

# High-Quality 3D Reconstruction With Depth Super-Resolution and Completion

JIANWEI LI<sup>1,2</sup>, WEI GAO<sup>1,2</sup>, AND YIHONG WU<sup>1,2</sup>

<sup>1</sup>National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

<sup>2</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Wei Gao (wgao@nlpr.ia.ac.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0502002, and in part by the Natural Science Foundation of China under Grant 61872361 and Grant 61836015.

**ABSTRACT** The 3D reconstruction is an important topic in computer vision with many applications, such as robotics and augmented reality. Since the raw depth images captured by consumer RGB-D cameras are often low resolution (LR), noisy, and incomplete. How to obtain high-quality 3D models with a consumer RGB-D camera is still a challenge for the existing systems. In this paper, we propose a new depth super-resolution and completion method implemented in a deep learning framework and build a high-quality 3D reconstruction system. We first improve the resolution of LR depth image with a depth super-resolution network and remove the outliers in high-resolution (HR) depth image based on gradient saliency. To further enhance the quality of HR depth image with the guide of HR color image, we learn surface normal and occlusion boundary images from the corresponding HR color image through two deep fully convolutional networks. In particular, the blurriness of HR color image is also detected and pixel-wise quantized. Finally, we obtain a completed HR depth image by optimizing the HR depth image with the surface normal, occlusion boundary, and color image blurriness. We have carried out qualitative and quantitative evaluations with baseline methods on public datasets. The experimental results demonstrate that our method has better performance both on single depth image enhancement and 3D reconstruction.

**INDEX TERMS** Deep learning, super-resolution, image processing, depth completion, 3D reconstruction.

## I. INTRODUCTION

3D reconstruction aims to estimate the motion of a camera and to reconstruct the geometric structure of the real world objects or scenes. The emergence of RGB-D camera provides an opportunity to develop indoor scene 3D reconstruction systems conveniently. For consumer grade RGB-D cameras, there are two main approaches in depth sensing, i.e., structured-light (SL) and time-of-flight (ToF). They are popular because of good mobility, low cost, and high frame rate. However, the resolution and accuracy of depth images are always constrained by the depth sensing techniques used in RGB-D cameras. As shown in Fig. 2 (a), Kinect v2 can capture low-resolution (LR) depth images, e.g.,  $512 \times 424$ , while high-resolution (HR) color images, e.g.,  $1920 \times 1080$ . Besides, the raw depth images are often incomplete when surfaces are shiny, bright, transparent or far

from the camera. These issues make the applications based on RGB-D cameras very restricted.

With the popularity of inexpensive RGB-D cameras, enhancing the quality of depth image is an extensively studied topic. Most depth super-resolution (SR) methods [1]–[5] combined with the help of a corresponding HR color image. The most popular method is joint bilateral upsampling (JBU) [1], which applies a spatial filter to the LR image and a range filter on another HR image. Furthermore, interpolation [6]–[8], markov random fields (MRFs) [9]–[13], shape from shading (SFS) technologies [14]–[16], and deep learning-based methods [17]–[19] are also proposed to recover accurate HR depth image. To fill in the holes in raw depth image, researchers have proposed many methods, such as cross-bilateral filtering [20], ray casting with truncated signed distance function (TSDF) [21], [22], MRFs optimization [23], and patch-based image synthesis [24]–[26]. Some works also focus on sparse-to-dense depth estimation [27], [28] to produce a dense depth image.

The associate editor coordinating the review of this manuscript and approving it for publication was Naveed Akhtar.

**TABLE 1.** The main symbols and notations used in this paper. The subscript  $i$  represents the  $i$ -th frame of image sequence.

Data type	Symbol	Description
Camera parameter	$K_d$	Pre-calibrated intrinsic matrix of depth camera
	$K_c$	Pre-calibrated intrinsic matrix of color camera
	$K$	Intrinsic matrix of depth camera after registration
Depth image	$D_i^r$	Raw depth image (depth image captured by a RGB-D camera)
	$D_i^l$	LR depth image (depth image after RGB-D image registration)
	$D_i^h$	HR depth image (depth image after DlapRSN and outliers removing)
	$D_i^c$	Completed HR depth image (HR depth image after depth optimization)
Other image	$G_i$	Gradient image of HR depth image after super-resolution
	$M_i$	Mask image used in outliers removing
	$N_i$	Surface normal image of HR color image
	$C_i$	Corresponding HR color image of depth image
	$O_i$	Occlusion boundary image of HR color image
	$B_i$	Blurriness image of HR color image

Recent advances of depth estimation researches using deep learning have been applied in 3D reconstruction. DeepMVS [29] predicts high-quality disparity images for multi-view stereo reconstruction (MVS) with a deep convolutional neural network (ConvNet), which has to perform a structure from motion (SFM) reconstruction to recover the camera pose in advance. CodeSLAM [30] presents a compact but dense representation of scene geometry which is conditioned on the intensity data from a single image and generated from a code, but is only suitable for key frame-based monocular dense simultaneous localization and mapping (SLAM) system.

In this paper, we aim to achieve high-quality 3D reconstruction through depth enhancement for consumer RGB-D cameras, and propose a depth super-resolution and completion (SRC) method applied in volumetric-based 3D reconstruction system. In summary, the main contributions of our work are:

- 1) a new depth SRC method implemented in deep learning framework to obtain completed HR depth image;
- 2) an adaptive optimization method to complete HR depth images by information extracted from HR color image;
- 3) a high-quality 3D reconstruction pipeline based on the proposed depth SRC method for consumer RGB-D cameras.

The rest of the paper is organized as follows: Section II introduces the related work and motivation of our research. Section III gives an overview of our high-quality 3D reconstruction system. The details of the proposed depth enhancement method are presented in Section IV. Section V describes experiment results and discussions while Section VI presents some concluding remarks.

For convenience of presentation, the main symbols and notations used in this paper are summarized in Table 1.

## II. RELATED WORKS

### A. RGB-D SCENE 3D RECONSTRUCTION

The emergence of consumer RGB-D cameras, such as Microsoft Kinect and Structure Sensor, provides an opportunity to develop indoor scene reconstruction systems conveniently. Kinectfusion [31], [32] is an outstanding volumetric-based method to generate photorealistic dense 3D models with RGB-D images. It uses a volumetric representation by TSDF to represent the scenes, and in conjunction with fast iterative closest point (ICP) [33] pose estimation to provide a real-time fused dense model. Based on the pioneering work of Kinectfusion, variants of improved systems and methods have been proposed.

Existing state-of-the-art methods can be classified into two categories: online reconstructions [34]–[38], i.e., dense simultaneous localization and mapping (SLAM); and offline reconstructions [39]–[43], which have higher accuracy. To obtain high-quality 3D reconstruction, Bundlefusion [38] uses additional color features for registration and global bundle adjustment for obtaining precise scene geometry; Choi *et al.* [39] and Zeng *et al.* [40] reconstructed local smooth scene fragments and globally aligned them together with 3D features. However, due to low-quality depth, inaccurate registration, camera tracking drift and the lack of accurate surface details, 3D models reconstructed from consumer RGB-D cameras are not yet popularly used in applications.

In contrast to above methods, we explore depth enhancement method using both depth information and color information, and further improve the performance of 3D reconstruction in terms of camera pose estimation and surface reconstruction.

### B. DEPTH ENHANCEMENT WITH DEEP LEARNING

To enhance the quality of depth image, researchers have made lots of exploring, which can be divide into depth



**FIGURE 1.** The pipeline of high-quality 3D reconstruction based on depth super-resolution and completion.

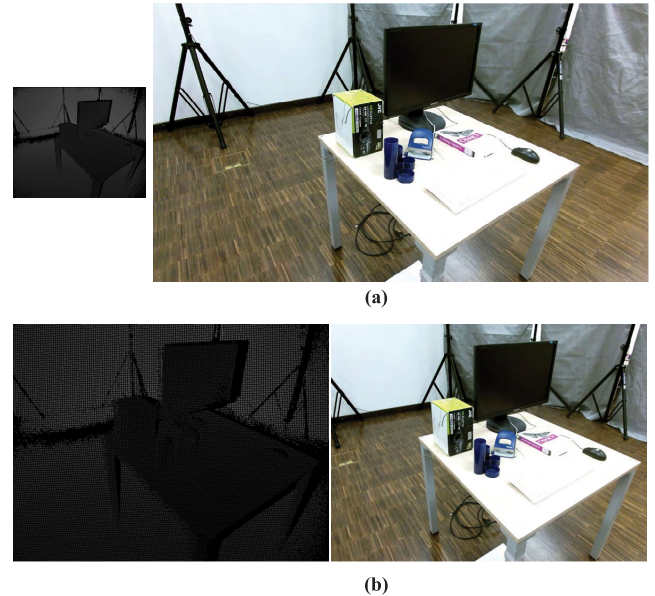
super-resolution and depth inpainting. Since deep learning has demonstrated highly successful in image processing, we focus in the related work on deep learning-based methods.

Super-resolution technique improves the observed LR image to the corresponding HR image. A common approach to refine a low quality depth image is to incorporate the concurrently captured high quality color image. Riegler *et al.* [17] increased the resolution of depth image with a HR color image to guide a deep fully convolutional network (FCN). MSG-Net [19] complements LR depth features with HR intensity features using a multi-scale fusion strategy. These methods train the networks use both depth image and color image. However, color image captured by a consumer RGB-D camera also suffers from blur and noise, which may degrade the performance of depth super-resolution. Researchers also explore depth super-resolution methods only using depth image as input. ATGV-Net [18] combines the benefits of deep learning-based single image super-resolution (SISR) with a variational method to recover accurate HR depth image, but it can not reduce noise and holes. Unlike above methods, we first make depth super-resolution and outliers removing only use depth information, and then complete the missing data in HR depth image with the guide of HR color image.

Deep learning-based methods for depth estimation from single color image and color image inpainting have received a lot of attention in recent years. For depth inpainting, it is still difficult due to the lack of training data and absence of depth feature. Sparse-to-Dense [27] estimates dense depth images from both sparse depth image and color image, which is suited for sensor fusion and sparse SLAM. In a recent work, Zhang and Funkhouser [44] trained a deep network that takes a color image as input to predict large missing areas in depth image, and created a dataset of RGB-D images paired with completed depth images. These methods force on enhancing single depth image, and assume that the corresponding color image has high quality. To maximize the enhancement of depth image from a consumer RGB-D camera, we first measure the quality of corresponding color image, and then adaptively optimize depth image by surface normal, occlusion boundary, and color image blurriness.

### III. SYSTEM OVERVIEW

The pipeline of the proposed system is shown in Fig. 1, which is composed of three parts: RGB-D image registration, depth super-resolution and completion, and 3D reconstruction.



**FIGURE 2.** An example of RGB-D image registration. (a) Original unregistered full resolution depth (512×424) and color (1920×1080) images. (b) Registered depth and color images in a unified camera coordinate system.

#### A. RGB-D IMAGE REGISTRATION

Since depth image and color image are captured by different sensors of a consumer RGB-D camera, we need to register them in a unified camera coordinate system. Fig. 2 shows a registration example of RGB-D image from CoRBS dataset [45], which is captured by Kinect v2. The resolutions of original unregistered depth and color images are 512×424 and 1920×1080 respectively. After image registration, the registered depth image and color image are in the same camera coordinate system.

For each input raw depth image  $D_i^r$  in depth camera coordinate system, we register it to the color camera coordinate system as follows:

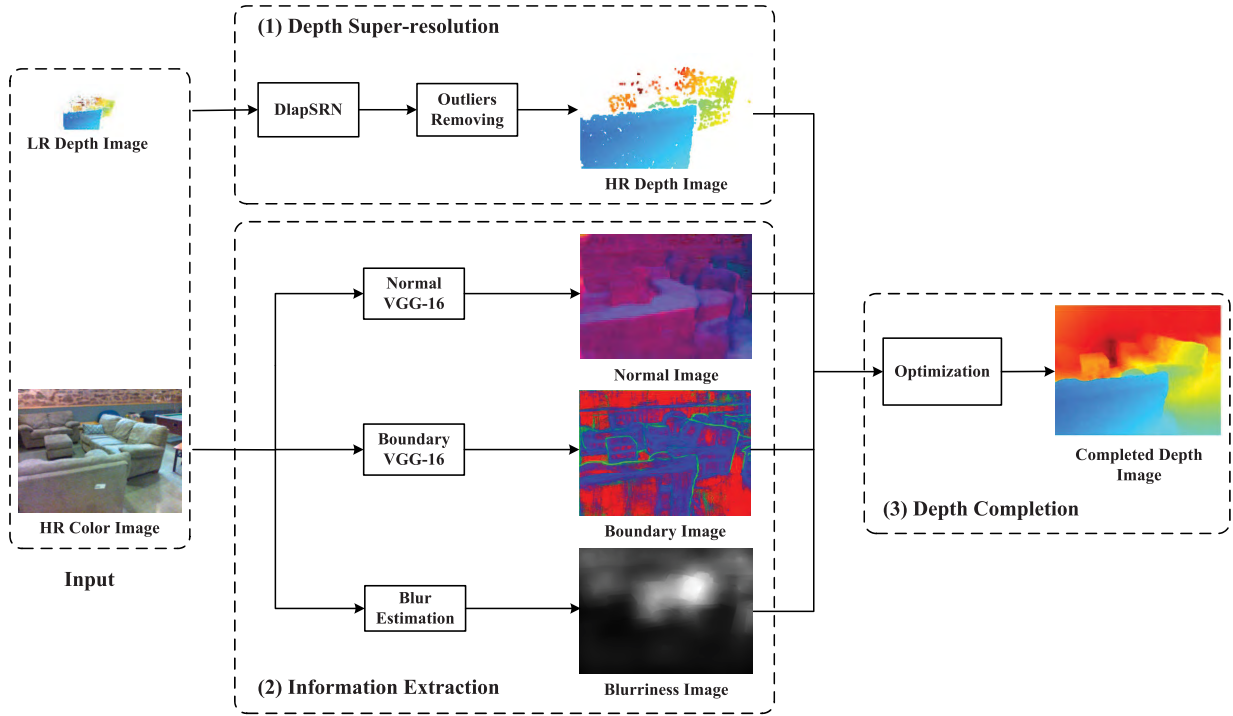
- First, we reconstruct a 3D point  $p(x)$  corresponding to the pixel  $x$  on the raw depth image as:

$$p(x) = D_i^r(x)K_d^{-1}[x, 1]^T, \quad (1)$$

where  $K_d$  is the pre-calibrated intrinsic matrix of depth camera.

- Second, we translate the 3D point to color camera coordinate system and obtain a new 3D point coordinates  $q(x)$ , which is reconstructed as:

$$Tp(x) = q(x) = D_i^l(x)K_c^{-1}[x, 1]^T, \quad (2)$$



**FIGURE 3.** The pipeline of the proposed depth super-resolution and completion (SRC) method. (1) The input LR depth image is enhanced to a HR depth image through DlapSRN and outlier removing units. (2) The input HR color image is used to learn surface normal and occlusion boundary images through two FCN networks, meanwhile the blurriness of HR color image is detected. (3) The completed HR depth image is adaptively optimized by HR depth image and information extracted from HR color image.

where  $T$  is the pre-calibrated transformation matrix from depth camera coordinate system to color camera coordinate system, and  $K_c$  is the pre-calibrated intrinsic matrix of color camera.

- Finally, the registered LR depth image  $D_i^l$  in color camera coordinate system is calculated as:

$$D_i^l(x) = D_i^c(x)TK_d^{-1}K_c \quad (3)$$

## B. DEPTH SUPER-RESOLUTION AND COMPLETION

In order to estimate a completed HR depth image  $D_i^c$  from a LR depth image  $D_i^l$  with corresponding HR color image  $C_i$ , we propose a depth SRC method implemented in deep learning framework. As shown in Fig. 3, it consists of three stages: depth super-resolution, information extraction, and depth completion.

### 1) DEPTH SUPER-RESOLUTION

The input LR depth image  $D_i^l$  is enhanced to a HR depth image  $D_i^h$  through depth laplacian super-resolution network (DlapSRN) and outlier removing units. As shown in Fig. 4, DlapSRN uses a laplacian pyramid network [46] as base architecture, and contains two branches: feature extraction and depth image reconstruction. At each pyramid layer: i) feature extraction branch consists of  $d$  convolutional layers and one transposed convolutional layer to upsample the feature by a scale of 2; ii) the input depth image is upsampled by a scale of 2 with a transposed convolutional layer; iii) the output

HR depth image at current level is fed into the depth image reconstruction branch of next level. Furthermore, the outliers in the learned HR depth image are detected and removed based on gradient saliency.

### 2) INFORMATION EXTRACTION

The input HR color image  $C_i$  is used to learn surface normal image  $N_i$  and occlusion boundary image  $O_i$  through two FCN networks, meanwhile the blurriness of HR color image is detected and pixel-wise quantized. We use two deep networks designed on the back-bone of VGG-16 with symmetry encoder and decoder [47], [48] to extract surface normal and occlusion boundary from HR color image. Normal VGG-16 and boundary VGG-16 are trained to learn occlusion boundary image and surface normal image respectively.

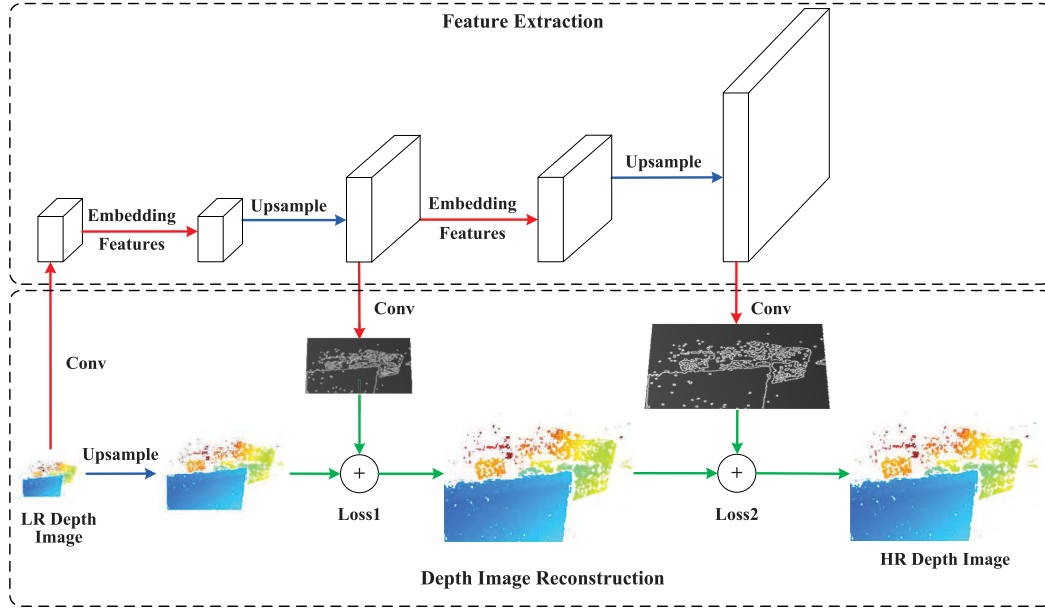
### 3) DEPTH COMPLETION

Depth completion is to obtain a completed HR depth image  $D_i^c$  with information extracted from HR color image. We construct an objective function with HR depth image  $D_i^h$ , surface normal image  $N_i$ , occlusion boundary image  $O_i$ , and blurriness image  $B_i$ , through which each HR depth image can be adaptively optimized by its corresponding HR color image.

## C. 3D RECONSTRUCTION

In accordance with what is widely used in previous works, we use the volumetric-based method to integrate





**FIGURE 4.** The architecture of depth super-resolution network (DlapSRN), which contains two branches: feature extraction and depth image reconstruction. Red arrows indicate convolutional layers, blue arrows indicate transposed convolutions, and the green arrows denote element-wise addition operators.

depth sequence. We implement 3D reconstruction based on ICP-based camera pose estimation, and in conjunction with volumetric representation by TSDF to provide a fused dense model.

#### 1) CAMERA POSE ESTIMATION

We estimate the camera pose through the well-known ICP algorithm to get a 6-DoF rigid relative transformation. For each pixel  $x = (u, v)^T$  at an enhanced depth image  $D_i^c$ , the vertex  $v_i(x)$  and normal vector  $n_i(x)$  are calculated as follows:

$$v_i(x) = z_i(x)K^{-1}[x, 1]^T, \quad (4)$$

$$n_i(x) = (v_i(u+1, v) - v_i(u, v)) \times (v_i(u, v+1) - v_i(u, v)), \quad (5)$$

where  $K$  is the pre-registered intrinsic matrix of depth camera,  $z_i(x)$  is the depth value.

The transformation  $T_{g,i}$  from current camera coordinate system to global coordinate system is calculated by minimizing the following point-to-plane error  $E(T_{g,i})$ :

$$E(T_{g,i}) = \sum_{x \in D_i} (\hat{n}_{i-1}^g(\hat{x}) \cdot (T_{g,i} \hat{v}_i(x) - \hat{v}_{i-1}^g(\hat{x})))^2, \quad (6)$$

where  $\hat{x}$  corresponds to  $x$  by projective data association,  $\hat{v}_i(x) = [v_i(x), 1]^T$  is the homogeneous form of vertex  $v_i(x)$  in current vertex map,  $\hat{v}_{i-1}^g(\hat{x})$  and  $\hat{n}_{i-1}^g(\hat{x})$  are the predicted vertex and normal at previous frame.

#### 2) SURFACE RECONSTRUCTION

After estimating the camera's global pose, depth images are integrated into a TSDF model. TSDF is discretized into a

voxel grid to represent a physical volume of space. For a given voxel  $c$  in the fused scene model  $F$ , the corresponding signed distance value  $F(c)$  is computed with  $n$  depth images:

$$F(c) = \frac{\sum_{i=1}^n sdf_i(c)w_i(c)}{W(c)}, \quad W(c) = \sum_{i=1}^n w_i(c). \quad (7)$$

Signed distance function  $sdf_i(c)$  is the projective distance between a voxel and  $i$ -th depth frame, and is defined as:

$$\max\{\min\{\Phi, |[K^{-1}z_i(x)[x, 1]^T]_z - [c]_z|, -\Phi\}, \quad (8)$$

where  $x = \pi(Kc)$  is the pixel into which the voxel center projects, and  $\Phi$  is the truncation threshold. We compute distance along the principal (Z) axis of the camera frame using the  $z$  component denoted as  $[.]_z$ .

Weighting function  $w_i(c)$  represents the confidence in the accuracy of the distance, and is fixed to 1 in our experiments.

Through above procedures, the enhanced depth sequence is fused into a TSDF representation. A predicted surface is ray-casted into the volumetric model, and aligned with the enhanced depth image in the camera pose estimation stage. The final mesh model is extracted with marching cubes algorithm [49].

### IV. DEPTH SUPER-RESOLUTION AND COMPLETION

The key points of our depth SRC method are elaborated in the following subsections.

#### A. NETWORKS IMPLEMENTATION

##### 1) TRAINING DATA

To train DlapSRN, we use 372 depth images of ICL-NUIM living room sequences (kt0, kt1 and kt3) [50] and 795 depth

images of NYU-v2 dataset [20]. The training data is sampled in three scales (1/2, 1/4 and 1/8) and augmented by scaling (random downscale between [0.5, 1.0]), rotation (random rotate images by 90°, 180°, or 270°), and flipping (flip images horizontally or vertically with a probability of 0.5). We use nearest neighbor interpolation in both scaling and rotation to avoid creating spurious sparse depth points.

To train normal VGG-16 and boundary VGG-16 networks, we use 547515 images of the synthetic dataset SUNCG-RGBD [48] and 59743 render depth images of Scan-Net dataset [44], [51]. The ground truth images of surface normal and occlusion boundary are provided by the datasets. We use color images as input to learn surface normal and occlusion boundary.

## 2) LOSS FUNCTION

Loss function used for HR depth estimation is defined in Eq. (9) with the Charbonnier penalty function [46], [53].

$$L_{SR}(D, D^*) = -\frac{1}{m} \sum_{j=1}^m \sum_{l=1}^k \sqrt{(D_j - D_j^*)^2 + \epsilon^2}, \quad (9)$$

where  $D$  and  $D^*$  are predicted and ground truth depth respectively,  $m$  is the number of training samples in each batch,  $k$  is the number of level in the pyramid.  $\epsilon$  is the parameter of the Charbonnier penalty function, which is set to  $1e - 3$ .

For surface normal and occlusion boundary estimations, we use the inverse of the dot product between the ground truth and the estimation as loss function, which is defined in Eq. (10).

$$\begin{aligned} L_{normal}(N, N^*) &= -\frac{1}{n} \sum_{j=1}^n N_j - N_j^* \\ L_{boundary}(O, O^*) &= -\frac{1}{n} \sum_{j=1}^n O_j - O_j^* \end{aligned} \quad (10)$$

where  $N$  and  $N^*$  are predicted and ground truth of surface normal,  $O$  and  $O^*$  are predicted and ground truth of occlusion boundary, and  $n$  is the number of valid pixels.

## B. OUTLIERS REMOVING

Since the raw depth image captured by a consumer RGB-D camera usually contains large error on the occlusion boundary. This error may be enlarged in the process of depth super-resolution. In order to remove the unreliable depth boundaries before depth optimization, the outliers is determined by the distance discontinuities in HR depth image.

For each learned HR depth image, we construct the outliers removing method as follows:

- First, we normalize the learned HR depth image, and apply a gradient operator  $Sobel(x)$  on each pixel  $x$ .
- Second, we generate a mask image  $M_i$  based on gradient saliency. The value  $M_i(x)$  of each pixel in mask

image  $M_i$  is defined as:

$$\begin{cases} M_i(x) = 0, & G_i(u) \geq g_h \\ M_i(x) = 1, & G_i(u) < g_h, \end{cases} \quad (11)$$

where  $G_i(x) = Sobel(x)$  is the gradient of pixel  $x$ .  $g_h$  is the sensitive threshold, which is set to 0.1~0.3 in our experiments.

- Finally, the outliers in HR depth image are corroded with the mask image  $M_i$ .

After above processing, the outliers on occlusion boundaries (structure edges) can be effectively removed. Besides, the texture edges of the objects will not be incorrectly removed by our method, because their corresponding depth gradients are not sensitive.

## C. HYBRID COEFFICIENT MATRIX

Given the corresponding HR color image captured by a consumer RGB-D camera may suffer from strong defocus blur and large noises in occlusion boundary. Information extracted from low-quality color image will affect the depth optimization results and introduce unnecessary noise. To maximize the quality of the depth, we detect the blurriness of HR color image and introduce a hybrid coefficient in the objective function of depth optimization. The optimization procedure will be stated in Section IV-D.

For each HR color image  $C_i$ , the hybrid coefficient matrix  $H_i$  is based on corresponding occlusion boundary image and defocus blur detection image, which is constructed as follows:

- First, we detect the blurriness of each HR color image and get pixel-wise defocus blur annotations using BTBNet [54].
- Second, we normalize the defocus blur value of each pixel in the range of [0,1], and obtain a blurriness image  $B_i$ . A smaller value denotes a higher blurriness.
- Finally, the hybrid coefficient matrix  $H_i$  is calculated through:

$$H_i(x) = P_i(x)B_i(x), \quad (12)$$

where  $P_i(x)$  is the predicted probability that a pixel is on the occlusion boundary image, and  $B_i(x)$  is the blurriness of pixel  $x$  on HR color image. The value of  $H_i(x)$  is between [0, 1].

## D. DEPTH OPTIMIZATION

The goal of depth optimization is to complete HR depth image with the predicted surface normal image, occlusion boundary image, and HR color image blurriness. Surface normals and occlusion boundaries capture local properties of the surface geometry, while image blurriness denotes the qualities of surface normals and occlusion boundaries extracted from HR color image. We combine them through global optimization to complete the depth image for all pixels in a consistent solution.

**TABLE 2.** Quantitative evaluation (RMSE in pixel disparity) with two SR factors on Middlebury dataset with synthetic noise proposed by Ferstl et al. [52]. Bold shows the best results, and underlined shows the second best results.

Methods	$\times 2$				$\times 4$			
	Art	Books	Moebius	Average	Art	Books	Moebius	Average
NN	6.55	6.16	6.59	6.43	7.48	6.31	6.78	6.86
Bilinear	4.55	3.95	4.20	4.23	5.62	4.31	4.56	4.83
He et.al. [3]	3.55	2.37	2.48	2.80	4.41	2.74	2.83	3.33
Ferstl et.al. [4]	3.19	1.52	1.47	2.06	4.06	2.21	2.03	2.77
ATGV-Net [18]	<u>1.84</u>	<b>1.13</b>	<u>1.24</u>	<u>1.40</u>	<u>2.98</u>	<b>1.72</b>	<u>1.95</u>	<u>2.22</u>
Our method	<b>1.20</b>	<u>1.31</u>	<b>1.23</b>	<b>1.25</b>	<b>1.69</b>	<u>1.82</u>	<b>1.74</b>	<b>1.75</b>

**TABLE 3.** Quantitative evaluation (RMSE in millimeter) on ToFMark dataset [4]. Bold shows the best results, and underlined shows the second best results.

Methods	Factors	RMSE (in millimeter)			
		Books	Devil	Shark	Average
NN		30.46	27.53	38.21	32.07
Bilinear		29.11	25.34	36.34	30.26
Kopf et.al. [1]	$\sim \times 5$	27.82	24.30	34.79	28.97
He et.al. [3]		27.11	23.45	33.26	27.94
Ferstl et.al. [4]		24.00	23.19	29.89	25.69
ATGV-Net [18]		24.67	21.74	28.51	24.97
Our method	$\times 8$	25.56	24.58	27.95	26.03
	$\times 4$	<u>21.82</u>	<u>19.23</u>	<u>24.38</u>	<u>21.81</u>
	$\times 2$	<b>21.39</b>	<b>19.26</b>	<b>23.47</b>	<b>21.37</b>

For each HR depth image  $D_t^h$ , the objective function is defined in Eq. (13). It is a weighted sum of squared errors with three terms: i)  $E_D$  measures the distance between the estimated depth  $D(x)$  and the HR depth  $D^h(x)$  at pixel  $x$ , and forces the global solution to maintain the correct scale by enforcing consistency with the HR depth; ii)  $E_S$  encourages adjacent pixels to have the same depths; iii)  $E_C$  measures the consistency between the estimated depth and the predicted surface normal  $N(x)$ .

$$\begin{aligned}
 E &= \lambda_D E_D + \lambda_S E_S + \lambda_C E_C \\
 E_D &= \sum_{x \in T_{obs}} \|D(x) - D^h(x)\|^2 \\
 E_S &= \sum_{x, q \in N} \|D(x) - D(q)\|^2 \\
 E_C &= \sum_{x, q \in N} \| \langle t(x, q), N(x) \rangle \|^2 H(x) \quad (13)
 \end{aligned}$$

where  $t(x, q)$  is the tangent vector, and  $H(x)$  is the hybrid coefficient. Like in [44],  $\lambda_D$ ,  $\lambda_S$ , and  $\lambda_C$  are fixed to be [1000, 1, 0.001] in our experiments.

As the matrix form of the system of equations in Eq. (13) is sparse and symmetric positive definite, we solve it with a sparse Cholesky factorization. The final solution is a global minimum to the approximated objective function.

**TABLE 4.** Accuracy of estimated camera pose (RMSE in centimeters) on ICL-NUIM living room kt2 sequences [50]. RMSE of baseline is 0.10 cm. X denotes the fail of camera pose estimation. Bold shows the best results, and underlined shows the second best results.

Inputs	RMSE		
	(in centimeters)		
LR depth sequence	1/2	1/4	1/8
	0.81	32.08	x
SRC depth sequence	$\times 2$	$\times 4$	$\times 8$
	without OR	<u>0.23</u>	<u>0.37</u>
	with OR	<b>0.21</b>	<b>0.35</b>

Since hybrid coefficient  $H(x)$  is used in the depth optimization, each HR depth image can be adaptively completed with the guide of its corresponding HR color image.

## V. EXPERIMENTS

We have carried out experiments to evaluate the quantitative and qualitative performance of our method. For all experiments, we run our system on a standard desktop PC with Intel Core i7-4790 3.6GHz CPU and GeForce GTX 1080 Ti 11GB GPU.

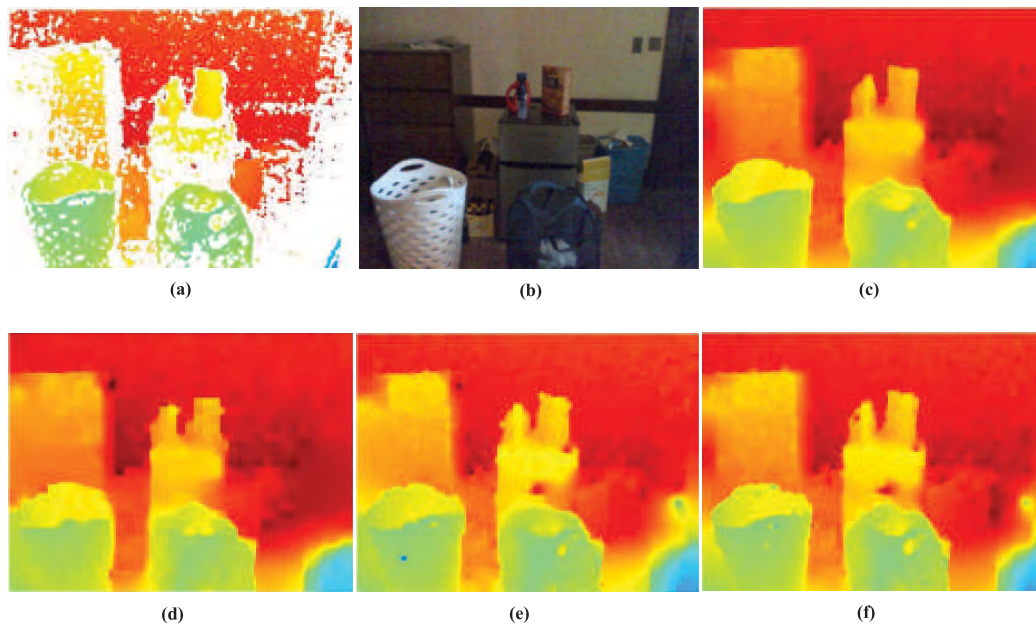
### A. SINGLE IMAGE EVALUATION

For single depth image, we first evaluate the depth super-resolution performance of our method on hole-filled RGB-D datasets [4], [52], and then evaluate the depth completion performance on RGB-D datasets [44], [51] with missing data in depth images.

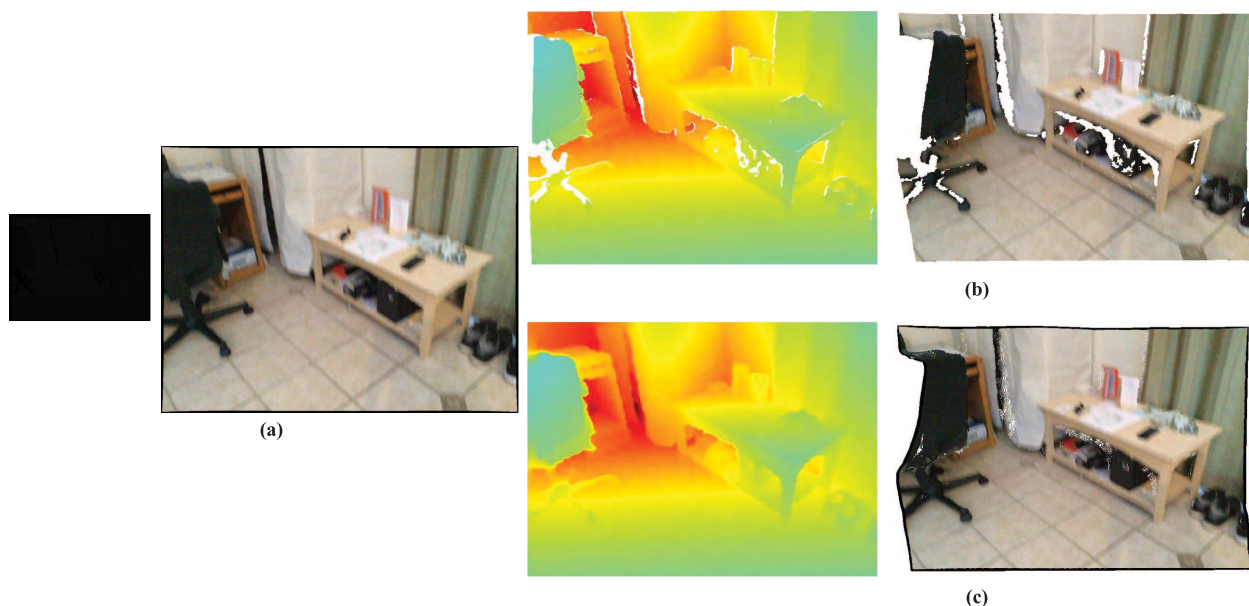
#### 1) DEPTH SUPER-RESOLUTION EVALUATION

The quantitative comparisons of super-resolution accuracy are performed with a series of baseline super-resolution methods: nearest neighbor (NN) interpolation, bilinear interpolation, Kopf et al. [1], He et al. [3], Ferstl et al. [4], and ATGV-Net [18].

Table 2 reports the root mean squared error (RMSE in pixel disparity) on Middlebury dataset with synthetic noise proposed by Ferstl et al. [52]. Note that the results of other methods are quoted from corresponding paper [18]. To have



**FIGURE 5.** One example of depth enhancement for RGB-D image captured by RealSense. (a-b) Depth image and color image with resolution of  $320 \times 240$ . (c) Completed depth image ( $320 \times 240$ ) with the method of Zhang and Funkhouser [44]. (d-f) Completed depth images with super-resolution factors of 8 ( $60 \times 40$  to  $320 \times 240$ ), 4 ( $80 \times 60$  to  $320 \times 240$ ), and 2 ( $160 \times 120$  to  $320 \times 240$ ) by our method.



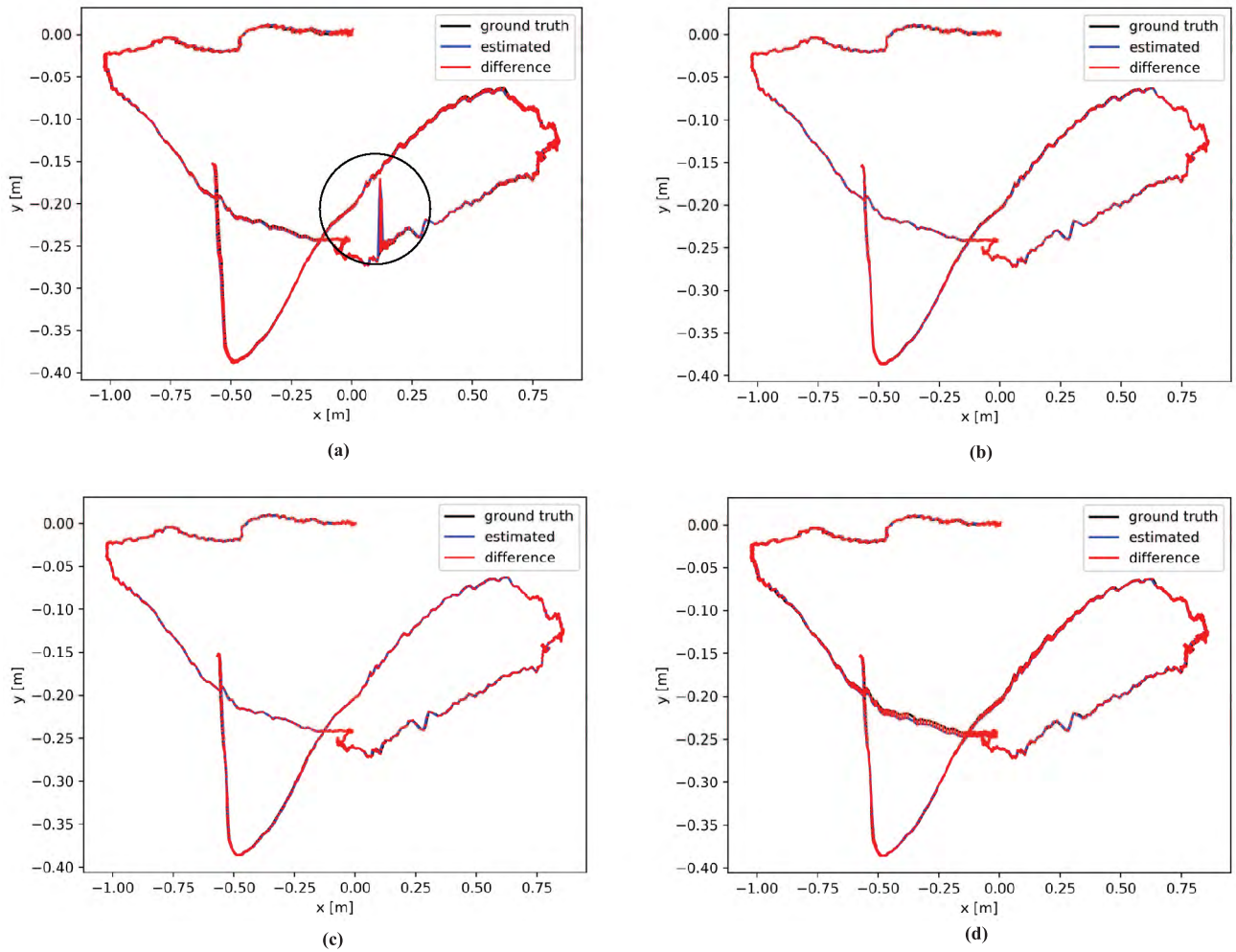
**FIGURE 6.** One example of depth enhancement for RGB-D image captured by Kinect. (a) The input depth image and color image. (b) The raw depth image and the corresponding point cloud calculated with it. (c) Our enhanced depth image and the corresponding point cloud calculated with it.

a fair evaluation, we convert all recovered HR depth images in the same data type (8-bit) since the ground truths are quantized to 8-bit. The results indicate that the average accuracy of our method on Middlebury dataset is higher than others.

Table 3 reports the quantitative results (RMSE in millimeter) on real ToF data from ToFMark dataset [4]. Note that

the results of other methods are quoted from corresponding papers with an approximately super-resolution factor 5 ( $160 \times 120$  to  $810 \times 610$ ). The best RMSE for each evaluation is in bold, whereas the second best one is underlined. Our method achieves the best results on  $2 \times \text{SR}$  and  $4 \times \text{SR}$ , and a comparable performance on  $8 \times \text{SR}$ .





**FIGURE 7.** Estimated camera trajectories through the propose 3D reconstruction system on ICL-NUIM living room kt2 sequences [50]. (a) Camera trajectories of 1/2 LR depth sequence compared with ground truth. (b-d) Camera trajectories of 2×SRC, 4×SRC and 8×SRC depth sequences compared with ground truth.

## 2) DEPTH COMPLETION EVALUATION

Since the depth from Intel RealSense contains more missing data, we use RGB-D images captured by RealSense [44] to evaluate depth completion performance. The resolution of depth and color images provided by the dataset is  $320 \times 240$ . We generate LR depth images through sampling a depth image with factors of 1/2, 1/4 and 1/8. Fig. 5 shows a comparison of original depth image and depth images enhanced by 2×SRC, 4×SRC and 8×SRC with our method. In order to compare our depth completion performance after depth super-resolution, the completed depth image ( $320 \times 240$ ) with the method of Zhang and Funkhouser [44] is used as a baseline. Peak signal-to-noise ratio (PSNR) compared to original depth image are also reported (higher is better) in the figure. The results indicate that our method performs well on 2×SRC, 4×SRC and 8×SRC depth images.

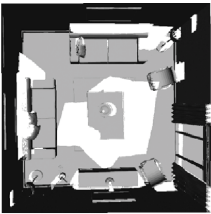
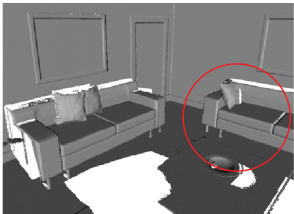
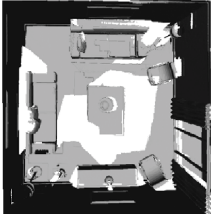
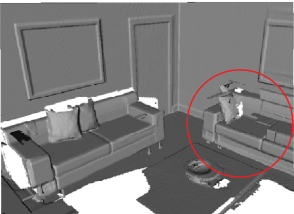
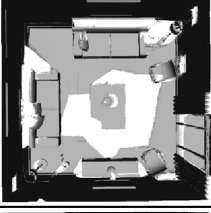
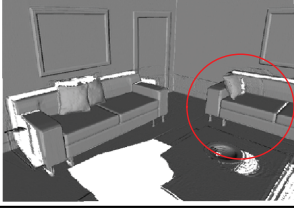
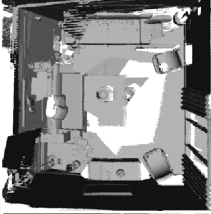
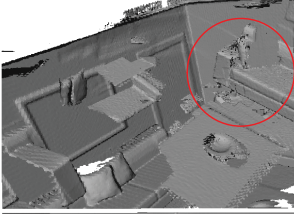
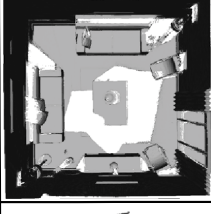
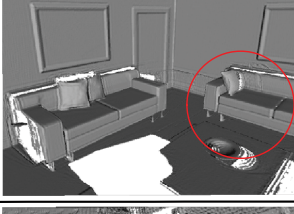
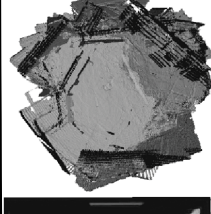
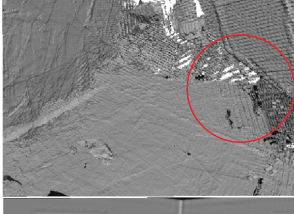

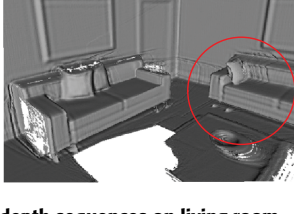
Fig. 6 gives a completion example of a holey depth image enhanced with the help of a blur color image from ScanNet dataset. The corresponding point clouds are calculated by the depth images, and the color information are just used for

display. It can be seen that, point cloud from raw depth image has many holes, while point cloud from our enhanced depth image is complete and accuracy.

## B. 3D RECONSTRUCTION EVALUATION

To evaluate the quantitative and qualitative performance of 3D reconstruction with enhanced depth sequence, we use the living room kt2 sequence (880 RGB-D images) of ICL-NUIM dataset. The ground truth (GT) depth sequence is noise-free with a resolution of  $640 \times 480$ . We generate LR depth sequences through sampling the GT depth images by factors of 1/2, 1/4 and 1/8, and enhance them by 2×SRC, 4×SRC and 8×SRC with our method. In our experiments, we only use depth images in 3D reconstruction after depth enhancement.

Table 4 reports the accuracy of camera pose estimation using the root mean square error (RMSE in centimeters) metric described by Handa *et al.* [50]. The camera poses are estimated through the proposed 3D reconstruction system. Both of the accuracies of SRC depth sequences with and

Inputs	Top View	General View
GT depth sequence 640 × 480		
1/2 LR depth sequence 320 × 240		
2 × SRC depth sequence 320 × 240 to 640 × 480		
1/4 LR depth sequence 160 × 120		
4 × SRC depth sequence 160 × 120 to 640 × 480		
1/8 LR depth sequence 80 × 60		
8 × SRC depth sequence 80 × 60 to 640 × 480		

**FIGURE 8.** 3D reconstruction results with different depth sequences on living room kt2 sequence from ICL-NUIM dataset [50]. Note that we do not complete large missing areas outside the camera’s field of view. Details of the differences are marked with red circles.

**TABLE 5.** Average computational time (seconds per frame) on living room kt2 sequence from ICL-NUIM dataset [50].

Operation	GPU		CPU	
	Depth super-resolution	Depth optimization	Information extraction	3D Reconstruction
Time (in seconds)	0.13	11.21	117	0.24

without outliers removing (OR) is reported. It can be clearly seen that the accuracy of camera pose estimation is greatly improved after depth enhancement with our method. Compared to  $4\times$ SRC depth sequence, RMSE for  $1/4$  LR depth sequence is very larger. The experiment with  $1/8$  LR depth sequence is failed due to severe camera tracking lost. The estimated camera trajectories compared with ground truth are also shown in Fig. 7. We can see an obvious camera pose error on  $1/2$  LR depth sequence, which is marked with a black circle in Fig. 7 (a). However, camera trajectories of  $2\times$ SRC,  $4\times$ SRC and  $8\times$ SRC depth sequences are all close to the ground truth.

Fig. 8 shows 3D scene models reconstructed with different depth sequences compared to the baseline. The first column represents the input depth sequences, and reconstructed 3D models are shown in the latter two columns. The first row shows the baseline model, which is reconstructed from GT depth sequence with a camera pose error of 0.10 cm RMSE. Note that we do not complete large missing areas outside the camera's field of view. There are large missing areas in the center of room models because both depth camera and color camera have not captured these regions. We have compared the reconstruction results with LR depth sequences and enhanced SRC depth sequences in top view and general view. Details of the differences are marked with red circles in general view. We can obviously see that 3D reconstruction using enhanced SRC depth sequences performs better than using LR depth sequences directly. Scene models reconstructed with  $1/2$  and  $1/4$  LR depth sequences looks good in top view, while the geometric details of them are inaccurate in general view. Due to serious camera pose error, scene model reconstructed with  $1/4$  LR depth sequence has some distortions and layering in general view. However, reconstruction results with our enhanced depth sequences are close to the baseline with enough geometric details.

### C. RUNTIME ANALYSIS

To evaluate the efficiency of our method, Table 5 shows the average computational time (seconds per frame) on living room kt2 sequences of ICL-NUIM dataset [50] for different procedures of our method. In our experiments, depth super-resolution and depth optimization are run on GeForce GTX 1080 Ti 11GB GPU, while information extraction and 3D reconstruction are run on Intel Core i7-4790 3.6GHz CPU. There exists a speed up of information extraction and depth optimization in the future.

### VI. CONCLUSION

In this paper, we propose a new depth SRC method implemented in deep learning framework, which can be applied to enhance low quality depth images captured by consumer

RGB-D cameras for 3D reconstruction. Experimental results demonstrate that our approach has better performance both on single depth image enhancement and 3D reconstruction. Our method can be extended for high-quality 3D reconstruction with higher resolution images. Furthermore, overcoming the noise in learned depth is the direction of our future work.

### REFERENCES

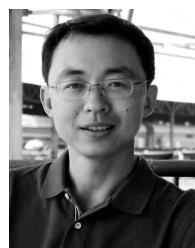
- [1] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, 2007.
- [2] M. Kiechle, S. Hawe, and M. Kleinsteuber, "A joint intensity and depth Co-sparse analysis model for depth map super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1545–1552.
- [3] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1397–1409, Jun. 2013.
- [4] D. Ferstl, C. Reinbacher, R. Ranftl, M. R  ther, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.
- [5] J. Park, H. Kim, Y. Tai, M. S. Brown, and I. S. Kweon, "High-quality depth map upsampling and completion for RGB-D cameras," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5559–5572, Dec. 2014.
- [6] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.
- [7] K. T. Gribbon and D. G. Bailey, "A novel approach to real-time bilinear interpolation," in *Proc. IEEE 2nd Int. Workshop Electron. Design, Test Appl.*, Jan. 2004, pp. 126–131.
- [8] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001.
- [9] J. Lu, D. Min, R. S. Pahwa, and M. N. Do, "A revisit to MRF-based depth map super-resolution and enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 985–988.
- [10] O. M. Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. Eur. Conf. Comput. Vis. Firenze, Italy: Springer*, 2012, pp. 71–84.
- [11] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1623–1630.
- [12] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 291–298.
- [13] E. Shabaninia, A. R. Naghsh-Nilchi, and S. Kasaei, "High-order Markov random field for single depth image super-resolution," *IET Comput. Vis.*, vol. 11, no. 8, pp. 683–690, Dec. 2017.
- [14] L.-F. Yu, S.-K. Yeung, Y.-W. Tai, and S. Lin, "Shading-based shape refinement of RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1415–1422.
- [15] Y. Han, J.-Y. Lee, and I. S. Kweon, "High quality shape from a single RGB-D image under uncalibrated natural illumination," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1617–1624.
- [16] B. Haefner, Y. Qu  au, T. M  llenhoff, and D. Cremers, "Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 164–174.
- [17] G. Riegler, D. Ferstl, M. R  ther, and H. Bischof. (2016). "A deep primal-dual network for guided depth super-resolution." [Online]. Available: <https://arxiv.org/abs/1607.08569>
- [18] G. Riegler, M. R  ther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Proc. Eur. Conf. Comput. Vis. Amsterdam, The Netherlands: Springer*, 2016, pp. 268–284.
- [19] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis. Amsterdam, The Netherlands: Springer*, 2016, pp. 353–369.



- [20] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGB-D images," in *Proc. Eur. Conf. Comput. Vis. Firenze, Italy: Springer*, 2012, pp. 746–760.
- [21] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 303–312.
- [22] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1625–1632.
- [23] Q. Chen and V. Koltun, "Fast MRF optimization with application to depth reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3914–3921.
- [24] J. Gautier, O. Le Meur, and C. Guillemot, "Depth-based image completion for view synthesis," in *Proc. 3DTV Conf., True Vis.-Capture, Transmiss. Display 3D Video (3DTV-CON)*, May 2011, pp. 1–4.
- [25] D. Doria and R. J. Radke, "Filling large holes in LiDAR data by inpainting depth gradients," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2012, pp. 65–72.
- [26] M. Ciotta and D. Andrououts, "Depth guided image completion for structure and texture synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1199–1203.
- [27] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2018, pp. 1–8.
- [28] Z. Chen, V. Badrinarayanan, G. Drozdov, and A. Rabinovich. (2018). "Estimating depth from RGB and sparse sensing." [Online]. Available: <https://arxiv.org/abs/1804.02771>
- [29] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2821–2830.
- [30] M. Bloesch, J. Czarowski, R. Clark, S. Leutenegger, and A. J. Davison, "CodeSLAM—Learning a compact, optimisable representation for dense visual SLAM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2018, pp. 1–10.
- [31] R. A. Newcombe et al., "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. 10th IEEE Int. Symp. Mixed Augmented Reality*, Oct. 2011, pp. 127–136.
- [32] S. Izadi et al., "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. UIST*, 2011, pp. 559–568.
- [33] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proc. 3DIM*, May/Jun. 2001, pp. 145–152.
- [34] T. Whelan, M. Kaess, J. J. Leonard, and J. McDonald, "Deformation-based loop closure for large scale dense RGB-D SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2013, pp. 548–555.
- [35] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2014, pp. 2100–2106.
- [36] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger, "Elasticfusion: Real-time dense SLAM and light source estimation," *Int. J. Robot. Res.*, vol. 35, no. 14, pp. 1697–1716, 2016.
- [37] V. A. Prisacariu et al. (2017). "InfiniTAM v3: A framework for large-scale 3D reconstruction with loop closure." [Online]. Available: <https://arxiv.org/abs/1708.00783>
- [38] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "BundleFusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 24.
- [39] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5556–5565.
- [40] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser. (2016). "3DMatch: Learning local geometric descriptors from RGB-D reconstructions," vol. 3. [Online]. Available: <https://arxiv.org/abs/1603.08182>
- [41] J.-W. Li, W. Gao, and Y.-H. Wu, "Elaborate scene reconstruction with a consumer depth camera," *Int. J. Autom. Comput.*, vol. 15, no. 4, pp. 443–453, 2018.
- [42] A. Kadambi, V. Taamazy, B. Shi, and R. Raskar, "Polarized 3D: High-quality depth sensing with polarization cues," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3370–3378.
- [43] W. Dong, Q. Wang, X. Wang, and H. Zha. (2018). "PSDF fusion: Probabilistic signed distance function for on-the-fly 3D data fusion and scene reconstruction." [Online]. Available: <https://arxiv.org/abs/1807.11034>
- [44] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 175–185.
- [45] O. Wasenmüller, M. Meyer, and D. Stricker, "CoRBS: Comprehensive RGB-D benchmark for SLAM using Kinect v2," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–7.
- [46] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5835–5843.
- [47] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [48] Y. Zhang et al., "Physically-based rendering for indoor scene understanding using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5057–5065.
- [49] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," in *Proc. 14th Annu. Conf. Comput. Graph. Interact. Techn.*, vol. 21, no. 4, 1987, pp. 163–169.
- [50] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. IEEE Int. Conf. Robot. Automat.*, May/Jun. 2014, pp. 1524–1531.
- [51] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–12.
- [52] D. Ferstl, M. Ruther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 513–521.
- [53] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade Meets Horn/Schunck: Combining local and global optic flow methods," *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 211–231, 2005.
- [54] W. Zhao, F. Zhao, D. Wang, and H. Lu, "Defocus blur detection via multi-stream bottom-top-bottom fully convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3080–3088.



images and simultaneous localization and mapping technology.



3D reconstruction from images and simultaneous localization and mapping technology.



**JIANWEI LI** received the B.Sc. degree in measurement and control technology and instrument and the M.Sc. degree in detection technology and automatic equipment from Beijing Jiaotong University, China, in 2008 and 2011, respectively. She is currently pursuing the Ph.D. degree in pattern recognition and intelligent system with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests include 3D reconstruction from

**WEI GAO** received the B.Sc. degree in computational mathematics, in 2002, the M.Sc. degree in pattern recognition and intelligent system from Shanxi University, in 2005, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, in 2008. Since 2008, he has been with the Robot Vision Group, National Laboratory of Pattern Recognition, where he is currently an Associate Professor. His research interests include

**YIHONG WU** received the Ph.D. degree from the Institute of Systems Science, Chinese Academy of Sciences, in 2001. She is currently a Professor with the Institute of Automation, Chinese Academy of Sciences. Her research interests include vision geometry, image matching, camera calibration, camera pose determination, simultaneous localization and mapping, and their applications.