IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 9, 2021, accepted April 25, 2021, date of publication April 30, 2021, date of current version May 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076853

Depth Map Super-Resolution Using Guided Deformable Convolution

JOON-YEON KIM, SEOWON JI, SEUNG-JIN BAEK, SEUNG-WON JUNG, (Senior Member, IEEE), AND SUNG-JEA KO, (Fellow, IEEE)
School of Electrical Engineering, Korea University, Seoul 02841, South Korea

Corresponding author: Seung-Jin Baek (sjinbaek@korea.ac.kr)

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2017-0-00250, Intelligent Defense Boundary Surveillance Technology Using Collaborative Reinforced Learning of Embedded Edge Camera and Image Analysis).

ABSTRACT Depth maps acquired by low-cost sensors have low spatial resolution, which restricts their usefulness in many image processing and computer vision tasks. To increase the spatial resolution of the depth map, most state-of-the-art depth map super-resolution methods based on deep learning extract the features from a high-resolution guidance image and concatenate them with the features from the depth map. However, such simple concatenation can transfer unnecessary textures, known as texture copying artifacts, of the guidance image to the depth map. To address this problem, we propose a novel depth map super-resolution method using guided deformable convolution. Unlike standard deformable convolution, guided deformable convolution obtains 2D kernel offsets of the depth features from the guidance features. Because the guidance features are not explicitly concatenated with the depth features but are used only to determine the kernel offsets for the depth features, the proposed method can significantly alleviate the texture copying artifacts in the resultant depth map. Experimental results show that the proposed method outperforms the state-of-the-art methods in terms of qualitative and quantitative evaluations.

INDEX TERMS Convolutional neural network, depth map, super-resolution.

I. INTRODUCTION

Depth maps are essential in various vision applications such as 3D reconstruction, virtual reality, and autonomous driving. In general, the depth map of a scene can be obtained using a passive or active method. The passive method [1] mainly obtains the depth map by estimating the correspondence between two or more images from different viewpoints. However, accurately identifying points of correspondence in textureless and occluded regions is fundamentally difficult, requiring a more robust solution for depth map acquisition.

The active method offers such a solution because it can robustly acquire the depth map from sensors such as time-of-flight (ToF) [2] or structured light [3] cameras. Although the depth map can be captured in real-time by these depth sensors, the low spatial resolution of the depth map can limit its applicability to various vision tasks. Therefore, it is necessary to increase the resolution of the depth map to facilitate the practical use of the active method.

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

The simplest approach to generate a high-resolution (HR) depth map from a low-resolution (LR) depth map is to apply interpolation, such as bilinear or bicubic interpolation, to the depth map. Despite their advantage of being simple, these methods cannot render sharp object boundaries. To address this problem, advanced super-resolution (SR) techniques, including optimization-based [4], [5], and learning-based methods [6], [7], have been proposed in the literature. In addition, because most commodity depth sensors are embodied with HR color sensors and because many applications, such as image refocusing and depth image-based rendering, require the color and depth images to have the same resolution, SR techniques that employ the corresponding HR color image as guidance have also received significant attention [8]–[15].

Recently, owing to the success of deep learning, several color-guided depth map SR techniques [16]–[21] using convolutional neural networks (CNNs) have been proposed. One of the common characteristics of these techniques is to extract the features from the depth map and guidance image separately and then concatenate them such that the depth decoding

or upsampling layers can inherit the HR details from the guidance features. However, it is difficult to control the influence of the guidance features after the two features have been concatenated. Excessive use of the guidance features on flat surfaces could result in unnecessary HR texture details being transferred from the guidance image to the depth map, which is known as texture copying artifacts.

To overcome the aforementioned problem, this paper proposes a novel SR network that uses guided deformable convolution. We adopt the deformable convolution [22] that augments the convolution kernel grid using learnable spatial offsets. Unlike the standard deformable convolution, the proposed network learns the spatial offsets for convolution of the depth features from the features of the HR guidance image. In other words, the proposed network does not directly use the features obtained from the HR guidance image when extracting and refining the features of the depth map. This approach enables us to obtain HR depth map without texture copying artifacts. The experimental results show that the proposed method outperforms state-of-the-art methods.

The remainder of this paper is organized as follows. We review related works in Section II. In Section III, we explain the proposed depth map SR method in detail. The experimental results are presented in Section IV. In Section V, we conclude this paper.

II. RELATED WORKS

Depending on the input data, depth map SR methods can be divided into two categories: single depth map SR and color-guided depth map SR. In this section, we briefly review conventional methods in each category. Furthermore, we introduce the deformable convolution, which is adopted by the proposed method.

A. SINGLE DEPTH MAP SR

SR techniques for color images have been widely researched in recent years. However, because the depth map has different characteristics from the color image, applying color image SR methods directly to an LR depth map could result in a suboptimal outcome. Based on the observation that real-world scenes exhibit repetitions of geometric primitives and objects with symmetries, Hornacek *et al.* [23] proposed a depth map SR algorithm that exploits the scene self-similarity. Some methods addressed the depth map SR as a Markov random field (MRF) optimization problem. Mac Aodha *et al.* [4] increased the resolution of the depth map by finding the appropriate HR candidate patch in the collected database by solving the MRF labeling problem. Xie *et al.* [5] constructed an HR edge map from LR depth maps using MRF optimization. Then, the HR depth map was obtained by employing a modified joint bilateral filter using the HR edge map as guidance.

Dictionary learning has also been employed to address the SR problem associated with single depth maps. Ferstl *et al.* [6] estimated edge priors from a given LR depth map and a learned dictionary. Then, the estimated edge prior was utilized to guide the regularization term when solving the variational SR problem. Xie *et al.* [7] proposed a coupled dictionary learning method with locality coordinate constraints to upsample the LR depth map.

B. COLOR-GUIDED DEPTH MAP SR

The HR color image of the same scene can easily be acquired with the depth map. In particular, most commodity depth sensors are already embodied with an HR color sensor, and thus one can expect that the HR color image can be used to upsample the LR depth map as a guidance. To this end, Kopf *et al.* [8] proposed a joint bilateral filter to produce a more precise HR map by obtaining a range kernel from the HR guidance image. Yang *et al.* [9] iteratively applied a joint bilateral filter to refine the depth map. Based on the fact that discontinuities in the depth map and the color image often co-occur, Diebel and Thrun [10] employed an MRF to integrate both data sources. Liu *et al.* [11] utilized the geodesic distance to upsample the LR depth map by using a registered HR color image. Jung and Choi [12] trained a classifier to select an effective upsampling filter for each depth pixel. Lu and Forsyth [14] segmented the HR color image and used the smoothing methods to determine the depth values in each color segment. Kwon *et al.* [15] proposed a refinement method based on a data-driven depth map by using the multi-scale dictionary learning.

Recently, CNNs were applied to various low-level vision tasks and their performance has been impressive. Hui *et al.* [16] first applied a deep CNN to color-guided depth map SR and proposed a multi-scale guided convolutional network (MSG-Net). MSG-Net extracts the HR features via the intensity branch and complements the LR depth features in the depth branch by using a multi-scale fusion strategy. Ni *et al.* [17] proposed a dual-stream CNN to upsample the LR depth map. They used an edge map inferred from the HR color image as network input to learn the relationship between the depth map and edge map. Zuo *et al.* [18] proposed a deep network with global and local residual learning to progressively upsample the LR depth map. In addition, batch normalization layers were used to improve the performance of depth map denoising. Inspired by the residual U-Net architecture [24], Guo *et al.* [19] proposed DepthSR-Net for depth map SR. DepthSR-Net learns the residual between the interpolated depth map and the ground truth HR depth map by using rich hierarchical features extracted from the network. Wen *et al.* [20] introduced a coarse-to-fine CNN to approximate the ideal filter for depth map SR. Li *et al.* [21] proposed a recumbent Y network (RYNet) for depth map SR. They built the network based on the residual channel attention blocks and utilized spatial attention based feature fusion blocks to suppress the texture copying and depth bleeding artifacts. Despite the potential benefits, the aforementioned methods still suffer from the texture copying artifacts owing to the inappropriate use of the guidance features. Accordingly, we aimed to solve this problem by designing a novel module that uses the guidance features more effectively.

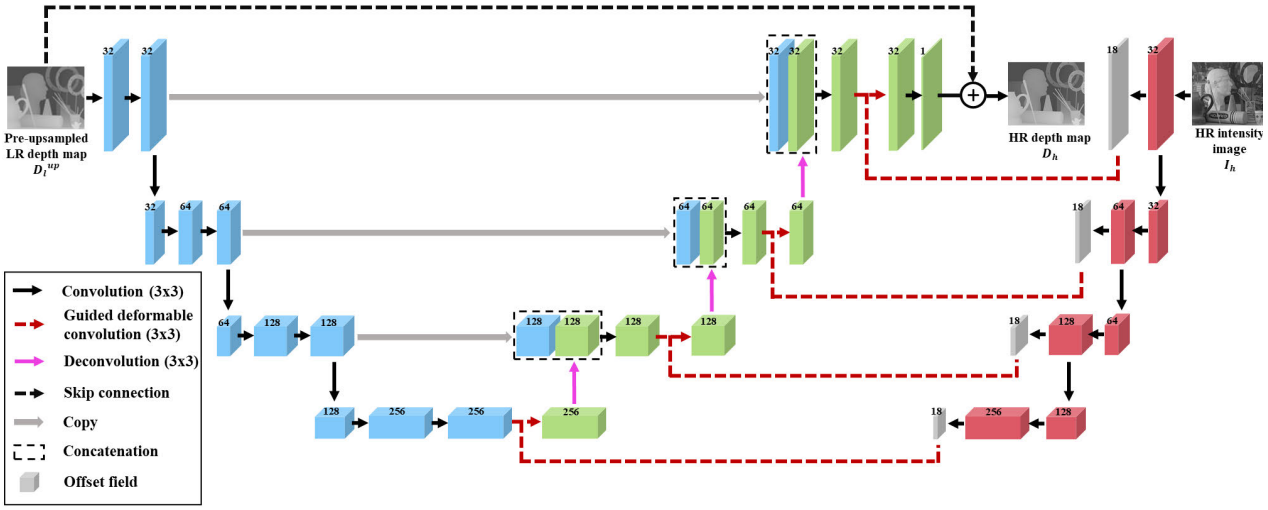


FIGURE 1. Overall architecture of the proposed color-guided depth map SR network. Our network consists of two encoders and a decoder. The two encoders extract the features from the HR intensity image and the pre-upsampled LR depth map, respectively. The HR depth map is generated by the decoder. The red, blue, and green boxes represent the feature maps of the HR guidance encoder, depth encoder, and depth decoder, respectively, and the number of channels is denoted above each box.

C. DEFORMABLE CONVOLUTION

Because the plain convolution module performs the convolution operation using fixed geometric structures, it is difficult to accommodate various geometric transformations of objects when using a CNN. Accordingly, the performance of a CNN with respect to visual recognition tasks tends to be limited. To enhance the modeling capability of CNNs, Dai *et al.* [22] proposed deformable convolution by introducing 2D kernel offsets to the regular sampling locations of the standard convolution. The offsets are learned through additional convolution layers without extra supervision. The deformable convolution module improved the performance of high-level vision tasks such as object detection [25] and semantic segmentation [26] when used as a replacement for the plain convolution module of existing CNNs. Deformable convolution was also used to align multiple frames for video restoration tasks [27], [28]. We thus surmised that the performance of depth SR can be improved by appropriately adopting deformable convolution to a depth SR network.

III. PROPOSED METHOD

A. PROBLEM FORMULATION

The objective of color-guided depth SR is to construct the HR depth map D_h, given the LR depth map D_l and the HR color intensity image I_h of the same scene. We pre-upsample D_l to the desired resolution by using the bicubic interpolation and obtain ${D_l}^{up}$, following Guo's method [19]. Then, the pre-upsampled depth map ${D_l}^{up}$ is used as input to the network. The final HR depth map D_h is obtained as:

$$D_h = F({D_l}^{up}, I_h; \theta) + {D_l}^{up}, \tag{1}$$

where F represents the nonlinear residual mapping function that estimates the residual between ${D_l}^{up}$ and D_h, and θ is a set of network parameters.

B. OVERALL NETWORK ARCHITECTURE

The overall network architecture of the proposed color-guided depth map SR is illustrated in Fig. 1. Similar to [16], [19], the proposed network consists of a depth encoder, guidance encoder, and depth decoder. The depth and guidance encoders extract the features from ${D_l}^{up}$ and I_h, respectively. Following the U-Net principle [24], the features of the depth encoder are concatenated with their corresponding features of the depth decoder. Then, before applying deconvolution, we insert the proposed guided deformable convolution module such that the features from the guidance encoder can facilitate depth feature refinement. After feature refinement, the decoder generates the residual [29], and thus D_h is obtained by adding the residual to ${D_l}^{up}$.

C. GUIDED DEFORMABLE CONVOLUTION MODULE

Fig. 2(a) illustrates the principles of deformable convolution [22]. First, 2D kernel offsets are obtained by applying a convolutional layer to the same input feature map. Specifically, these 2D kernel offsets are obtained as follows:

$$f^{OI} = M\left(f^{I}\right), \tag{2}$$

where f^I represents the input feature and M is a general function consisting of convolution layers. In our work, we used a single 3×3 convolution layer. The f^{OI} contains 2D kernel offsets for every position, and thus it requires 18 channels for 3×3 convolution. Given f^{OI}, the output feature f^O is obtained

as

$$f^O(\mathbf{x}) = \delta\left(\sum_{\mathbf{r}\in\Omega_R} W(\mathbf{r}) \cdot f^I\left(\mathbf{x}+\mathbf{r}+f^{OI}(\mathbf{x},\mathbf{r})\right)\right), \quad (3)$$

where Ω_R represents the regular grid for the convolution kernel, and δ represents the rectified linear unit (ReLU) as the activation function. $f^{OI}(\mathbf{x},\mathbf{r})$ denotes the 2D offset of element $\mathbf{r}$ in Ω_R at position $\mathbf{x}$.

Fig. 2(b) illustrates the proposed guided deformable convolution. Similar to the deformable convolution, 2D kernel offsets are obtained for convolution of the input feature. However, these offsets are obtained from the guidance feature as follows:

$$f^{OG} = M\left(f^G\right), \quad (4)$$

where f^G represents the guidance feature and f^{OG} is the offset field for the deformable convolution of f^I. The guided deformable convolution is operated as the deformable convolution in (3) with f^{OG} in the place of f^{OI}. Through the guided deformable convolution, the input feature is refined using rich HR information inherent in the guidance feature. Also, since the guidance feature is not directly concatenated to the depth feature and is only used to transform the kernel, we can effectively prevent the texture copying problem.

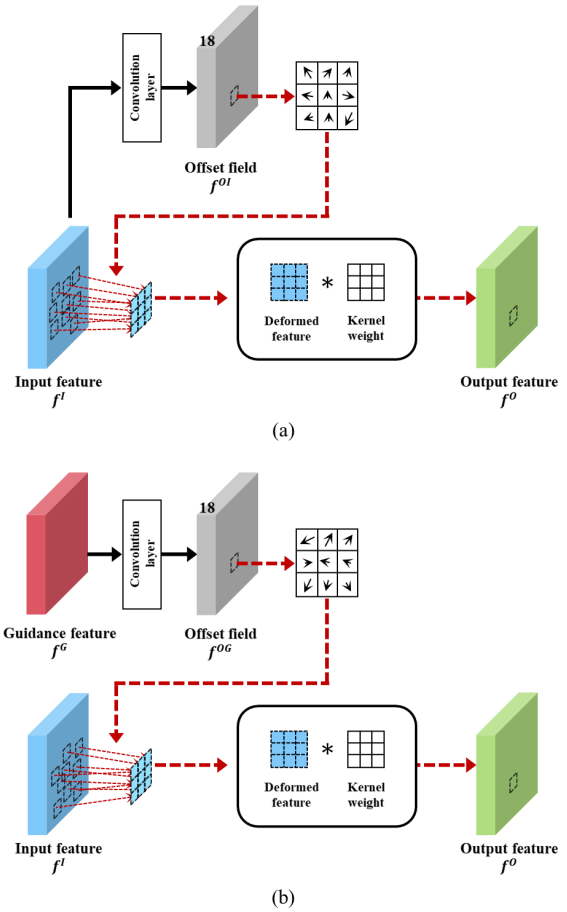


FIGURE 2. Comparison of (a) deformable convolution [22] and (b) guided deformable convolution. Unlike the deformable convolution module, which obtains the offset field from the same input feature map, the guided deformable convolution module generates an offset field from the guidance feature map.

D. LOSS FUNCTION

Following the previous color-guided depth SR methods [16]–[19], we minimize the mean squared error (MSE) loss to train the proposed network as follows:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left\| F((D_l{}^{up})_i, (I_h)_i; \theta) + (D_l{}^{up})_i - (D_h)_i \right\|^2, \quad (5)$$

where N represents the total number of training samples. We noticed that the performance of the network improved slightly when the MSE loss was utilized instead of the l_1 loss.

E. IMPLEMENTATION DETAILS

We used 58 RGB-D images from the MPI Sintel depth dataset [30] and 34 RGB-D images (6, 10, and 18 images from the 2001, 2006, and 2014 datasets, respectively) from the Middlebury dataset [1], [31], [32]. Among them, 82 images were used for training, and the remainder was used for validation. From the dataset, we sampled 96 × 96 patches with a stride of 48 for training. To obtain the LR depth maps, we downsampled the collected depth patches using the bicubic interpolation. Rotation and flip were randomly applied to the training dataset for its augmentation. To train the proposed network, we used the Adam optimizer [33] with a batch size of 64. The ReLU was used as an activation function after the convolution layers except for the layers generating the offset fields and residual map. We initialized the filter weights for the offset layer to zero and the remaining filter weights were initialized by He initialization method [34]. We used a constant learning rate of 10^{-4} in the training process.

IV. EXPERIMENTAL RESULTS

Our experiments were conducted with the PyTorch framework [35] on a PC with an Intel(R) Core(TM) i7-8700K CPU @3.70GHz and an NVIDIA TITAN Xp GPU. We compared the proposed method with the baseline bicubic interpolation, the CNNs for the single image SR (SRCNN [36] and SAN [37]), and the recent color-guided depth map SR networks (MSG-Net [16], MFR-SR [18], DepthSR-Net [19], and RYNet [21]). We evaluated the performance of the conventional and proposed methods by conducting qualitative and quantitative comparisons for noise-free and noisy test data with various scaling factors (i.e., 2×, 4×, 8×, 16×). To generate noisy test data, we added Gaussian noise with mean = 0 and variance = 25 to the noise-free LR depth map. Gaussian noise was also added to the LR depth map during the training process.

The results of DepthSR-Net on the noise-free test data were obtained by directly applying the author-provided trained model to the test images. For the noisy test data, the DepthSR-Net model was re-trained. We implemented

SRCNN, MSG-Net, MFR-SR, and RYNet. The former two networks, SRCNN and MSG-Net, were trained by applying the same training strategy that was used to train the proposed method. MFR-SR was trained using a strategy as described in [18]. The batch size was set to 32 for training the RYNet due to the GPU memory. We trained the SAN using the source code provided by the authors, and evaluated the performance only for the scaling factors 8 and 16 because of the limitation of GPU memory.

A. ABLATION STUDY

To demonstrate the effectiveness of the guided deformable convolution in the depth map SR, we performed three comparative experiments by changing the feature map refinement modules shown in Fig. 1. First, as shown in Fig. 3(a), the depth feature map is directly refined by performing the conventional deformable convolution [22], which is referred to as "Case 1". This is equivalent to a single depth map SR because it does not use the features extracted from the guidance image. Second, as shown in Fig. 3(b), the depth feature is extracted after concatenating the depth and guidance feature maps, and the extracted feature map is then refined using the conventional deformable convolution, which is referred to as "Case 2". Third, as shown in Fig. 3(c), to show the effectiveness of deformable convolution in depth SR, "Case 3" performs feature map refinement through the self-attention mechanism. Among the various attention modules, we utilized convolutional block attention module (CBAM) [38] which sequentially applies channel and spatial attention modules.

We evaluated the performance of the proposed method and the aforementioned methods for the test data. Table 1 shows the average performance of each method in terms of the peak signal-to-noise ratio (PSNR) and the root mean square error (RMSE). The best and second-best results are bold-faced and underlined, respectively. From the performance gap between "Case 1" and "Case 2", it can be seen that the use of guidance, despite simple concatenation, enhances the performance of depth map SR using the conventional deformable convolution. From a comparison of the proposed method with "Case 2", we can confirm that the performance gap is occurred according to the different use of guidance feature and deformable convolution. Moreover, comparing the performances of "Case 2" and "Case 3", we can see that the feature map refinement through attention module is inefficient compared with deformable convolution. The results of the proposed method demonstrate that the utilization of guidance features by the proposed guided deformable convolution is superior to the other approaches.

TABLE 1. Average performance of each method on the test data in terms of PSNR(dB)/RMSE.

Methods	Scaling factor	
	8×	16×
Case 1	46.90 / 1.24	41.14 / 2.41
Case 2	47.05 / 1.21	42.29 / 2.09
Case 3	46.70 / 1.27	41.84 / 2.21
Proposed method	**47.25 / 1.19**	**42.43 / 2.06**

B. QUANTITATIVE EVALUATION

Fig. 4 shows the average PSNR and processing time of the color-guided depth SR methods for the color images with the size 1312 × 1072 and the scaling factor of 16. The radius of circle is proportional to the number of network parameters which are indicated below each circle. The RYNet and the proposed method show the best and second-best average PSNRs of 42.54 dB and 42.43 dB, respectively. Although RYNet shows a slightly higher PSNR than the proposed method, it requires approximately ×18.4 more parameters and ×11.5 more processing time. The proposed method, which only utilizes 3 × 3 kernels, shows the fastest run-

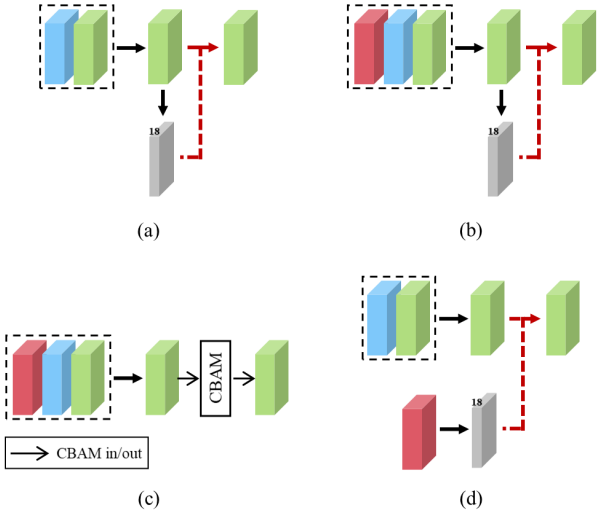


FIGURE 3. Depth feature refinement modules for (a) "Case 1", (b) "Case 2", (c) "Case 3", and (d) the proposed method.

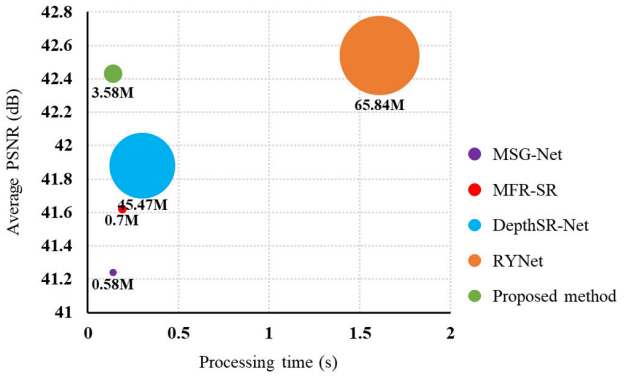


FIGURE 4. Average PSNR and processing time of color-guided depth SR methods for the scaling factor of 16. The radius of each circle is proportional to the number of network parameters.

TABLE 2. Quantitative comparison of the proposed method with state-of-the-art methods on the noise-free dataset in terms of PSNR(dB)/RMSE.

Image	Scaling factor	Methods							
		Bicubic	SRCNN	SAN	MSG-Net	MFR-SR	DepthSR-Net	RYNet	Proposed method
Art	2×	39.71 / 2.64	52.34 / 0.62	-	56.02 / 0.40	55.19 / 0.44	53.70 / 0.53	**60.33 / 0.25**	57.67 / 0.33
	4×	36.37 / 3.87	43.71 / 1.66	-	46.06 / 1.27	46.22 / 1.25	46.51 / 1.21	**48.06 / 1.01**	47.35 / 1.09
	8×	33.38 / 5.46	37.85 / 3.27	39.79 / 2.61	40.55 / 2.49	41.03 / 2.26	41.21 / 2.22	**42.10 / 2.00**	41.93 / 2.04
	16×	29.90 / 8.16	31.57 / 6.73	35.43 / 4.32	36.37 / 3.87	36.65 / 3.75	36.30 / 3.90	**37.79 / 3.29**	37.06 / 3.58
Books	2×	47.72 / 1.05	59.70 / 0.26	-	61.41 / 0.22	61.23 / 0.22	55.61 / 0.42	**63.94 / 0.16**	62.70 / 0.19
	4×	44.03 / 1.06	52.72 / 0.59	-	54.70 / 0.47	54.66 / 0.47	52.46 / 0.61	**56.91 / 0.36**	56.44 / 0.38
	8×	40.79 / 2.33	45.52 / 1.35	50.99 / 0.72	48.60 / 0.95	49.48 / 0.86	49.14 / 0.89	51.03 / 0.72	**51.42 / 0.68**
	16×	37.69 / 3.33	38.85 / 2.91	42.35 / 1.95	43.20 / 1.77	43.34 / 1.74	44.50 / 1.52	44.48 / 1.52	**45.14 / 1.41**
Moebius	2×	49.43 / 0.86	58.05 / 0.32	-	59.80 / 0.26	59.83 / 0.26	55.36 / 0.44	**61.88 / 0.21**	61.06 / 0.23
	4×	45.75 / 1.31	51.65 / 0.67	-	53.07 / 0.57	53.26 / 0.55	51.69 / 0.66	**54.32 / 0.49**	54.26 / 0.49
	8×	42.35 / 1.95	46.21 / 1.25	48.89 / 0.92	48.50 / 0.96	49.38 / 0.87	48.62 / 0.95	49.99 / 0.81	**50.15 / 0.79**
	16×	39.06 / 2.84	40.19 / 2.49	42.70 / 1.87	44.02 / 1.61	44.33 / 1.55	44.28 / 1.56	45.11 / 1.42	**45.42 / 1.37**
Reindeer	2×	42.40 / 1.94	54.49 / 0.48	-	57.70 / 0.33	57.42 / 0.34	53.95 / 0.51	**60.70 / 0.24**	59.19 / 0.28
	4×	39.14 / 2.81	46.52 / 1.20	-	48.79 / 0.93	48.76 / 0.93	48.48 / 0.96	**50.45 / 0.77**	49.65 / 0.84
	8×	36.09 / 4.00	40.38 / 2.44	42.84 / 1.84	43.77 / 1.65	44.21 / 1.57	44.17 / 1.58	**45.31 / 1.38**	44.56 / 1.51
	16×	32.79 / 5.85	34.28 / 4.93	38.15 / 3.16	39.04 / 2.85	39.91 / 2.58	40.29 / 2.47	**40.97 / 2.28**	40.41 / 2.43
Laundry	2×	44.09 / 1.59	56.49 / 0.38	-	59.25 / 0.28	58.47 / 0.30	55.35 / 0.44	**62.06 / 0.20**	60.41 / 0.24
	4×	40.59 / 2.38	48.84 / 0.92	-	51.14 / 0.71	51.09 / 0.71	50.36 / 0.77	**52.85 / 0.58**	51.95 / 0.64
	8×	37.47 / 3.41	41.94 / 2.04	46.14 / 1.26	45.19 / 1.40	45.93 / 1.29	45.87 / 1.30	46.82 / 1.16	**47.05 / 1.13**
	16×	34.06 / 5.05	35.35 / 4.36	39.58 / 2.68	40.58 / 2.39	41.28 / 2.20	41.40 / 2.17	**42.05 / 2.02**	41.57 / 2.13
Dolls	2×	48.96 / 0.91	55.75 / 0.42	-	58.74 / 0.30	57.21 / 0.35	54.38 / 0.49	**61.43 / 0.22**	59.16 / 0.28
	4×	45.85 / 1.30	49.69 / 0.84	-	51.22 / 0.70	51.29 / 0.70	50.13 / 0.79	**52.58 / 0.60**	52.11 / 0.63
	8×	42.82 / 1.84	45.30 / 1.39	47.34 / 1.10	47.47 / 1.08	47.61 / 1.06	47.33 / 1.10	**48.64 / 0.94**	48.38 / 0.97
	16×	39.79 / 2.61	40.80 / 2.33	42.17 / 1.99	44.21 / 1.57	44.22 / 1.57	44.48 / 1.52	44.83 / 1.46	**44.96 / 1.44**
Cones	2×	40.10 / 2.52	48.14 / 1.00	-	50.25 / 0.78	50.79 / 0.74	50.58 / 0.75	**55.78 / 0.41**	53.60 / 0.53
	4×	36.42 / 3.85	38.53 / 3.02	-	39.64 / 2.66	41.04 / 2.26	40.86 / 2.31	**42.45 / 1.92**	41.44 / 2.16
	8×	33.31 / 5.51	34.33 / 4.90	35.89 / 4.09	36.92 / 3.63	37.70 / 3.33	37.17 / 3.53	**38.63 / 2.99**	36.95 / 3.62
Teddy	2×	42.40 / 1.94	50.38 / 0.77	-	51.57 / 0.67	51.35 / 0.69	50.06 / 0.80	**53.74 / 0.52**	53.06 / 0.57
	4×	39.08 / 2.84	43.60 / 1.69	-	44.02 / 1.61	44.75 / 1.48	45.05 / 1.35	**46.95 / 1.15**	45.93 / 1.29
	8×	36.06 / 4.01	38.83 / 2.92	39.26 / 2.78	39.76 / 2.62	40.22 / 2.49	40.24 / 2.48	**41.87 / 2.06**	40.74 / 2.34
Tsukuba	2×	32.85 / 5.81	41.41 / 2.17	-	43.91 / 1.63	43.78 / 1.65	45.61 / 1.34	**51.01 / 0.72**	46.29 / 1.24
	4×	29.48 / 8.56	33.67 / 5.29	-	35.92 / 4.08	36.96 / 3.62	37.86 / 3.26	**39.23 / 2.79**	37.22 / 3.51
	8×	26.31 / 12.33	27.78 / 10.41	30.44 / 7.66	29.69 / 8.36	30.68 / 7.46	31.37 / 6.88	**31.91 / 6.48**	30.75 / 7.40
Venus	2×	45.81 / 1.31	61.58 / 0.21	-	61.06 / 0.23	60.98 / 0.23	55.02 / 0.45	**68.91 / 0.09**	68.60 / 0.10
	4×	42.51 / 1.91	52.22 / 0.63	-	53.45 / 0.54	54.81 / 0.46	52.22 / 0.62	**61.87 / 0.21**	61.10 / 0.23
	8×	39.35 / 2.75	45.33 / 1.38	47.33 / 1.10	48.40 / 0.97	47.53 / 1.07	47.63 / 1.06	**50.45 / 0.77**	50.33 / 0.78

TABLE 3. Quantitative comparison of the proposed method with state-of-the-art methods on noisy dataset A in terms of PSNR(dB)/RMSE.

Image	Scaling factor	Methods							
		Bicubic	SRCNN	SAN	MSG-Net	MFR-SR	DepthSR-Net	RYNet	Proposed method
Art	8×	31.45 / 6.82	34.78 / 4.65	36.12 / 3.99	37.21 / 3.52	37.60 / 3.36	37.32 / 3.47	37.48 / 3.41	**37.73 / 3.31**
	16×	28.93 / 9.12	30.26 / 7.83	30.70 / 7.44	34.10 / 5.03	**34.79 / 4.65**	34.06 / 5.05	34.58 / 4.76	34.55 / 4.77
Books	8×	34.76 / 4.66	40.01 / 2.55	43.28 / 1.75	42.60 / 1.89	43.02 / 1.80	43.28 / 1.75	43.40 / 1.73	**43.56 / 1.69**
	16×	33.70 / 5.30	36.11 / 3.99	33.59 / 5.33	39.10 / 2.83	39.65 / 2.65	39.71 / 2.64	39.45 / 2.72	**40.33 / 2.46**
Moebius	8×	35.03 / 4.52	39.75 / 2.62	42.03 / 2.02	42.54 / 1.90	43.22 / 1.76	43.12 / 1.78	43.30 / 1.75	**43.43 / 1.72**
	16×	34.14 / 5.01	36.43 / 3.85	35.40 / 4.33	39.06 / 2.84	**39.61 / 2.67**	39.47 / 2.71	39.36 / 2.75	39.58 / 2.68
Reindeer	8×	32.95 / 5.74	37.03 / 3.59	38.34 / 3.09	39.45 / 2.72	39.78 / 2.62	39.60 / 2.67	39.80 / 2.61	**39.93 / 2.57**
	16×	31.07 / 7.13	32.69 / 5.91	32.31 / 6.18	36.47 / 3.83	37.28 / 3.49	37.16 / 3.54	37.07 / 3.57	**37.40 / 3.44**
Laundry	8×	33.60 / 5.33	37.63 / 3.35	39.99 / 2.55	40.35 / 2.45	41.02 / 2.27	40.72 / 2.35	**41.29 / 2.20**	41.28 / 2.20
	16×	31.95 / 6.44	33.51 / 5.38	33.49 / 5.39	37.44 / 3.43	**37.71 / 3.32**	36.89 / 3.65	37.11 / 3.56	37.60 / 3.36
Dolls	8×	35.07 / 4.50	39.68 / 2.65	40.91 / 2.30	42.18 / 1.98	42.58 / 1.90	42.44 / 1.93	**43.03 / 1.80**	42.58 / 1.89
	16×	34.31 / 4.91	36.73 / 2.72	32.86 / 5.80	39.31 / 2.76	39.54 / 2.69	39.67 / 2.65	**40.05 / 2.54**	39.88 / 2.59

ning time with MSG-Net which uses the fewest number of parameters.

Table 2 presents the performance for each test image using PSNR and RMSE in the noise-free case. The best and second-best results are boldfaced and underlined, respectively. For the last four low resolution test images, the maximum scaling factor was set to 8. The performance of the SRCNN is comparable to that of the state-of-the-art methods

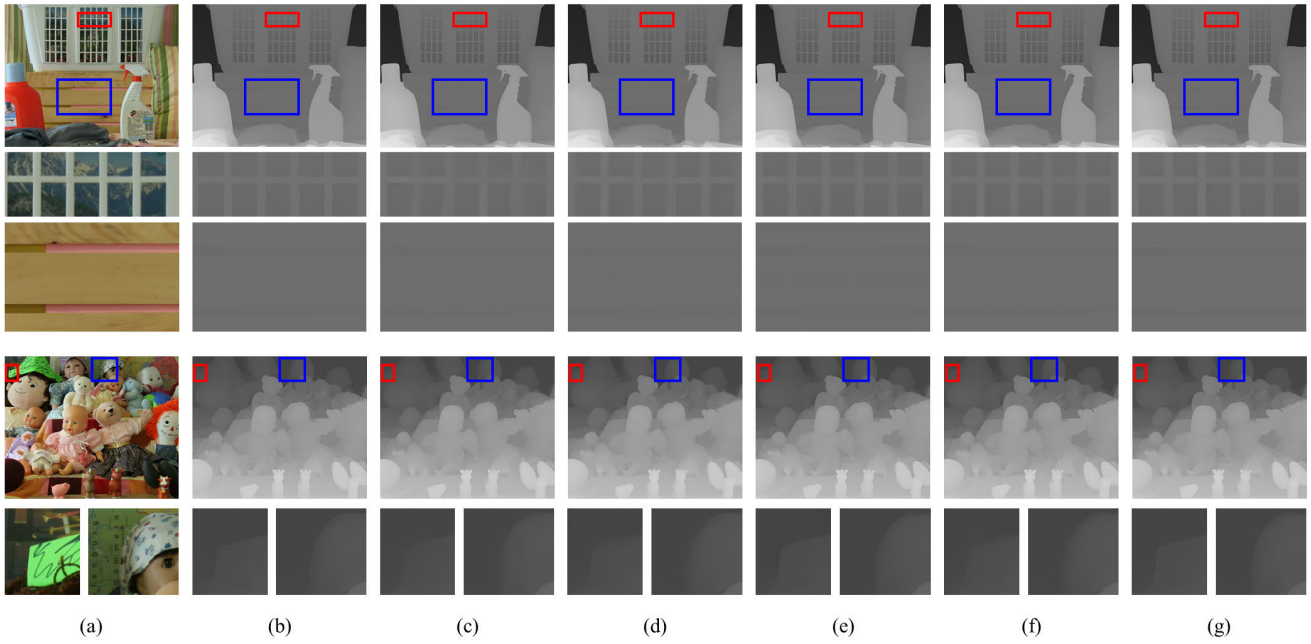


FIGURE 5. Visual comparison of the noise-free test images "Laundry" (16×) and "Dolls" (16×): (a) HR color image, (b) ground truth, upsampled depth maps obtained by (c) MSG-Net, (d) MFR-SR, (e) DepthSR-Net, (f) RYNet, and (g) the proposed method. Magnified regions are shown below each image.

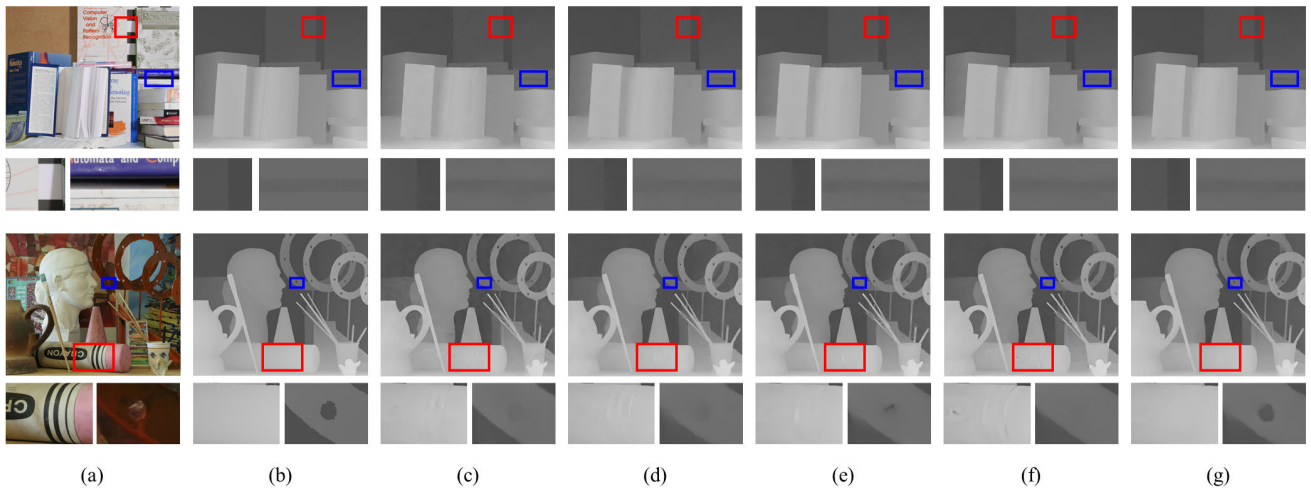


FIGURE 6. Visual comparison of the noisy test images "Books" (8×) and "Art" (16×): (a) HR color image, (b) ground truth, upsampled depth maps obtained by (c) MSG-Net, (d) MFR-SR, (e) DepthSR-Net, (f) RYNet, and (g) the proposed method. Magnified regions are shown below each image.

with small scaling factors (i.e., 2×, 4×). However, because the SRCNN upsamples the depth map without the guidance image, its performance with the high scaling factors is unsatisfactory. The state-of-the-art single image SR method, SAN, shows superior performance over SRCNN, but still exhibits unsatisfactory performance compared with the color-guided depth SR methods. Among the color-guided depth map SR methods, the proposed method and RYNet exhibit the second-best and the best performance for most test images, respectively. In addition, the results of each method on the noisy dataset are presented in Table 3. The authors of MFR-SR [18] adopted batch normalization layers to improve the performance of depth map denoising. Although our proposed network does not contain layers for depth map denoising, it outperforms the other methods in most cases.

C. QUALITATIVE EVALUATION

Fig. 5 shows the upsampled results of each method for the noise-free test images. Since the laundry basket in the test image "Laundry" has detailed structures with many edges, it is hard to construct the HR depth map. As shown in Figs. 5(c) and (d), the MSG-Net and MFR-SR produce the HR depth map with blurred edges. On the other hand, Fig. 5(g) shows that the proposed method generates the sharp

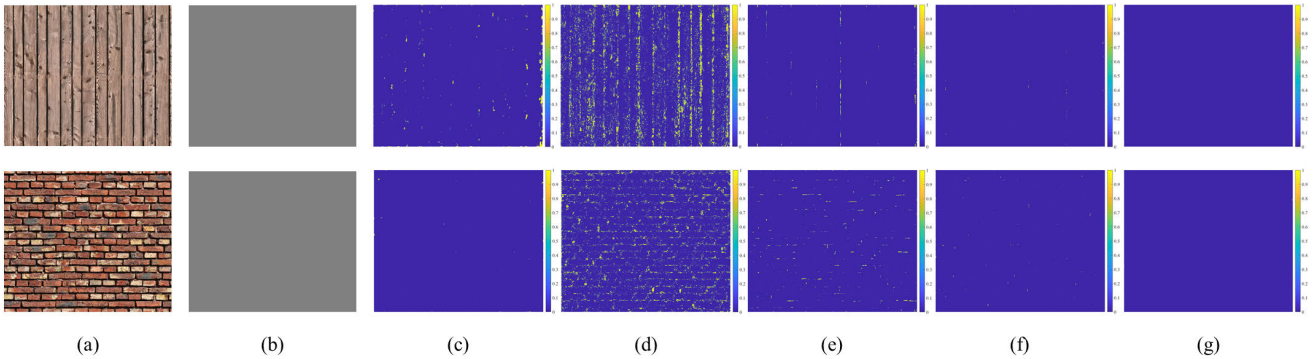


FIGURE 7. Visual comparison of the robustness against the texture copying artifacts: (a) HR color image, (b) ground truth, difference maps between ground truth and upsampled (16×) depth maps obtained by (c) MSG-Net, (d) MFR-SR, (e) DepthSR-Net, (f) RYNet, and (g) the proposed method. The first and second rows show the experimental results for the test images "Wall1" and "Wall2", respectively.

edges. The upsampled results for the test image "Dolls" are shown at the bottom of Fig. 5. It can be seen that the proposed method restores sharp depth boundaries compared to the conventional methods.

We also show the upsampled results of the noisy LR depth maps. The resultant HR depth maps for the noisy test image "Books" are presented at the top of Fig. 6. Compared with the conventional methods, the proposed method successfully preserves sharp depth discontinuities. The resultant HR depth maps of each method for the noisy test image "Art" are shown at the bottom of Fig. 6. Because the conventional methods directly concatenate the features of the HR intensity image to the depth features, texture copying artifacts occur, as shown in the red rectangles in Figs. 6(c)–(f). In contrast, as seen in Fig. 6(g), the proposed method effectively prevents the texture copying artifacts and produces a homogeneous depth map. In addition, it can be seen that the proposed method reconstructs the fine details better than the other methods.

To clearly demonstrate the advantages of using the proposed method to solve the texture copying artifacts, additional experimental results are shown in Fig. 7. We prepared two images of walls containing repetitive texture patterns and assumed that each wall has a homogeneous depth map (i.e., all pixels in the depth map have the same depth value). The differences between the ground truth HR depth map and the results of each method are presented in Figs. 7(c)–(g). Although MSG-Net alleviates the texture copying artifacts owing to its high-frequency domain training approach, a few artifacts are still generated in the resultant depth map, as shown in Fig. 7(c). Figs. 7(d) and (e) show that the MFR-SR and DepthSR-Net experience the texture copying problem because the features of the HR intensity image affect the constructed depth map directly. Although RYNet includes the spatial attention based feature fusion blocks to suppress the texture copying artifacts, its results still suffer from the undesired artifacts, as shown in Fig. 7(f). Compared with the other methods, the proposed method successfully avoids undesired artifacts and produces the closest result to the ground truth, as shown in Fig. 7(g).

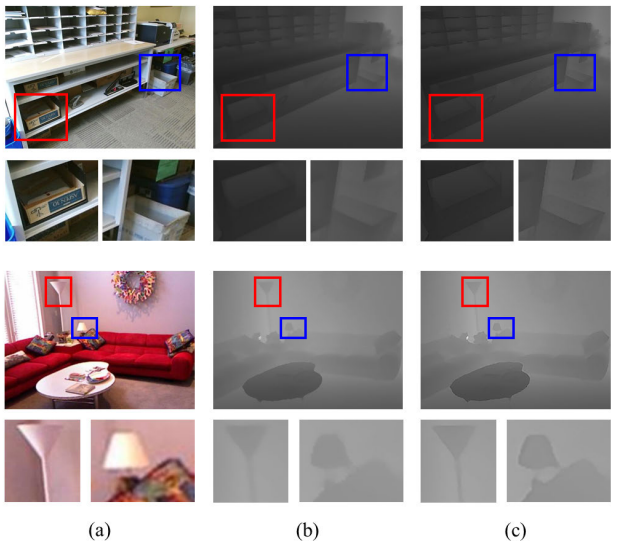


FIGURE 8. Experimental results on real data: (a) HR color image, (b) LR depth map, and (c) HR depth map generated by the proposed method.

D. REAL DATA

To find out the performance of the proposed method in the real environment, we utilized the real data [39], [40] collected using the Microsoft Kinect. The depth map is pre-upsampled by the authors and provided in the same size as the color image. We directly applied the proposed network learned with the scaling factor of 4 to the depth map. The HR depth map generated by proposed method is shown in Fig. 8(c). It can be seen that the depth discontinuities become sharper by the proposed method.

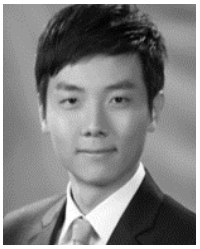
V. CONCLUSION

In this paper, we presented a deep network for color-guided depth map SR. Observing that recent color-guided depth map SR networks produce unwanted texture copying artifacts because of the excessive use of guidance features, we proposed guided deformable convolution, which uses the guidance features only for choosing the sampling locations of the convolution kernel of depth features. Consequently, the

proposed method successfully restores sharp depth boundaries and prevents texture copying artifacts from coming into existence. The experimental results demonstrate that the proposed network significantly outperforms conventional networks.

REFERENCES

[1] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.
[2] L. Li, "Time-of-flight camera—An introduction," Texas Instrum., Dallas, TX, USA, White Paper SLOA190B, May 2014.
[3] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2003, pp. 1–8.
[4] O. Mac Aodha, N. D. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 71–84.
[5] J. Xie, R. S. Feris, and M.-T. Sun, "Edge-guided single depth image super resolution," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 428–438, Jan. 2016.
[6] D. Ferstl, M. Ruther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 513–521.
[7] J. Xie, R. S. Feris, S.-S. Yu, and M.-T. Sun, "Joint super resolution and denoising from a single depth image," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1525–1537, Sep. 2015.
[8] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, 2007.
[9] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[10] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 291–298.
[11] M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 169–176.
[12] S.-W. Jung and O. Choi, "Learning-based filter selection scheme for depth image super resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1641–1650, Oct. 2014.
[13] O. Choi and S.-W. Jung, "A consensus-driven approach for structure and texture aware depth map upsampling," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3321–3335, Aug. 2014.
[14] J. Lu and D. Forsyth, "Sparse depth super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2245–2253.
[15] H. Kwon, Y.-W. Tai, and S. Lin, "Data-driven depth map refinement via multi-scale sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 159–167.
[16] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 353–369.
[17] M. Ni, J. Lei, R. Cong, K. Zheng, B. Peng, and X. Fan, "Color-guided depth map super resolution using convolutional neural network," *IEEE Access*, vol. 5, pp. 26666–26672, 2017.
[18] Y. Zuo, Q. Wu, Y. Fang, P. An, L. Huang, and Z. Chen, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 297–306, Feb. 2020.
[19] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, May 2019.
[20] Y. Wen, B. Sheng, P. Li, W. Lin, and D. D. Feng, "Deep color guided coarse-to-fine convolutional network cascade for depth image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 994–1006, Feb. 2019.
[21] T. Li, X. Dong, and H. Lin, "Guided depth map super-resolution using recumbent y network," *IEEE Access*, vol. 8, pp. 122695–122708, 2020.
[22] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[23] M. Hornacek, C. Rhemann, M. Gelautz, and C. Rother, "Depth super resolution by rigid body self-similarity in 3D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1123–1130.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[27] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, "Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3367–3376.
[28] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1954–1963.
[29] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
[30] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.
[31] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[32] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[35] A. Paszke, S. Gross, F. Massa, and A. Lerer, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[36] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[37] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11065–11074.
[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[39] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 567–576.
[40] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.



JOON-YEON KIM received the B.S. degree in electrical engineering from Korea University, Seoul, Korea, in 2014, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include image processing, computer vision, and deep learning.

SEOWON JI received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in electrical engineering. His interests include image processing, computer vision, and deep-learning.

SEUNG-JIN BAEK received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2007 and 2013, respectively. He joined the Digital Media and Communications Research and Development Center, Samsung Electronics Company Ltd., Suwon, South Korea, in 2013, where he was a Senior Engineer, from 2014 to 2015. From 2015 to 2020, he was a Staff Engineer with the Visual Display Business Division, Samsung Electronics Company Ltd., Suwon. He is currently a Research Professor with the Research Institute of Information and Communication Technology, Korea University. His current research interests include deep learning, computer vision, and image processing applications.

SEUNG-WON JUNG (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2005 and 2011, respectively. From 2011 to 2012, he was a Research Professor with the Research Institute of Information and Communication Technology, Korea University. From 2012 to 2014, he was a Research Scientist with Samsung Advanced Institute of Technology, Yongin, South Korea. From 2014 to 2020, he was an Assistant Professor with the Department of Multimedia Engineering, Dongguk University, Seoul. In 2020, he joined the Department of Electrical Engineering, Korea University, where he is currently an Associate Professor. He has published over 60 peer-reviewed articles in international journals. His current research interests include image processing and computer vision. He received the Hae-Dong Young Scholar Award from the Institute of Electronics and Information Engineers, in 2019.

SUNG-JEA KO (Fellow, IEEE) received the B.S. degree in electronic engineering from Korea University, in 1980, and the M.S. and Ph.D. degrees in electrical and computer engineering from The State University of New York at Buffalo, in 1986 and 1988, respectively.

From 1988 to 1992, he was an Assistant Professor with the Department of Electrical and Computer Engineering, University of Michigan-Dearborn. In 1992, he joined the Department of Electronic Engineering, Korea University, where he is currently a Professor. He has published over 210 international journal articles. He also holds over 60 registered patents in fields, such as video signal processing and computer vision.

Dr. Ko is a member of the National Academy of Engineering of Korea. He was a Recipient of the Best Paper Award from the IEEE Asia Pacific Conference on Circuits and Systems, in 1996; the Hae-Dong Best Paper Award from the Institute of Electronics and Information Engineers (IEIE), in 1997; the 1999 LG Research Award; the Research Excellence Award from Korea University, in 2004; the Technical Achievement Award from the IEEE Consumer Electronics (CE) Society, in 2012; the 15-Year Service Award from the TPC of ICCE, in 2014; the Chester Sall Award (First Place Transaction Paper Award) from the IEEE CE Society, in 2017; and the Science and Technology Achievement Medal from the Korean Government, in 2020. He served as the General Chairman for ITC-CSCC 2012 and the General Chairman for IEICE 2013. He was the President of the IEIE, in 2013; the Vice President of the IEEE CE Society, from 2013 to 2016; and the Distinguished Lecturer of the IEEE, from 2015 to 2017. He is also an Editorial Board Member of the IEEE TRANSACTIONS ON CONSUMER ELECTRONICS.