Towards Fast and Accurate Real-World Depth Super-Resolution: Benchmark Dataset and Baseline

Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong,
Chunjie Zhang, Chunyu Lin, Meiqin Liu, Yao Zhao*
Institute of Information Science, Beijing Jiaotong University
Beijing Key Laboratory of Advanced Information Science and Network, Beijing, 100044, China
{lingzhihe, hongguang, l1feng, hhbai, rmcong, cjzhang, cylin, mqliu, yzhao}@bjtu.edu.cn

Abstract

Depth maps obtained by commercial depth sensors are always in low-resolution, making it difficult to be used in various computer vision tasks. Thus, depth map super-resolution (SR) is a practical and valuable task, which upscales the depth map into high-resolution (HR) space. However, limited by the lack of real-world paired low-resolution (LR) and HR depth maps, most existing methods use downsampling to obtain paired training samples. To this end, we first construct a large-scale dataset named "RGB-D-D", which can greatly promote the study of depth map SR and even more depth-related real-world tasks. The "D-D" in our dataset represents the paired LR and HR depth maps captured from mobile phone and Lucid Helios respectively ranging from indoor scenes to challenging outdoor scenes. Besides, we provide a fast depth map super-resolution (FDSR) baseline, in which the high-frequency component adaptively decomposed from RGB image to guide the depth map SR. Extensive experiments on existing public datasets demonstrate the effectiveness and efficiency of our network compared with the state-of-the-art methods. Moreover, for the real-world LR depth maps, our algorithm can produce more accurate HR depth maps with clearer boundaries and to some extent correct the depth value errors.

1. Introduction

As a supplement of the RGB modality, the depth map can provide useful depth information, which has been applied in bokeh rendering [25], AR modeling [26], face recognition [3], gesture recognition [27], *etc*. Meanwhile, the low-power depth sensors equipped on mobile consumer electronics, such as Huawei and Samsung, have been popular in our daily life. However, the resolution of depth maps cannot match the resolution of RGB images, limiting practical applications to some extent. Therefore, investigating the depth map SR is an effective solution for this issue. Furthermore, downsampling as a straightforward strategy has been widely used in the existing depth map SR algorithms [16, 23] to construct paired training samples. But the downsampling manner fails to comprehensively simulate the real-world complex correspondences between the LR and HR depth maps. To bridge this gap, we construct the first benchmark dataset towards the real-world depth map SR. Furthermore, to meet the actual application requirements, we provide a fast and accurate depth map SR baseline model.

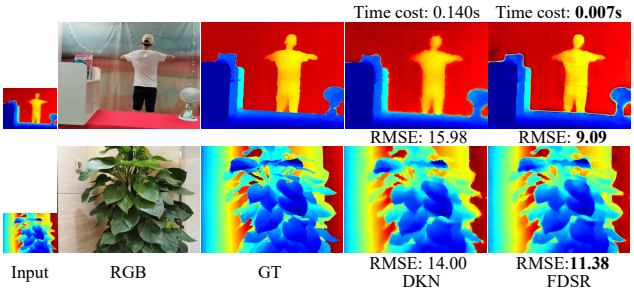


Figure 1. RGB-D-D dataset display and depth map SR results comparison. Depth map SR results given by DKN [16] and FDSR is shown in the last two columns. The quantitive results in terms of RMSE is shown blow each row (lower is better). The runing time of DKN [16] and FDSR is shown on the top of the figure.

The existing "RGB-D" depth map SR datasets [37, 5, 13] mainly focus on using a single HR depth map to generate paired LR and HR depth map correspondences through the downsampling strategy. In the real applications, the depth map SR task is more challenging and complicated because the real LR depth maps captured by depth sensors generally contain some noise even depth holes. Therefore, for the real scenes and real correspondences, we construct a large-scale

*Corresponding author: yzhao@bjtu.edu.cn

paired depth map SR dataset named "RGB-D-D", which includes 4811 paired samples ranging from indoor scenes to challenging outdoor scenes. The "D-D" in our dataset represents the paired LR and HR depth maps captured from the mobile phone (LR sensors) and Lucid Helios (HR sensors) [1], respectively. The dataset can offer two LR depth maps as input to evaluate the depth map SR task: LR depth map downsampled from the HR ground truth like previous research, and the raw LR depth map captured by LR sensor facing the real application scene. Besides, our dataset can contribute to many popular application scenarios of mobile phone and other depth-related tasks, such as portrait photography [25], object modeling [17], depth estimation [20], depth completion [44], *etc*.

Although numerous algorithms have been proposed for depth map SR and presented impressive performance, there are still some unsatisfactory points in detail preserving, computation complexity, and real-world application. Firstly, the sharp boundaries and elaborate details in the depth map SR are hard to recover especially when the scaling factor is large. Therefore, color image guided methods [41, 43] are introduced to solve this problem. Different from them, we design a high-frequency guided multi-scale dilated structure to introduce the color guidance in an image decomposition manner and exploit the contextual information under different receptive fields. Secondly, to be applied on the platforms of the mobile devices and embedded systems, the depth map SR algorithms should take into account both the efficiency and accuracy. Inspired by [6], we design a high-frequency layer in our network, where the high-frequency features from RGB image are only used as the guidance in the early stages of depth map reconstruction branch, and the low-frequency components are suppressed to reduce the parameters. Lastly, the existing methods use downsample operation to get paired HR and LR depth maps for training which fails to simulate the real correspondences between HR and LR depth maps. We use the paired depth maps in RGB-D-D dataset to train a model, which greatly improve the value accuracy and visual effects in the real-world depth map SR task.

Focused on the real-world applications and the practical demands, we construct a large-scale and real-world depth map SR benchmark dataset and provide a fast solution for the depth map SR task. The contributions are highlighted in the following aspects:

- We build the first and large-scale depth map SR benchmark dataset named RGB-D-D dataset[1], towards the real scenes and real correspondences. This dataset bridges the gap between theoretical research and real-world applications, and also flourishes the depth-related tasks in terms of benchmark dataset.
- We design a fast depth map super-resolution (FDSR) baseline, in which a high-frequency guided multi-scale structure is introduced to provide the frequency guidance and exploit the contextual information. Such decomposition strategy can improve the efficiency while retaining the reconstruction performance.
- Our network achieves the superior performance on the public datasets and our RGB-D-D benchmark dataset in terms of the speed and accuracy. Moreover, for the real-world depth map SR task, our algorithm can generate more accurate results with clearer boundaries and to some extent correct the value errors.

[1]Refer to http://mepro.bjtu.edu.cn/resource.html for the RGB-D-D dataset download link.

2. Related Work

In this section, we will briefly introduce the related benchmark datasets and algorithms for depth map SR.

Benchmark Datasets. There are various RGB-D datasets used for training and evaluating the depth map SR task. These datasets can be roughly divided into the synthetic datasets and the real-scene datasets. The synthetic datasets are built by synthetic computer graphic techniques and offer relatively high-quality data, such as New Tsukuba [32], Sintel [5] and ICL [11]. Limited by the discrepancy of virtual scenes and real scenes, some datasets towards real scene are constructed. Middlebury dataset [36, 35, 13, 34] provides a few samples containing high-quality and noise-free depth maps. Focusing on the indoor scenes, NYU v2 dataset [28] and SUN RGBD dataset [37] are built, where NYU v2 [28] includes 1449 RGB-D pairs, and SUN RGBD dataset [37] consists of three different RGB-D datasets (*i.e*., NYU v2 [28], B3DO [15] and SUN3D [2]). However, the mentioned datasets only face to the real scenes but fail to build the real correspondences between HR and LR depth maps, which are important for real-world depth map SR. To this end, we construct the first real-world depth map SR dataset, which not only faces the real scenes in practical applications, but also meets the real correspondences of LR and HR depth maps.

Algorithm Models. According to the characteristics of input data, depth map SR algorithms can be categorized into two categories: non RGB-guided depth map SR and RGB-guided depth map SR. Non RGB-guided methods [33, 41] only used LR depth maps as input to produce HR ones. These methods do not fully utilize the color information that may induce unsatisfying performance. By contrast, RGB-guided methods [16, 23, 40, 14] have become the mainstream of this task. For the unsupervised methods [31, 42, 9, 24], the depth map SR task is generally modeled as an optimization problem. As for the learning-based methods, the RGB information is used as features directly or used to convert and produce other types of guidance in-

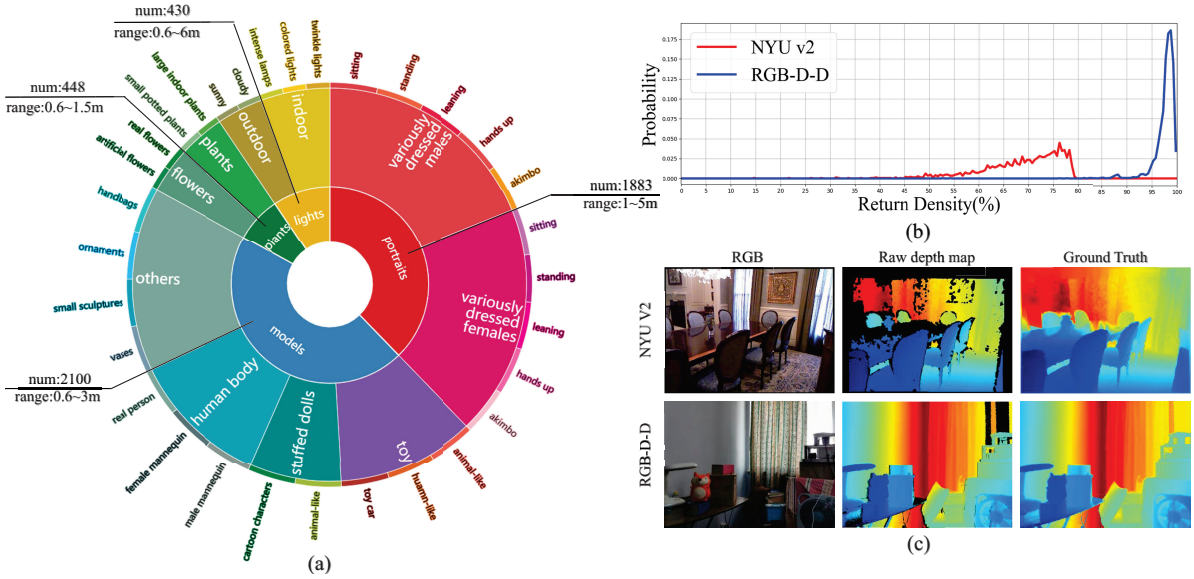


Figure 2. **Dataset statistic.** (a) The scenes and corresponding hierarchical content structure of RGB-D-D. The inner ring represents the classification of scenes. The legends express the number of samples and the depth range of corresponding subsets. (b) The fitted probability density curves of return density for the raw depth maps from NYU v2 [28] and RGB-D-D. (c) Examples of the raw depth map and ground truth from NYU v2 [28] and RGB-D-D. Black region indicates the missing depth value.

formation. In [19, 14, 40, 45, 16], the authors extract features of multi-level from RGB image to solve the depth map SR task. While [22, 30, 23] concern about the RGB images and try to use the HR RGB images to produce more useful features which achieves better performance. However these methods only put more efforts on the accuracy improvement which may lead to high computation complexity. Thus, we propose a fast depth map SR method to well balance efficiency and accuracy. Though, there are some works[10] use the real-scene LR depth maps to evaluate their methods, they fail to solve the problem at source. We are the first to use the paired depth maps in RGB-D-D dataset to train and simulate the correspondences which greatly improve the effectiveness of depth map SR.

3. RGB-D-D Dataset

We collect the first real-world depth map SR dataset which contains a total of 4811 RGB-D-D pairs. Each image pair contains the HR color images from mobile phone, the real-world LR depth maps captured by the low-power Time of Flight (ToF) camera on mobile phone, and the HR depth maps captured by industrial ToF camera.

3.1. Dataset Collection

Acquisition Devices. We use the Huawei P30 Pro to collect color images and LR depth maps. The Huawei P30 Pro has a 40 million pixels Quad RYYB sensor which can capture 3648×2736 HR color image, and a ToF camera with 240×180 resolution. The HR depth maps are captured by Helios ToF camera [1] produced by LUCID vision labs. They use the same depth acquisition principle which ensures the depth values captured by them are almost the same. Meanwhile, we guarantee little missing values of LR depth maps by limiting the farthest distance of backgrounds.

Data Processing. We calibrate the primary camera of the phone with the Helios ToF camera, and align them on the 640×480 resolution color image by the intrinsic and extrinsic parameters. Due to the different field of view (FOV) between them, the 640×480 raw point cloud of Helios is projected on the center area of the corresponding 640×480 resolution color image, and finally generate a dense and high-quality depth map which is smaller than 640×480. Then we crop it as the 512×384 HR depth map, which corresponds to the central 192×144 area of the LR depth map with the same scale variations. Towards the depth holes caused by the occlusion effect of projection processing and some low-reflection objects (such as glass surface and infrared absorbing surface), we firstly use the over-segmentation algorithm [29] to get plentiful boundary information of color image. And the colorization method [21] is used to fill holes in the supervision of boundary information. Clear depth edges can be filled according to the obvious border between foreground and background of color image, especially for strip-shaped holes on background caused by the occlusion effect of projection processing.

User Study. The four-round different people evaluation scores the filled HR depth maps to judge whether they could be the ground truth (the full mark of every round is 10). Af-

ter four-round evaluation, the filled depth maps with total scores beyond 35 become the part of ground truth samples. Besides, some of the eliminated depth maps have high evaluation scores (beyond 30) and few repairable defects. We further introduce the manual scribbles [39] on them to get convincing depth maps by means of user intervention. After the additional four-round blind selection, these remaining depth maps which achieve good visual effects on edge serve as the other part of ground truth samples.

3.2. Dataset Statistic

Real Scenes. We collect the paired "RGB-D-D" samples in various scenes as shown in Figure 2 (a). The RGB-D-D dataset is divided into four main categories: portraits, models, plants, and lights. The portraits category is mainly for the real applications of depth-of-field blur [25] in portrait photography. The traditional stereo camera cannot acquire satisfying depth information to simulate the large aperture in the repeated or weak texture scenes, so we collect plentiful samples containing the human body with pose variation as the foreground in different backgrounds. The models category can be used to optimize the edge of objects in depth maps when modeling the object by LR depth maps in sequential views [26]. We collect different depth maps of the relatively static objects on a rotating booth by discretely extracting video frames. The plants category which has dense branches and leaves contains a lot of structural and hierarchical details that LR depth sensors cannot capture. It is a challenge to the depth enhancement algorithms to infer these details by low-quality depth map and RGB. So we capture images containing various kinds of luxuriant plants and flowers in close range. In addition, strong indoor light sources and outdoor light have a great impact on the quality of depth maps, especially for the low-power sensors. Therefore, the lights category can be used to improve the quality of depth maps as close to the high-performance camera as possible, and to explore the effect of complex illumination environment to the color guided depth SR algorithms.

Real Correspondences. Actually there are only LR depth maps when applying the depth map SR algorithm in real-world applications. The relationship between LR and HR depth maps can not be simulated by traditional downsampling operations. Hence, it is necessary to capture the paired LR and HR depth maps by devices with different resolution. Both the LR depth map and the color image captured by the phone have aligned in 192×144 resolution. Meanwhile, the projected HR depth map from Helios [1] also align with color image in 512×384 resolution. The alignment of these three depth maps at different scales guarantee the real correspondences of LR and HR depth map.

High Quality. Because of the inevitable error values in the filled depth maps, the quality of preprocessed depth maps often degrade. Facing the practical applications, we capture the depth maps under the suggested distance to obtain depth maps with low rate of missing values and clear boundaries. Besides, the existence of many-to-one relationship in the process of raw point cloud projects to the image plane ensures that our raw depth maps contain dense depth values. All these can guarantee the high quality of our raw depth map. Figure 2 (b) shows the comparison between RGB-D-D and NYU v2 [28] on the statistic distribution of return density of which most of our raw HR depth maps are more than 90%. As shown in Figure 2 (c), benefiting by the higher return density, our ground truth has less depth values errors and better boundaries than NYU v2 [28].

4. Proposed Framework

4.1. Problem Formulation

Given a LR depth map $D_L \in \mathbb{R}^{M \times N \times 1}$ and the corresponding HR RGB image $G \in \mathbb{R}^{sM \times sN \times 3}$, the purpose of this work is to recover a HR depth map $D_H \in \mathbb{R}^{sM \times sN \times 3}$ with the guidance of G, where M and N denote the height and width of D_L, respectively, s is the scaling factor. We use bicubic interpolation to upscale D_L to HR space, which results in $D_U \in \mathbb{R}^{sM \times sN \times 1}$. As shown in Figure 3, we feed the paired D_U and G into our network to learn the nonlinear mapping from D_U to D_H through residual learning. Such process can be formulated as

$$D_H = D_U + \mathcal{F}(D_U, \mathcal{H}(G); \theta) \tag{1}$$

where $\mathcal{F}(\cdot)$ is a function to learn the residual mapping between D_U and D_H, the G is embedded into a high-frequency extractor $\mathcal{H}(\cdot)$ to provide high-frequency guidance for depth map SR, and θ is the learned weights set.

4.2. Network Architecture

Figure 3 outlines the whole architecture of our fast depth super-resolution network called FDSR, which consists of a high-frequency guidance branch (HFGB) and a multi-scale reconstruction branch (MSRB). Our framework progressively equip with four multi-scale reconstruction blocks to exploit the contextual information under different receptive fields in MSRB, meanwhile, the high-frequency guidance extracted from the HFGB is integrated with the multi-scale contextual information to enhance the ability of detail recovery for depth map SR. Finally, the comprehensive and discriminative reconstruction features are fed into a residual mapping function to generate HR depth map.

High-Frequency Guidance Branch. Motivated by previous methods [23, 45], we design a high-frequency layer (HFL) to adaptively highlight the high-frequency components and suppress the low-frequency component. Different from existing methods, we put more efforts on the following two aspects (1) a direct high-frequency decomposition

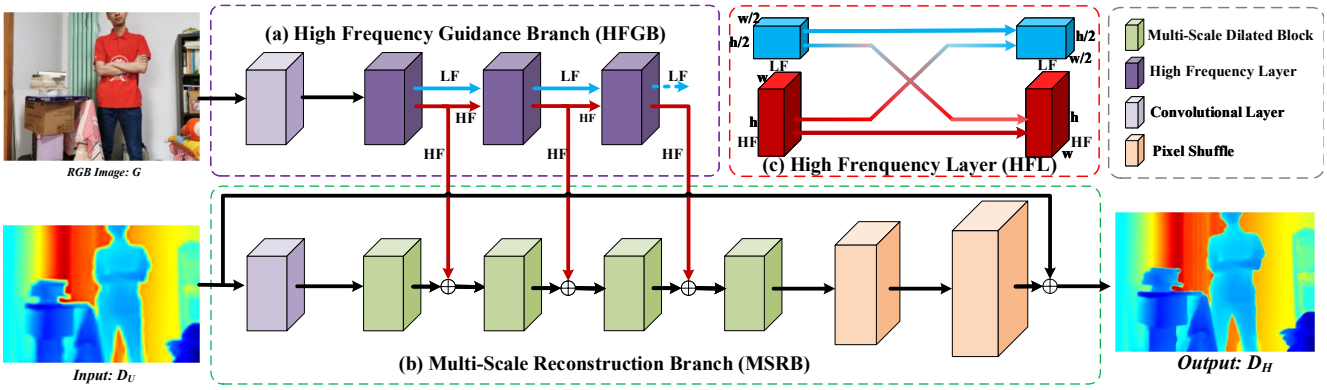


Figure 3. Overview of FDSR architecture. MSRB uses input depth map, high-frequency components extracted from RGB image to generate HR depth map. The blue lines and red lines in (a) indicate the low- and high-frequency components splited by HFL, respectively. (c) shows how the HFL works.

method is designed, where the octave convolution [6] is utilized to decompose the RGB features into high- and low-frequency components. (2) the high-frequency components are effectively used to guide depth map SR. Such design focuses on the useful high-frequency detail information to improve the performance, while it reduces the computation complexity due to the low-frequency components are not used in the MSRB. As shown in the Figure 3 (c), the previous high- and low-frequency feartures are embedded into the HFL to generate the current ones, which can be formulated as follows:

$$\begin{aligned} Y_{i+1}^{H} &= f(Y_i^H; W_i^{H\to H}) + up(f(Y_i^L; W_i^{L\to H}), 2) \\ Y_{i+1}^{L} &= f(Y_i^L; W_i^{L\to L}) + f(down(Y_i^H, 2); W_i^{H\to L}) \end{aligned} \tag{2}$$

where Y_{i+1}^H and Y_{i+1}^L denotes the high- and low-frequency features, respectively, the W is convolutional kernel, $f(Y; W)$ is a convolutional operation for Y with the kernel W, $up(Y, 2)$ is an upsample operation by a factor of 2 via nearest interpolation, and $down(Y, 2)$ represents $2\times$ downsampling for Y by using average pooling operation.

Multi-Scale Reconstruction Branch. This branch aims to progressively recover HR depth map through utilizing mulit-scale contextual information. We first use one 3×3 convolution layer to initial feature extraction. Then, to exploit the contextual information under different receptive fields, we combine two dilated convolutions to form a multi-scale dilated block (MSDB), and one convolution layer is used to integrate the concatenated features:

$$M_i(F_i) = W_i^j(\sum_{j\in K}(W_{M_i}^j * F_i + b_{M_i}^j)) + b_i^j \tag{3}$$

where M_i is the ith multi-scale block, F_i is the input of M_i, $K = \{1, 2\}$, $*$ denotes dilated convolution operation, W and b are parameters which the convolution layer should learn. Our MSDB not only enlarges the receptive field, but also enriches the diversity of convolutions, which results in an ensemble of convolutions with different receptive regions and dilation rates.

As for feature combination, three levels of high-frequency features extracted by HLFs are fused with different MSDBs respectively in the early stage of MSRB. What we have to emphasize is that, we split the output of each HLF and only use the extracted high-frequency component as guidance. Each stage of feature fusion F_{add}^i can be described as follows:

$$F^{i+1} = Y_i^H \oplus M_i(F_i) \tag{4}$$

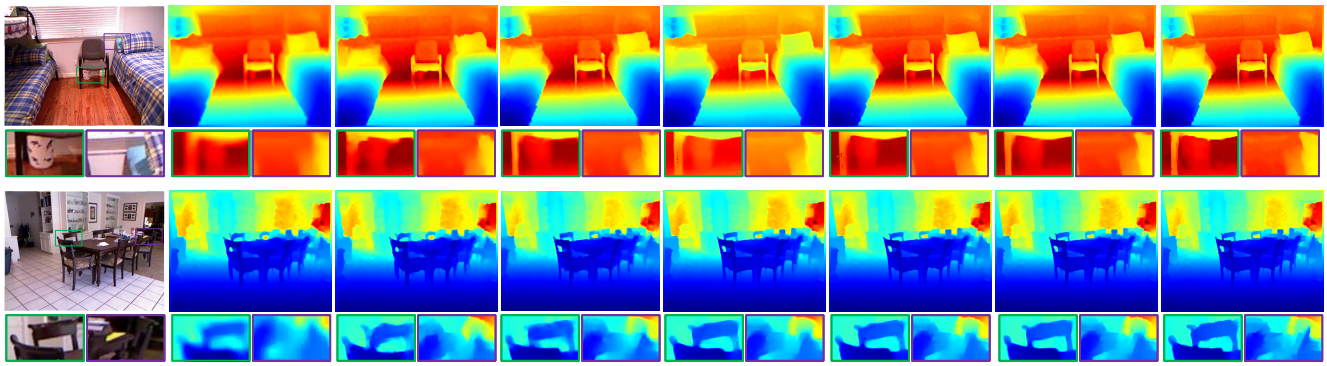
where $i = 1, 2, 3$, and $\oplus$ is concat operation. Then, the final fusion features F^4 is embedded into the last MSDB to generate reconstruction features.

In order to better apply FDSR in the platform of mobile devices and embedded systems, we mainly use two key operations to make our network faster. First, the designed HFL adaptively extracts the high-frequency features what we should focus on. Therefore, the parameters are reduced proportionately while ensuring effective performance. Second, in the stage of data preparation, we resample the input data by transform the color image to gray scale. Then we split the gray image and the input depth map to r^2 pieces of blocks of size $h/r \times w/r$ and stack them together. In the end, we use pixel shuffle operation to recover the HR depth map to the original size of $h \times w \times 1$.

4.3. Loss Function

We train our model by minimizing the L_1 norm between the output of our method $\mathcal{F}(\cdot)$ and ground truth as follows:

$$L(\hat{\mathcal{F}}, \mathcal{F}^{gt}) = \sum_P \|\mathcal{F}_p^{gt} - \hat{\mathcal{F}}_p\|_1 \tag{5}$$



(a) RGB	(b) SDF	(c) SVLRM	(d) DJFR	(e) FDKN	(f) DKN	(g) FDSR	(h) GT
Running Time (s)	25 (CPU)	0.10 (GPU)	0.01 (GPU)	0.01 (GPU)	0.17 (GPU)	0.01 (GPU)	

Figure 4. Visual comparison of $\times 8$ depth map SR results on NYU v2 [28]. (a) RGB images. (b) SDF [22]. (c) SVLRM [30]. (d) DJFR [23]. (e) FDKN [16]. (f) DKN [16]. (g) FDSR (trained on NYU v2). (h) GT. The GPU time is tested on a NVIDIA GTX TITAN XP GPU.

RMSE	Bicubic	MRF [7]	GF [12]	JBU [18]	TGV [8]	Park [31]	SDF [22]	FBS [4]	DMSG [14]	PAC [38]	DJF [22]	DJFR [23]	DKN [16]	FDKN [16]	FDSR
$\times 4$	8.16	7.84	7.32	4.07	4.98	5.21	5.27	4.29	3.02	2.39	3.54	3.38	1.62	1.86	**1.61**
$\times 8$	14.22	13.98	13.62	8.29	11.23	9.56	12.31	8.94	5.38	4.59	6.2	5.86	3.26	3.58	**3.18**
$\times 16$	22.32	22.2	22.03	13.35	28.13	18.1	19.24	14.59	9.17	8.09	10.21	10.11	6.51	6.96	**5.86**

Table 1. Comparisons with the state-of-the-art methods in terms of RMSE on NYU v2 [28]. The depth values are measured in centimeter.

where $\hat{\mathcal{F}}$ and $\mathcal{F}^{gt}$ denote the depth SR result and ground truth, respectively, $\|\cdot\|_1$ computes the L_1 norm, P is the set of all pixels and p represents a pixel in an image.

5. Experiments

5.1. Datasets and Implementation Details

To evaluate the performance of different methods, we conduct sufficient experiments on the public NYU v2 dataset [28] to and our real-world RGB-D-D dataset.

As for public dataset, we choose the widely used depth map SR dataset NYU v2 [28], and evaluate ours and other methods on it. Following [16], we sample 1000 RGB-D image pairs of size 640×480 from the NYU v2 dataset for training and the rest 449 image pairs for testing. As for RGB-D-D dataset, we randomly split 1586 portraits, 380 plants, 249 models for training and 297 portraits, 68 plants, 40 models for testing. Our FDSR is implemented in PyTorch on a PC with an NVIDIA GTX TITAN XP GPU. A MindSpore implementation version is also provided. Limited by the length of the paper, more details of experimental settings can be found in the supplemental materials.

5.2. Experiments on NYU v2 Dataset

As for training on NYU v2 dataset [28], we obtain the LR depth maps from ground truth by using bicubic downsampling operation. The initial learning rate is 0.0005 and reduce to half every 80k iterators and the training is stopped after 100 epochs since more epochs do not provide more improvement. We compare our FDSR with other methods with the scaling factors of 4, 8, 16. The quantitative results are shown in Table 1. It can be observed that our method achieves the best performance on NYU v2.

Percentage	Value Errors (in 10 m)			Edge Errors		
	$\times 4$	$\times 8$	$\times 16$	$\times 4$	$\times 8$	$\times 16$
SDF [22]	0.42	1.28	3.52	4.20	10.19	25.06
SVLRM [30]	1.08	2.56	5.76	6.04	24.28	49.26
DJF [22]	1.05	2.74	6.25	9.87	30.38	55.35
DJFR [23]	1.04	2.72	6.25	6.78	25.01	53.98
FDKN [16]	0.04	0.24	1.00	0.83	3.27	13.03
DKN [16]	0.05	0.20	1.10	0.95	2.95	13.78
FDSR	**0.04**	**0.18**	**0.69**	**0.78**	**2.60**	**9.44**

Table 2. Value errors and edge errors on NYU v2 [28].

To further analyze the robustness of our method, we conduct two extra experiments on NYU v2: (1) depth value errors to inflect the global depth map SR accuracy, (2) edge errors to measure the local accuracy. We report the value errors which is calculated by the percentage of value errors over 10% between ground truth and output. As for edge errors, we report the percentage of errors over 1.2% in the edge area. The details of calculation process for value errors and edge errors will be described in supplement material. Observing Table 2, FDSR has both less value errors and edge errors, which means our method produces more accurate results globally and locally.

As for qualitative results, we show the visual comparison for $\times 8$ depth map SR in Figure 4. The overall and details of the results demonstrate that the proposed method FDSR can

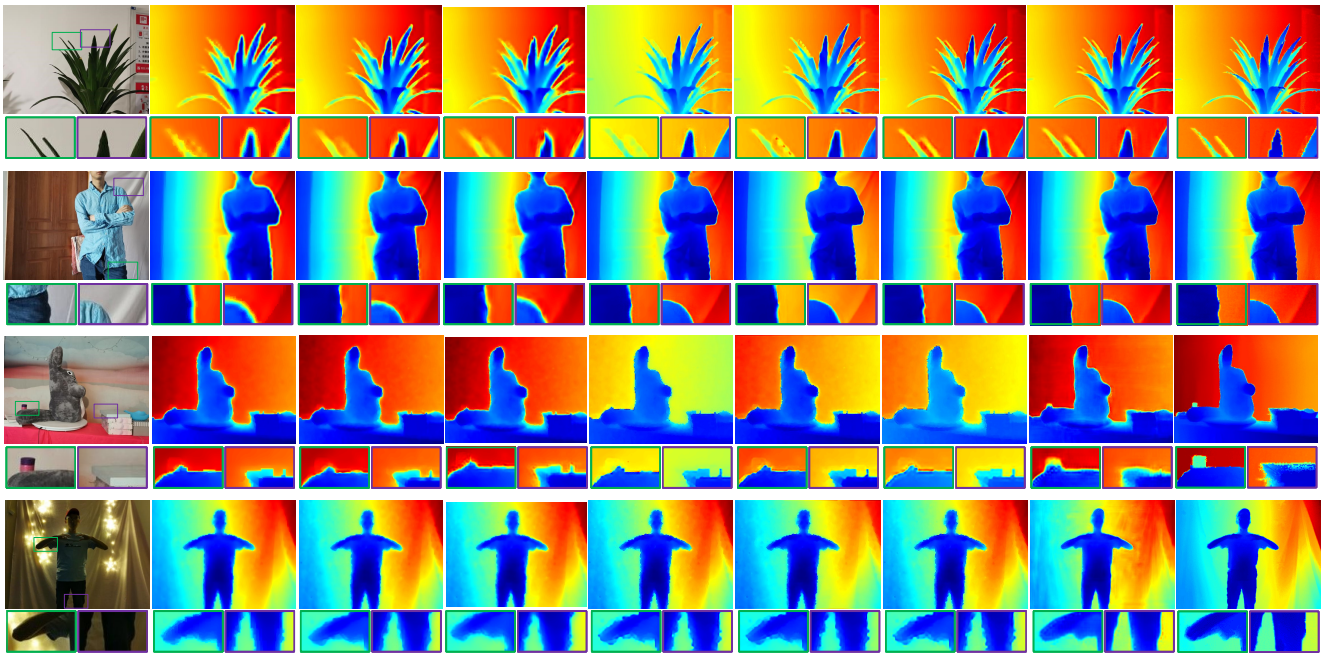


Figure 5. Visual comparison of $\times 8$ depth map SR results on RGB-D-D. The first two and last two rows are the results of FDSR^{+} and FDSR^{++} respectively (a) RGB images. (b) SDF [22]. (c) SVLRM [30]. (d) DJFR [23]. (e) FDKN [16]. (f) DKN [16]. (g) FDSR (trained on NYU v2 [28]). (h) FDSR^{+} / FDSR^{++} (Trained in downsampling manner / Trained in real-world manner on RGB-D-D). (i) GT.

RMSE	SDF [22]	SVLRM [30]	DJF [22]	DJFR [23]	FDKN [16]	DKN [16]	FDSR	FDSR^{+}
$\times 4$	2.00	3.39	3.41	3.35	1.18	1.30	1.16	**1.11**
$\times 8$	3.23	5.59	5.57	5.57	1.91	1.96	1.82	**1.71**
$\times 16$	5.16	8.28	8.15	7.99	3.41	3.42	3.06	**3.01**

Table 3. Quantitative depth map SR results on RGB-D-D. FDSR^{+} is trained in downsampling manner on RGB-D-D)

obtain more accurate depth map values. Our results show finer boundaries and more visual pleasant details without the texture-copy artifacts and extra noise introduced .

The running time is also shown in Figure 4. The size of input is 640×480. Our FDSR method achieves the comparable efficiency with DJFR [23] and FDKN [16] while the performance of FDSR is better than anyone of them.

5.3. Experiments on RGB-D-D Dataset

To verify the generalizability of RGB-D-D, we conduct sufficient experiments on it. The experiments on our dataset also further demonstrate the performance of our algorithm.

Testing without Retraining on RGB-D-D. Firstly, to make a fair comparison with other algorithms on our dataset, we conduct experiments among models trained on NYU v2 [28] without retraining. The quantitative results in terms of RMSE are shown in Table 3. The value errors and edge errors on RGB-D-D given by each algorithm are also reported in Table 5. The smaller RMSE value, value errors and edge errors of FDSR among evaluated methods demonstrate the accuracy and effectiveness of our algorithm. Figure 5 illustrates the qualitative results. The evaluated methods, SDF [22], SVLRM [30] and DJFR [23] cannot recover clear boundaries and fine details. Though DKN [16] and FDKN [16] produce clear boundaries, they have larger global errors and even some noises are brought in.

Training in Downsampling Manner on RGB-D-D. We retrain our models on the training set of RGB-D-D dataset to demonstrate the effectiveness of our dataset and our model. We use downsampled LR depth maps as input. When training on RGB-D-D dataset, the initial learning rate is 0.0005 and reduce to half every 40k iterators and the training for each scaling factor model is stopped after 40 epochs.

The results are appended in Table 3, Table 5 and Figure 5, which are obviously improved by training on our training data. Benefiting by the more clearly and sharper boundaries in our training and testing data, our model can achieve better performance, especially on the boundaries of

	SDF [22]	SVLRM [30]	DJF [22]	DJFR [23]	FDKN [16]	DKN [16]	FDSR	FDSR^{++}
RMSE	7.16	8.05	7.90	8.01	7.50	7.38	7.50	**5.49**
Value Errors	2.86	3.62	3.62	3.67	2.85	2.83	2.90	**1.71**
Edge Errors	52.78	51.87	50.56	52.28	51.73	51.90	51.89	**42.89**

Table 4. RMSE, value errors and edge errors of depth SR results. FDSR^{++} is trained on RGB-D-D in real-world training manner.

Percentage	Value Errors (in 3 m)			Edge Errors		
	×4	×8	×16	×4	×8	×16
SDF [22]	0.33	0.90	2.37	3.22	8.74	20.71
SVLRM [30]	0.80	2.11	4.58	5.08	15.18	34.30
DJF [22]	0.82	2.19	4.89	5.65	17.07	35.32
DJFR [23]	0.79	2.15	4.78	5.26	15.66	34.54
FDKN [16]	0.11	0.28	0.94	1.39	3.41	11.73
DKN [16]	0.14	0.33	1.54	2.11	3.55	12.93
FDSR	0.10	0.26	0.76	1.38	3.09	12.47
FDSR^{+}	**0.09**	**0.21**	**0.67**	**1.15**	**2.79**	**11.68**

Table 5. Value errors and edge errors of depth SR results on RGB-D-D. FDSR^{+} is trained in downsampling training manner.

objects and accuracy of depth values.

Training in Real-World Manner on RGB-D-D. To make full use of our proposed RGB-D-D dataset, we train our model on the training set by utilizing the LR depth maps as input. Before evaluating, the missing holes of raw LR depth maps are filled by the colorization method [21]. The size of the LR depth map is 192×144 and the target resolution is 512×384. The settings and training strategies are as same as we trained on HR depth maps of RGB-D-D. We test all the evaluated methods on the filled LR depth maps via using the existing $\times 4$ models and the results can be seen in Table 4. We append our results obtained by the model trained on the paired LR and HR depth maps. It can be observed that, all the evaluated methods have bad performance facing the real-world depth map SR task, which means the traditional downsample training strategy fails to model the real-correspondence between LR and HR depth maps. Observing the last two rows of Figure 5, after retraining FDSR on the paired LR and HR depth map in RGB-D-D dataset, the visual effects and value accuracy are greatly improved, which demonstrates that our dataset reflects the real-correspondences characteristics between LR and HR depth maps. Thus, the RGB-D-D dataset has great potential to promote the development of real-world depth map SR.

We also conduct experiments on the group of lights in our RGB-D-D dataset. It is a very challenging set of data, because the illumination intensity is complicated and the LR depth maps are in lower quality with bigger missing holes. Observing the last two rows in Figure 5, we obtain a better result with good boundaries, more accuracy depth values and more pleasant visual effects, while other algorithms fail to recover good HR depth maps.

5.4. Ablation Study

To demonstrate the effectiveness of the designed architecture of our depth map SR baseline, we conduct serveral ablation studies. For such an ablation study, the basic setup refers to the experiments above. The results in Table 6 clearly demonstrates that both the HFL and HFGB can be used to improve the performance of FDSR. What's more, the improvement of FDSR implies that the employing HFL components that HFGB to great extent.

Methods	NYU v2 [28]			RGB-D-D		
	×4	×8	×16	×4	×8	×16
w/o HFGB	2.02	3.90	7.58	1.16	1.88	3.47
w/o HFL	1.68	3.21	5.89	1.13	1.85	3.20
FDSR	**1.61**	**3.18**	**5.86**	**1.11**	**1.71**	**3.01**

Table 6. RMSE evaluation of HFL and HFGB.

6. Conclusion

Towards the real-world depth map SR, we build the first benchmark dataset which satisfy both real scene and real coorespondence. The dataset contains paired LR and HR depth maps in multiple scenarios, and contributes the completely new benchmark dataset for real-world depth map SR research. Furthermore, the "RGB-D-D" triples not only can complete the traditional depth-related tasks, such as depth estimation, depth completion, *etc*. but also have significant potential to promote the application of depth maps on portable intelligent electronics. We also provide a fast and accurate depth map SR baseline adaptively focusing on the high-frequency components of the guidance and suppress the low-frequency components. Our algorithm achieves the competitive performance on public datasets and our proposed dataset, what's more, it has an ability to cope with the task of real-world depth map SR.

Acknowledgements: This work was supported by the National Key Research and Development of China (No. 2018AAA0102100), the National Natural Science Foundation of China (No. U1936212, 61972028), the Beijing Natural Science Foundation (No. JQ20022), the Beijing Nova Program under Grant (No. Z201100006820016) and the CAAI-Huawei MindSpore Open Fund.

References

[1] Helios Time of Flight(ToF) camera. https://thinklucid.com/product/helios-time-of-flight-imx556/. 2, 3, 4

[2] Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1625–1632, 2013. 2

[3] Andrea F Abate, Michele Nappi, Daniel Riccio, and Gabriele Sabatino. 2d and 3d face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007. 1

[4] Jonathan T. Barron and Ben Poole. The fast bilateral solver. 2016. 6

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 611–625. Springer, 2012. 1, 2

[6] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3435–3444, 2019. 2, 5

[7] James Diebel and Sebastian Thrun. An application of markov random fields to range sensing. In *Advances in neural information processing systems*, pages 291–298, 2006. 6

[8] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Ruether, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 6

[9] 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 (ICCV)*, pages 993–1000, 2013. 2

[10] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing*, 28(5):2545–2557, 2018. 3

[11] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014. 2

[12] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 6

[13] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 1, 2

[14] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 353–369. Springer, 2016. 2, 3, 6

[15] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer depth cameras for computer vision*, pages 141–165. Springer, 2013. 2

[16] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, pages 1–22, 2020. 1, 2, 3, 6, 7, 8

[17] Kalin Kolev, Petri Tanskanen, Pablo Speciale, and Marc Pollefeys. Turning mobile phones into 3d scanners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3946–3953, 2014. 2

[18] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics*, 26(3):96, 2007. 6

[19] HyeokHyen Kwon, Yu-Wing Tai, and Stephen Lin. Data-driven depth map refinement via multi-scale sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 159–167, 2015. 3

[20] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 2

[21] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, pages 689–694. 2004. 3, 8

[22] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–169. Springer, 2016. 3, 6, 7, 8

[23] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1909–1923, 2019. 1, 2, 3, 4, 6, 7, 8

[24] Yanjie Li, Tianfan Xue, Lifeng Sun, and Jianzhuang Liu. Joint example-based depth map super-resolution. In *2012 IEEE International Conference on Multimedia and Expo*, pages 152–157. IEEE, 2012. 2

[25] Chenchi Luo, Yingmao Li, Kaimo Lin, George Chen, Seok-Jun Lee, Jihwan Choi, Youngjun Francis Yoo, and Michael O. Polley. Wavelet synthesis net for disparity estimation to synthesize dslr calibre bokeh effect on smartphones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 4

[26] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2633–2651, 2015. 1, 4

[27] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. Hand gesture recognition with 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–7, 2015. 1

[28] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from

rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 2, 3, 4, 6, 7, 8

[29] Hieu Tat Nguyen, Marcel Worring, and Rein Van Den Boomgaard. Watersnakes: Energy-driven watershed segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3):330–342, 2003. 3

[30] Jinshan Pan, Jiangxin Dong, Jimmy S Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang. Spatially variant linear representation models for joint filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1702–1711, 2019. 3, 6, 7, 8

[31] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon. High quality depth map upsampling for 3d-tof cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1623–1630. IEEE, 2011. 2, 6

[32] Martin Peris, Sara Martull, Atsuto Maki, Yasuhiro Ohkawa, and Kazuhiro Fukui. Towards a simulation driven stereo vision system. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1038–1042. IEEE, 2012. 2

[33] Gernot Riegler, Matthias Rüther, and Horst Bischof. Atgv-net: Accurate depth super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284. Springer, 2016. 2

[34] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014. 2

[35] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2

[36] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003. 2

[37] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1, 2

[38] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[39] Oliver Wang, Manuel Lang, Matthias Frei, Alexander Hornung, Aljoscha Smolic, and Markus Gross. Stereobrush: interactive 2d to 3d conversion using discontinuous warps. In *Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling*, pages 47–54, 2011. 4

[40] Yang Wen, Bin Sheng, Ping Li, Weiyao Lin, and David Dagan Feng. Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution. *IEEE Transactions on Image Processing*, 28(2):994–1006, 2018. 2, 3

[41] Jun Xie, Rogerio Schmidt Feris, and Ming-Ting Sun. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing*, 25(1):428–438, 2015. 2

[42] Jingyu Yang, Xinchen Ye, Kun Li, and Chunping Hou. Depth recovery using an adaptive color-guided auto-regressive model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 158–171. Springer, 2012. 2

[43] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[44] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 175–185, 2018. 2

[45] Yifan Zuo, Qiang Wu, Yuming Fang, Ping An, Liqin Huang, and Zhifeng Chen. Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):297–306, 2019. 3, 4