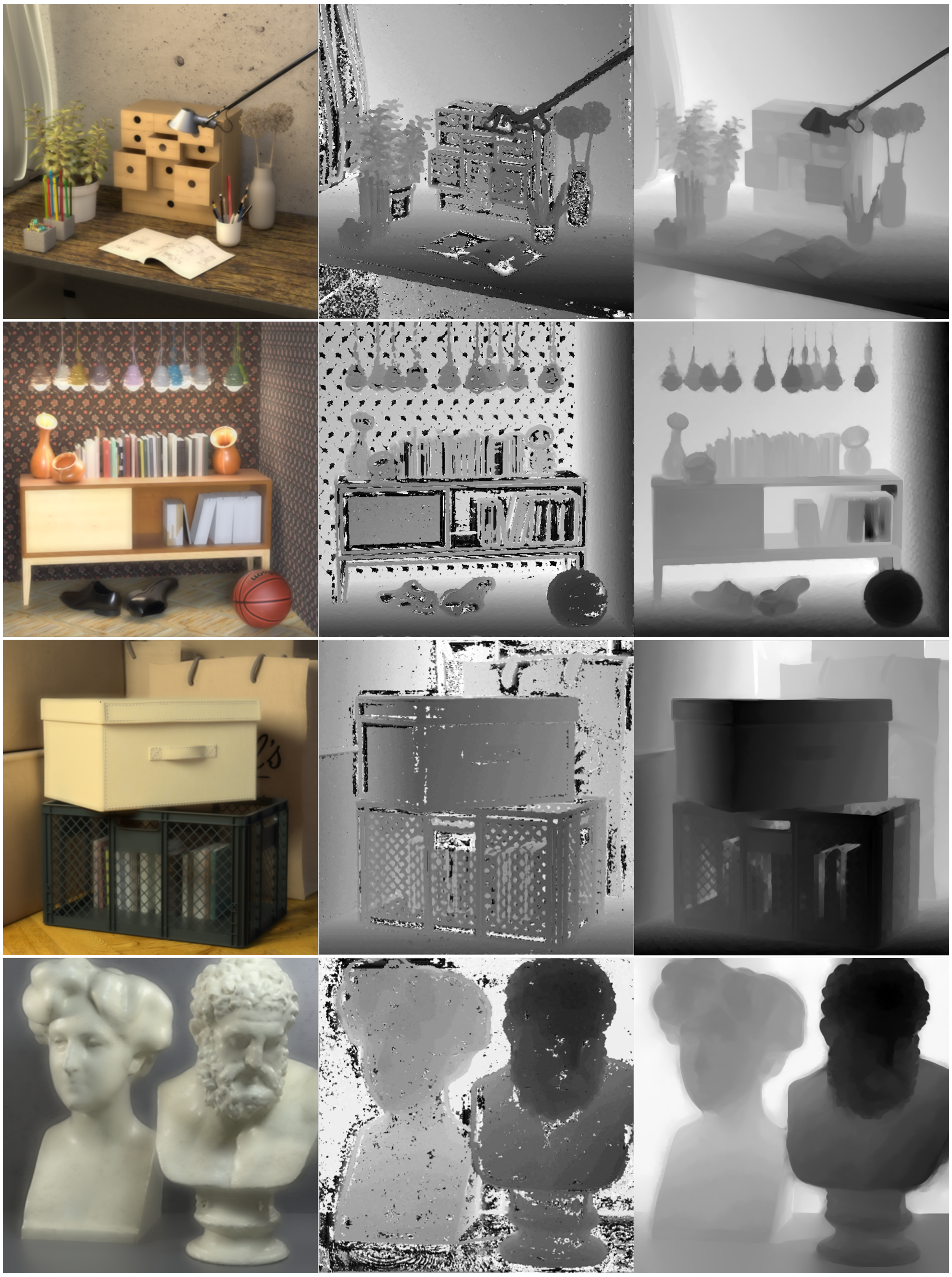

Figure 3.9: Reconstruction results using proposed method with nonlocal matting Laplacian prior. The left column images are the all-in-focus images derived from the proposed method. The middle column images are initial depth images using SMLAP operator after Gaussian interpolation. The right column images are reconstructed depth images.

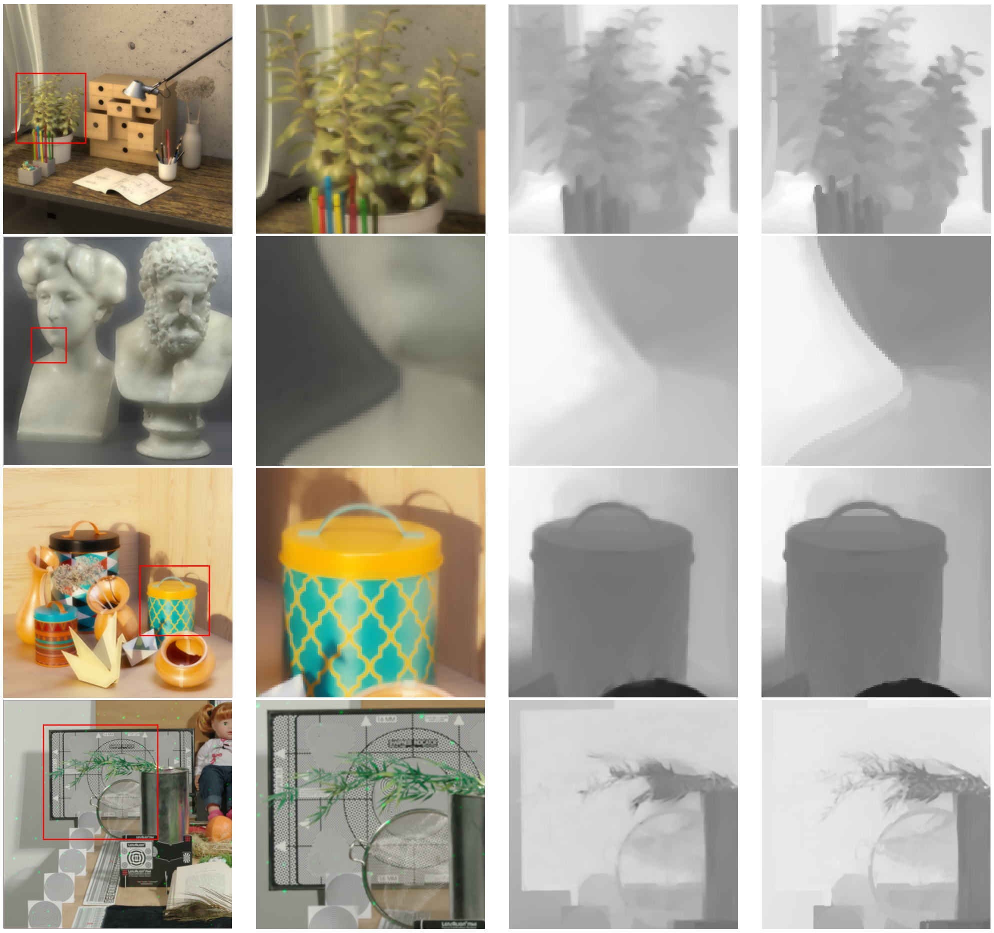

Figure 3.10: Comparison of our reconstructed depth image using traditional local matting Laplacian and modified nonlocal matting Laplacian. The first three rows show the reconstruction results using the synthetic datasets from [56], and the last row shows the reconstruction result using real scene images from [3]. From left to right: all-in-focus images, cropped images, our reconstructed depth images using traditional local matting Laplacian, and reconstructed depth images using modified nonlocal matting Laplacian, respectively.

proposed scheme, both local and nonlocal methods exhibit the ability to deal with texture-less regions.

3.5.5 Spatial consistency analysis

As aforementioned, the nonlocal method has better performance local method in terms of object structure and fine details preservation. On the contrary, the local method which can be considered as a spatial consistency model could have better performance in spatial consistency. This experiment is to verify this prediction on the spatial consistency. Figure 3.11 presents an example of spatial consistency analysis between the local and nonlocal method. The profiles of the averaged depth values with respect to the marked regions are plotted in the bottom image. As shown, the profile marked red in top-left depth image derived using local matting Laplacian method shows less fluctuation compared with the blue one in the top-right depth image derived from the proposed method, which means that the local method exhibits better performance in maintaining spatial consistency compared to nonlocal method. In other words, the nonlocal method suffers from texture-copy artifacts much more serious than local method.

3.5.6 Parameter setting and analysis

Table 3.1: Parameter setting used in depth reconstruction process

Parameter	r	ϵ	C_s	t_{RM}	$coef_l$	K_{NN}
Value	1	10^{-5}	1/3	20	0.1	6

Empirical parameters used in the depth reconstruction process are summarized in Table 3.1. The analysis of these parameters are described as follows:

1. r corresponds to the window size for the focus measure computation using the SMLAP operator. As using a large window size could seriously blur

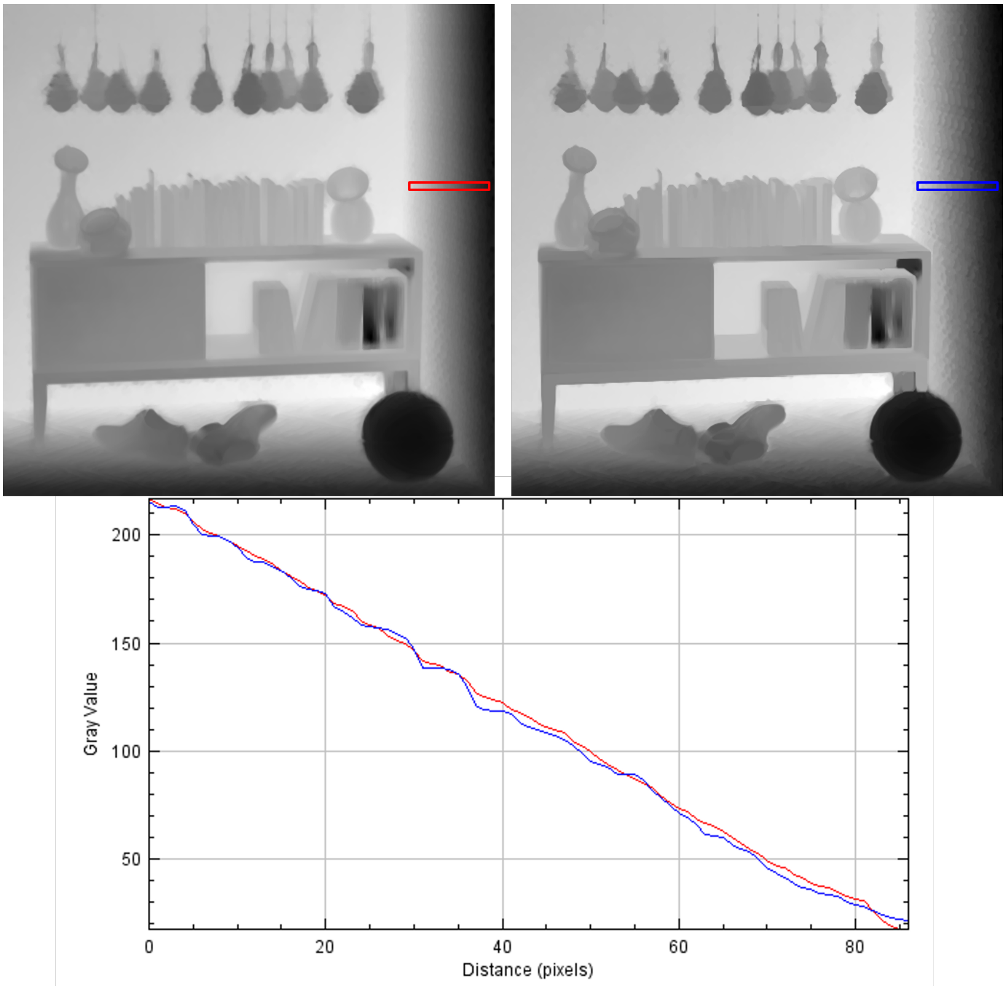

Figure 3.11: Spatial consistency analysis between reconstructed images using traditional local matting Laplacian and modified nonlocal matting Laplacian. The top-left depth image is the reconstructed depth image using traditional local matting Laplacian. The top-right depth image is the reconstructed depth image using proposed method. The bottom plot shows the depth profiles with respect to marked regions in above depth images.

depth edges, r is set to one for all synthetic datasets and two for the noisy real scene datasets.

2. ϵ is the weight of the regularization term to derive a numerically stable

solution. If ϵ is extremely small, our reconstructed depth would be sensitive to image noise. On the contrary, the depth edges could not be well preserved for a large ϵ.

3. C_s is the scale factor of the RGB values in our feature space. If C_s is too small, the reconstruction results would be similar to those using local method. A large C_s value would seriously break the spatial consistency, thus resulting in inaccurate depth estimation.

4. t_{RM} is the threshold of the reliability measure to determine the unreliable depth regions. If t_{RM} is too large, it may remove critical data and decrease the accuracy of depth reconstruction. By contrast, a small value of t_{RM} would generate noisy and inaccurate depth image.

5. $coef_l$ is a constant coefficient of confidence level λ. If $coef_l$ is too large, the reconstructed depth image would be similar to the initial depth image. If $coef_l$ is too small, more texture-copy artifacts would be introduced.

6. K_{NN} is the number of nearest neighbors in our defined feature space to construct the matting Laplacian matrix. For a small number of K_{NN}, there is insufficient information to recover the depth correctly. A large K_{NN} would produce a dense Laplacian matrix and thus introduce speed and memory problems. In addition, the nonlocal color smoothness assumption would not be valid for a large number of nearest neighbors. Note that, K_{NN} is set to nine for the noisy real scene datasets.

3.6 Summary

Texture-copy artifacts suffering and depth discontinuities blurring are two main issues in 3D depth reconstruction. As color image contains rich and useful information about object structures and details, a color-guided depth reconstruction method is proposed. As it is hard to avoid texture-copy artifacts using color-guided method, the purpose of this reconstruction stage changes to effectively preserve object structure and fine details with the ignorance of texture-copy artifacts. In this dissertation, a depth reconstruction method using matting Laplacian prior is presented. Experimental results show that the proposed depth reconstruction utilizing matting Laplacian as a prior can effectively recover depth over texture-less regions. Considering that the nonlocal principle has drawn significant attention owing to its excellent edge preservation property in image denoising and image matting, the nonlocal principle is adopted in the construction of matting Laplacian matrix to preserve object structure and fine details. Experimental results also demonstrate the effectiveness and superiority of the proposed nonlocal method compared to the local method.

Chapter 4

Closed-form MRF-based Depth Refinement

4.1 Overview

The proposed depth reconstruction algorithm can extract clear depth edges and preserve fine details owing to the adoption of the nonlocal principle in the matting Laplacian matrix construction, whereas the method using local matting Laplacian can provide more spatially consistent and less noisy depth image because the nonlocal principle breaks the spatial consistency. Additionally, our results suffer from the texture-copy artifacts caused by the edge inconsistency between the initial depth image and all-in-focus image. Figure 4.1 illustrates an example of texture-copy artifacts introduced in depth reconstruction process. As shown, the estimated depth images are noisy and the texture-copy artifacts are introduced in the smooth depth regions when the corresponding color image regions are highly textured.

In this chapter, we will describe an algorithm for depth image smoothing and texture-copy artifacts reduction. As edges are critical features in various depth-based applications such as object segmentation and measurement, an edge-preserving image denoising algorithm is required to smooth noise while preserving depth edges and fine details.

Over the decades, a number of researchers have dedicated their efforts in searching for efficient image denoising algorithms and various image denoising approaches have been developed. Traditional image denoising algorithms such as the mean filter [59], Gaussian filter [60] and Wiener filter [61] are the typically used linear local filters owing to their excellent properties in noise removal and fast computation. However, those linear filters cannot maintain sharp edges and preserve image details. In the last few years, many nonlinear filters have been proposed to resolve the issues above. Most of the popular nonlinear denoising methods are based on partial differential equations such as the anisotropic filter [32, 62, 63] that utilizes the anisotropic diffusion equation to denoise images. The nonlinear anisotropic filter is highly effective in smoothing noise while maintaining fine image details across sharp edges. However, it is implemented in an iterative process in which the iteration number is a critical parameter for the denoising performance of anisotropic diffusion methods. A bilateral filter [23, 64] is an alternative nonlinear filter to iterative filters. Unlike other local filters that consider only the pixels' geometric closeness, a bilateral filter enforces both geometric closeness in the spatial domain and gray level similarity in the denoising operation. He [65] proposed an efficient and

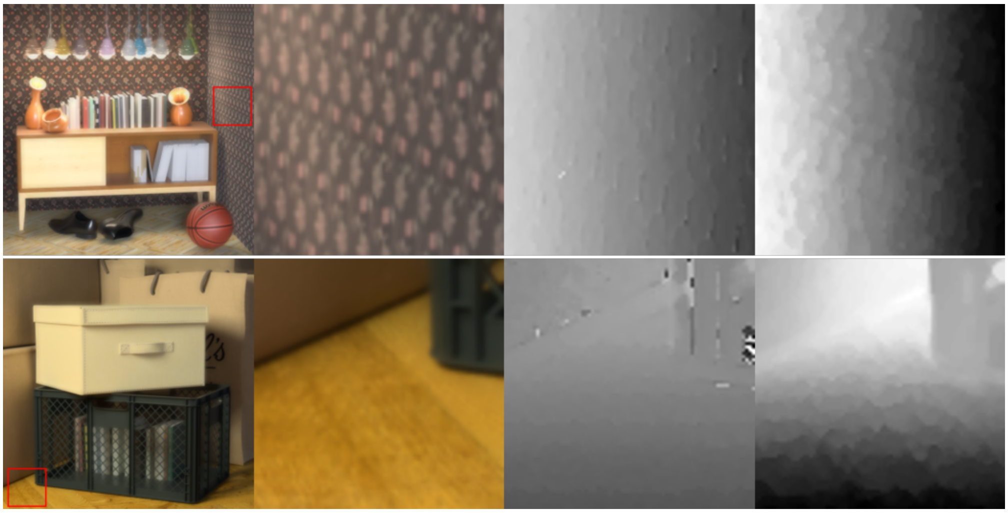

Figure 4.1: Illustration of texture-copy artifacts caused by structure inconsistency between depth image and all-in-focus image. From left to right: all-in-focus images, cropped images, initial depth images, and the texture-copy artifacts in our reconstructed depth images using nonlocal matting Laplacian, respectively.

effective edge-preserving image smoothing filtering-guided filter, that performs image denoising by considering the content of a guidance image. Shen et al. in [41] proposed a normalized cross correlation (NCC) based joint filtering using mutual structure, which utilizes the common structure between reference and target images to suppress noise while preserving the edges. In this chapter, a closed-form MAP-MRF edge-preserving algorithm is proposed to smooth the noisy depth images and reduce texture-copy artifacts.

4.2 Problem statement

Markov random fields were first introduced in computer vision in [66], and have proven to be useful for various computer vision problems such as image segmentation [67], image restoration [66], and stereo vision correspondence [68]. In the context of image denoising, we represent the components of MRFs with $D = [d_1, d_2, ..., d_n]^T$ and $F = [f_1, f_2, ..., f_n]^T$, where D and F denote the observed depth image with noise and the denoised depth image, respectively. n is the total number of pixels in the noisy depth image. Here d_i and f_i denote the depth values of pixel i, where $1 <= i <= n$. Similar with depth reconstruction, we formulate depth refinement as an optimization problem that maximizes the posterior probability $P(F|D)$. According to the Bayesian rule, we have

$$F^* = \arg\max_F P(D|F)P(F), \tag{4.1}$$

The first term $P(D|F)$ in (4.1) is a likelihood function of observing the data given a certain hidden state F and can be represented with the sensor noise model. Here, we assume that the noise can be modeled as additive Gaussian white noise and only neighboring pixels are statistically dependent. Subsequently, the negative logarithmic likelihood function can be expressed as

$$\begin{aligned} -\ln(P(D|F)) &\propto -\ln\Big(\prod_{i=1}^{n} \exp\left(-V_i\left(d_i, f_i\right)\right)\Big) \\ &= \sum_{i=1}^{n} V_i\left(d_i, f_i\right) \end{aligned} \tag{4.2}$$

where $\exp(-V_i(d_i, f_i)$ is the noise model and $V_i(d_i, f_i)$ is the data fidelity term that penalizes the inconsistency between the pixels of hidden depth image and observed depth image. Herein, we use squared-distance to define the data fidelity term which is defined as

$$V_i\left(d_i, f_i\right) = |f_i - d_i|^2. \tag{4.3}$$

According to the Hammersley-Clifford theorem [69] and our independent identical distribution assumption, the prior probability of an MRF can be factorized as the product of the summation over all cliques in the neighborhood system. Hence, we have

$$\begin{aligned} -\ln(P(F)) &\propto -\ln\Big(\prod_{i=1}^{n}\Big(\exp\Big(-\sum_{j\in N(i)} U_{ij}(f_i, f_j)\Big)\Big)\Big) \\ &= \sum_{i=1}^{n}\sum_{j\in N(i)} U_{ij}\left(f_i, f_j\right), \end{aligned} \tag{4.4}$$

where $N(i)$ is defined as the four-connected neighborhoods of the element i, and U_{ij} is the clique potential that is also known as a smoothness term to enforce the depth spatial consistency in our depth refinement process. Herein, we formulate our smoothness term in a quadratic form as it is a better prior for slanted surfaces and can facilitate in deriving our energy function with a closed-form solution based on using the Laplacian matrix. Here, we have

$$U_{ij} = w_{ij} \cdot |f_i - f_j|^2. \tag{4.5}$$

The variable w_{ij} is denoted as the affinity value between each neighborhood pair that is utilized to control the degree of smoothness based on local statistical information. Herein, w_{ij} is defined as

$$\exp(-\frac{|d_i - d_j|^2}{2{\sigma_1}^2}) \exp(-\frac{|i-j|^2}{2{\sigma_2}^2}) \exp(-\frac{\sum\limits_{c \in C} |\mathit{AIF}^c_i - \mathit{AIF}^c_j|^2}{|C| \times 2{\sigma_3}^2}), \tag{4.6}$$

where σ_1, σ_2, and σ_3 are three user defined constant to balance the contribution of w_{ij} to U_{ij}. $C = \{R, G, B\}$ represents different channels of the all-in-focus image and $|C|$ is the number of color channels. Our weight w_{ij} consists of three terms: depth range filter, spatial filter and color range filter. Qualitatively, the depth term would give a large value if d_i is close to d_j. This term is designed to avoid incorrect depth prediction owing to the structure inconsistency between input depth image and all-in-focus image. The effect of the spatial term on the smoothness penalty U_{ij} would be decreased as the distance between pixels i and j increases. The color term is designed to make use of consistent edges in both input depth image and all-in-focus image, and it would give a large value as the color of AIF_i and AIF_j is similar. From (4.1), (4.2), and (4.4), our MAP-MRF based depth refinement is equivalent to minimizing the following energy function:

$$E(F) = \sum_{i=1}^{n} \Big(V_i\left(d_i, f_i\right) + \sum_{j \in N(i)} U_{ij}\left(f_i, f_j\right) \Big). \tag{4.7}$$

4.3 Closed-form solution

With the definitions of the data term and smoothness term in (4.3) and (4.5), respectively, our final energy function (4.7) can be rewritten as

$$
\begin{aligned}
E(F) &= \sum_{i=1}^{n} \Big(\tau_i V_i(d_i, f_i) + \sum_{j \in N(i)} U_{ij}(f_i, f_j) \Big) \\
&= \sum_{i=1}^{n} \Big(\tau_i \cdot |d_i - f_i|^2 + \sum_{j \in N(i)} w_{ij} \cdot |f_i - f_j|^2 \Big).
\end{aligned}
\tag{4.8}
$$

In the final energy function, a confidence level τ is adopted to balance the importance between the data and smoothness term, which is defined as

$$
\tau_i = \mathit{coef}_m \cdot \exp\left(-\sigma_D(i)\right), \tag{4.9}
$$

where coef_m is a user defined coefficient of confidence level τ; σ_D is the normalized variance of the reconstructed depth image. Herein, τ is set inversely proportional to the depth variance value. Large τ values are assigned for the pixels with small variances to maintain the spatial consistency; and small τ values for the noisy pixels.

With the definition of the smoothness term in (4.8), our proposed MRF is isotropic and we define an undirected weighted graph $G = (V, E)$, in which the vertices V represent the depth image pixels and the edge E is a set of weighted edges representing the affinities between the corresponding depth image pixels.

Subsequently, the adjacent matrix of G is W, whose elements are defined as

$$W_{ij} = \begin{cases} w_{ij} & i \neq j, j \in N(i), \\ 0 & i \neq j, j \notin N(i), \\ C_w & i = j, \end{cases} \tag{4.10}$$

where C_w is a positive and small constant indicating the constant edge between vertex i and itself. Let Dia be an $n \times n$ diagonal matrix with the entry $Dia_{ii} = \sum_{j=1}^{n} W_{ij}$. If vertex i is isolated in the graph, Dia_{ii} becomes zero, thereby resulting in a singularity in the adjacent matrix W. Using constant C_w can help avoid the singularity problem and achieve an accurate and numerically stable solution. With the construction of our energy function $E(F)$ and the corresponding undirected weighted graph G, our proposed MAP-MRF model yields a closed form solution according to the following theorem.

Theorem 1. *Let $D = [d_1, d_2, \ldots, d_n]^T$ and $F = [f_1, f_2, \ldots, f_n]^T$. Therefore, the $Dia^{-1/2}(T_M Dia^{-1} + 2\overline{L_M})^{-1} T_M Dia^{-1/2} D$ is the closed form solution for the following energy function:*

$$E(F) = \sum_{i=1}^{n} \Big(\tau_i \cdot |d_i - f_i|^2 + \sum_{j \in N(i)} W_{ij} \cdot |f_i - f_j|^2 \Big), \tag{4.11}$$

where T_M is a $n \times n$ diagonal matrix, in which the element $T_M(i,i)$ equals τ_i, and $\overline{L_M}$ is the normalized Laplacian matrix of the undirected weighted graph G constructed above, which can be expressed as

$$\overline{L_M} = Dia^{-1/2}(Dia - W)Dia^{-1/2}. \tag{4.12}$$

Proof. Let $R = [r_1, r_2, ..., r_n]^T$, where $r_i = \sqrt{(Dia_{ii})} \cdot f_i$, $1 \leq i \leq n$, for a set of medium variables, we have

$$F = Dia^{-1/2}R. \tag{4.13}$$

Subsequently, the energy function (4.11) can be rewritten as

$$E(F) = \overline{E}(R) = \sum_{i=1}^{n} \tau_i \cdot \left| \frac{r_i}{\sqrt{Dia_{ii}}} - d_i \right|^2 + \sum_{i,j=1}^{n} W_{ij} \cdot \left| \frac{r_i}{\sqrt{Dia_{ii}}} - \frac{r_j}{\sqrt{Dia_{jj}}} \right|^2. \tag{4.14}$$

A compact matrix form can be expressed as

$$\overline{E}(R) = (Dia^{-1/2}R - D)^T T_M (Dia^{-1/2}R - D) + 2(R^T \overline{L_M} R), \tag{4.15}$$

where

$$(Dia^{-1/2}R - D)^T T_M (Dia^{-1/2}R - D) = \sum_{i=1}^{n} \tau_i \cdot \left| \frac{r_i}{\sqrt{Dia_{ii}}} - d_i \right|^2, \tag{4.16}$$

and

$$2(R^T \overline{L_M} R) = \sum_{i,j=1}^{n} W_{ij} \cdot \left| \frac{r_i}{\sqrt{Dia_{ii}}} - \frac{r_j}{\sqrt{Dia_{jj}}} \right|^2. \tag{4.17}$$

To minimize $\overline{E}(R)$, we set the first derivative with respect to R to zero, which yields

$$\frac{\partial \overline{E}(R)}{\partial R} = 2T_M Dia^{-1/2}(Dia^{-1/2}R - D) + 4\overline{L_M}R = 0$$

$$\Rightarrow (T_M Dia^{-1} + 2\overline{L_M})R = T_M Dia^{-1/2}D.$$

It is noteworthy that $(T_M Dia^{-1} + 2\overline{L_M})$ is positive semi-definite because $\overline{L_M}$ is positive semi-definite and Dia is a diagonal matrix. The closed form solution with respect to the medium variable R can be derived as

$$R = (T_M Dia^{-1} + 2\overline{L_M})^{-1} T_M Dia^{-1/2} D. \tag{4.18}$$

Thus, the optimal solution with respect to F is

$$F = Dia^{-1/2} R = Dia^{-1/2} (T_M Dia^{-1} + 2\overline{L_M})^{-1} T_M Dia^{-1/2} D. \tag{4.19}$$

□

4.4 Edge preservation

As our Gaussian MRFs tends to over-smooth depth image and blur real depth edges, additional operations on these edges are needed. By exploiting the fact that real depth edges often coincide with color edges, we first aim to find the common edges in both initial depth image and all-in-focus image, and then subsequently increase confidence level τ values for these edges. Since the edges in initial depth image are not perfectly consistent with color edges owing to the noise and edge bleeding problem present in initial depth image, the common edges herein are detected as the color edges located in dilated edge regions of initial depth image. The dilation operation is further employed to enhance the these edges.

Figure 4.2 illustrates some examples of our detected common edges. By

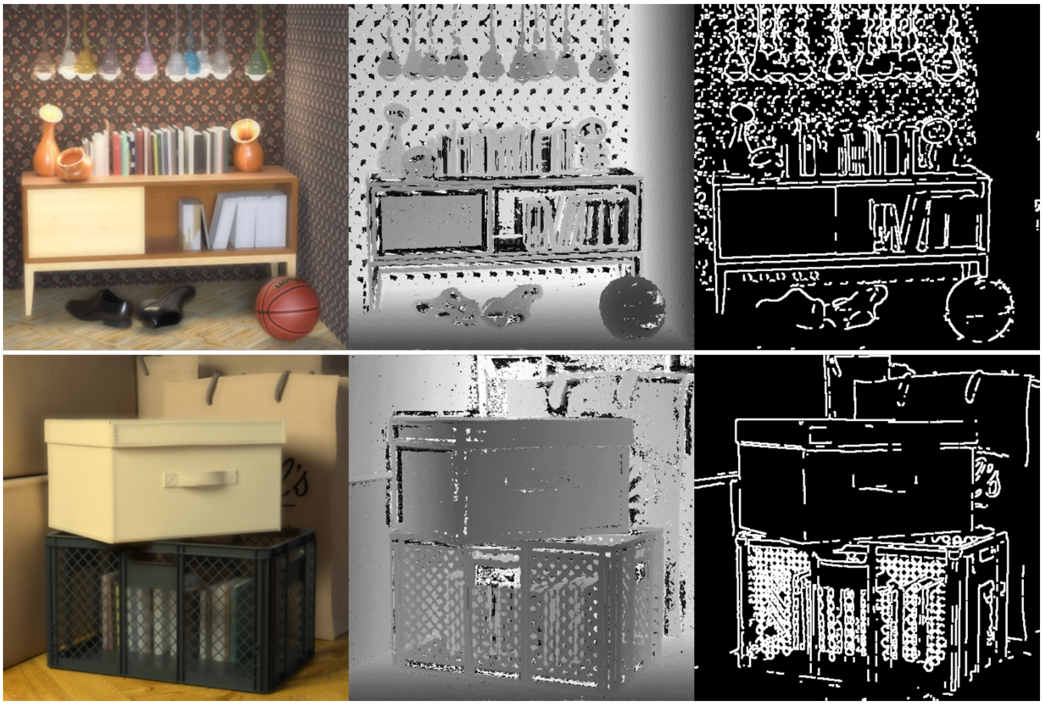

Figure 4.2: Common edges between all-in-focus image and initial depth image. From left to right: all-in-focus images, initial depth images, and common edges, respectively.

taking the pre-estimated edges information into consideration, the real depth edges be well preserved.

4.5 Texture-copy artifacts suppression

As our algorithm is designed for edge-preserving image denoising, it is still difficult to suppress texture-copy artifacts especially in the artifacts regions with strong edges. To suppress these artifacts, our strategy is to detect texture-copy artifacts and subsequently increase the effect of the smoothness penalty function U_{ij} over the detected artifacts regions. Typically, the texture-copy artifacts

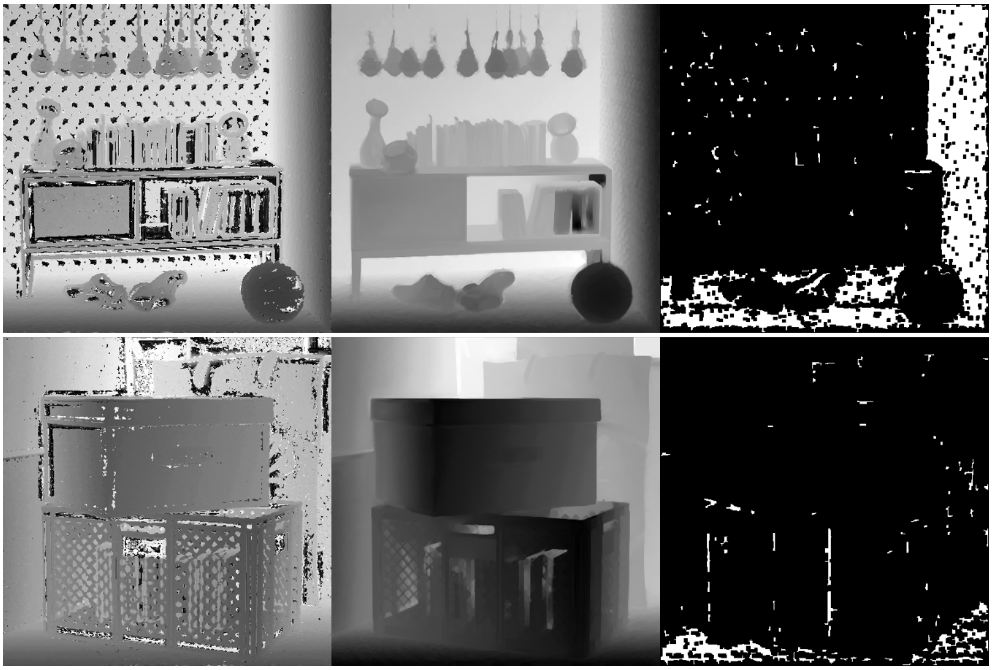

Figure 4.3: Texture-copy artifacts detection using mutual structure between initial depth image and reconstructed depth image. From left to right: initial depth image, reconstructed depth image, and detected texture-copy artifacts, respectively.

appear in the smooth depth regions when the corresponding all-in-focus image regions are highly textured, thus implying that the structure of texture-copy regions in initial depth image $\tilde{D}$ and our reconstructed depth image D are inconsistent. In this dissertation, we utilize the mutual structure proposed in [41] to detect texture-copy artifacts. The structure similarity measure between corresponding patches in $\tilde{D}$ and D is defined as

$$S(\tilde{D}_p, D_p) = \left(\sigma(\tilde{D}_p)^2 + \sigma(D_p)^2\right)\left(1 - \rho(\tilde{D}_p, D_p)^2\right)^2, \tag{4.20}$$

where $\sigma(\tilde{D}_p)$ and $\sigma(D_p)$ denote the variances of the patch depth values;

$\rho(\tilde{D}_p, D_p)$ is the NCC of the corresponding patch in the initial depth image $\tilde{D}$ and reconstructed depth image D, which is defined as

$$\rho(\tilde{D}_p, D_p) = \frac{cov(\tilde{D}_p, D_p)}{\sqrt{\sigma(\tilde{D}_p) + \sigma(D_p)}}, \tag{4.21}$$

where $cov(\tilde{D}_p, D_p)$ is the covariance of the patch depth values. When two patches contain same edges, $|\rho(\tilde{D}_p, D_p)| = 1$. Otherwise, $|\rho(\tilde{D}_p, D_p)|$ is small when patch structures are different. In texture-copy regions where the edges appear in the reconstructed depth image but not in the initial depth image, $\sigma(D_p)$ is large and $\sigma(\tilde{D}_p)$ is small. $S(\tilde{D}_p, D_p)$ therefore outputs a relatively large number. On the other hand, when common edges appear in two patches or when both patches do not contain any significant edges, $S(\tilde{D}_p, D_p)$ would be a small value. In the initial depth image, the reliable regions with a large structure similarity measure and small variance will be detected as the texture-copy regions, which can be expressed as

$$TC_i = \begin{cases} 1 & S(i) \geq t_{ssm} \,\&\, \sigma_{\tilde{D}}(i) \leq t_{var} \,\&\, RM_i \leq t_{RM}, \\ 0 & others, \end{cases} \tag{4.22}$$

where t_{ssm} is the threshold to the structure similarity measure $S(\tilde{D}, D)$. The definitions of t_{var} and t_{RM} are the same as those used in depth reconstruction process. The morphological operations are further employed to enhance the extracted texture-copy regions. Figure 4.3 illustrates some examples of our detected texture-copy regions. with this information, we can effectively suppress the texture-copy artifacts by increasing the effect of the smoothness penalty function U_{ij}, through increasing the σ_1 used in depth kernel construction for all of the marked pixels in TC.

4.6 Experimental results

Our depth refinement algorithm is designed for edge-preserving depth image denoising and texture-copy artifacts suppression. In this section, we evaluate the performance of the proposed depth refinement algorithm after the depth image is reconstructed in Chapter 3. To demonstrate the performance of the proposed algorithm, five edge-preserving smoothing methods are compared: fast bilateral solver (FBS) [70], anisotropic diffusion (AD) [32], nonlocal means filter (NLM) [34], mutual structure for joint filtering (MS) [41] and robust color guided filtering (RCG) [71].

Figure 4.4 gives some examples of depth refinement results from datasets[56]. As shown, the proposed method can effectively preserve the edges and details while maintaining spatial consistency. Figs 4.5 and 4.6 present our refined depth images and results from five edge-preserving smoothing algorithms. FBS is fast and can preserve object details, but the false edges are enhanced due to the structure inconsistency between color image and depth image. AD method has good performance in noise suppression while preserving depth edges; however, it is difficult to obtain satisfactory result in the texture-copy artifacts regions with strong edges. The MS and NLM methods have better performance in noise removing and texture-copy artifacts suppression compared with the FBS, AD,

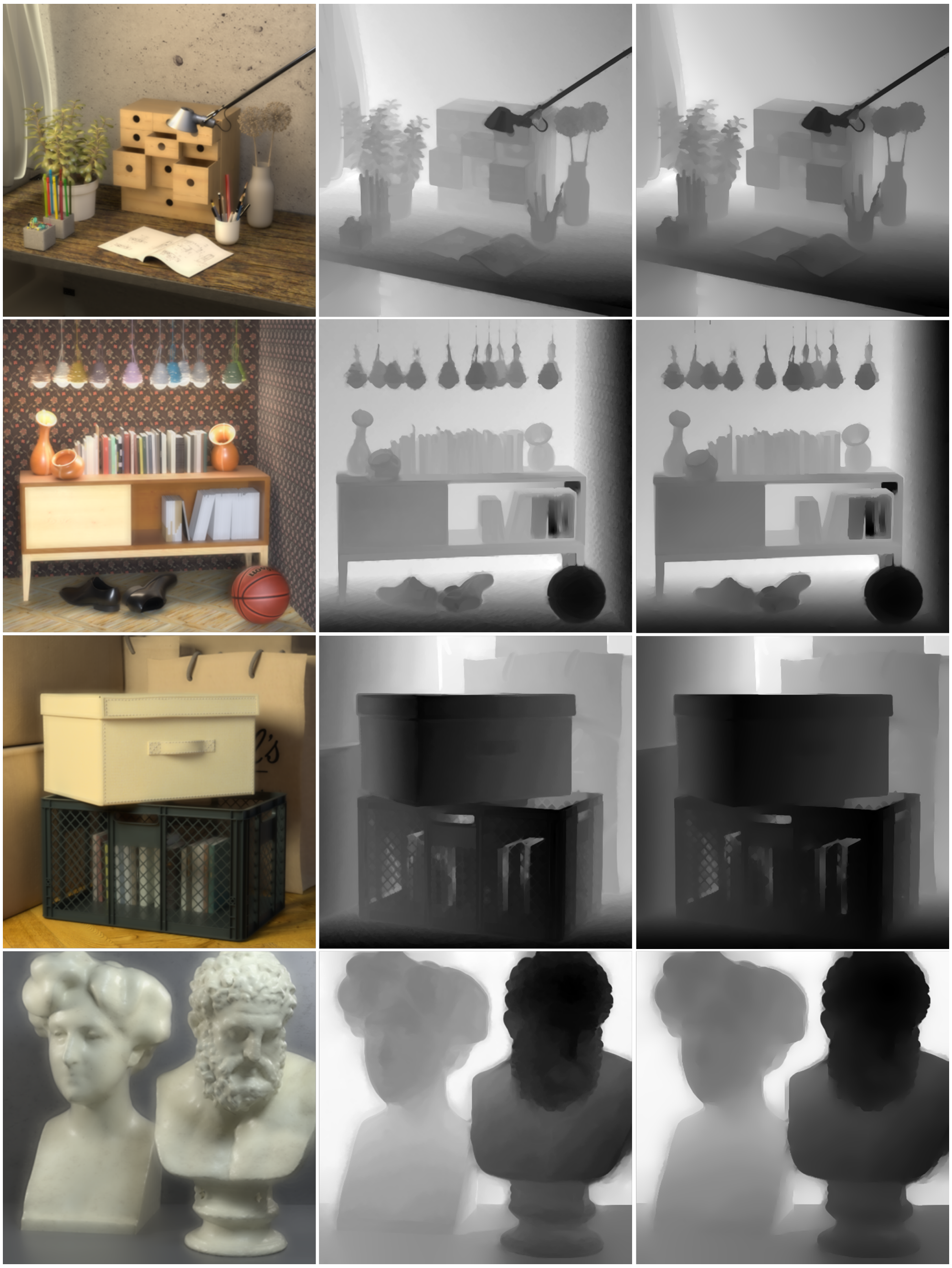

Figure 4.4: Refinement results using proposed edge-preserving method. The left column images are the all-in-focus images. The middle column images are reconstructed depth images using the proposed reconstruction method. Images on the right column are refinement results.

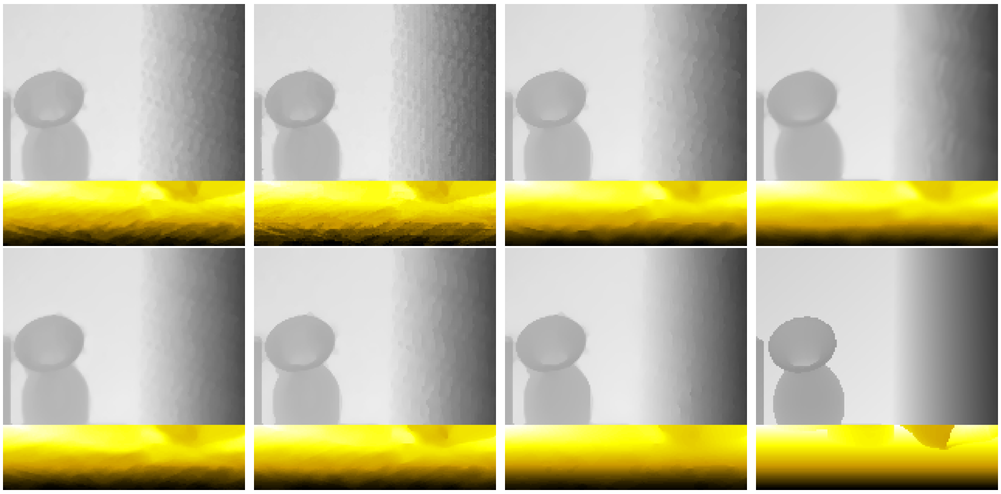

Figure 4.5: Comparison of depth refinement results using various edge-preserving image denoising algorithms. From top left to bottom right: input noisy depth image, and the results by FBS, AD, NLM, MS, RCG, our algorithm, and ground truth, respectively. Bottom parts are color-mapped to clearly show the better performance of our refinement algorithm.

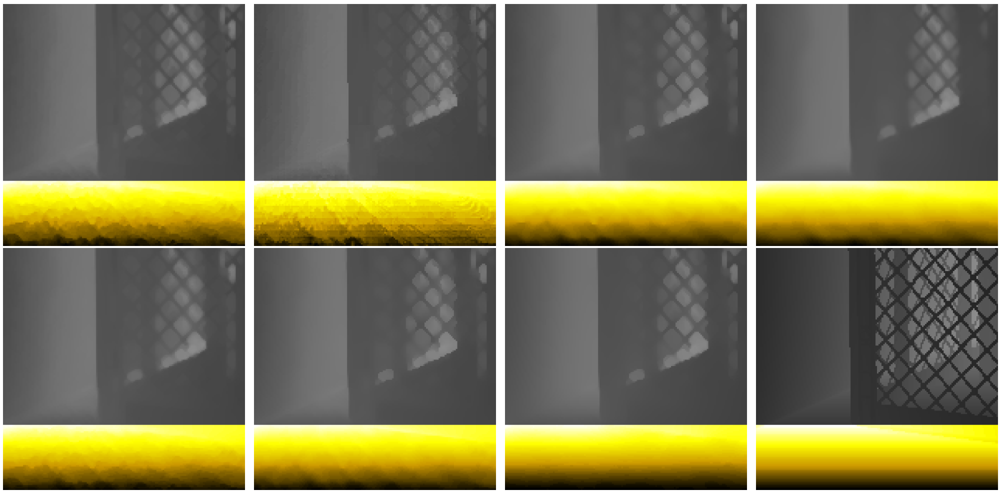

Figure 4.6: Comparison of depth refinement results using various edge-preserving image denoising algorithms. From top left to bottom right: input noisy depth image, and the results by FBS, AD, NLM, MS, RCG, our algorithm, and ground truth, respectively. Bottom parts are color-mapped to clearly show the better performance of our refinement algorithm.

and RCG methods, but at the expense of slightly blurring sharp edges. The RCG method exhibits a comparable performance in preserving edges and better performance in texture-copy artifacts suppression compared with FBS and AD methods. As shown, with the incorporation of edges and mutual structure information into our formulation, our method achieves the competitive and better performance in edge preservation and texture-copy artifacts suppression compared to these methods.

Table 4.1: Parameter setting used in depth refinement process

Parameter	σ_1	σ_2	σ_3	t_{ssm}	$coef_m$
Value	1	0.35	0.1	10^{-3}	0.1

Table 4.1 shows the empirical parameter setting used in the depth refinement process. The analysis of those parameters are described as follows:

1. σ_1 is the parameter controlling the fall-off weight in depth domain. If σ_1 is too large, the tolerance for two different depth values to be considered close enough would be large too, thus implying that the depth edges would be smoothed.

2. σ_2 and σ_3 are the parameters controlling the decay rate of the spatial and color range filter, and adjust the importance of spatial difference and intensity difference, respectively. Increasing σ_2 results in large features being degraded. Too small a value of σ_3 fails to suppress noise, while too large a value would result in blurring depth discontinuities.

3. t_{ssm} is the threshold of structure similarity measure between initial depth image and reconstructed depth image. If t_{ssm} is too small, depth regions

would be mis-detected as texture-copy artifacts, thus resulting in over-smoothing. On the contrary, texture-copy artifacts could not be well suppressed if t_{ssm} is too large.

4. $coef_m$ is a constant coefficient of confidence level τ of the input noisy depth image. As $coef_m$ increases, the effect of τ on data term becomes stronger, and thus the refined depth image would be more close to the input noisy depth image.

4.7 Summary

Depth refinement acting as a post processing is commonly used in depth recovery tasks. Different with the purpose of general depth refinement methods that only aim at edge-preserving smoothing, the purpose of depth refinement in this dissertation needs to take texture-copy artifacts suppression into consideration. The traditional edge-preserving filters such as bilateral filter, anisotropic fusion and nonlocal means filters are the most methods to suppress noise while preserving edges. However, it is difficult for these filters to suppress texture-copy artifacts at the same time. Even worse, the texture-copy artifacts could be enhanced when the artifacts have large edge gradients. To suppress texture-copy artifacts, enhance spatial smoothness while preserving edges, a MAP-MRF based edge-preserving depth refinement algorithm is proposed.

As the proposed method is based on the Gaussian MRF model, it would over-smooth depth discontinuities if no additional operations are taken. Based on the fact that real depth edges are generally coincide with color edges, a specially designed smoothness weight containing the information of common edges is proposed to preserve truth depth edges. Similar with the idea to preserve real depth edges, the strategy of suppressing texture-copy artifacts is firstly to find texture-copy artifacts regions, and then improve the important ratio of smoothness term in the proposed energy function over the detected artifacts regions. Typically, the texture-copy artifacts generally appear in the regions with inconsistent structures between initial depth image and color image. Considering that the mutual structure based method can effectively find mutual structures between images, in this dissertation, mutual structure based method is inversely utilized to detect such texture-copy artifacts regions. Additionally, the proposed method can obtain an global optimum by utilizing the Laplacian matrix based on the undirected weighted graph representing the energy function.

Chapter 5

Evaluation

5.1 Overview

To validate the effectiveness and superiority of the proposed depth recovery method herein, experiments over synthetic and real scene datasets are conducted. The synthetic image sequences used in our experiments are the same with those used in depth reconstruction process. To evaluate our proposed algorithm on real scene image sequences, we test the DDFF 12-Scene datasets [14] for the quantitative comparison. Additionally, more experiments on real scene image sequences from [72] and [3] are conducted for the qualitative comparison. The real scene images are downsampled by a factor of two before our depth reconstruction to reduce the amount of data. the proposed method is qualitatively and quantitatively compared with the related approaches.

5.2 Evaluation metrics

The comparison results of depth recovery are evaluated using mean square error (MSE) and the structural similarity index measure (SSIM) [73]. The MSE is used as a integral error criteria between the recovered depth image and ground truth, which is defined as:

$$MSE(F,G) = \frac{1}{n}\sum_{i=1}^{n}(f_i - g_i)^2, \tag{5.1}$$

where f_i and g_i are the recovered pixel depth value ground truth value at position i. n is the pixel number of the all-in-focus image.

The SSIM is used as a metric to measure structural accuracy between recovered depth image and ground truth, which is defined as:

$$SSIM\left(F,G\right) = L\left(F,G\right)C\left(F,G\right)S\left(F,G\right), \tag{5.2}$$

where

$$\begin{cases} L\left(F,G\right) = \frac{2\mu_F\mu_G + C_1}{{\mu_F}^2 + {\mu_G}^2 + C_1} \\ C\left(F,G\right) = \frac{2\sigma_F\sigma_G + C_2}{{\sigma_F}^2 + {\sigma_G}^2 + C_2} \\ S\left(F,G\right) = \frac{\sigma_{FG} + C_3}{\sigma_F\sigma_G + C_3} \end{cases} . \tag{5.3}$$

The first term $L(F,G)$ in 5.3 is the luminance comparison function measuring the difference between the recovered depth image F and ground truth G, in which μ represents the images' mean luminance. The second term $C(F,G)$ is the contrast comparison function measuring contrast difference between F and G, in which σ denotes the the images' standard deviation. The third term $S(F,G)$ is the structure comparison function measuring the correlation coefficient between F and G, in which σ_{FG} denotes the covariance between F and G.

5.3 Evaluation on synthetic datasets

In this section, the evaluation of the proposed method over synthetic datasetss is firstly presented. To do this, the comparison between the proposed approach against the previous state-of-the-art DFF reconstruction algorithms such as those by Tseng [12], Moeller [3], Javidnia [13], and Hazibras [14] is conducted. Note that the pretrained DDFFNet-CC3 neural network provided by authors is utilized for disparity image predictions, whose parameter settings are recommended and reported in [14].

Figure 5.1 illustrates some depth reconstruction results from the synthetic datasets [56]. As shown, the proposed method is robust, and performs best in preserving sharp depth edges and fine details while maintaining spatial consistency. Tseng's method can present more details compared with other methods, but suffers seriously from the texture-copy artifacts. This is because that Tseng's method over relies on the color information owing to its inaccurate entropy based confidence level calculations. By contrast, our variance based confidence level calculation considers both the independent pixel and its local spatial consistency, and this makes our reconstruction method more robust over texture-less regions. Another issue in Tseng's method is the edge blurring problem. It is because the matting Laplacian matrix is constructed only in spatial domain. In our method, the nonlocal principle is adopted during the construc-

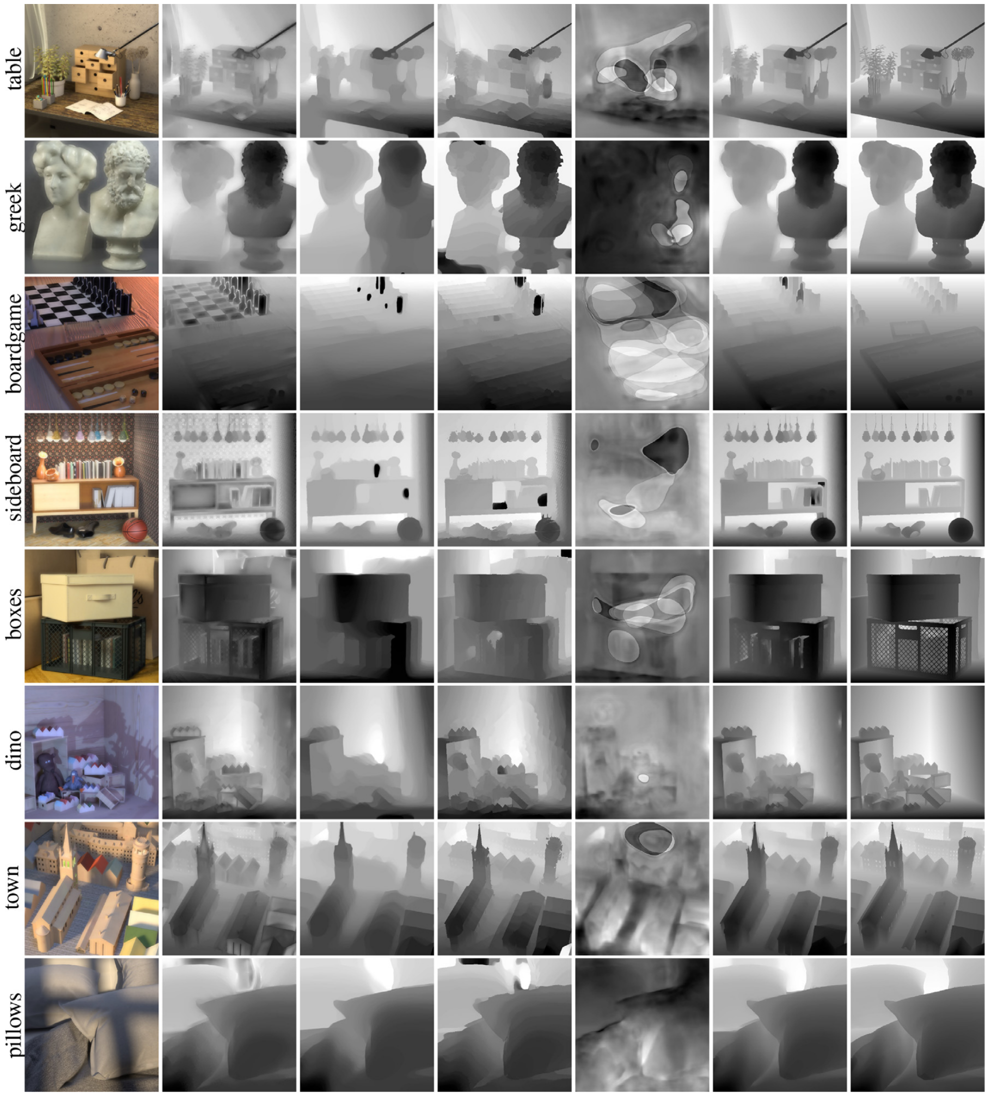

Figure 5.1: Comparison results of recovered depth images. From left to right: all-in-focus images, the results by Tseng [12], Moeller [3], Javidnia [13], our algorithm, and ground truth, respectively.

Table 5.1: Quantitative comparison of depth reconstruction for synthetic datasets [56] in MSE.

Method	table	greek	b.g.	s.b.	boxes	dino	town	pillows
Nayar[2]	0.120	0.367	0.139	1.537	0.021	0.104	6.852	0.036
Tseng[12]	0.055	0.123	0.101	0.395	0.004	0.012	1.210	0.013
Moeller[3]	0.087	0.186	0.027	0.333	0.005	0.013	0.417	0.006
Javidnia[13]	0.049	0.113	0.015	0.321	0.009	0.015	0.949	0.018
Hazirbas[14]	0.156	1.528	0.406	4.008	0.016	0.133	13.815	0.188
Ours	**0.048**	**0.057**	**0.007**	**0.184**	**0.004**	**0.006**	**0.337**	**0.002**

Table 5.2: Quantitative comparison of depth reconstruction for synthetic datasets [56] in SSIM.

Method	table	greek	b.g.	s.b.	boxes	dino	town	pillows
Nayar[2]	0.539	0.458	0.373	0.433	0.457	0.537	0.399	0.841
Tseng[12]	0.852	0.874	0.813	0.760	0.732	0.914	0.876	0.938
Moeller[3]	0.822	0.810	0.879	0.816	0.704	0.901	0.894	0.941
Javidnia[13]	0.835	0.866	0.923	0.804	0.627	0.886	0.893	0.921
Hazirbas[14]	0.694	0.440	0.529	0.596	0.550	0.800	0.601	0.540
Ours	**0.890**	**0.913**	**0.964**	**0.886**	**0.774**	**0.945**	**0.904**	**0.973**

tion stage of matting Laplacian matrix, and thus can obtain better clustering especially when consistent edges exist in both depth image and all-in-focus image. Moeller's method can yield good results in terms of spatial consistency. However, as it uses a large kernel to generate initial depth image and no additional information such as color information utilized in the optimization framework, numerous fine details are lost and depth edges are blurred. Even though Javidnia's method can preserve high structural accuracy for some cases, it is less robust to texture-less regions, and suffers much more from texture-copy artifacts compared to the proposed method. Additionally, to improve the robustness, all images are downscaled by a factor of three before applying the PADMM, and thus numerous fine details would be lost for the images with low

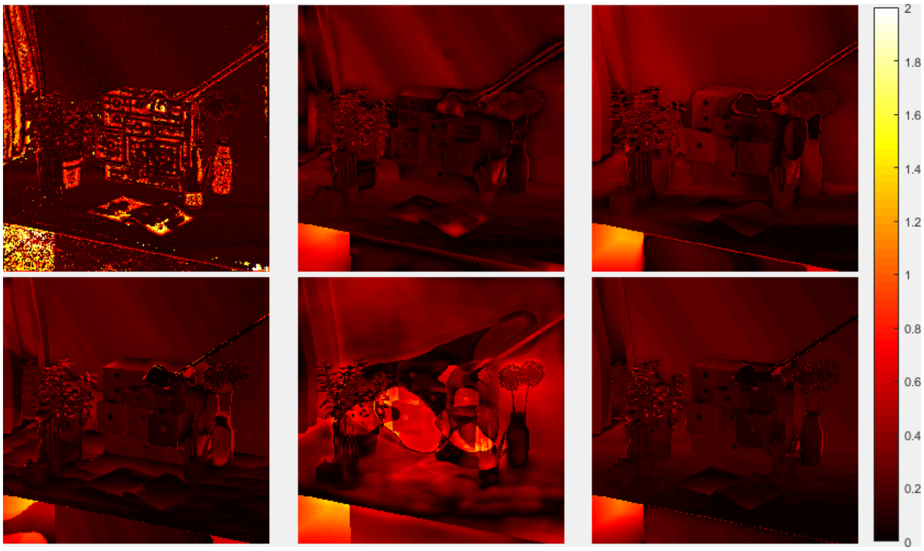

Figure 5.2: Absolute difference comparison between recovered depth images derived from the-state-of-the art methods and the proposed method. From top left to bottom-right: absolute difference between ground truth and depth images using methods from Nayar[2], Tseng [12], Moeller [3], Javidina [13], Hazirbas [14], and ours, respectively. Brighter pixel intensities indicate larger differences.

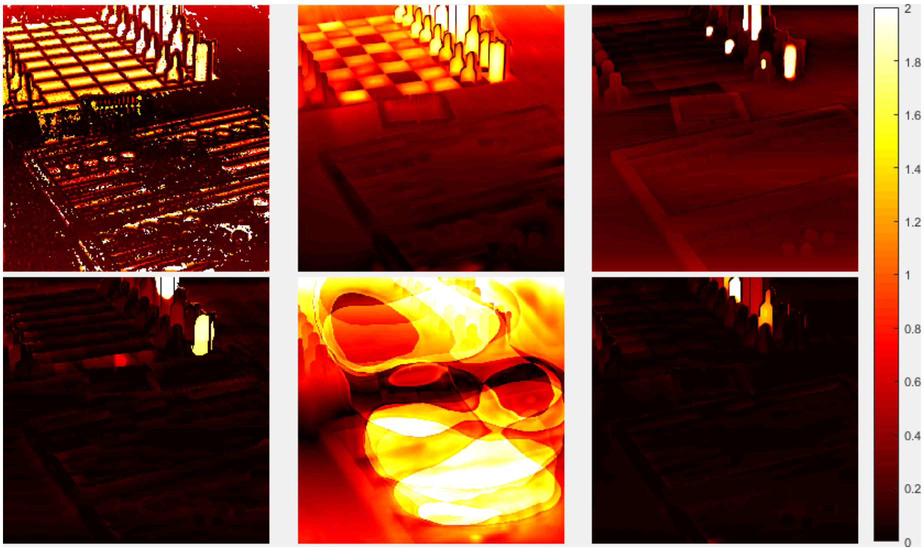

Figure 5.3: Absolute difference comparison between recovered depth images derived from the-state-of-the art methods and the proposed method. From top left to bottom-right: absolute difference between ground truth and depth images using methods from Nayar[2], Tseng [12], Moeller [3], Javidina [13], Hazirbas [14], and ours, respectively. Brighter pixel intensities indicate larger differences.

spatial resolution. Hazirbas's method fails to predict correct depth values, and can not preserve edges and structural details.

The quantitative comparisons between the proposed method and these related approaches in terms of the mean square error (MSE) and the structural similarity index measure (SSIM) [73] are shown in Tables 5.1 and 5.2, respectively. Table 5.1 presents the comparison of the overall error of the recovered depth image, in which the lower is better. As shown, the proposed method has the minimum difference in reference depth image compared to the related methods. Table 5.2 gives the evaluation of structure similarity between the recovered depth images and ground truth, in which the higher is better. As shown, the proposed method exhibits the highest structure similarity to the ground truth. To intuitively demonstrate the performance of the proposed method, a comparison of absolute errors between ground truth and recovered images using related methods are given in Figs 5.2 and 5.3. The darker intensities show lower difference between ground truth. As shown, the proposed method exhibits better performance over texture-less regions, and the recovered depth images are spatially smoother. Additionally, the recovered depth image derived from the proposed method show less errors in real depth edges. It is clear that the proposed method performs better than these state-of-the-art approaches in terms of overall accuracy of depth recovery and the ability to effectively recover the object structures.

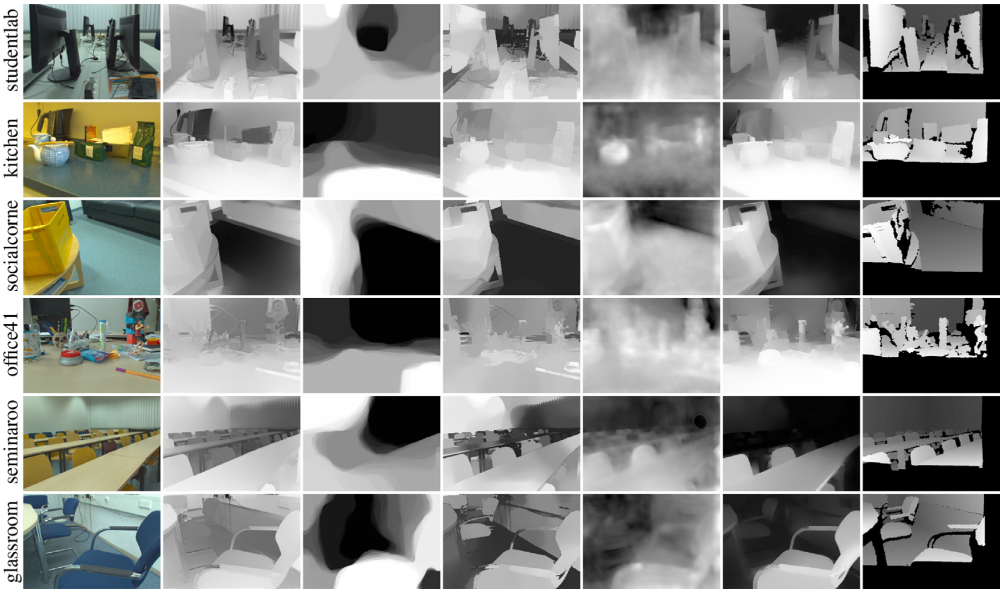

Figure 5.4: Comparison of recovered disparity maps. From left to right: all-in-focus image, the results by Tseng [12], Moeller [3], Javidina [13], Hazirbas [14], ours and ground truth, respectively. Brighter pixel intensities indicate closer distances.

5.4 Evaluation on real scene datasets

To demonstrate the effectiveness of our proposed method, additional experiments on real scene datasets are further conducted. The comparison here is performed on disparity images, which can be achieved by mapping the depth values from traditional optimization methods to their corresponding disparity values.

Figure 5.4 illustrates some results from DDFF 12-Scene datasets [14]. As shown, the proposed method is robust over texture-less regions, and can preserve edges and fine details while maintaining spatial consistency. Moreover, our approach can effectively suppress texture-copy artifacts compared to Tseng's

Table 5.3: Quantitative results of the proposed method for DDFF 12-Scene datasets [14] in MSE. Metrics are computed on the recovered and ground truth disparity maps.

Method	studentlab	kitchen	socialcorner	office41	seminaroom	glassroom
Tseng[12]	5.764e-3	2.293e-3	3.505e-3	5.669e-3	1.743e-3	1.206e-3
Moeller[3]	2.477e-3	2.205e-3	2.967e-3	6.143e-3	2.341e-3	1.177e-3
Javidnia[13]	2.107e-3	2.024e-3	3.028e-3	4.915e-3	2.243e-3	1.227e-3
Hazirbas[14]	1.416e-3	1.828e-3	2.865e-3	3.991e-3	0.630e-3	0.997e-3
Ours	**0.947e-3**	**0.990e-3**	**1.063e-3**	**3.096e-3**	**0.504e-3**	**0.675e-3**

Table 5.4: Quantitative results of the proposed method for DDFF 12-Scene datasets [14] in SSIM. Metrics are computed on the recovered and ground truth disparity maps.

Method	studentlab	kitchen	socialcorner	office41	seminaroom	glassroom
Tseng[12]	0.790	0.925	0.837	0.834	0.787	0.911
Moeller[3]	0.883	0.945	0.792	0.845	0.659	0.898
Javidnia[13]	0.930	0.946	0.897	**0.870**	0.894	0.923
Hazirbas[14]	0.849	0.938	0.785	0.824	0.699	0.879
Ours	**0.933**	**0.957**	**0.908**	0.861	**0.904**	**0.937**

and Javidnia's methods. In comparison, Hazibras's method cannot effectively preserve object structure, even though it can alleviate texture-copy artifacts. The corresponding quantitative comparisons in terms of MSE and SSIM are reported in tables 5.3 and 5.4, respectively. Table 5.3 shows the comparison of the integral error of the recovered depth image, in which the lower is better. As shown, the proposed method achieves the lowest number compared to the previous state-of-the-art approaches. Additionally, as shown in Table 5.4, the proposed method also achieves the best performance in structure similarity measure compared to other related methods. The comparison results demonstrate that the proposed method outperforms the previous state-of-the-art algorithms in terms of robustness and accuracy on real scene datasets for most cases.

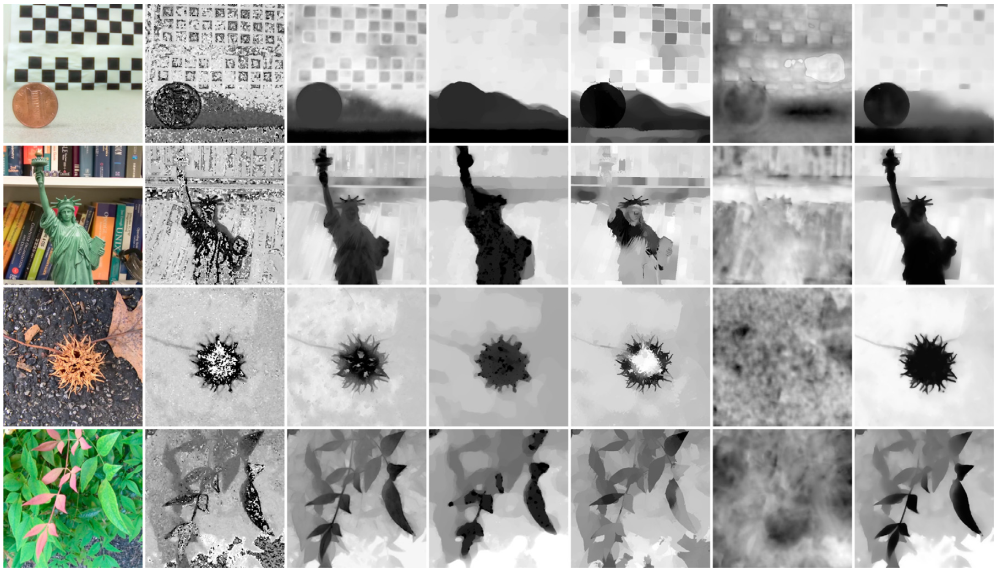

Figure 5.5: Comparison of recovered depth image. From top left to bottom right: all-in-focus image, the results by Nayar [2], Tseng [12], Moeller [3], Javidina [13], and ours, respectively.

The qualitative comparisons with the datasets from [72] and [3] are further carried out to comprehensively evaluate the performance of the proposed algorithm. It is noteworthy that the recovered depth here is relative depth among objects rather than the physical distance of the object from the camera. The experimental results of above datasets are shown in Figs. 5.5, 5.6, and 5.7. As there are only ten images in each datasets [72] used in the experiments, it is quite difficult for the related methods to derive satisfactory depth images. Experimental results shown in Figure 5.5 demonstrate the effectiveness and superiority of the proposed method even for a focal stack with very few images. Experimental results shown in Figs 5.6, and 5.7 clearly demonstrate the capability of the proposed method to recover accurate depth images under both

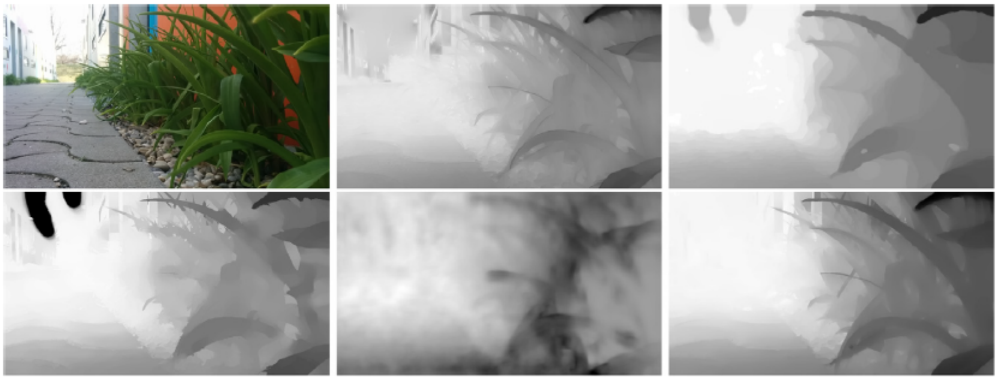

Figure 5.6: Comparison of recovered depth image. From top left to bottom right: all-in-focus image, the results by Tseng [12], Moeller [3], Javidina [13], Hazirbas [14], and ours, respectively.

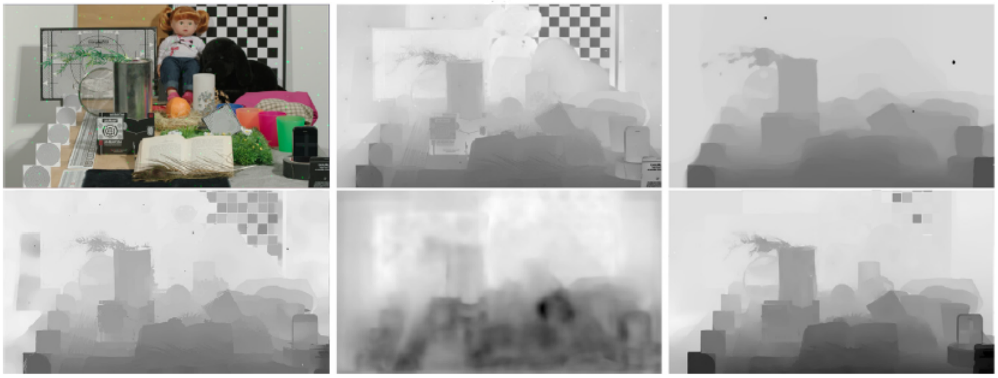

Figure 5.7: Comparison of recovered depth image. From top left to bottom right: all-in-focus image, the results by Tseng [12], Moeller [3], Javidina [13], Hazirbas [14], and ours, respectively.

indoor and outdoor environments.

5.5 Limitations

Even though our proposed method can significantly increase the accuracy of depth estimation, it still cannot effectively deal with transparent and reflected

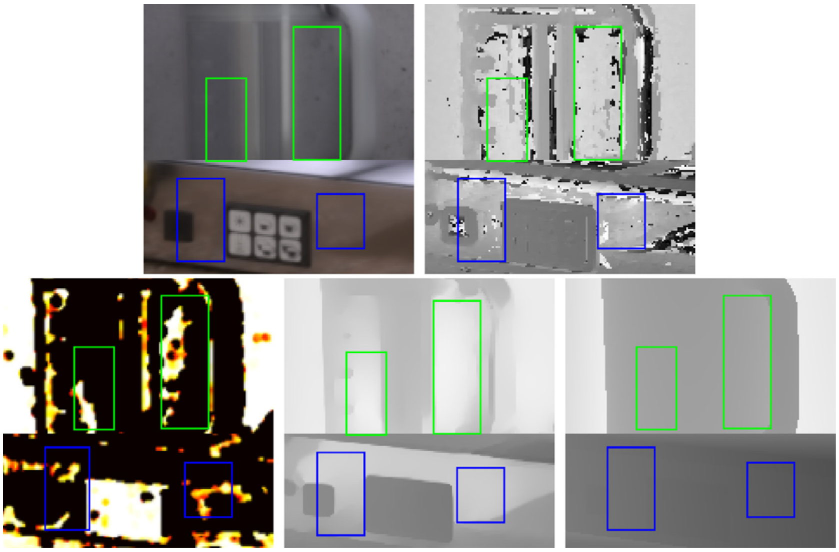

Figure 5.8: Illustration of depth reconstruction for transparent and reflected surfaces. From top left to bottom right: cropped all-in-focus image, initial depth image, confidence map, our reconstructed depth image, and ground truth, respectively.

surfaces. In such circumstances, inaccurate but reliable depth values in the initial depth image would be utilized for further depth reconstruction, and thus decrease the reconstruction accuracy. Figure 5.8 illustrates an example of this problem. As shown, in the transparent regions marked in green boxes and reflection regions marked in blue boxes, some incorrect initial depths with high confidence levels are used in the depth reconstruction, thus resulting in inaccurate depth estimation. To address these cases, more researches regarding advanced focus measure operators and reliability measure algorithms should be performed in the future.

5.6 Computational performances

All the experiments were performed on a desktop computer with Intel(R) Core(TM) i7-7700k CPU 4.2GHz and an NVIDIA GeForce GTX Titan X graphic card. Our method and other comparison algorithms were implemented in Matlab 2018b except for Moeller's and Hazirbas's methods which were implemented in parallel with CUDA. As the Laplacian-based methods require solving a large linear system, where the Laplacian matrix is derived from color guidance image, the computational performance of our and Tseng's methods is highly-related with the image resolution. Figure 5.9 shows the computational time of different methods on different datasets. The image resolutions of the tested datasets are 552×383, 512×512, 1080×1080, and 1920×1080, respectively. Note that all the real scene images were downsampled by a factor two for all but Moeller's and Hazirbas's methods. As shown, Hazirbas's deep learning method has the best computational performance, and it is not much affected by the image resolution. Moeller's method is faster and less highly-related to image resolution compared to the other traditional optimization methods. As the proposed framework contains two stages and each stage needs to solve such a large linear system, it is much slower than the other methods although it has better performance. Nevertheless, the computational time of our method can be further dramatically reduced by adopting GPU-based matting Laplacian solution [74] or cell-based framework [12].

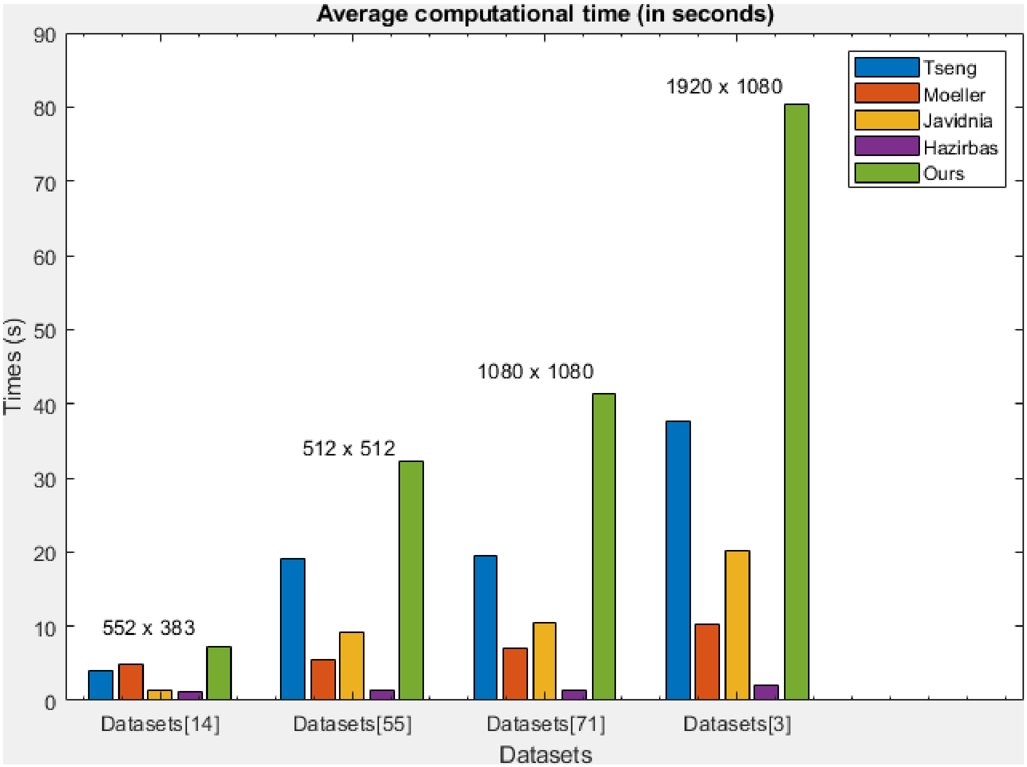

Figure 5.9: Average computational time (in seconds) comparison of different methods on different datasets.

Chapter 6

Conclusion

In this dissertation, we presented a robust depth recovery framework to recover 3D depth for a given sequence of multi-focus images. Our proposed depth recovery framework involved two processes: depth reconstruction and depth refinement. In the reconstruction process, we formulated the DFF problem as MAP estimation problem in the Bayesian network. Under the assumption of identically independent Gaussian distribution, our likelihood function or data term can be derived in L_2 norm with the pixel-wised confidence measure. The prior function or smoothness term was designed by including the matting Laplacian matrix which is often used as a prior to transfer the object structure in the color image to the target image. As it is very difficult to avoid the introduction of texture-copy artifacts, the global goal of texture-less artifacts reduction is temporarily neglected. In the reconstruction process, the nonlocal principle was adopted in the construction process of the matting Laplacian matrix to preserve the object structure and fine details as much as possible. With the adoption of the nonlocal matting Laplacian prior and the effective variance based confidence level computation, our proposed reconstruction approach is robust over texture-less regions, and can reconstruct a depth image with clear edges and fine details. However, as the nonlocal principle destroyed the spatial consistency, the reconstructed depth image was spatially inconsistent and suffered from the texture-copy artifacts caused by inconsistent structures between depth image and color image. To suppress noise and texture-copy artifacts, we then proposed a MAP-MRF based depth refinement algorithm. With the construction of undirected weighted graph representing the energy function, we utilized the Laplacian matrix corresponding to the graph to derive a closed form solution. Moreover, the depth edges and fine details can be well preserved, and the texture-copy artifacts can be effectively suppressed by incorporating the pre-estimated edges and mutual structure information into our formulation. Experiments over synthetic and real scene datasets demonstrated that the proposed framework outperformed the previous state-of-the-art methods.

Even though the performance of the proposed approach has been significantly improved, there still many limitations. As aforementioned in section *limitations*, it still cannot effectively deal with transparent and reflected object. That is because of the inaccurate reliability measures for unreliable depth pixels in initial depth image. Those unreliable initial depth values are further utilized in depth reconstruction process leading inaccurate reconstruction result. Apparently, the reliability measure by measuring the fit of focus measure function to Gaussian model is accurate. As the reliability measure is based on the focus measure, further researches on effective focus measure operator are necessary. One possible solution is the two-round focus measure using adaptive SMLAP for different regions. For the first round, small kernel size 3×3 SMLAP operator is utilized to generate a initial depth image and reliability map. In the second round, the summation for each pixel in the reliable KNN neighbors is re-calculated, such that the pixel depth in unreliable regions normally less textured regions can be recovered by relatively reliable pixel depth.

Even though we took various kinds of operations to suppress texture-copy artifacts, it still can not be perfectly solved. One possible solution to this issue is the precise segmentation aware depth reconstruction. For one object, precise structure information is just what we need. On the contrary, the color information inside the object not only can provide unuseful information, but possibly introduce the texture-copy artifacts. Once precise object segmentation is obtained, the object can be assigned with homogeneous color, such that the texture-copy problem can be fundamentally solved by using our depth reconstruction approach. Additionally, deep learning based segmentation methods can be utilized to obtain precise segmentation result. An alternative idea for texture-copy artifacts reduction is to precisely estimate depth values on all edge pixels. The remaining works can be done with image inpainting-like approaches.

Bibliography

[1] M. Subbarao and Tao Choi. Accurate recovery of three-dimensional shape from image focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):266–274, March 1995.

[2] S. K. Nayar and Y. Nakagawa. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, Aug 1994.

[3] M. Moeller, M. Benning, C. Schönlieb, and D. Cremers. Variational depth from focus reconstruction. *IEEE Transactions on Image Processing*, 24(12):5369–5378, Dec 2015.

[4] M. Muhammad and T. Choi. Sampling for shape from focus in optical microscopy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):564–573, March 2012.

[5] Tarkan Aydin and Yusuf Sinan Akgul. A new adaptive focus measure for shape from focus. In *BMVC*, pages 1–10.

[6] A. Thelen, S. Frey, S. Hirsch, and P. Hering. Improvements in shape-from-focus for holographic reconstructions with regard to focus operators,

neighborhood-size, and height value interpolation. *IEEE Transactions on Image Processing*, 18(1):151–157, Jan 2009.

[7] J. Surh, H. Jeon, Y. Park, S. Im, H. Ha, and I. S. Kweon. Noise robust depth from focus using a ring difference filter. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2444–2453, July 2017.

[8] P. Sakurikar and P. J. Narayanan. Composite focus measure for high quality depth maps. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1623–1631, Oct 2017.

[9] Victor Gaganov and Alexey Ignatenko. Robust shape from focus via markov random fields. In *Proceedings of Graphicon Conference*, pages 74–80, 2010.

[10] RR Sahay and AN Rajagopalan. Shape extraction of low-textured objects in video microscopy. *Journal of Microscopy*, 245(3):252–264, 2012.

[11] C. Tseng and S. Wang. Maximum-a-posteriori estimation for global spatial coherence recovery based on matting laplacian. In *2012 19th IEEE International Conference on Image Processing*, pages 293–296, Sep. 2012.

[12] C. Tseng and S. Wang. Shape-from-focus depth reconstruction with a spatial consistency model. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(12):2063–2076, Dec 2014.

[13] Hossein Javidnia and Peter Corcoran. Application of preconditioned alter-

nating direction method of multipliers in depth from focal stack. *Journal of Electronic Imaging*, 27(2):023019, 2018.

[14] C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers. Deep depth from focus. In *Asian Conference on Computer Vision*, pages 525–541, 2018.

[15] Wei Huang and Zhongliang Jing. Evaluation of focus measures in multi-focus image fusion. *Pattern Recognition Letters*, 28(4):493 – 500, 2007.

[16] Ge Yang and B. J. Nelson. Wavelet-based autofocusing and unsupervised segmentation of microscopic images. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, volume 3, pages 2143–2148 vol.3, Oct 2003.

[17] Aamir Saeed Malik and Tae-Sun Choi. A novel algorithm for estimation of depth map using image focus for 3d shape recovery in the presence of noise. *Pattern Recognition*, 41(7):2200 – 2225, 2008.

[18] S. Lee, J. Yoo, Y. Kumar, and S. Kim. Reduced energy-ratio measure for robust autofocusing in digital camera. *IEEE Signal Processing Letters*, 16(2):133–136, Feb 2009.

[19] Rashid Minhas, Abdul A Mohammed, QM Jonathan Wu, and Maher A Sid-Ahmed. 3d shape from focus and depth map computation using steerable filters. In *International Conference Image Analysis and Recognition*, pages 573–583. Springer, 2009.

[20] Yu Sun, Stefan Duthaler, and Bradley J Nelson. Autofocusing in computer microscopy: selecting the optimal focus algorithm. *Microscopy research and technique*, 65(3):139–149, 2004.

[21] Andrés Santos, C Ortiz de Solórzano, Juan José Vaquero, Javier Márquez Pena, Norberto Malpica, and Francisco del Pozo. Evaluation of autofocus functions in molecular cytogenetic analysis. *Journal of microscopy*, 188(3):264–272, 1997.

[22] Ge Yang and Bradley J Nelson. Wavelet-based autofocusing and unsupervised segmentation of microscopic images. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 3, pages 2143–2148. IEEE, 2003.

[23] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *ICCV*, volume 98, page 2.

[24] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[25] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[26] Tao Pham Dinh, Hoai Minh Le, Hoai An Le Thi, and Fabien Lauer. A difference of convex functions algorithm for switched linear regression. *IEEE Transactions on Automatic Control*, 59(8):2277–2282, 2014.

[27] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[30] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[31] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[32] Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[33] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of

high-dynamic-range images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 257–266, 2002.

[34] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Nonlocal image and movie denoising. *International Journal of Computer Vision*, 76(2):123–139, 2008.

[35] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65, 2005.

[36] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon. High quality depth map upsampling for 3d-tof cameras. In *2011 International Conference on Computer Vision*, pages 1623–1630. IEEE, 2011.

[37] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000, 2013.

[38] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics (ToG)*, 26(3):96–es, 2007.

[39] Qiong Yan, Xiaoyong Shen, Li Xu, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Jiaya Jia. Cross-field joint image restoration via scale map. In

Proceedings of the IEEE International Conference on Computer Vision, pages 1537–1544, 2013.

[40] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *European Conference on Computer Vision*, pages 1–14. Springer, 2010.

[41] X. Shen, C. Zhou, L. Xu, and J. Jia. Mutual-structure for joint filtering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3406–3414, Dec 2015.

[42] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):228–242, Feb 2008.

[43] Kaiming He, Jian Sun, and Xiaoou Tang. Fast matting using large kernel matting laplacian matrices. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2165–2172. IEEE, 2010.

[44] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2175–2188, 2013.

[45] Xiaowu Chen, Dongqing Zou, Steven Zhiying Zhou, Qinping Zhao, and Ping Tan. Image matting with local and nonlocal smooth priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1902–1907, 2013.

[46] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 30(10):1699–1712, 2008.

[47] Wu-Chih Hu, Jia-Jie Jhu, and Cheng-Pin Lin. Unsupervised and reliable image matting based on modified spectral matting. *Journal of Visual Communication and Image Representation*, 23(4):665–676, 2012.

[48] Martin Eisemann, Julia Wolf, and Marcus A Magnor. Spectral video matting. In *VMV*, pages 121–126, 2009.

[49] Dongqing Zou, Xiaowu Chen, Guangying Cao, and Xiaogang Wang. Unsupervised video matting via sparse and low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1501–1514, 2019.

[50] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, Dec 2011.

[51] Shengyang Dai and Ying Wu. Motion from blur. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[52] Jialue Fan, Xiaohui Shen, and Ying Wu. Closed-loop adaptation for robust tracking. In *European Conference on Computer Vision*, pages 411–424. Springer, 2010.

[53] J. Li, M. Lu, and Z. Li. Continuous depth map reconstruction from light fields. *IEEE Transactions on Image Processing*, 24(11):3257–3265, Nov 2015.

[54] Said Pertuz, Domenec Puig, and Miguel Angel Garcia. Reliability measure for shape-from-focus. *Image and Vision Computing*, 31(10):725–734, 2013.

[55] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, Jan 1979.

[56] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2016.

[57] Q. Chen, D. Li, and C. Tang. Knn matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2175–2188, Sep. 2013.

[58] D Dansereau. Light field toolbox v0. 4. *Online].[Accessed May 2017]*, 2016.

[59] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[60] Michael Lindenbaum, M Fischer, and A Bruckstein. On gabor's contribution to image enhancement. *Pattern Recognition*, 27(1):1–8, 1994.

[61] Jacob Benesty, Jingdong Chen, and Yiteng Huang. Study of the widely linear wiener filter for noise reduction. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 205–208. IEEE.

[62] Michael J Black, Guillermo Sapiro, David H Marimont, and David Heeger. Robust anisotropic diffusion. *IEEE Transactions on Image Processing*, 7(3):421–432, 1998.

[63] Joachim Weickert, BMTH Romeny, and Max A Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Transactions on Image Processing*, 7(3):398–410, 1998.

[64] Stephen M Smith and J Michael Brady. Susan—a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78, 1997.

[65] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013.

[66] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, Nov 1984.

[67] Y. Y. Boykov and M. . Jolly. Interactive graph cuts for optimal boundary amp; region segmentation of objects in n-d images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112 vol.1, July 2001.

[68] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 377–384. IEEE.

[69] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings. 1998 IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition (Cat. No.98CB36231), pages 648–655, June 1998.

[70] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *European Conference on Computer Vision*, pages 617–632, 2016.

[71] Wei Liu, Xiaogang Chen, Jie Yang, and Qiang Wu. Robust color guided depth map restoration. *IEEE Transactions on Image Processing*, 26(1):315–327, 2017.

[72] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu. Saliency detection on light field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1605–1616, Aug 2017.

[73] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[74] Chunxia Xiao, Meng Liu, Donglin Xiao, Zhao Dong, and Kwan-Liu Ma. Fast closed-form matting using a hierarchical data structure. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(49-62):57, 2014.