

# High-Resolution Depth Maps Imaging via Attention-Based Hierarchical Multi-Modal Fusion

Zhiwei Zhong, Xianming Liu<sup>✉</sup>, *Member, IEEE*, Junjun Jiang<sup>✉</sup>, *Member, IEEE*, Debin Zhao<sup>✉</sup>, *Member, IEEE*, Zhiwen Chen<sup>✉</sup>, *Member, IEEE*, and Xiangyang Ji<sup>✉</sup>, *Member, IEEE*

**Abstract**—Depth map records distance between the viewpoint and objects in the scene, which plays a critical role in many real-world applications. However, depth map captured by consumer-grade RGB-D cameras suffers from low spatial resolution. Guided depth map super-resolution (DSR) is a popular approach to address this problem, which attempts to restore a high-resolution (HR) depth map from the input low-resolution (LR) depth and its coupled HR RGB image that serves as the guidance. The most challenging issue for guided DSR is how to correctly select consistent structures and propagate them, and properly handle inconsistent ones. In this paper, we propose a novel attention-based hierarchical multi-modal fusion (AHMF) network for guided DSR. Specifically, to effectively extract and combine relevant information from LR depth and HR guidance, we propose a multi-modal attention based fusion (MMAF) strategy for hierarchical convolutional layers, including a feature enhancement block to select valuable features and a feature recalibration block to unify the similarity metrics of modalities with different appearance characteristics. Furthermore, we propose a bi-directional hierarchical feature collaboration (BHFC) module to fully leverage low-level spatial information and high-level structure information among multi-scale features. Experimental results show that our approach outperforms state-of-the-art methods in terms of reconstruction accuracy, running speed and memory efficiency.

**Index Terms**—Depth map super-resolution, multi-modal attention, bi-directional feature propagation.

## I. INTRODUCTION

**D**EPTH information plays a critical role in a myriad of applications such as autonomous driving [1], virtual

Manuscript received March 10, 2021; revised August 18, 2021 and October 1, 2021; accepted November 7, 2021. Date of publication December 8, 2021; date of current version December 28, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFE0109600 and in part by the National Natural Science Foundation of China under Grant 61922027 and Grant 61932022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rafal K. Mantiuk. (*Corresponding author: Xianming Liu.*)

Zhiwei Zhong is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: zhzhong@hit.edu.cn).

Xianming Liu, Junjun Jiang, and Debin Zhao are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, and also with Peng Cheng Laboratory, Shenzhen 518052, China (e-mail: csxm@hit.edu.cn; jiangjunjun@hit.edu.cn; dbzhao@hit.edu.cn).

Zhiwen Chen is with Taobao (China) Software Company Ltd., Beijing 311121, China (e-mail: zhiwen.czw@alibaba-inc.com).

Xiangyang Ji is with the Department of Automation and BNRist, Tsinghua University, Beijing 100084, China (e-mail: xyji@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3131041

reality [2], 3D reconstruction [3] and scene understanding [4]. In recent years, with the progress of sensing technology, depth maps can be readily captured by consumer-grade depth cameras such as Time-of-Flight (ToF) and Microsoft Kinect. However, the depth map taken from these commercialized cameras usually suffers from low-resolution, which hinders the subsequent depth based applications. Therefore, depth map super-resolution (DSR) has raised a lively interest in the communities of academia and industry.

DSR is inherently an ill-posed problem, as there exist multiple high-resolution (HR) depth maps corresponding to the same low-resolution (LR) degradation. To solve this inverse problem, one popular approach is guided DSR, considering that in practice many smartphones and robots are equipped with a conventional RGB camera as well as a depth camera. The former acquires an intensity image with higher spatial resolution than the depth map. Since the targets they shoot are the same scene, it is thus natural to enhance the resolution of depth by transferring structure from the HR guidance image. Specifically, guided DSR aims to apply HR guidance image as a prior for reconstructing regions in depth where there is semantically-related and structure-consistent content, and fall back to a plausible reconstruction for regions in depth with inconsistent content of the guidance.

To achieve this goal, two major problems should be carefully addressed. Firstly, it is challenging to select reference structures and propagate them properly by defining hand-crafted rules. Secondly, the guidance based approach makes a basic assumption that the guidance image should contain correct mutual structural information. However, the guidance could be insufficient, or even wrong locally. It is challenging to handle the structure inconsistency problem. For regions with inconsistent structures, it is expected that guidance based approach could reduce the wrong influence of the guidance and predict HR reconstruction properly.

For the first issue, data-driven based strategies have been proposed to remedy the difficulty of hand-crafted design. For instance, Hui *et al.* [5] proposed a multi-scale guided convolutional network (DMSG) for DSR, which fuses rich hierarchical features at different levels to generate accurate HR depth map. Similarly, in [6], Guo *et al.* proposed a DSRNet to infer a HR depth map from its LR version by hierarchical features driven residual learning. Su *et al.* [7] proposed a pixel-adaptive convolution based network (PacNet), which

is actually a fine-grained filtering operation that can effectively learn to leverage guidance information. Similar strategy that uses pixel-wise transformation also appears in [8]. The work [9] presented a progressive multi-branch aggregation network (PMBAN) for depth SR, which consisted of stacked multi-branch aggregation blocks to progressively recover the degraded depth map.

For the second issue, recent efforts focus on learning-based selection strategies for the common structures existing in both the target and guidance images. For instance, Li *et al.* [10] proposed a joint image filtering with deep convolutional networks, which can selectively transfer salient structures that are consistent with multi-modal inputs. The network architecture of DKN [11] is similar to DJFR [10], but contains a weight and offset learning module to explicitly learn the sparse and spatially-variant kernels. Recently, Deng and Dragotti [12] proposed a common and unique information splitting network (CUNet) to automatically determine the common information among different modalities, according to which an adaptive fusion operation was performed.

Although significant progress has been achieved, learning-based DSR is still an open problem. The key challenge is how to achieve a good balance among performance, running time and network complexity, so as to promote its usage in practical scenarios. In this paper, we embrace this challenge and propose a novel attention-based hierarchical multi-modal fusion (AHMF) network to perform structure selection, propagation and prediction simultaneously for guided DSR. Specifically, to effectively explore and combine relevant information from LR depth and HR guidance, we propose a multi-modal attention based fusion (MMAF) for hierarchical convolutional layers. It consists of a feature enhancement block that is tailored to adaptively select useful information and filter out unwanted ones, such as texture information in guidance and noise in depth that would disturb the depth reconstruction; and a feature recalibration block that is designed to adaptively rescale enhanced features to unify the similarity metrics of modalities with different appearance characteristics. Furthermore, considering that in CNN shallower layers encode rich spatial details but lack semantic knowledge while deeper layers are more effective to capture high-level context and structure information but lose spatial information, we propose a bi-directional hierarchical feature collaboration (BHFC) module to fully leverage the complementarity of the hierarchical fused features. To verify the effectiveness of the proposed method, we conduct experiments on widely used benchmark datasets, and the experimental results demonstrate that our method achieves superior performance than the state-of-the-arts.

The main contributions of our method can be summarized as follows:

- We propose a multi-modal attention based fusion module, which can adaptively select and effectively fuse features extracted from the depth and guidance images. It contains a feature enhancement block to select valuable information and a feature recalibration block to rescale multi-modal features. Contrary to existing methods that fuse multi-modal features by simple concatenation or

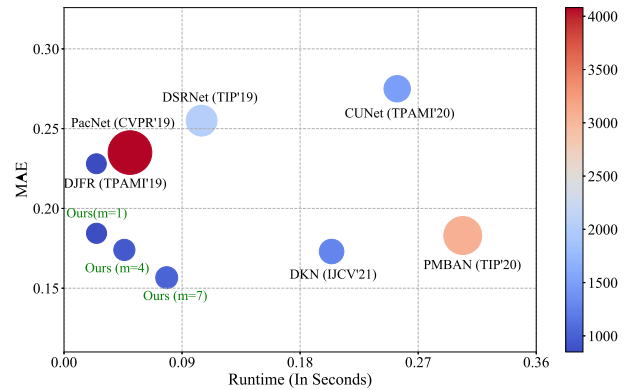


Fig. 1. Comparison of state-of-the-art methods for  $4\times$  DSR on Middlebury 2005 [13] dataset in terms of MAE (the lower the better), running time and peak GPU memory consumption that is indicated by radius of circles. Our method includes three cases with different  $m$ , which refers to the number of layers for multi-modal feature fusion. The experiments are evaluated on NVIDIA 1080ti GPU with depth map size  $480 \times 640$ .

summation, the proposed method can avoid transferring erroneous structures that are not existed in the depth image, *i.e.*, the texture-copying artifacts.

- We propose a bi-directional hierarchical feature collaboration module, which can facilitate the hierarchical fused features propagation and collaboration with each other.
- We propose an attention-based hierarchical multi-modal fusion framework (AHMF) for guided depth map super-resolution. With the proposed MMAF and BHFC, the proposed method can effectively explore the complementarity of multi-level and multi-modal features. As shown in Fig. 1, our method achieves better DSR performance, faster running speed and moderate memory consumption over state-of-the-art methods.

The remainder of this paper is organized as follows. Section II reviews related work in the literature. Section III introduces the overview of the proposed method. Then, we elaborate on the proposed multi-modal attention based fusion in Section IV and hierarchical feature collaboration strategy in Section V. Extensive experimental results are presented in Section V. Finally, Section VII concludes this paper.

## II. RELATED WORK

### A. Guided Depth Super-Resolution

In the literature, many works have been developed for guided DSR, which can be roughly divided into three categories: filtering based, optimization based and deep learning based. The filtering based approaches, such as [15], [16], reconstruct the depth map by means of the weighted average of target image values with a local filter. Although these filtering based methods are at low computational cost, they cannot maintain the global information as the weights of the filter are calculated by the local content in the guidance image. This would inevitably transfer inaccurate structures to the depth image when the assumption of structure similarity is invalid. Different from filtering based methods, optimization based methods formulate depth image super-resolution as a global optimization problem. They adopt various priors to

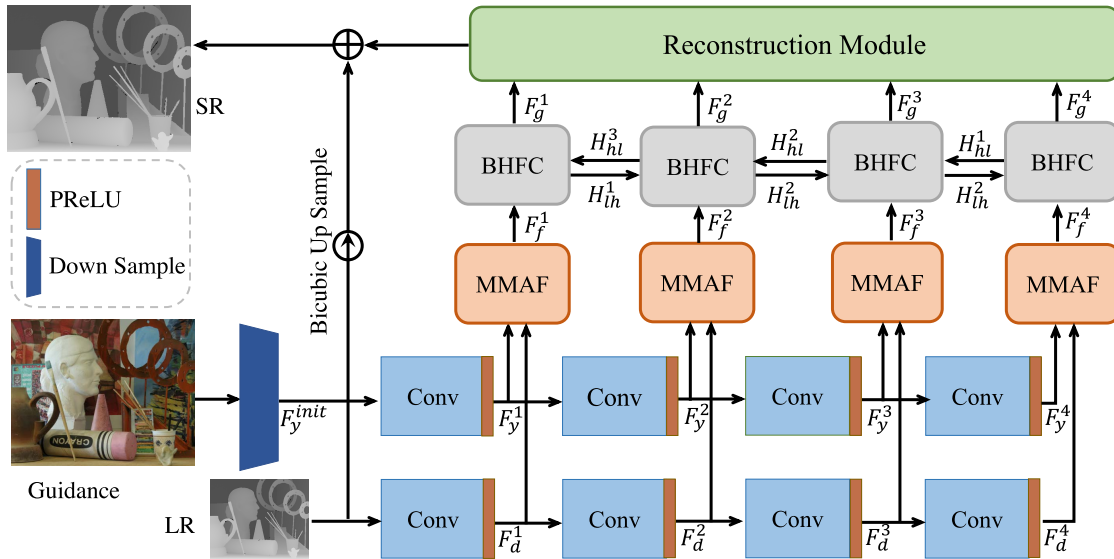


Fig. 2. The network architecture of our proposed attention-based hierarchical multi-modal fusion (AHMF) network, where MMAF represents the proposed multi-modal attention based fusion module and BHFC represents the bi-directional hierarchical feature collaboration module.

constrain the high-dimensional solution space for this ill-posed problem. For example, Ferstl *et al.* [17] formulated the DSR problem as a convex optimization problem and employed a high-order Total Generalized Variation (TGV) regularization to get a piece-wise smooth solution. Liu *et al.* [18] proposed a DSR method by combining both internal and external priors formulated in the graph domain. However, these hand-crafted priors suffer from limitations in modeling the real world image degradation process. Moreover, solving the optimization problem is usually time-consuming due to the iterative computation process.

Based on U-Net architecture, Guo *et al.* [6] proposed a hierarchical feature driven network and they claimed that their network could make full use of features extracted from depth and guided image compared to the existing methods. To learn the potential relationship between the guidance and the depth images, Zuo *et al.* [19] proposed a coarse-to-fine network by combining both global and local residual learning strategies. Li *et al.* [20] introduced a multi-scale symmetric network, which includes a symmetric unit to restore edge details and a correlation-controlled color guidance block to investigate the inner-channel correlation between the depth and guidance sub-network. Kim *et al.* [11] proposed a deformable kernel network for depth SR, they employed the kernel based method but the weights and offsets of the kernel were learned by the network automatically. Su *et al.* [7] argued that the convolutional operation was content-agnostic and proposed a pixel-adaptive convolution to address this problem, and they conducted plenty of experiments to show that their method could easily adapt to a lot of applications such as joint upsampling, semantic segmentation and CRF inference.

Besides the color-guided depth super-resolution considered in this work, in the field of 3D imaging, there is another approach to achieve high-quality depth imaging by fusing high-quality RGB image and noisy 3D single-photo avalanche detector (SPAD) arrays [21]–[24]. Our method can also be

applied to treat this problem by modifying accordingly to meet the requirements of RGB-SPAD fusion, including:

- 1) Replace the 2D convolution layers for depth image features extraction with 3D convolution layers.
- 2) Redesign the multi-modal feature fusion module (MMAF in our model) to make it capable of fusing 2D and 3D data, i.e., add a 2D-3D up-projection module as mentioned in [22] at the head of MMAF to expand the temporal dimension for 2D features.
- 3) Decrease the temporal dimensions to one before the final reconstruction module for the purpose of generating 2D depth image.

Since this is out of the scope of this paper, we leave the investigation of the effectiveness of our method for RGB-SPAD fusion in the future work.

### B. Attention Model

Attention mechanism has shown remarkable performance in deep convolutional neural networks and has been introduced in a large number of computer vision tasks. Zhang *et al.* [25] proposed a self-attention based GAN architecture for image generation problem, the experimental results showed that the self-attention module could effectively capture long-range dependencies. Inspired by the non-local mean method, Wang *et al.* [26] proposed a non-local neural network for video classification and the self-attention could be viewed as a special case of it. Hu *et al.* [27] proposed a squeeze-and-excitation module to let the network pay more attention to important feature maps by explicitly modelling channel-wise interrelations. Li *et al.* [28] proposed a selective kernel network (SKNet) to let the neural adjust its receptive field adaptively. Mei *et al.* [29] proposed a pyramid attention module for image restoration, their method could capture long range correspondences in a multi-level fashion. To accelerate the channel attention and decrease the model complexity,

Wang *et al.* [30] replaced the fully connected layer used in the channel attention module with a 1D convolution. Yu *et al.* [31] proposed a gated convolution to solve the problem of the traditional convolutional layer treating all input pixels equally for free-from image inpainting tasks, then Chang *et al.* [32] extended the 2D gated convolution to 3D part for free-from video inpainting. Among these works, the most related to our method is [28], however, there are still some differences between them, firstly the feature recalibration block in our method is to rescale the different modal data while [28] aims to dynamically select the receptive field for each neuron. Secondly, we use mean and standard deviation pooling instead of max pooling used in [28] to get global embedding, which is more suitable for low-level vision.

### C. Multi-Level Feature Fusion

In deep convolutional networks, the features of shallower layers usually contain low-level details while the ones of deeper layers are composed of high-level semantic information. In order to make the best use of both high and low level features, enormous methods are proposed. He *et al.* [33] introduced a skip-connection which was also called residual learning operator for training very deep convolutional networks, with the help of this operation, the deeper layer can directly access the low-level features. Instead of pixel-wise addition, Huang *et al.* [34] concatenated the low-level and high-level features to fuse the multi-level features. Gu *et al.* [35] proposed a self-guided network for image denoising, they used high-level features to progressively guide the low-level feature, which could enlarge the receptive field of the shallower layer and enhance the network representing capability. Lin *et al.* [36] proposed a refinement network in which the low-level features were refined by the high-level semantic levels. Although the performance of the network is greatly improved, there are still some problems, for example, in [33] and [34] the low-level features cannot directly contact with high-level information due to the feed-forward nature of convolutional neural network, on the contrary, in the methods of [36] and [35] only the low-level features can be refined by the high-level features. To solve these problems, in this paper, we propose a bi-directional hierarchical feature collaboration module, in which the low-level feature and high-level can propagate to each other effectively.

## III. OVERVIEW OF THE PROPOSED METHOD

In this section, we provide an overview about the proposed attention-based hierarchical multi-modal fusion network.

Our model takes a LR depth map  $D_{LR} \in \mathbb{R}^{H \times W \times 1}$  and a HR guidance image  $Y \in \mathbb{R}^{\alpha H \times \alpha W \times C}$  as inputs, where  $\alpha$  is the upscale factor,  $H$ ,  $W$  and  $C$  represent the height, width and the number of channels respectively. The pipeline of our network is illustrated in Fig. 2, consisting of five major modules that are marked by different colors, which are tailored to address the corresponding issues in color guided depth SR:

- **Module of guidance image downsampling.** Considering that the resolution of guidance  $Y$  is higher ( $\alpha \times$ ) than the corresponding depth  $D_{LR}$ , to facilitate the subsequent

processing, this module is tailored to downsample  $Y$  to achieve resolution consistency. Instead of using traditional downsampling strategy such as bicubic, which would result in information loss, we propose to leverage inverse pixel-shuffle [37] to progressively downsample  $Y$  with the upscale factor  $\alpha$ . In this way, it can not only preserve all the original information of  $Y$  but also achieve resolution balance of two inputs. More specifically, taking  $\alpha = 4$  as an example,  $Y$  needs to be downsampled by 2 ( $\log_2 \alpha$ ) times. We first employ a  $3 \times 3$  convolution layer to expand the feature channels of the guidance image to obtain  $\hat{Y}$ , then the downsampling process can be described as follows:

$$F_y^{init} = \text{Down}(\sigma(W_y^0 * \text{Down}(\hat{Y}) + b_y^0)), \quad (1)$$

where  $\text{Down}(\cdot)$  represents the inverse pixel-shuffle operator with downscale factor 2;  $W_y^0$  is a  $1 \times 1$  convolutional kernel;  $b_y^0$  is the bias term;  $*$  means convolutional operation;  $\sigma$  represents a Parametric Rectified Linear Unit (PReLU) [38] activation function.

- **Module of feature extraction.** This module is employed to fully extract meaningful features from both depth and guidance in a multi-level fashion:

$$F_d^1 = \sigma(W_d^1 * D_{LR} + b_d^1), \quad (2)$$

$$F_y^1 = \sigma(W_y^1 * F_y^{init} + b_y^1), \quad (3)$$

$$F_d^i = \sigma(W_d^i * F_d^{i-1} + b_d^i), \quad 1 < i \leq m, \quad (4)$$

$$F_y^i = \sigma(W_y^i * F_y^{i-1} + b_y^i), \quad 1 < i \leq m, \quad (5)$$

where  $W_d^i$  and  $W_y^i$  are convolutional kernels of  $i$ -th layer that are used for depth and guidance feature extraction respectively;  $b_d^i$  and  $b_y^i$  are the bias terms;  $m$  refers to the number of layers for feature extraction.

- **Module of multi-modal attention based fusion.** After extracting multi-modal features, the following question is how to fuse them properly. We do this by the proposed multi-modal attention based fusion (MMAF):

$$F_f^i = \text{MMAF}_i(F_y^i, F_d^i), \quad 1 \leq i \leq m, \quad (6)$$

which considers the difference between guidance and depth and adaptively combines multi-modal features by the attention mechanism. In this way, it optimally preserves consistent structures and suppresses inconsistent components in a learning manner.

- **Module of bi-directional hierarchical feature collaboration.** MMAF is followed by the proposed bi-directional hierarchical feature collaboration (BHFC), which is tailored to further jointly leverage low-level spatial and high-level semantic information:

$$F_g^i = \text{BHFC}_i([F_f^i, H_{lh}^{i-1}, H_{hl}^{m-i}]), \quad 1 \leq i \leq m, \quad (7)$$

where  $H$  is hidden state and the subscripts  $lh$  and  $hl$  denote the information propagation direction.  $H^0$  is the initial state and is set as zero.

- **Module of final HR depth reconstruction.** Finally, with the refined features  $\{F_g^i\}_{i=1}^m$ , we arrive at the HR depth

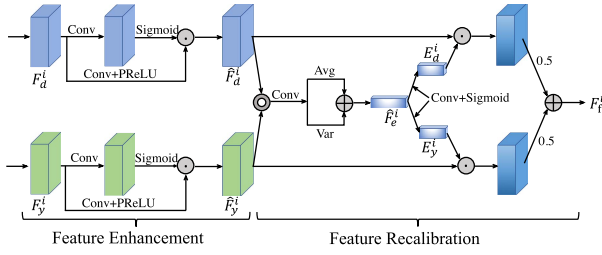


Fig. 3. Multi-modal attention based fusion (MMAF), where  $\odot$  is pixel-wise multiplication,  $\oplus$  is pixel-wise summation and  $\odot$  denotes concatenation. Avg and Var mean the average and variance pooling. The MMAF takes the depth and guidance features as inputs and outputs the fused features.

reconstruction module, which produces the upsampled depth image by pixel-shuffle [14]:

$$\mathbf{F}_{up} = \text{Up} \left( \sigma(\mathbf{W}_{cr} * [\mathbf{F}_g^1, \mathbf{F}_g^2, \dots, \mathbf{F}_g^m] + b_{cr}) \right) \quad (8)$$

$$\mathbf{D}_{SR} = \sigma(\mathbf{W}_{out} * \mathbf{F}_{up} + b_{out}) + \mathbf{D}_{LR}^\uparrow, \quad (9)$$

where  $\mathbf{W}_{cr}$  and  $b_{cr}$  are the weight and bias of a  $1 \times 1$  convolutional layer for channel reduction;  $\text{Up}(\cdot)$  is denoted as the up-sampling module which employ the pixel-shuffle operation to progressive up-sample the input features;  $[\cdot]$  is used to concatenate the refined features;  $\mathbf{W}_{out}$  and  $b_{out}$  are convolution kernels and bias, respectively.  $\mathbf{D}_{LR}^\uparrow$  is the bicubic upsampled version of the input LR depth image;  $\mathbf{D}_{SR} \in \mathbb{R}^{\alpha H \times \alpha W \times 1}$  is the final reconstructed HR depth image.

The network is trained using the following loss function:

$$\mathbf{L}(\theta) = \frac{1}{N} \sum_{i=1}^N \|\mathcal{F}(\mathbf{D}_{LR}, \mathbf{Y}; \theta) - \mathbf{D}_{GT}\|_1, \quad (10)$$

where  $\mathcal{F}(\cdot)$  represents the overall network architecture and  $\theta$  are network parameters.  $\mathbf{D}_{GT}$  is the ground-truth HR depth map,  $N$  is the number of training samples.

#### IV. MULTI-MODAL ATTENTION-BASED FUSION

In this section, we introduce the proposed multi-modal attention based fusion strategy. In guided DSR, the core is the act of extracting and combining relevant information from LR depth and HR guidance so as to derive superior performance over using only depth. One can perform concatenation or summation of the outputs of two separate branches for depth and guidance, respectively, and then use uni-modal CNNs to get the final reconstruction, as done in [10]. Some works suggest mid-level feature fusion could benefit reconstruction. For instance, [5], [6], [9] propose to fuse the hierarchical features of convolutional layers of two branches, which is still conducted by concatenation. However, the intermediate level features of depth and guidance have different semantic meanings, making the intermediate fusion more challenging. The simple concatenation is not effective for this purpose.

Different from existing guided DSR methods, we propose to hierarchically combine multi-modal features by attention mechanism. We attempt to preserve consistent structures and suppress inconsistent components in a learning manner. As shown in Fig. 3, our proposed fusion scheme consists of a feature enhancement block and a feature recalibration block, which are elaborated in the following.

#### A. Feature Enhancement Block

The guidance image contains rich texture information, which would disturb the depth reconstruction, while depth itself contains noise due to the limitation of sensors. In view of these, we propose to leverage feature enhancement block (FEB) to adaptively select useful information and filter out unwanted one. Specifically, inspired by [31], we leverage gated units as the FEB. Given the extracted multi-modal features  $\mathbf{F}_d^i$  and  $\mathbf{F}_y^i$  ( $1 \leq i \leq m$ ) as inputs, the process of FEB can be formulated as follows:

$$\hat{\mathbf{F}}_d^i = \sigma(\mathbf{W}_{d,1}^i * \mathbf{F}_d^i + b_{d,1}^i) \odot \phi(\mathbf{W}_{d,2}^i * \mathbf{F}_d^i + b_{d,2}^i), \quad (11)$$

$$\hat{\mathbf{F}}_y^i = \sigma(\mathbf{W}_{y,1}^i * \mathbf{F}_y^i + b_{y,1}^i) \odot \phi(\mathbf{W}_{y,2}^i * \mathbf{F}_y^i + b_{y,2}^i), \quad (12)$$

where  $\mathbf{W}_{d,1}^i$ ,  $\mathbf{W}_{d,2}^i$  and  $\mathbf{W}_{y,1}^i$ ,  $\mathbf{W}_{y,2}^i$  are convolutional kernels for depth and guidance respectively, the subscripts 1 and 2 represent two different convolutional operations;  $b_{d,1}^i$ ,  $b_{d,2}^i$  and  $b_{y,1}^i$ ,  $b_{y,2}^i$  are the corresponding bias terms;  $\odot$  denotes element-wise multiplication;  $\phi$  is sigmoid function to limit the output of the gating operation within the range of 0 and 1;  $\sigma$  is PReLU activation function;  $\hat{\mathbf{F}}_d^i$  and  $\hat{\mathbf{F}}_y^i$  are the enhanced features of depth and guidance respectively.

#### B. Feature Recalibration Block

Considering depth and guidance images exhibit notably different appearance characteristics due to the difference in imaging principle, similar pixels in guidance image may have quite different values in depth and vice versa. To unify the similarity metrics, we employ multi-modal feature recalibration block (FRB) to recalibrate the enhanced multi-modal features by FEB. The FRB consists of two units: *multi-modal feature squeeze unit*, which is tailored to learn global joint knowledge from all modalities, and *feature excitation unit*, which uses the learned joint knowledge to adaptively emphasize useful features. Specifically, the multi-modal squeeze unit can be formulated as follows:

$$\hat{\mathbf{F}}_c^i = \sigma(\mathbf{W}_c^i * ([\hat{\mathbf{F}}_d^i, \hat{\mathbf{F}}_y^i]) + b_c^i), \quad (13)$$

$$\hat{\mathbf{F}}_e^i = \text{AvgPool}(\hat{\mathbf{F}}_c^i) + \text{VarPool}(\hat{\mathbf{F}}_c^i), \quad (14)$$

where  $[\cdot, \cdot]$  denotes concatenation operation;  $\mathbf{W}_c^i$  is a  $3 \times 3$  convolutional kernel;  $\text{AvgPool}(\cdot)$  and  $\text{VarPool}(\cdot)$  mean average pooling and variance pooling respectively.  $\hat{\mathbf{F}}_e^i$  is the global joint knowledge learned from all modalities, which is further fed into the feature excitation unit:

$$\mathbf{E}_d^i = \phi(\mathbf{W}_d^i * \hat{\mathbf{F}}_e^i + b_d^i), \quad (15)$$

$$\mathbf{E}_y^i = \phi(\mathbf{W}_y^i * \hat{\mathbf{F}}_e^i + b_y^i), \quad (16)$$

$$\mathbf{F}_f^i = 0.5 \times \mathbf{E}_d^i \odot \hat{\mathbf{F}}_d^i + 0.5 \times \mathbf{E}_y^i \odot \hat{\mathbf{F}}_y^i, \quad (17)$$

where  $\mathbf{W}_d^i$ ,  $\mathbf{W}_y^i$  and  $b_d^i$ ,  $b_y^i$  are  $1 \times 1$  convolutional kernels and bias terms for depth and guidance respectively;  $\phi$  is the sigmoid function;  $\mathbf{E}_d^i$  and  $\mathbf{E}_y^i$  are the excitation signals for depth feature and guidance feature respectively;  $\mathbf{F}_f^i$  denotes the fused multi-modal feature.

TABLE I

MAE PERFORMANCE COMPARISON FOR SCALE FACTORS 4×, 8× AND 16× WITH BICUBIC DEGRADATION ON MIDDLEBURY DATASET. THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND SECOND BEST PERFORMANCE IS THE UNDERScoreD ONES (LOWER MAE VALUES, BETTER PERFORMANCE)

Method	Art			Books			Dools			Laundry			Mobeius			Reindeer			Average		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	1.10	2.10	4.02	0.35	0.67	1.29	0.36	0.68	1.25	0.60	1.12	2.13	0.36	0.70	1.32	0.58	1.09	2.18	0.558	1.060	2.032
CLMF1 [39]	0.74	1.44	2.87	0.28	0.51	1.02	0.34	0.66	1.01	0.50	0.80	1.67	0.29	0.51	0.97	0.51	0.84	1.55	0.443	0.793	1.515
ATGV [40]	0.65	0.81	1.42	0.43	0.51	0.79	0.41	0.52	0.56	0.37	0.89	0.94	0.38	0.45	0.80	0.41	0.58	1.01	0.442	0.627	0.920
DMSG [5]	0.46	0.76	1.53	0.15	0.41	0.76	0.25	0.51	0.87	0.30	0.46	1.12	0.21	0.43	0.76	0.31	0.52	0.99	0.280	0.515	1.005
DGDIE [41]	0.48	1.20	2.44	0.30	0.58	1.02	0.34	0.63	0.93	0.35	0.86	1.56	0.28	0.58	0.98	0.35	0.73	1.29	0.350	0.763	1.370
DEIN [42]	0.40	0.64	1.34	0.22	0.37	0.78	0.22	0.38	0.73	0.23	0.36	0.81	0.20	0.35	0.73	0.26	0.40	0.80	0.255	0.417	0.865
DSRNet [6]	0.63	1.24	2.44	0.23	0.45	0.84	0.24	0.47	0.85	0.36	0.70	1.35	0.24	0.45	0.87	0.38	0.69	1.19	0.347	0.667	1.257
CCFN [43]	0.43	0.72	1.50	0.17	0.36	0.69	0.25	0.46	0.75	0.24	0.41	0.71	0.23	0.39	0.73	0.29	0.46	0.95	0.268	0.467	0.888
GSPRT [8]	0.48	0.74	1.48	0.21	0.38	0.76	0.28	0.48	0.79	0.33	0.56	1.24	0.24	0.49	0.80	0.31	0.61	1.07	0.308	0.543	1.023
DJFR [10]	0.35	0.76	1.88	0.17	0.34	0.74	0.22	0.41	0.79	0.21	0.48	1.10	0.19	0.37	0.75	0.23	0.44	0.99	0.228	0.467	1.042
PacNet [7]	0.34	1.12	2.13	0.19	0.62	1.13	0.23	0.70	1.18	0.22	0.74	1.25	0.19	0.51	0.92	0.24	0.74	1.20	0.235	0.738	1.302
CUNet [12]	0.38	0.99	2.34	0.23	0.54	1.41	0.28	0.60	1.24	0.30	0.72	1.85	0.21	0.47	1.08	0.25	0.59	1.37	0.275	0.652	1.548
MDSR [44]	0.46	0.62	1.87	0.24	0.37	0.73	0.29	0.51	0.79	0.32	0.53	1.11	0.19	0.37	0.74	0.41	0.55	0.95	0.318	0.492	1.032
PMBAN [9]	0.26	<b>0.51</b>	<b>1.22</b>	0.15	<u>0.26</u>	0.59	0.19	<b>0.32</b>	<b>0.59</b>	0.17	0.34	<b>0.71</b>	0.16	<b>0.26</b>	0.67	<u>0.17</u>	0.34	0.74	0.183	<u>0.338</u>	<u>0.753</u>
DKN [11]	<u>0.25</u>	<u>0.52</u>	1.34	0.14	0.27	<u>0.58</u>	<u>0.17</u>	<b>0.32</b>	0.61	0.17	<u>0.33</u>	0.85	<u>0.14</u>	<u>0.27</u>	<u>0.53</u>	<u>0.17</u>	<u>0.33</u>	<u>0.73</u>	0.173	0.340	0.773
CTKT [45]	<u>0.25</u>	0.53	1.44	<b>0.11</b>	<u>0.26</u>	0.67	<b>0.16</b>	<u>0.36</u>	0.65	<u>0.16</u>	0.36	0.76	<b>0.13</b>	<u>0.27</u>	0.69	<u>0.17</u>	0.35	0.77	<u>0.163</u>	0.355	0.830
AHMF	<b>0.22</b>	<b>0.51</b>	<u>1.26</u>	<u>0.13</u>	<b>0.25</b>	<b>0.48</b>	<u>0.17</u>	<b>0.32</b>	<u>0.62</u>	<b>0.15</b>	<b>0.32</b>	<u>0.74</u>	<u>0.14</u>	<b>0.26</b>	<b>0.52</b>	<b>0.15</b>	<b>0.31</b>	<b>0.62</b>	<b>0.157</b>	<b>0.327</b>	<b>0.706</b>

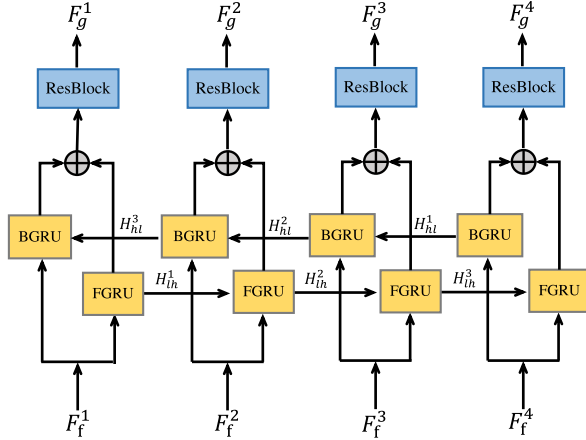


Fig. 4. Hierarchical feature collaboration with four BHFCs, where  $\oplus$  means pixel-wise summation. It takes the fused multi-modal features as inputs and sends them to a bi-directional GRU for multi-level features collaboration.

## V. HIERARCHICAL FEATURE COLLABORATION

It is known to all that in CNN shallower layers encode rich spatial details but lack semantic knowledge, while deeper layers are more effective to capture high-level context and structure information but lose spatial information. Motivated by this observation, we argue that low-level spatial details and high-level structure information should collaborate to boost by each other. Accordingly, we propose a bi-directional hierarchical feature collaboration (BHFC) module to fully leverage the hierarchical fused features.

Specifically, as shown in Fig. 4, we propose to use bi-directional convolutional gated recurrent units (GRU) for hierarchical feature collaboration by propagating information from one layer to all other layers. BHFC takes the fused multi-modal features  $\{F_f^1, \dots, F_f^m\}$  as inputs. Without loss

of generality, we take the  $i$ -th input  $F_f^i$  as an example. It is processed by forward GRU (FGRU) and backward GRU (BGRU), whose results are added together and passed through a residual block [46] for further boosting the hierarchical features. This procedure is formulated as follows:

$$F_g^i = \text{Res}(\text{FGRU}(F_f^i, H_{lh}^{i-1}) + \text{BGRU}(F_f^i, H_{hl}^{m-i})), \quad (18)$$

where  $\text{Res}(\cdot)$  represents a residual block;  $H_{lh}$  and  $H_{hl}$  are the hidden states of FGRU and BGRU respectively. The mathematical model of FGRU can be formulated as follows:

$$Z_{lh}^i = \sigma(W_z^i * [H_{lh}^{i-1}, F_f^i] + b_z^i), \quad (19)$$

$$R_{lh}^i = \sigma(W_r^i * [H_{lh}^{i-1}, F_f^i] + b_r^i), \quad (20)$$

$$\hat{H}_{lh}^i = \tanh(W_h^i * [R_{lh}^i \odot H_{lh}^{i-1}, F_f^i] + b_h^i), \quad (21)$$

$$H_{lh}^i = Z_{lh}^i \odot \hat{H}_{lh}^i + (1 - Z_{lh}^i) \odot H_{lh}^{i-1}, \quad (22)$$

where  $W$  denotes  $3 \times 3$  convolutional kernel, the subscripts of which represent different convolutional layers;  $b$  is the bias term;  $Z_{lh}^i$ ,  $R_{lh}^i$  and  $H_{lh}^i$  represent update gate, reset gate and hidden states, respectively. We set the initial states to zero. BGRU( $\cdot$ ) is with the same formulation as FGRU( $\cdot$ ). FGRU( $\cdot$ ) controls the information flow from low-to-high layers, while BGRU( $\cdot$ ) does it from high-to-low layers.

## VI. EXPERIMENTS

In this section, we conduct several experiments to evaluate the performance of the proposed method. Firstly, we introduce the datasets and evaluation metrics used in our experiments in subsection VI-A. Besides, the experimental settings are presented in subsection VI-B. Moreover, we compare the proposed method with other state-of-the-art depth image SR approaches in subsection VI-C. Then, ablation studies are presented in subsection VI-D to analyze the design choices

of our method. At last, the network complexity comparisons and discussion on limitations of our method are presented in subsection VI-E and subsection VI-F, respectively.

### A. Datasets and Metrics

1) *Middlebury Dataset*: Middlebury dataset is a widely used dataset to evaluate the performance of depth super-resolution algorithms. Following [9], we use 36 RGB-D image pairs (6, 21, and 9 pairs from 2001 [49], 2006 [50] and 2014 [51] datasets, respectively) from Middlebury dataset to train our model. 6 RGB-D image pairs (*Art*, *Books*, *Dools*, *Laundry*, *Mobeius*, *Reindeer*) from Middlebury 2005 [13] are utilized as testing dataset to evaluate the performance of the proposed method. To evaluate the generalization ability of the proposed method, we select 6 RGB-D image pairs from Lu [52] dataset as another test dataset for our method. For middlebury dataset, we utilize the hole-filled depth maps collected by [5], [49]. As suggested by [49], the whole process to fill the depth holes can be divided into three steps: 1) detect the occluded regions by using cross-checking [53], [54], 2) apply a median filter to remove spurious mismatches, 3) fill the holes by surface fitting or by distributing neighboring disparity estimates [55], [56]. We train two kinds of models for two different tasks: (1) depth map super-resolution and (2) joint depth map super-resolution and denoising. Similar to other works [6], [9], [45], [48], we generate the low-resolution depth maps by using Bicubic interpolation and employ Mean Absolute Error (MAE) and root mean squared error (RMSE) to evaluate the objective performance of the proposed method. Following [6], [9], [48], we quantize all recovered depth maps to 8-bits before calculating the MAE or RMSE values for fair evaluation. For both metrics, lower values indicate better performance.

2) *NYU v2 Dataset*: NYU v2 dataset [57] consists of 1449 RGB-D image pairs captured by Microsoft Kinect sensors. Following the similar settings of previous depth map super-resolution methods [10], [11], our method is trained on the first 1000 RGB-D image pairs, and tested on the remained 449 RGB-D images pairs. Following the experimental protocol of Kim *et al.* [11], we use the Bicubic and direct down-sampling to generate the low-resolution depth maps, and utilize RMSE as the default metric to evaluate the performance of the proposed method. To show the robustness of the proposed method, we also conduct experiments on depth maps which are captured by different sensors, such as Lu dataset [52] and Middlebury dataset [13]. Since the acquired depth maps usually have missing values, the depth maps are in-painted by the official toolbox<sup>1</sup> which employs the colorization framework [58] to fill the missing values.

### B. Implementation Details

Our model has only one specific hyper-parameter  $m$ , which is used to control the number of convolutional layers for multi-modal feature extraction. To balance the efficiency and network performance, we set  $m = 4$  as default. We set the channel number of all intermediate layers as 64. The ablation

experiments presented below will verify the effectiveness of our configurations. The kernel size of a convolutional layer is set as  $3 \times 3$ , except for those in the upsampling and downsampling module that are set as  $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$  with a stride size of 0, 1, 2 in  $4\times$ ,  $8\times$ ,  $16\times$  super-resolution, respectively. We use PReLU [38] as the default activation function. During training, we randomly select 32 HR depth maps with the size of  $256 \times 256$  as the ground truth, and the LR depth maps are generated by using the down-sampling operator. Our model is optimized by Adam [59] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 1e - 8$ ; the initial learning rate is  $2 \times 10^{-4}$  and decreased by multiplying 0.5 for every 100 epochs. It takes about 25 hours to train our model. The proposed model is implemented by PyTorch [60] and trained on a NVIDIA V100 GPU.

### C. Comparison With the State-of-the-Arts

1) *Experimental Results on Middlebury Dataset (Noiseless Case)*: For Middlebury dataset, we compare the proposed method with 15 state-of-the-art DSR methods: cross-based local multipoint filtering (CLMF1) [39], anisotropic total generalized variation network (ATGV) [40], deep multi-scale guidance network (DMSG) [5], learned dynamic guidance network (DEDIE) [41], deep edge inference network (DEIN) [42], depth super-resolution network (DSRNet) [6], color-guided coarse-to-fine network (CCFN) [43], guided super-resolution via pixel-to-pixel transformation (GSPRT) [8], deep joint image filter (DJFR) [10], pixel-adaptive convolution neural network (PacNet) [7], common and unique network (CUNet) [12], multi-direction dictionary learning (MDSR) [44], progressive multi-branch aggregation network (PMBAN) [9], deformable kernel network (DKN) [11] and cross-task knowledge network (CTKT) [45]. For fair comparison, we re-train DJFR [10], Pacnet [7], CUNet [12] DKN [11] and PMBAN [9] with the same datasets as ours. The results of other compared methods are obtained from the authors. We report MAE and RMSE values for  $4\times$ ,  $8\times$  and  $16\times$  depth map super-resolution in Table I and Table II, respectively. As can be seen from Table I, with respect to the average MAE, the proposed method achieves the best performance for all scaling factors, especially for  $16\times$  case that is most difficult to recover. The superior performance benefits from the proposed attention-based hierarchical multi-modal fusion strategy, which can recover structure-consistent content and make a plausible prediction for regions with inconsistent contents of the guidance.

We also show the zoomed results of various methods in Fig. 5, from which we can see that most of the existing approaches cannot generate clear boundaries and suffer from various artifacts. Take *Art* as an example, the results of CLMF1 [39] and DEDIE [41] are blur. The results of DEIN [42] suffer from broken edge artifacts. The results of DJFR [10] and DKN [11] cannot generate continuous boundaries of the highlighted pencil region. In contrast, the proposed method can produce sharp boundaries with less artifacts. We attribute this to the proposed feature fusion

<sup>1</sup>[http://cs.nyu.edu/~silberman/code/toolbox\\_nyu\\_depth\\_v2.zip](http://cs.nyu.edu/~silberman/code/toolbox_nyu_depth_v2.zip)

TABLE II  
RMSE PERFORMANCE COMPARISON FOR SCALE FACTORS 4 $\times$ , 8 $\times$  AND 16 $\times$  WITH BICUBIC DEGRADATION ON MIDDLEBURY DATASET.  
THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND SECOND BEST PERFORMANCE IS THE UNDERScoreD  
ONES (LOWER MAE VALUES, BETTER PERFORMANCE)

Method	Art			Books			Dools			Laundry			Mobeius			Reindeer		
	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$
Bicubic	3.87	5.46	8.17	1.27	2.34	3.34	1.31	1.86	2.62	2.06	3.45	5.07	1.33	1.97	2.85	2.42	3.99	5.86
DMSG [5]	1.47	2.46	4.57	0.67	1.03	1.60	<u>0.69</u>	<u>1.05</u>	<u>1.60</u>	<u>0.79</u>	1.51	2.63	0.66	1.02	1.63	0.98	1.76	2.92
DGDIE [41]	2.00	3.84	6.16	0.91	1.68	2.67	0.84	1.54	2.34	1.30	3.37	4.14	0.85	1.86	2.31	1.52	3.88	4.30
DEIN [42]	3.26	4.20	6.40	1.38	2.12	3.38	1.20	1.64	2.27	2.00	2.59	4.07	1.20	1.75	2.97	2.27	2.95	4.21
DSRNet [6]	1.20	2.22	<b>3.90</b>	0.60	0.89	<u>1.51</u>	0.84	1.14	<b>1.52</b>	0.78	<u>1.31</u>	2.26	0.96	1.19	<u>1.58</u>	0.96	<u>1.57</u>	<b>2.47</b>
DJFR [10]	1.62	3.08	5.81	0.54	1.11	2.24	0.78	1.27	2.02	0.90	1.83	3.65	0.68	1.22	2.21	1.25	2.38	4.22
PacNet [7]	1.66	2.92	6.02	0.58	1.03	2.43	0.81	1.25	2.38	0.92	1.79	3.61	0.70	1.18	2.25	1.16	2.12	4.11
CUNet [12]	1.57	2.60	4.74	0.56	0.97	1.96	0.76	1.14	1.97	0.88	1.50	3.00	0.67	1.04	2.00	1.10	1.97	3.25
CGN [47]	1.50	2.69	<u>4.14</u>	0.60	0.97	1.73	0.88	1.20	1.80	0.98	1.57	2.57	0.69	1.06	1.69	1.22	2.02	3.60
DSR_N [48]	1.77	3.32	6.08	0.73	1.39	2.38	0.81	1.35	2.05	0.98	2.06	4.09	0.74	1.37	2.21	1.24	2.39	4.08
MDAR [44]	2.57	3.20	4.87	1.33	1.46	2.51	1.07	1.19	1.90	2.00	2.11	4.07	0.85	1.10	1.98	1.07	1.19	3.44
MFR-SR [18]	1.54	2.71	4.35	0.63	1.05	1.78	0.89	1.22	1.74	1.23	2.06	3.74	0.72	1.10	1.73	1.23	2.06	3.74
PMBAN [9]	1.19	2.47	4.37	<u>0.43</u>	1.10	<u>1.51</u>	0.66	<u>1.08</u>	1.75	0.80	1.54	2.72	<u>0.55</u>	1.13	<u>1.62</u>	<u>0.92</u>	<u>1.76</u>	2.86
DKN [11]	<u>1.12</u>	<u>2.46</u>	4.61	0.44	<u>0.82</u>	1.71	0.71	1.14	1.75	<u>0.79</u>	1.46	<u>2.23</u>	0.59	<u>0.97</u>	1.68	<u>0.92</u>	1.83	3.30
AHMF	<b>1.09</b>	<b>2.14</b>	4.20	<b>0.38</b>	<b>0.72</b>	<b>1.49</b>	<b>0.62</b>	<b>1.03</b>	<u>1.66</u>	<b>0.64</b>	<b>1.22</b>	<b>2.14</b>	<b>0.54</b>	<b>0.88</b>	<b>1.53</b>	<b>0.85</b>	<b>1.56</b>	<u>2.84</u>

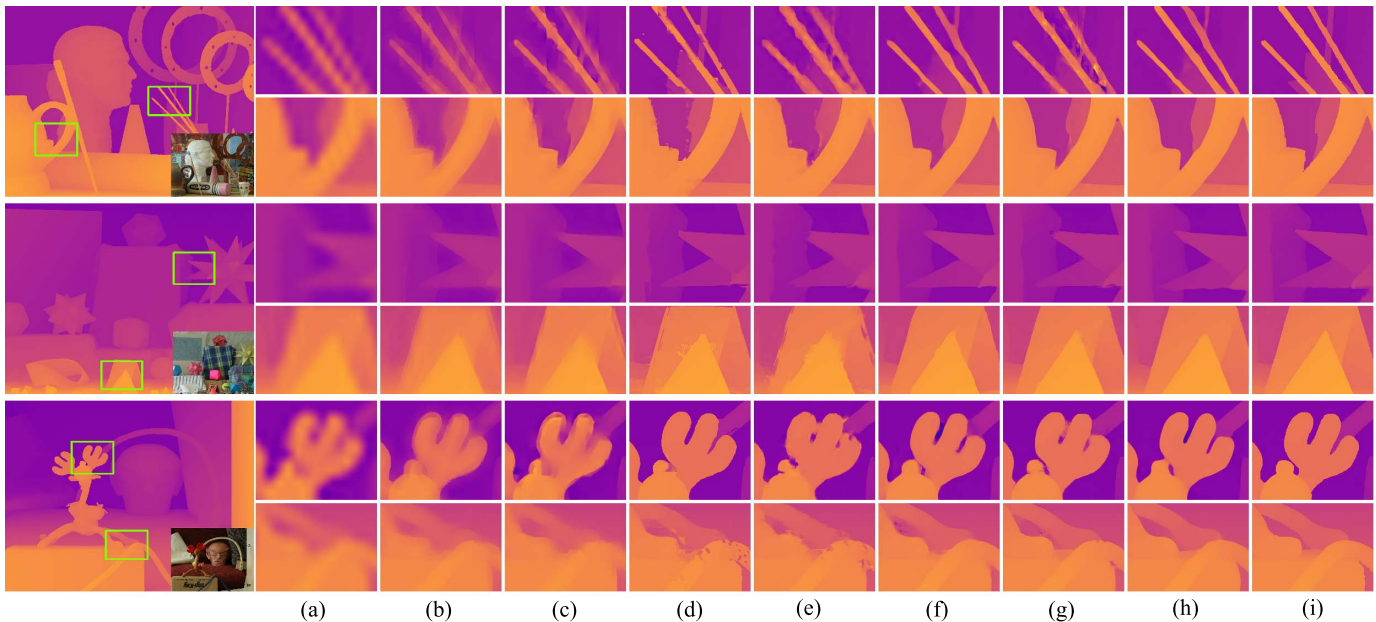


Fig. 5. Visual comparison of 16 $\times$  upsampling results on Art, Moebius and Reindeer from Middlebury dataset [13]: (a): Bicubic, (b): CLMF1 [39], (c): DGDIE [41], (d): DEIN [42], (e): DJFR [10], (f): PMBAN [9], (g): DKN [11], (h): Ours, (i): GT. Please enlarge the PDF for more details.

and collaboration modules, which fully exploit the relevant information from LR depth and HR guidance.

2) *Experimental Results on Middlebury Dataset (Noisy Case)*: Following [6], we simulate the ToF-like degradation by adding Gaussian noise with standard deviation of 5 to the LR depth maps. We list the MAE and RMSE values for 4 $\times$ , 8 $\times$  and 16 $\times$  upsampling on Middlebury dataset in Table III and Table IV, respectively. From these tables, it can be found that the proposed method can better deal with the effect of noise when upsampling the depth maps, even compared to the recently proposed deep learning methods, *i.e.*, CUNet [12], PMBAN [9] and DKN [11]. We further present the qualitative results of 16 $\times$  downsampling and noisy

degradation in Fig. 6. We can see that our proposed method achieves the clearest and sharpest results. Both CUNet [12] and DKN [11] generate competitive MAE scores, but suffer from edge diffusion artifacts.

3) *Evaluation on Generalization Ability Across Datasets*: To verify the generalization ability of the proposed method, we test our method on Lu dataset, which is from different resources to the training dataset. The comparison study is conducted with DepthSR [6], CUNet [12], PacNet [7], DJFR [10], PMBAN [9] and DKN [11], which are all trained on the same datasets for fair comparison. The quantitative results are illustrated in Table V (MAE) and Table VI (RMSE), from which we can see that the proposed method obtains the best

TABLE III

MAE COMPARISON FOR SCALE FACTORS 4 $\times$ , 8 $\times$  AND 16 $\times$  WITH TOF-LIKE DEGRADATION ON MIDDLEBURY DATASET. THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND THE SECOND BEST PERFORMANCE IS THE UNDERScored ONES (LOWER MAE VALUES, BETTER PERFORMANCE)

Method	Art			Books			Dools			Laundry			Mobeius			Reindeer		
	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$
Bicubic	3.23	4.00	5.58	2.62	2.83	3.27	2.61	2.75	3.10	2.79	3.96	4.14	2.63	2.84	3.21	2.84	3.18	4.05
GF [61]	1.91	2.87	4.90	1.13	1.84	2.80	1.14	1.86	2.69	1.31	2.12	3.39	1.17	1.87	2.77	1.33	2.10	3.56
DMSG [5]	0.84	1.57	2.98	0.62	1.18	1.48	0.84	1.12	1.78	0.78	1.03	1.89	0.66	1.13	1.76	0.57	1.12	1.87
DGDIE [41]	0.99	1.84	3.34	0.81	1.29	2.04	0.95	1.39	2.05	1.10	1.73	2.67	0.84	1.37	2.16	0.79	1.33	2.19
DSRNet [6]	2.19	2.57	3.69	1.81	1.82	2.16	1.84	1.91	2.21	1.93	2.07	2.66	1.81	1.84	2.26	1.95	2.09	2.59
GSPRT [8]	0.68	1.33	2.47	0.52	0.87	1.37	0.78	1.26	2.03	0.76	1.24	1.86	0.65	1.03	1.68	0.55	1.04	1.70
DJFR [10]	0.82	1.44	2.77	0.62	1.03	1.67	0.73	1.12	1.66	0.72	1.23	2.02	0.66	1.05	1.71	0.65	1.11	1.88
PacNet [7]	0.72	1.40	2.45	0.49	0.83	1.47	0.71	1.15	1.64	0.64	1.22	1.92	0.56	0.93	1.86	0.57	1.03	1.38
CUNet [12]	0.81	1.25	2.35	0.54	0.85	1.47	0.74	1.07	1.78	0.66	1.05	2.00	0.60	0.93	1.64	0.61	0.96	1.73
PMBAN [9]	<b>0.59</b>	0.98	1.89	<u>0.44</u>	<b>0.71</b>	1.23	<b>0.64</b>	1.01	1.56	<u>0.54</u>	<u>0.89</u>	<b>1.62</b>	<u>0.48</u>	0.81	<u>1.30</u>	<b>0.47</b>	<u>0.78</u>	1.52
DKN [11]	0.68	<u>0.95</u>	<u>1.86</u>	0.51	<u>0.80</u>	<u>1.21</u>	<u>0.67</u>	<u>0.99</u>	<b>1.46</b>	0.62	1.00	1.72	0.55	<u>0.80</u>	1.37	<u>0.54</u>	<u>0.78</u>	<u>1.50</u>
AHMF (Ours)	<u>0.62</u>	<b>0.93</b>	<b>1.85</b>	<b>0.41</b>	<b>0.71</b>	<b>1.19</b>	0.68	<b>0.94</b>	<u>1.47</u>	<b>0.52</b>	<b>0.88</b>	<u>1.63</u>	<b>0.46</b>	<b>0.79</b>	<b>1.28</b>	<b>0.47</b>	<b>0.76</b>	<b>1.48</b>

TABLE IV

RMSE PERFORMANCE COMPARISON FOR SCALE FACTORS 4 $\times$ , 8 $\times$  AND 16 $\times$  WITH TOF-LIKE DEGRADATION ON MIDDLEBURY DATASET. THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND SECOND BEST PERFORMANCE IS THE UNDERScored ONES (LOWER MAE VALUES, BETTER PERFORMANCE)

Method	Art			Books			Dools			Laundry			Mobeius			Reindeer		
	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$	4 $\times$	8 $\times$	16 $\times$
Bicubic	5.74	6.89	9.12	4.57	4.86	5.28	4.50	4.69	5.02	4.91	5.47	6.48	4.48	4.72	5.23	5.11	5.80	7.12
DMSG [5]	3.06	4.48	6.44	1.90	3.13	4.63	1.99	3.17	4.51	2.25	3.60	5.39	1.97	3.22	4.50	2.32	3.75	5.35
DGDIE [41]	4.29	5.94	9.78	1.99	2.65	4.16	1.74	2.28	3.18	3.00	4.21	6.80	1.69	2.37	3.54	3.40	4.47	7.86
DSRNet [6]	–	–	6.24	–	–	3.36	–	–	–	–	–	4.95	–	–	–	–	–	4.65
DJFR [10]	2.96	4.81	7.30	1.41	2.58	3.47	1.91	2.57	4.33	2.55	3.07	5.43	2.43	2.62	4.35	2.60	3.21	5.66
PacNet [7]	2.95	4.41	7.80	1.37	2.05	3.80	1.76	2.45	3.87	2.01	3.09	4.99	1.61	2.26	3.84	2.07	3.24	5.79
CUNet [12]	2.85	<u>3.84</u>	6.12	1.42	1.95	2.95	1.71	2.24	<u>2.93</u>	1.93	2.82	4.23	1.55	2.08	<b>2.98</b>	2.01	3.01	<u>4.20</u>
PMBAN [9]	<b>2.58</b>	3.92	6.67	<u>1.25</u>	<u>1.82</u>	2.94	<u>1.52</u>	2.18	3.10	<u>1.69</u>	2.71	<b>3.67</b>	<u>1.38</u>	<b>1.93</b>	3.08	<u>1.81</u>	<b>2.72</b>	4.58
DKN [11]	<u>2.77</u>	<b>3.79</b>	<u>6.01</u>	1.36	1.91	<u>2.93</u>	1.68	<u>2.16</u>	<b>2.85</b>	1.89	<u>2.69</u>	4.15	1.49	2.10	<u>2.99</u>	1.93	2.99	4.24
AHMF (Ours)	<b>2.58</b>	3.85	<b>6.00</b>	<b>1.21</b>	<b>1.80</b>	<b>2.75</b>	<b>1.47</b>	<b>2.04</b>	<b>2.85</b>	<b>1.63</b>	<b>2.45</b>	<u>3.79</u>	<b>1.33</b>	<u>2.03</u>	<b>2.98</b>	<b>1.78</b>	<u>2.80</u>	<b>4.18</b>

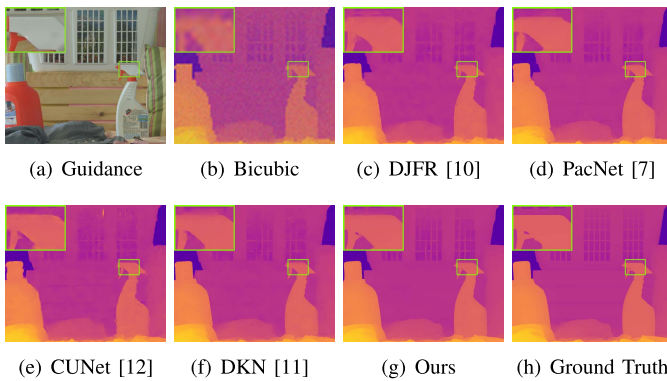


Fig. 6. Visual comparison of 16 $\times$  results on *Laundry* from Middlebury [13] dataset. Please enlarge the PDF for more details.

performance among compared methods for most upsampling factors. We show the visual comparison of the compared methods in Fig 7. From Fig. 7(a), it can be found that, in the initial bicubic upsampling, there are holes in a large region. The method DGDIE [41] works well in handling the holes due to that it performs guided filtering iteratively. The deep neural networks based methods, such as CUNet [12], PacNet [7],

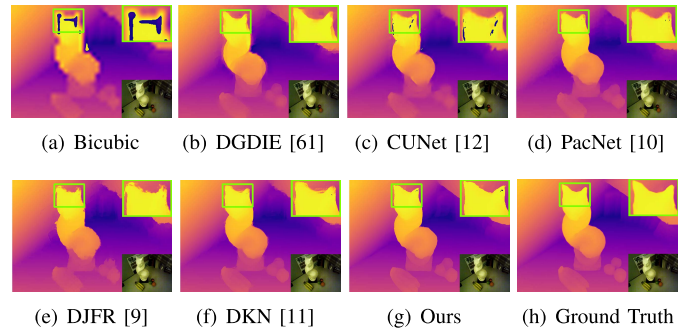


Fig. 7. Visual comparison of 16 $\times$  results on *Image\_05* from Lu dataset [52]. Please enlarge the PDF for more details.

DJFR [10], DKN [11] and ours, work in an end-to-end learning manner, which all suffer from the hole artifact more or less. Compared with other deep neural networks based methods, our scheme produces much sharper edges, as shown in the highlighted region.

4) *Experimental Results on NYU v2 Dataset*: To evaluate the effectiveness of the proposed method, we conduct experiments on NYU v2 dataset. We compare our method with

TABLE V

MAE VALUES FOR SCALE FACTORS 4×, 8× AND 16× WITH BICUBIC DEGRADATION ON LU DATASET. THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND SECOND BEST PERFORMANCE IS THE UNDERScored ONES (LOWER MAE VALUES, BETTER PERFORMANCE)

Method	Image_01			Image_02			Image_03			Image_04			Image_05			Image_06		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	0.57	1.32	2.46	0.82	2.06	4.34	0.40	0.93	1.85	0.73	1.73	3.44	0.56	1.47	3.06	0.52	1.24	2.58
DSRNet [6]	0.57	1.01	1.89	0.82	1.48	3.08	0.40	0.83	1.56	0.73	1.21	2.52	0.56	1.10	2.12	0.52	0.85	1.85
DEDIE [41]	0.53	1.19	2.11	0.83	1.81	3.57	0.45	1.02	1.73	0.56	1.61	2.81	0.55	1.30	2.39	0.48	1.19	2.19
CUNet [12]	0.57	1.01	1.89	0.82	1.48	3.08	0.4	0.83	1.56	0.73	1.21	2.52	0.56	1.10	2.12	0.52	0.85	1.85
PacNet [7]	0.61	1.34	2.77	0.84	1.65	3.52	0.52	1.14	2.22	0.61	1.32	2.55	0.57	1.12	2.38	0.52	1.03	2.15
DJFR [10]	0.53	1.11	2.35	0.74	1.47	3.08	0.41	0.86	1.80	0.55	1.09	2.36	0.53	0.93	1.94	0.43	0.75	1.53
PMBAN [9]	0.54	<b>0.79</b>	1.83	0.74	<u>1.17</u>	2.83	0.42	<u>0.72</u>	<u>1.36</u>	0.57	<u>0.84</u>	<b>1.86</b>	0.55	<b>0.74</b>	<u>1.61</u>	0.49	<u>0.64</u>	<u>1.13</u>
DKN [11]	<u>0.48</u>	<u>0.80</u>	<u>1.80</u>	<u>0.69</u>	<b>1.16</b>	<u>2.86</u>	<u>0.38</u>	0.73	1.75	<u>0.50</u>	0.92	2.34	<u>0.50</u>	0.78	1.84	<u>0.40</u>	<u>0.64</u>	1.26
AHMF (Ours)	<b>0.42</b>	0.82	<b>1.76</b>	<b>0.64</b>	1.19	<b>2.78</b>	<b>0.36</b>	<b>0.67</b>	<b>1.32</b>	<b>0.45</b>	<b>0.79</b>	<u>1.90</u>	<b>0.46</b>	<u>0.76</u>	<b>1.53</b>	<b>0.39</b>	<b>0.61</b>	<b>1.07</b>

TABLE VI

RMSE VALUES FOR SCALE FACTORS 4×, 8× AND 16× WITH BICUBIC DEGRADATION ON LU DATASET. THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND SECOND BEST PERFORMANCE IS THE UNDERScored ONES (LOWER RMSE VALUES, BETTER PERFORMANCE)

Method	Image_01			Image_02			Image_03			Image_04			Image_05			Image_06		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	2.18	3.93	6.10	3.55	6.72	11.19	1.36	2.42	4.18	2.71	4.76	7.55	2.39	4.92	7.88	2.32	4.53	7.37
DGDIE [41]	1.90	3.33	5.50	3.33	6.09	10.41	1.21	2.46	4.20	1.81	4.69	7.23	2.17	4.20	6.57	2.03	4.34	7.01
DSRNet [6]	1.47	<u>2.34</u>	5.25	<b>2.01</b>	<b>3.45</b>	8.18	0.92	<b>1.60</b>	3.36	1.53	2.28	4.92	1.65	<b>2.29</b>	3.65	<u>1.39</u>	<b>1.83</b>	<u>2.97</u>
CUNet [12]	1.55	2.47	<u>5.23</u>	3.08	4.04	7.98	1.15	2.11	3.95	1.73	2.17	4.93	2.31	2.80	3.98	2.12	2.43	3.35
PacNet [7]	1.81	2.83	5.66	3.54	4.29	8.69	1.38	2.08	4.42	1.98	2.35	5.27	2.63	2.79	4.53	2.46	2.37	3.79
DJFR [10]	2.10	2.98	7.46	3.53	4.69	10.60	1.92	2.14	5.81	3.46	2.97	9.21	2.96	2.99	7.20	2.48	2.65	5.34
PMBAN [9]	<u>1.35</u>	2.51	5.35	2.37	3.98	<u>7.60</u>	<u>0.85</u>	1.87	<u>3.04</u>	1.37	<b>2.12</b>	5.07	1.66	2.74	3.58	1.58	2.39	3.03
DKN [11]	1.37	2.67	5.46	2.65	4.38	7.80	0.87	1.82	4.24	<u>1.36</u>	2.46	<u>4.75</u>	<u>1.44</u>	2.94	<u>3.56</u>	1.57	2.57	3.70
AHMF (Ours)	<b>1.23</b>	<b>2.22</b>	<b>4.92</b>	<u>2.21</u>	<u>3.97</u>	<b>7.47</b>	<b>0.72</b>	<u>1.65</u>	<b>3.03</b>	<b>1.11</b>	<u>2.14</u>	<b>4.63</b>	<b>1.36</b>	<u>2.69</u>	<b>3.51</b>	<b>1.34</b>	<u>2.36</u>	<b>2.87</b>

TABLE VII

AVERAGE RMSE PERFORMANCE COMPARISON FOR SCALE FACTORS 4×, 8× AND 16× WITH DIRECT DOWNSAMPLING DEGRADATION. THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND SECOND BEST PERFORMANCE IS THE UNDERScored ONES. FOR NYU v2 [57] WE CALCULATE IN CENTIMETER, FOR OTHER DATASETS WE CALCULATE RMSE WITH DEPTH VALUE SCALED TO [0, 255] (LOWER RMSE VALUES, BETTER PERFORMANCE)

Method	Middlebury [13]			Lu [52]			NYU v2 [57]			Average		
	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
Bicubic	4.44	7.58	11.87	5.07	9.22	14.27	8.16	14.22	22.32	5.89	10.34	16.15
MRF [62]	4.26	7.43	11.80	4.90	9.03	14.19	7.84	13.98	22.20	5.67	10.15	16.06
GF [61]	4.01	7.22	11.70	4.87	8.85	14.09	7.32	13.62	22.03	5.40	9.90	15.94
TGV [16]	3.39	5.41	12.03	4.48	7.58	17.46	6.98	11.23	28.13	4.95	8.07	19.21
Park [63]	2.82	4.08	7.26	4.09	6.19	10.14	5.21	9.56	18.10	4.04	6.61	11.83
Ham [64]	3.14	5.03	8.83	4.65	7.73	11.52	5.27	12.31	19.24	4.35	8.36	13.20
JBU [14]	2.44	3.81	6.13	2.99	5.06	7.51	4.07	8.29	13.35	3.17	5.72	9.00
DGF [65]	3.92	6.04	10.02	2.73	5.98	11.73	4.50	8.98	16.77	3.72	7.00	12.84
DJF [66]	2.14	3.77	6.12	2.54	4.71	7.66	3.54	6.20	10.21	2.74	4.89	8.00
DMSG [5]	2.11	3.74	6.03	2.48	4.74	7.51	3.37	6.20	10.05	2.65	4.89	7.86
PacNet [7]	1.91	3.20	5.60	2.48	4.37	6.60	2.82	5.01	8.64	2.40	4.19	6.95
DJFR [10]	1.98	3.61	6.07	<u>2.21</u>	<b>3.75</b>	7.53	3.38	5.86	10.11	2.52	4.41	7.90
DSRNet [6]	2.08	3.26	5.78	2.57	4.46	6.45	3.49	5.70	9.76	2.71	4.47	7.30
FDKN [11]	2.21	3.64	6.15	2.64	4.55	7.20	2.63	4.99	8.67	2.49	4.39	7.34
DKN [11]	<u>1.93</u>	<u>3.17</u>	<u>5.49</u>	2.35	4.16	<u>6.33</u>	<u>2.46</u>	<u>4.76</u>	<u>8.50</u>	<u>2.25</u>	<u>4.03</u>	<u>6.77</u>
AHMF (Ours)	<b>1.79</b>	<b>2.81</b>	<b>5.02</b>	<b>2.00</b>	<u>3.83</u>	<b>6.21</b>	<b>2.25</b>	<b>4.50</b>	<b>8.10</b>	<b>2.01</b>	<b>3.71</b>	<b>6.41</b>

DJFR [10], DKN [11] and PacNet [7]. For fairness, we re-train PacNet [7] and DSRNet [6] with the same training dataset as ours. Since these methods are evaluated by RMSE index in their original papers, we report the RMSE performance for direct downsampling and Bicubic downsampling in Table VII and Table VIII, respectively. Obviously, our method

outperforms all compared methods with a large margin for average RMSE values. To further analyze the performance of our method, we show visual results at 8× upsampling for *Image\_1241*, *Image\_1242*, *Image\_1339* and *Image\_1360* in Fig. 8. As can be seen, the results of GF [61] are over-smoothed due to that the local filter cannot capture global

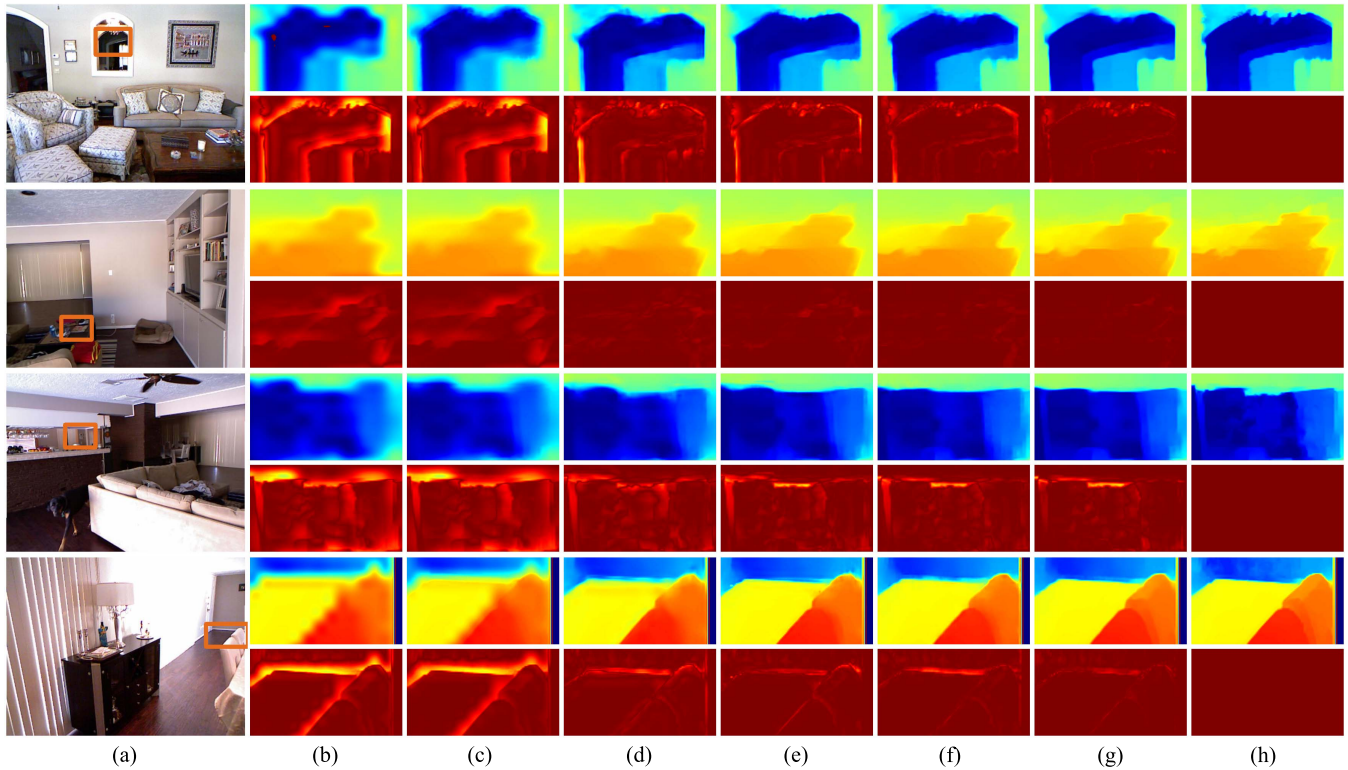


Fig. 8. Visual comparison of  $8\times$  results on *Image\_1339*, *Image\_1242*, *Image\_1241* and *Image\_1360* from NYU v2 [57] dataset: (a): Guidance Image, (b): Bicubic, (c): GF [61] (d): DJFR [10], (e): DKN [11], (f): PacNet [7], (g): Ours, (h): GT. For each sample, the second row shows the error map between the results and ground truth. In the error map, brighter area means the larger error. Please enlarge the PDF for more details.

TABLE VIII

AVERAGE RMSE PERFORMANCE COMPARISON FOR SCALE FACTORS  $4\times$ ,  $8\times$  AND  $16\times$  WITH BICUBIC DEGRADATION. THE BEST PERFORMANCE IS SHOWN IN **BOLD** AND SECOND BEST PERFORMANCE IS THE UNDERScoreD ONES. FOR NYU v2 [57] WE CALCULATE IN CENTIMETER, FOR OTHER DATASETS WE CALCULATE RMSE WITH DEPTH VALUE SCALED TO [0, 255] (LOWER RMSE VALUES, BETTER PERFORMANCE)

Method	Middlebury [13]			Lu [52]			NYU v2 [57]			Average		
	$4\times$	$8\times$	$16\times$	$4\times$	$8\times$	$16\times$	$4\times$	$8\times$	$16\times$	$4\times$	$8\times$	$16\times$
Bicubic	2.47	4.65	7.49	2.63	5.23	8.77	4.71	8.29	13.17	3.27	6.06	9.81
GF [61]	3.24	4.36	6.79	4.18	5.34	8.02	5.84	7.86	12.41	4.42	5.85	9.07
TGV [16]	1.87	6.23	17.01	1.98	6.71	18.31	3.64	10.97	39.74	2.50	7.97	25.02
DGF [65]	1.94	3.36	5.81	2.45	4.42	7.26	3.21	5.92	10.45	2.53	4.57	7.84
DJF [66]	1.68	3.24	5.62	1.65	3.96	6.75	2.80	5.33	9.46	2.04	4.18	7.28
DMSG [5]	1.88	3.45	6.28	2.30	4.17	7.22	3.02	5.38	9.17	2.40	4.33	7.17
DJFR [10]	1.32	3.19	5.57	1.15	3.57	6.77	2.38	4.94	9.18	1.62	3.90	7.17
DSRNet [6]	1.77	3.05	4.96	1.77	3.10	<u>5.11</u>	3.00	5.16	8.41	2.18	3.77	6.16
PacNet [7]	1.32	2.62	4.58	1.20	2.33	5.19	1.89	3.33	6.78	1.47	2.76	5.53
FDKN [11]	<u>1.08</u>	2.17	4.50	<b>0.82</b>	2.10	5.05	1.86	3.58	6.96	1.25	2.62	5.50
DKN [11]	1.23	<u>2.12</u>	4.24	0.96	<u>2.16</u>	<u>5.11</u>	<u>1.62</u>	<u>3.26</u>	<u>6.51</u>	1.27	2.51	5.29
AHMF (Ours)	<b>1.07</b>	<b>1.63</b>	<b>3.14</b>	<u>0.88</u>	<b>1.66</b>	<b>3.71</b>	<b>1.40</b>	<b>2.89</b>	<b>5.64</b>	<b>1.12</b>	<b>2.06</b>	<b>4.16</b>

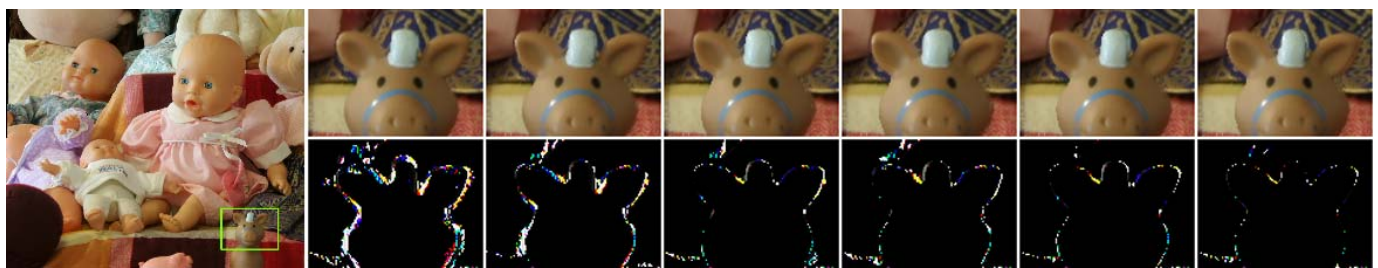


Fig. 9. Visual comparisons of depth image based rendering on *Dolls* from Middlebury [13] dataset. The second row shows the error maps between the results and ground truth, in which brighter areas mean larger errors. Please enlarge the PDF for more details.

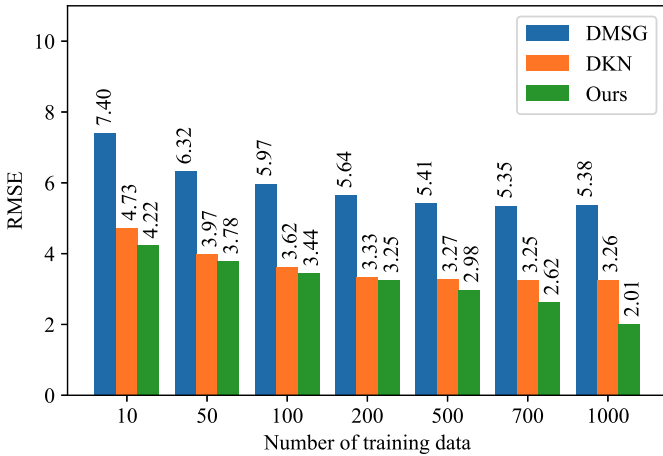


Fig. 10. RMSE comparison by varying the number of training data on NYU v2 [57] dataset for 4 $\times$  depth map super-resolution.

TABLE IX  
PSNR VALUES FOR DEPTH IMAGE BASED RENDERING

Method	Art	Books	Dolls	Laundry	Moebius	Reindeer
Bicubic	25.43	32.42	34.33	26.19	32.53	29.76
DSRNet [6]	27.49	33.71	34.65	28.10	33.60	30.00
CUNet [12]	29.78	33.77	35.44	28.00	34.09	33.58
PacNet [7]	31.52	35.08	35.50	30.50	34.85	35.45
DJFR [10]	32.31	35.95	36.72	30.55	35.54	34.37
PMBAN [9]	<u>33.01</u>	<u>37.44</u>	<u>37.14</u>	<u>31.82</u>	<u>35.96</u>	<u>35.98</u>
Ours	<b>34.57</b>	<b>37.87</b>	<b>37.85</b>	<b>32.64</b>	<b>37.22</b>	<b>38.49</b>

information. The results of DJFR [10] and DKN [11] suffer from diffusion artifacts. The result of PacNet [7] can preserve the local details, but cannot reconstruct the boundary well. On the contrary, our method generates depth maps with smaller reconstruction errors than other compared methods. Besides, to demonstrate the robustness of our method for different size of training data, we train our model by varying the number of training data, and report the RMSE values in Fig. 10. As the result shows, our method achieves the best performance for all case. In particular, the proposed method trained with only 500 images obtains the similar performance with DKN [11] which is the state-of-the-art approach on NYU v2 [57] dataset.

5) *Experimental Results on Depth Image-Based Rendering*: Similar to [67], we take depth image based rendering (DIBR) as a measurement to evaluate the performance of the depth SR methods. The stereo depth map used for DIBR is down-sampled by bicubic with 8 $\times$  scaling factor, then the super-resolved depth maps are used to perform DIBR. Table IX tabulates the PSNR values for the compared methods on Middlebury dataset. It can be observed that the proposed method achieves the best performance for all test images. We present the visual comparison results and error maps for image *Dolls* in Fig. 9. Clearly, the proposed AHMF method can produce visually appealing results with smaller error maps.

#### D. Ablation Study

In this subsection, we conduct ablation studies to verify influence of different configurations to the final performance and the effectiveness of the proposed multi-modal attention fusion (MMAF) and bi-directional hierarchical feature collaboration (BHFC) module.

1) *Influence of Different Configurations to the Final Performance*: We first investigate effect of various parameter settings and effect of different loss functions to the final performance.

*Effect of Various Parameter Settings*: To train the proposed deep model, there are some critical parameters needed to carefully set, including the number of feature extraction layers  $m$  and the number of feature channels  $n$ . We provide experimental analysis about the influence of various settings to the final performance. Fig. 12 shows the MAE trend with the layer number  $m$ . It can be found that the MAE values decrease rapidly before  $m = 4$ , and the larger  $m$  tends to lead to the better reconstruction performance, which however is at the expense of network complexity. For example, the model's parameters are 4.20M when  $m = 8$ , which are about twice larger than that of  $m = 4$  (2.54M). In the proposed method, we empirically choose  $m = 4$  to obtain a good trade-off between the network complexity and reconstruction performance. Moreover, we plot the MAE curve with the feature channel number  $n$  in Fig. 13. It can be found that the MAE values decrease rapidly before  $n = 64$ . With similar purpose as  $m$ , we set  $n$  as 64 in our model.

*Effect of Different Loss Functions*: We investigate the influence of different loss functions to the final performance. The comparison study group includes  $L_1$ ,  $L_2$  and perceptual loss. The experimental analysis is conducted on Middlebury [13], NYU v2 [57] dataset and Lu [52] dataset. The comparison results are provided in Table XI, from which it can be found that, the  $L_1$  loss achieves the best results with respect to MAE for all test datasets, and achieves the best results with respect to RMSE for two datasets. This is mainly because  $L_1$  is more robust to outliers, thus the depth boundaries are preserved better than  $L_2$  and perceptual loss. According to this analysis, we choose  $L_1$  as our loss function.

2) *Influence of Different Modules to the Final Performance*: We further investigate the role of multi-modal attention fusion and hierarchical feature collaboration to the final performance. As shown in Table X, the experimental analysis is conducted on Middlebury 2005 dataset [13] with five different variants. For fair comparison, we carefully adjust the feature channels of the networks to guarantee that different variants have roughly the same size of parameters as that of our full model. Specifically, for 4 $\times$ , 8 $\times$  and 16 $\times$  super-resolution, the sizes of parameters of all variants are around 2.54M, 3.36M and 5.75M, respectively.

*Effect of Multi-Modal Attention Fusion*: The multi-modal attention fusion model is used to fuse the extracted depth and guidance features. To validate the effectiveness of our proposed strategy, we compare it with two alternatives: 1) `Module1`, we replace the MMAF by addition operation to fuse the extracted features but keep the BHFC unchanged, 2) `Module2`, in which we use concatenation operation to fuse the extracted features but keep the BHFC unchanged. We report the results in Table X. Compared with the proposed MMAF, simply fusing the multi-modal features by addition or concatenation significantly worsen the results, especially for the large up-sampling factor. We can observe that the result of `Module2` is a slightly better than that of `Module1`. This is easy to understand, since the features of depth and guidance are heterogeneous, and the same pixel values may represent

TABLE X

ABLATION STUDY OF MULTI-MODAL ATTENTION BASED FUSION (MMAF) AND BI-DIRECTIONAL HIERARCHICAL FEATURE COLLABORATION (BHFC). THE EXPERIMENTS ARE CONDUCTED ON MIDDLEBURY 2005 [13] DATASET. WE REPORT THE AVERAGE MAE VALUES FOR ALL VARIANTS

Model	Multi-Modal Fusion	Hierarchical Feature Collaboration	MAE		
			4×	8×	16×
Model1	Addition	Forward+Backward	0.1621	0.3391	0.7204
Model2	Concatenation	Forward+Backward	0.1586	0.3307	0.7126
Model3	MMAF	w/o Feature Collaboration	0.1640	0.3329	0.7411
Model4	MMAF	w/o Backward	0.1592	0.3315	0.7337
Model5	MMAF	w/o Forward	0.1584	0.3305	0.7318
Model6 (our full model)	MMAF	Forward+Backward (BHFC)	<b>0.1574</b>	<b>0.3269</b>	<b>0.7058</b>

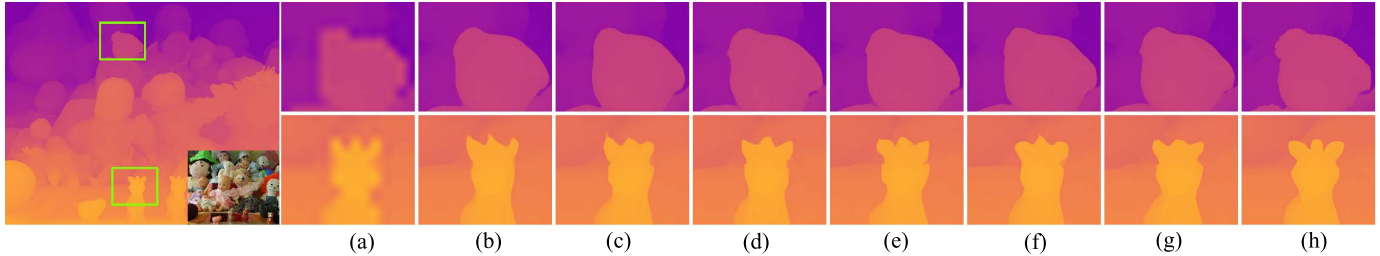


Fig. 11. Visual comparison of 16× depth map super-resolution on *Dools* from Middlebury 2005 dataset [13]: (a): Bicubic, (b): Model1, (c): Model2, (d): Model3, (e): Model4, (f): Model5, (g): Model6 (ours full models, AHMF). Please enlarge the PDF for more details.

TABLE XI

PERFORMANCE COMPARISON (MAE / RMSE) FOR 4× DEPTH MAP SUPER-RESOLUTION WITH DIFFERENT LOSS FUNCTIONS

Loss	Middlebury [13]	NYU v2 [57]	Lu [52]
$L_1$	<b>0.1574 / 0.7021</b>	<b>0.4533 / 1.4013</b>	<b>0.4467 / 1.3283</b>
$L_2$	0.2160 / <b>0.6972</b>	0.5514 / 1.4844	0.4492 / 1.4843
$L_1$ +Perceptual Loss	0.2237 / 0.7594	0.5834 / 1.6131	0.4837 / 1.5190

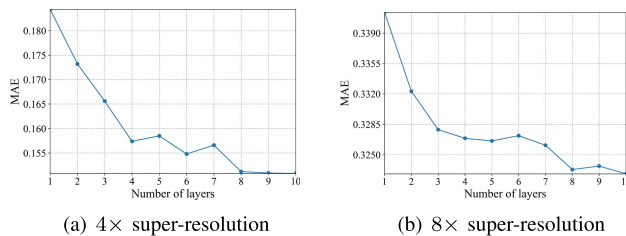


Fig. 12. Ablation study of the number of feature extraction layers  $m$  on Middlebury dataset [13]. We report the average MAE for all variants.

totally different objects. Directly fuse them by addition would destroy the useful information. Fig. 11 further shows the visual comparisons of different variants for 16× depth map super-resolution. As can be seen, Model1 and Model2 cannot generate accurate depth boundaries for small objects, *e.g.*, head of the toy, since the input low-resolution depth image is severely damaged. In contrast, our method clearly reconstructs the depth boundaries and obtains the best performance.

Moreover, we visualize the input and output feature maps of Fusion by Concatenation (the first row) and MMAF (the second row) in Fig. 14. Taking the first row as an example, Fig. 14 (a) and Fig. 14 (b) are extracted guidance and depth feature maps, respectively. Fig. 14 (c) ( $F_f^1$  in Eq. 6) and Fig. 14 (d) ( $F_f^2$  in Eq. 6) show the features

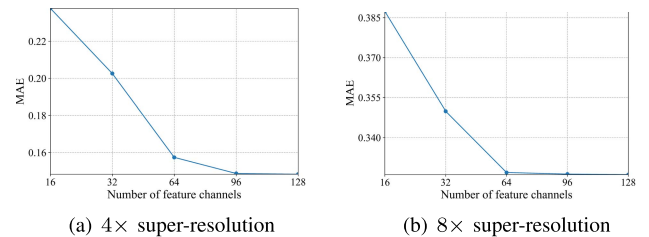


Fig. 13. Ablation study of the number of feature channels  $n$  on Middlebury dataset [13]. We report the average MAE for all variants.

fused by Concatenation. From Fig. 14, we can find that: 1) the extracted guidance feature contains a lot of redundant texture information; 2) both of Concatenation and MMAF can fully fuse the multi-modal features of the consistent regions, *e.g.*, the depth boundaries of the large-scale objects are enhanced; 3) for the inconsistent components, as highlighted in Fig. 14 (d) and Fig. 14 (h), the proposed MMAF can sufficiently filter out the redundant texture information while Concatenation suffers from texture-copying artifacts. This is mainly because that the proposed MMAF selects useful information and neglects the erroneous structures by the attention mechanism.

*Effect of Hierarchical Feature Collaboration:* In deep convolution neural networks, the low-level features typically contain rich spatial details, while the high-level features usually contain sufficient structure information, and the hierarchical features are complementary to each other. To fully leverage these hierarchical features, we propose a bi-directional hierarchical feature fusion module (BHFC) and it contains a forward and a backward process to facilitate the hierarchical features propagation and collaboration with each other. To validate this, we compare it with three different models, 1): Model3,

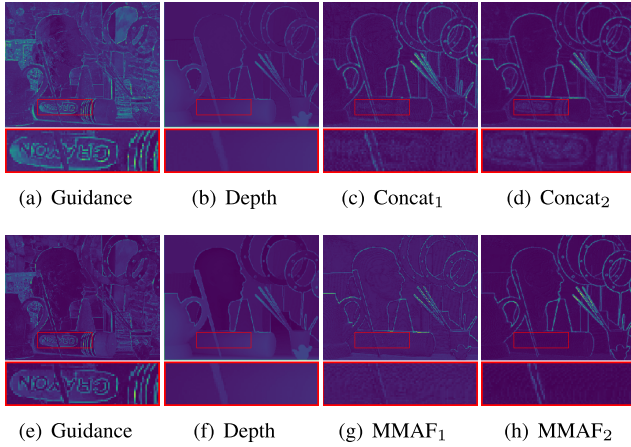


Fig. 14. Visualization of feature maps (*Art* for  $4\times$  depth map super-resolution). (a) and (e) features maps extracted from guidance image; (b) and (f) features maps extracted from depth image; (c) and (g) feature maps of the first layer fused by concatenation (the first row) and the proposed MMAF (the second row), respectively; (d) and (h) feature maps of the second layer fused by concatenation (the first row) and the proposed MMAF (the second row), respectively. Please enlarge the PDF for more details.

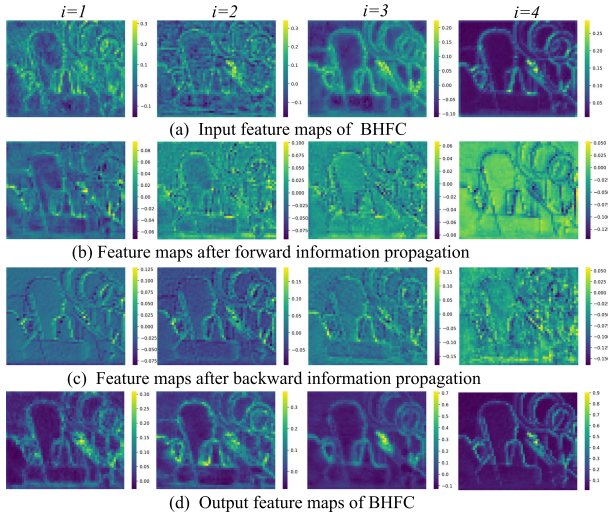


Fig. 15. Average feature maps before and after BHFC,  $i$  is the  $i$ -th layer.

we remove the BHFC in our model; 2) Model4, we remove the backward feature propagation process for this model, 3) Model5, we remove the forward feature propagation process for this model. The quantitative results are illustrated in Table X and the qualitative results are presented in Fig. 11. The performance drop of the Model3 verifies our motivation that fuse hierarchical features is necessary in depth map super-resolution. In Model4, only the low-level features can propagate to the high-level features, suffer from a slightly performance drop. The same conclusion can be drawn for the model Model5. Compare with these three variants, our full model Model6 with both forward and backward process achieves the best performance, especially for large scaling factors which is more to recover. This phenomenon further reinforces the hypothesis that the proposed BHFC plays a significant role in our model.

To further clarify the mechanism of hierarchical feature collaboration, we visualize the average feature maps of BHFC in Fig. 15. Two observations can be obtained. First, the features

TABLE XII

AVERAGE RUNNING TIME, GPU MEMORY AND NETWORK PARAMETERS COMPARISON FOR  $4\times$  SUPER-RESOLUTION OF DIFFERENT METHODS

Method	Time (ms)	Memory (MB)	Paras (M)	MAE
DMSG [5]	32.59	1215	0.33	0.280
DJFR [10]	24.87	851	0.08	0.228
DSRNet [6]	105.38	2039	45.49	0.255
PacNet [7]	50.33	4081	0.18	0.235
CUNet [12]	104.91	1497	0.49	0.275
DKN [11]	204.11	1353	1.16	0.173
PMBAN [9]	476.61	3145	25.06	0.183
AHMF	78.32	1035	2.54	0.157

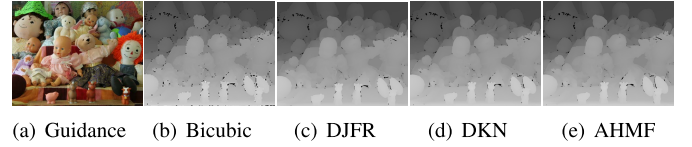


Fig. 16. A failure case of  $8\times$  depth map super-resolution. (a) guidance image, (b) input depth map with missing data, (c): DJFR [10], (d): DKN [11], (e): Ours. Please enlarge the PDF for more details.

encoded by different layers have different attributes: shallower features contain rich details while deeper features contain clear structure information (Fig. 15 (a)); Second, our proposed BHFC can propagate information from one layer to all other layers (Fig.15 (b)-(c)), thus the output features of BHFC can maintain both of fine-grained details and clear structure information (Fig.15 (d)). Thanks to this hierarchical feature collaboration mechanism, our proposed method achieves superior performance than the state-of-the-arts.

### E. Comparison of Network Complexity

In this subsection, we provide experimental analysis about the comparison of our method with other state-of-the-art methods with respect to running time, memory cost and parameter size. In Table XII, we report the average running time (ms), GPU memory consumption (MB) and network parameter size (M) for  $4\times$  depth map super-resolution on the Middlebury [13]. The compared methods are run on the same server with an Intel Core i7 3.6GHz CPU and a NVIDIA GTX 1080ti GPU. To obtain the average running time and memory consumption, we crop the test images into the size of  $480 \times 640$ , and run these methods on these images for 500 times to calculate the average values. From Table XII, it can be found that the proposed method achieves the best MAE performance, while with less running time and memory cost than PMBAN [9], DKN [11] and CUNet [12], which are the most recently published methods. The size of parameters of our method is also much smaller than PMBAN [9]. This result shows that our scheme achieves a better trade-off between reconstruction performance and network complexity.

### F. Limitations

Although the proposed is capable of generating lower errors and visual appealing results for depth map super-resolution,

some cases still pose challenges to our approach, which are listed as follows:

- The proposed method tends to generate unsatisfactory results when dealing with raw depth maps with missing data. We show a failure case in Fig. 16, from which we can see that all methods produce unnatural artifacts, and our method fails to reconstruct the missing boundaries. Since our model is trained on the hole-filled data, and the reconstructed results may be sub-optimal to the raw data. However, in the real case, the depth map obtained by consumer cameras suffers from both low-resolution and missing value artifacts. In the future, we will consider how to extend our idea for joint depth map super-resolution and completion.
- The upsampling factors of the proposed method are fixed instead of arbitrary scales, which may limit the potential applications of our method. This problem will also be considered in our future work.

## VII. CONCLUSION

In this paper, we presented a novel attention-based hierarchical multi-modal fusion (AHMF) network for guided depth map super-resolution. It consists of a multi-modal attention based fusion (MMAF) and a bi-directional hierarchical feature collaboration (BHFC) module. The MMAF can effectively select and combine relevant information from multi-modal features extracted from input depth and guidance images in a learning manner. The BHFC is designed for optimizing the use of hierarchical features fused by MMAF with the proposed bi-directional feature propagation and collaboration mechanism. Extensive experiments on widely used benchmark datasets demonstrate that the proposed method can achieve state-of-the-art performance in terms of reconstruction accuracy, inference speed as well as peak GPU memory consumption.

## REFERENCES

- [1] H. Caesar *et al.*, “NuScenes: A multimodal dataset for autonomous driving,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.
- [2] A. Meuleman, S.-H. Baek, F. Heide, and M. H. Kim, “Single-shot monocular RGB-D imaging using uneven double refraction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2465–2474.
- [3] J. Hou, A. Dai, and M. NieBner, “3D-SIS: 3D semantic instance segmentation of RGB-D scans,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4421–4430.
- [4] D. Lin and H. Huang, “Zig-zag network for semantic segmentation of RGB-D images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2642–2655, Oct. 2020.
- [5] 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.
- [6] 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.
- [7] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, “Pixel-adaptive convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11166–11175.
- [8] R. D. Lutio, S. D’aronco, J. D. Wegner, and K. Schindler, “Guided super-resolution as pixel-to-pixel transformation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8829–8837.
- [9] X. Ye *et al.*, “PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution,” *IEEE Trans. Image Process.*, vol. 29, pp. 7427–7442, 2020.
- [10] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Joint image filtering with deep convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1909–1923, Aug. 2019.
- [11] B. Kim, J. Ponce, and B. Ham, “Deformable kernel networks for joint image filtering,” *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 579–600, Feb. 2021.
- [12] X. Deng and P. L. Dragotti, “Deep convolutional neural network for multi-modal image restoration and fusion,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2021.
- [13] D. Scharstein and C. Pal, “Learning conditional random fields for stereo,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
- [14] W. Shi *et al.*, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
- [15] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral upsampling,” *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.
- [16] 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.
- [17] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.
- [18] X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao, “Depth restoration from RGB-D data via joint adaptive regularization and thresholding on manifolds,” *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1068–1079, Mar. 2019.
- [19] 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.
- [20] T. Li, H. Lin, X. Dong, and X. Zhang, “Depth image super-resolution using correlation-controlled color guidance and multi-scale symmetric network,” *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107513.
- [21] D. B. Lindell, M. O’Toole, and G. Wetzstein, “Single-photon 3D imaging with deep sensor fusion,” *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–12, 2018.
- [22] Z. Sun, D. B. Lindell, O. Solgaard, and G. Wetzstein, “SPADnet: Deep RGB-SPAD sensor fusion assisted by monocular depth estimation,” *Opt. Exp.*, vol. 28, no. 10, pp. 14948–14962, May 2020.
- [23] S. Chan *et al.*, “Long-range depth imaging using a single-photon detector array and non-local data fusion,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.
- [24] A. Ruget, S. McLaughlin, R. K. Henderson, I. Gyongy, A. Halimi, and J. Leach, “Robust super-resolution depth imaging via a multi-feature fusion deep network,” *Opt. Exp.*, vol. 29, no. 8, pp. 11917–11937, 2021.
- [25] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 7354–7363.
- [26] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 7132–7141.
- [28] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 510–519.
- [29] Y. Mei *et al.*, “Pyramid attention networks for image restoration,” 2020, *arXiv:2004.13824*.
- [30] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-Net: Efficient channel attention for deep convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
- [31] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, “Free-form image inpainting with gated convolution,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4470–4479.
- [32] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, “Free-form video inpainting with 3D gated convolution and temporal PatchGAN,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9065–9074.
- [33] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

- [34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
- [35] S. Gu, Y. Li, L. Van Gool, and R. Timofte, "Self-guided network for fast image denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2511–2520.
- [36] G. Lin, F. Liu, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1228–1242, May 2020.
- [37] J. Kwak and D. Son, "Fractal residual network and solutions for real super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2114–2121.
- [38] 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.
- [39] J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, "Cross-based local multipoint filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 430–437.
- [40] G. Riegler, M. R  ther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Proc. 14th Eur. Conf.*, 2016, pp. 268–284.
- [41] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, "Learning dynamic guidance for depth image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3769–3778.
- [42] X. Ye, X. Duan, and H. Li, "Depth super-resolution with deep edge-inference network and edge-guided depth filling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1398–1402.
- [43] 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.
- [44] J. Wang, W. Xu, J.-F. Cai, Q. Zhu, Y. Shi, and B. Yin, "Multi-direction dictionary learning based depth map super-resolution with autoregressive modeling," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1470–1484, Jun. 2020.
- [45] B. Sun, X. Ye, B. Li, H. Li, Z. Wang, and R. Xu, "Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7792–7801.
- [46] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.
- [47] Y. Zuo, Y. Fang, P. An, X. Shang, and J. Yang, "Frequency-dependent depth map enhancement via iterative depth-guided affine transformation and intensity-guided refinement," *IEEE Trans. Multimedia*, vol. 23, pp. 772–783, 2021.
- [48] Z. Wang, X. Ye, B. Sun, J. Yang, R. Xu, and H. Li, "Depth upsampling based on deep edge-aware learning," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107274.
- [49] 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.
- [50] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
- [51] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.
- [52] S. Lu, X. Ren, and F. Liu, "Depth enhancement via low-rank matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3390–3397.
- [53] S. D. Cochran and G. Medioni, "3-D surface description from binocular stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 10, pp. 981–994, Oct. 1992.
- [54] P. Fua, "A parallel stereo algorithm that produces dense depth maps and preserves image features," *Mach. Vis. Appl.*, vol. 6, no. 1, pp. 35–49, 1993.
- [55] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 401–406, Apr. 1998.
- [56] D. Scharstein, *View Synthesis Using Stereo Vision* (Lecture Notes in Computer Science), vol. 1583. Springer, 1999.
- [57] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGB-D images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.
- [58] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, Aug. 2004.
- [59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
- [60] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop Autodiff Submission*, 2017, pp. 1–4.
- [61] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.
- [62] 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.
- [63] J. Park, H. Kim, Y. W. Tai, M. S. Brown, and I. Kweon, "High quality depth map upsampling for 3D-TOF cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1623–1630.
- [64] B. Ham, M. Cho, and J. Ponce, "Robust image filtering using joint static and dynamic guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4823–4831.
- [65] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1838–1847.
- [66] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.
- [67] Y. Zhang *et al.*, "Color-guided depth image recovery with adaptive data fidelity and transferred graph Laplacian regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 320–333, Feb. 2020.