High-resolution Depth Maps Imaging via Attention-based Hierarchical Multi-modal Fusion

Zhiwei Zhong, Xianming Liu, Member, IEEE, Junjun Jiang, Member, IEEE, Debin Zhao, Member, IEEE, Zhiwen Chen, Member, IEEE, Xiangyang Ji, 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 (AHMFM) 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

Depth information plays a critical role in a myriad of applications such as autonomous driving [1], virtual 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.

This work was supported by National Natural Science Foundation of China under Grants 61922027 and 61932022. Z. Zhong, X. Liu, J. Jiang and D. 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: {zhzhong, cxsx, jiangjunjun}@hit.edu.cn.
Z. Chen is with Taobao (China) Software Co., Ltd., Beijing, China. E-mail: zhiwen.czw @alibaba-inc.com.
X. Ji is with the Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: xyji@tsinghua.edu.cn.
Correspondence to: Xianming Liu (cxsx@hit.edu.cn).

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 (DSGAN) 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 et al. [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 (MAAF) 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 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 MAAF 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 [14], [15], 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 constrain the high-dimensional solution space for this ill-posed problem. For example, Ferstl et al. [16] 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. [17] 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. [18] proposed a coarse-to-fine network by
combining both global and local residual learning strategies. Li et al. [19] 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 [20]–[23]. 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 [21] 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.

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. [32] introduced a skip-connection which was also called

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.
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. [33] concatenated the low-level and high-level features to fuse the multi-level features. Gu et al. [34] 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. [35] 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 [32] and [33] 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 [35] and [34] 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 [36] 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 \(2 \cdot \log_2\alpha\) 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:

\[
F^{\text{init}}_y = \text{Down}(\sigma(W^0_y \ast \text{Down}(\hat{Y}) + b^0_y)),
\]

where \(\text{Down}(\cdot)\) represents the inverse pixel-shuffle operator with downscale factor 2; \(W^0_y\) is a \(1 \times 1\) convolutional kernel; \(b^0_y\) is the bias term; \(\ast\) means convolutional operation; \(\sigma\) represents a Parametric Rectified Linear Unit (PReLU) [37] 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:

\[
F^1_d = \sigma\left(W^1_d \ast D_{LR} + b^1_d\right), \quad (2)
\]

\[
F^1_y = \sigma\left(W^1_y \ast F^{\text{init}}_y + b^1_y\right), \quad (3)
\]

\[
F^i_d = \sigma\left(W^i_d \ast F^{i-1}_d + b^i_d\right), \quad 1 < i \leq m, \quad (4)
\]

\[
F^i_y = \sigma\left(W^i_y \ast F^{i-1}_y + b^i_y\right), \quad 1 < i \leq m, \quad (5)
\]

where \(W^i_d\) and \(W^i_y\) are convolutional kernels of \(i\)-th layer that are used for depth and guidance feature extraction respectively; \(b^i_d\) and \(b^i_y\) 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):

\[
F^i = \text{MMAF}_i(F^i_y, F^i_d), \quad 1 \leq i \leq m, \quad (6)
\]

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:

\[
F_i^g = \text{BHFC}_i([F^i_f, H^i_{lh}, H^m_{hl}]), \quad 1 \leq i \leq m, \quad (7)
\]

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 \(\{F^m_g\}_{i=1}^m\), we arrive at the HR depth reconstruction module, which produces the upsampled depth image by pixel-shuffle [38]:

\[
F_{up} = \text{Up}(\sigma(W_{cr} \ast [F^1_g, F^2_g, \cdots, F^m_g] + b_{cr})) \quad (8)
\]

\[
D_{SR} = \sigma(W_{out} \ast \text{Up}(F_{up}) + b_{out}) + D_{LR}^\dagger, \quad (9)
\]

where \(W_{cr}\) 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; \(\lceil \cdot \rceil\) is used to concatenate the refined features; \(W_{out}\) and \(b_{out}\) are convolution kernels and bias, respectively. \(D_{LR}^\dagger\) is the bicubic upsampled version of the input LR depth image; \(D_{SR} \in \mathbb{R}^{\alpha H \times \alpha W \times 1}\) is the final reconstructed HR depth image.

The network is trained using the following loss function:

\[
L(\theta) = \frac{1}{N} \sum_{i=1}^{N} ||F(D_{LR}, Y; \theta) - D_{GT}||_1, \quad (10)
\]

where \(F(\cdot)\) represents the overall network architecture and \(\theta\) are network parameters. \(D_{GT}\) is the ground-truth HR depth map, \(N\) is the number of training samples.
output of the gating operation within the range of 0 and 1; and
tics due to the difference in imaging principle, similar pixels in
ance images exhibit notably different appearance characteris-
\( \hat{y} \) represents two different convolutional operations;
2
\( W \)
inputs, the process of FEB can be formulated as follows:
and Var mean the average and variance pooling. The MMAF takes the depth
and guidance features as inputs and outputs the fused features.

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
generate 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.

**Feature Enhancement Block:** The guidance image con-
tains 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 [30], we leverage gated units as the FEB. Given the
extracted multi-modal features \( F_d \) and \( F_y \) (1 ≤ \( i \) ≤ \( m \)) as
inputs, the process of FEB can be formulated as follows:

\[
\hat{F}_d^i = \sigma(W_{d,1}^i \ast F_d^i + b_{d,1}^i) \circ \phi(W_{d,2}^i \ast F_d^i + b_{d,2}^i), \tag{11}
\]

\[
\hat{F}_y^i = \sigma(W_{y,1}^i \ast F_y^i + b_{y,1}^i) \circ \phi(W_{y,2}^i \ast F_y^i + b_{y,2}^i), \tag{12}
\]

where \( W_{d,1}^i \), \( W_{d,2}^i \), \( W_{y,1}^i \), \( W_{y,2}^i \), \( b_{d,1}^i \), \( b_{d,2}^i \), \( b_{y,1}^i \), \( b_{y,2}^i \) are the corresponding bias terms; \( \circ \) 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; \( \hat{F}_d^i \) and \( \hat{F}_y^i \) are the enhanced
features of depth and guidance respectively.

**Feature Recalibration Block:** Considering depth and guid-
ance images exhibit notably different appearance characteris-
tics 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:

\[
\tilde{F}_d^i = \sigma(W_d^i \ast (\hat{F}_d^i \circ \hat{F}_y^i)) + b_d^i, \tag{13}
\]

\[
\tilde{F}_y^i = \text{AvgPool}(\hat{F}_d^i) + \text{VarPool}(\hat{F}_y^i), \tag{14}
\]

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

\[
E_d^i = \phi(W_d^i \ast \tilde{F}_d^i + b_d^i), \tag{15}
\]

\[
E_y^i = \phi(W_y^i \ast \tilde{F}_y^i + b_y^i), \tag{16}
\]

\[
F_f^i = 0.5 \times E_d^i \odot \hat{F}_d^i + 0.5 \times E_y^i \odot \hat{F}_y^i, \tag{17}
\]

where \( W_d^i \), \( W_y^i \) and \( b_d^i \), \( b_y^i \) are 1×1 convolutional kernels
and bias terms for depth and guidance respectively; \( \phi \) is the
sigmoid function; \( E_d^i \) and \( E_y^i \) are the excitation signals for
depth feature and guidance feature respectively; \( F_f^i \) denotes
the fused multi-modal feature.

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
TABLE I
MAE PERFORMANCE COMPARISON FOR SCALE FACTORS 4x, 8x AND 16x WITH BICUCIC 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     | 4x  | 8x  | 16x | 4x  | 8x  | 16x | 4x  | 8x  | 16x | 4x  | 8x  | 16x |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Art        | 1.10| 2.10| 4.02| 0.35| 0.67| 1.29| 0.36| 0.68| 1.25| 0.60| 1.12| 2.13|
| Books      | 0.74| 1.44| 2.87| 0.28| 0.51| 1.02| 0.34| 0.66| 1.01| 0.50| 0.80| 1.67|
| Dools      | 0.65| 0.81| 1.42| 0.43| 0.51| 0.79| 0.41| 0.52| 0.56| 0.37| 0.89| 0.94|
| Laundry    | 0.46| 0.76| 1.53| 0.15| 0.41| 0.76| 0.25| 0.51| 0.87| 0.30| 0.46| 1.12|
| Mobeius    | 0.48| 1.20| 2.44| 0.30| 0.58| 1.02| 0.34| 0.63| 0.93| 0.35| 0.86| 1.56|
| Reindeer   | 0.40| 0.64| 1.34| 0.22| 0.37| 0.78| 0.22| 0.38| 0.73| 0.23| 0.36| 0.81|
| Average    | 0.66| 1.30| 2.60| 0.43| 0.84| 1.87| 0.50| 1.02| 1.87| 0.58| 1.09| 2.18|

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

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], 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.

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
TABLE II

| Method       | Art 4x | Books 4x | Dools 4x | Laundry 4x | Mobeius 4x | Reindeer 4x |
|--------------|--------|---------|----------|------------|------------|-------------|
|              | 4x     | 8x      | 16x      | 4x         | 8x         | 16x         |
| Bicubic      | 3.87   | 5.46    | 8.17     | 1.27       | 2.34       | 3.34        |
| DMSG [5]     | 1.47   | 2.46    | 4.57     | 0.67       | 1.03       | 1.60        |
| DGDIE [41]   | 2.00   | 3.84    | 6.16     | 0.91       | 1.68       | 2.67        |
| DEIN [42]    | 3.26   | 4.20    | 6.40     | 1.38       | 2.12       | 3.38        |
| DSRNet [6]   | 1.20   | 2.22    | 3.90     | 0.60       | 0.89       | 1.51        |
| DJFR [10]    | 1.62   | 3.08    | 5.81     | 0.54       | 1.11       | 2.24        |
| PacNet [7]   | 1.66   | 2.92    | 6.02     | 0.58       | 1.03       | 2.43        |
| CUNet [12]   | 1.57   | 2.60    | 4.74     | 0.56       | 0.97       | 1.96        |
| CGN [47]     | 1.50   | 2.69    | 4.44     | 0.60       | 0.97       | 1.73        |
| DSR_N [48]   | 1.77   | 3.32    | 6.08     | 0.73       | 1.39       | 2.38        |
| MFR-SR [18]  | 2.57   | 3.20    | 4.87     | 1.33       | 1.46       | 2.51        |
| PMBAN [9]    | 1.54   | 2.71    | 4.35     | 0.63       | 1.05       | 1.78        |
| DKN [11]     | 1.19   | 2.47    | 4.37     | 0.43       | 1.10       | 1.51        |
| AHMF         | 1.09   | 2.14    | 4.20     | 0.38       | 0.72       | 1.49        |

RMSE performance comparison for scale factors 4x, 8x and 16x 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).

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 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 3×3, except for those in the upsampling and downsampling module that are set as 1×1, 3×3 and 5×5 with a stride size of 0, 1, 2 in 4x, 8x, 16x super-resolution, respectively. We use PReLU [37] as the default activation function. During training, we randomly select 32 HR depth maps with the size of 256×256 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 β1 = 0.9, β2 = 0.999, and ε = 1e−8; the initial learning rate is 2×10−4 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 4x, 8x and 16x 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 16x 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 and collaboration modules, which fully exploit the relevant information from LR depth and HR guidance.
TABLE III
MAE COMPARISON FOR SCALE FACTORS 4×, 8× AND 16× 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 4×8×16× | Books 4×8×16× | Dools 4×8×16× | Laundry 4×8×16× | Mobius 4×8×16× | Reindeer 4×8×16× |
|----------|--------------|---------------|---------------|----------------|----------------|-----------------|
| Bicubic  | 3.23         | 4.00          | 5.58          | 2.62           | 2.83           | 3.27            |
| GF [61]  | 1.91         | 2.87          | 4.90          | 1.13           | 1.84           | 2.80            |
| DMSG [5] | 0.84         | 1.57          | 2.98          | 0.62           | 1.18           | 1.48            |
| DGDIE [41]| 0.99         | 1.84          | 3.34          | 0.81           | 1.29           | 2.04            |
| DSRNet [6]| 2.19         | 2.57          | 3.69          | 1.81           | 1.82           | 2.16            |
| GSPRT [8]| 0.68         | 1.33          | 2.47          | 0.52           | 0.87           | 1.37            |
| DJFR [10]| 0.82         | 1.44          | 2.77          | 0.62           | 1.03           | 1.67            |
| PacNet [7]| 0.72         | 1.40          | 2.45          | 0.49           | 0.83           | 1.47            |
| CUNet [12]| 0.81         | 1.25          | 2.35          | 0.54           | 0.85           | 1.47            |
| PMBAN [9]| 0.59         | 0.98          | 1.89          | 0.44           | 0.71           | 1.23            |
| DKN [11]| 0.68         | 0.95          | 1.86          | 0.51           | 0.80           | 1.21            |
| AHMF (Ours)| 0.62         | 0.93          | 1.85          | 0.41           | 0.71           | 1.19            |

TABLE IV
RMSE PERFORMANCE COMPARISON FOR SCALE FACTORS 4×, 8× AND 16× 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 4×8×16× | Books 4×8×16× | Dools 4×8×16× | Laundry 4×8×16× | Mobius 4×8×16× | Reindeer 4×8×16× |
|----------|--------------|---------------|---------------|----------------|----------------|-----------------|
| Bicubic  | 5.74         | 6.89          | 9.12          | 4.57           | 4.86           | 5.28            |
| DMSG [5] | 3.06         | 4.48          | 6.44          | 1.90           | 3.13           | 4.63            |
| DGDIE [41]| 4.29         | 5.94          | 9.78          | 1.99           | 2.65           | 4.16            |
| DSRNet [6]| 4.29         | 5.94          | 9.78          | 1.99           | 2.65           | 4.16            |
| DJFR [10]| 2.96         | 4.81          | 7.30          | 1.41           | 2.58           | 3.47            |
| PacNet [7]| 2.95         | 4.41          | 7.80          | 1.37           | 2.05           | 3.80            |
| CUNet [12]| 2.85         | 3.84          | 6.12          | 1.42           | 1.95           | 2.95            |
| PMBAN [9]| 2.58         | 3.92          | 6.67          | 1.25           | 1.82           | 2.94            |
| DKN [11]| 2.77         | 3.79          | 6.01          | 1.36           | 1.91           | 2.93            |
| AHMF (Ours)| 2.58         | 3.85          | 6.00          | 1.21           | 1.80           | 2.75            |

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×, 8× and 16× 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× 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.

Table V

| Method       | Image_01 | Image_02 | Image_03 | Image_04 | Image_05 | Image_06 |
|--------------|----------|----------|----------|----------|----------|----------|
| Bicubic      | 0.57     | 1.32     | 2.46     | 0.82     | 2.06     | 4.34     |
| DSRNet [6]   | 0.57     | 1.01     | 1.89     | 0.82     | 1.48     | 3.08     |
| DEDIE [41]   | 0.53     | 1.19     | 2.11     | 0.83     | 1.81     | 3.57     |
| CUNet [12]   | 0.57     | 1.01     | 1.89     | 0.82     | 1.48     | 3.08     |
| PacNet [7]   | 0.61     | 1.34     | 2.77     | 0.84     | 1.65     | 3.52     |
| DJFR [10]    | 0.53     | 1.11     | 2.35     | 0.74     | 1.47     | 3.08     |
| PMBAN [9]    | 0.54     | **0.79** | 1.83     | 0.74     | **1.17** | 2.83     |
| DKN [11]     | 0.48     | 0.80     | 1.80     | 0.69     | **1.16** | 2.86     |
| AHMF (Ours)  | **0.42** | 0.82     | 1.76     | **0.64** | 1.19     | **2.78** |

Table VI

| Method       | Image_01 | Image_02 | Image_03 | Image_04 | Image_05 | Image_06 |
|--------------|----------|----------|----------|----------|----------|----------|
| Bicubic      | 2.18     | 3.93     | 6.10     | 3.55     | 6.72     | 11.19    |
| DSGDIE [41]  | 1.90     | 3.33     | 5.50     | 3.33     | 6.09     | 10.41    |
| DSRNet [6]   | 1.47     | **2.34** | 5.25     | **2.01** | **3.45** | 8.18     |
| CUNet [12]   | 1.55     | 2.47     | 5.23     | 3.08     | 4.04     | 7.98     |
| PacNet [7]   | 1.81     | 2.38     | 5.66     | 3.54     | 4.29     | 8.69     |
| DJFR [10]    | 2.10     | 2.98     | 7.46     | 3.53     | 4.69     | 10.60    |
| PMBAN [9]    | 1.35     | 2.51     | 5.35     | 2.37     | 3.98     | 7.60     |
| DKN [11]     | 1.37     | 2.67     | 5.46     | 2.65     | 4.38     | 7.80     |
| AHMF (Ours)  | **1.23** | **2.22** | **4.92** | **2.21** | **3.97** | **7.47** |

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 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], 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 DJFR [10], DKN [11] and PacNet [7]. For fairness, we retrain 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...
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 [63] | 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 [64] | 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 [65]  | 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 [66] | 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 [67] | 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  | 2.21 | 3.75 | 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] | 1.93 | 3.17 | 5.49  | 2.35 | 4.16 | 6.33  | 2.46 | 4.76  | 8.50  | 2.25 | 4.03  | 6.77  |
| AHMF (Ours) | 1.79 | 2.81 | 5.02  | 2.00 | 3.83 | 6.21  | 2.25 | 4.50  | 8.10  | 2.01 | 3.71  | 6.41  |

Fig. 8. Visual comparison of 8× 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.

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.

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 the local filter cannot capture global 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 [68], 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× 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],
As shown in Table X, the experimental analysis is conducted and hierarchical feature collaboration to the final performance. We further investigate the role of multi-modal attention fusion and perceptual loss. According to this analysis, 1) the extracted guidance features are more robust to outliers, thus the depth boundaries are preserved better than L2 and perceptual loss. According to this analysis, we choose L1 as our loss function.

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) Model1, 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 totally different objects. Directly fuse them by addition would destroy the useful information. Fig. 11 further shows the visual comparisons of different variants for 16x depth map super-resolution with different loss functions.

Table X

| Model                  | Multi-Modal Fusion | Hierarchical Feature Collaboration | MAE       |
|------------------------|--------------------|------------------------------------|-----------|
| Model1                 | Addition           | Forward+Backward                   | 0.1621    |
| Model2                 | Concatenation      | Forward+Backward                   | 0.1586    |
| Model3                 | MMAF               | w/o Feature Collaboration          | 0.1640    |
| Model4                 | MMAF               | w/o Backward                       | 0.1592    |
| Model5                 | MMAF               | w/o Forward                        | 0.1584    |
| Model6 (our full model)| MMAF               | Forward+Backward (BHFC)            | 0.1574    |

NYU v2 [57] dataset and Lu [52] dataset. The comparison results are provided in Table XI, from which it can be found that, the L1 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 L1 is more robust to outliers, thus the depth boundaries are preserved better than L2 and perceptual loss. According to this analysis, we choose L1 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 4x, 8x and 16x super-resolution, the sizes of parameters of all variants are around 2.54M, 3.36M and 5.75M, respectively.

Table XI

| Loss             | Middlebury [13] | NYU v2 [57] | Lu [52] |
|------------------|-----------------|-------------|---------|
| L1               | 0.1574 / 0.7021 | 0.4533 / 1.4013 | 0.4467 / 1.3283 |
| L2               | 0.2160 / 0.6972 | 0.5514 / 1.4844 | 0.4492 / 1.4843 |
| L1+Perceptual Loss | 0.2237 / 0.7594 | 0.5830 / 1.6131 | 0.4837 / 1.5190 |

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) (F1 in Eq. 6) and Fig. 14 (d) (F2 in Eq. 6) show the features 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, 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 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 × 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 × 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

---

**Fig. 14.** Visualization of feature maps (Art for 4× 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.

**TABLE XII**

| 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  |
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, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 11 621–11 631.
[2] A. Meuleman, S.-H. Baek, F. Heide, and M. H. Kim, “Single-shot monocular rgb-d imaging using uneven double refraction,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 2465–2474.
[3] J. Hou, A. Dai, and M. Niessner, “3d+six: 3d semantic instance segmentation of rgb-d scans,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2019, pp. 4421–4430.
[4] D. Lin and H. Huang, “Zig-zag network for semantic segmentation of rgb-d images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, pp. 2642–2655, 2020.
[5] T.-W. Hui, C. C. Loy, and X. Tang, “Depth map super-resolution by deep multi-scale guidance,” in Proceedings of the European Conference on Computer Vision, 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 Transactions on Image Processing, vol. 28, no. 5, pp. 2545–2557, 2019.
[7] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, “Pixel-adaptive convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11 166–11 175.
[8] R. d. Lutio, S. D’arono, J. D. Wegner, and K. Schindler, “Guided super-resolution as pixel-to-pixel transformation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8829–8837.
[9] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li, “Pmbunet: Progressive multi-branch aggregation network for scene depth super-resolution,” IEEE Transactions on Image Processing, 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 Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 8, pp. 1909–1923, 2019.
[11] B. Kim, J. Ponce, and B. Han, “Deformable kernel networks for joint image filtering,” International Journal of Computer Vision, vol. 129, no. 2, pp. 579–600, 2021.
[12] X. Deng and P. L. Dragotti, “Deep convolutional neural network for multi-modal image restoration and fusion,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2020.
[13] D. Scharstein and C. Pal, “Learning conditional random fields for stereo,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007, pp. 1–8.
[14] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral upsampling,” ACM Transactions on Graphics, vol. 26, no. 3, pp. 96.1–96.4, 2007.
[15] M. Liu, O. Tuzel, and Y. Taguchi, “Joint geodesic upsampling of depth images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 169–176.
[16] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruheter, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 993–1000.
[17] 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 Transactions on Image Processing, vol. 28, no. 3, pp. 1068–1079, 2019.
[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 Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 297–306, 2019.
[19] T. Li, H. Lin, X. Dong, and X. Zhang, “Depth image super-resolution using correlation-controlled color guidance and multi-scale symmetric network,” Pattern Recognition, vol. 107, p. 107513, 2020.
[20] 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. 113–1, 2018.
[21] Z. Sun, D. B. Lindell, O. Solgaard, and G. Wetzstein, “Spadnet: deep rgb-spad sensor fusion assisted by monocular depth estimation,” Optics Express, vol. 28, no. 10, pp. 14 948–14 962, 2020.
[22] S. Chan, A. Halimi, F. Zhu, I. Gyongy, R. K. Henderson, R. Bowman, S. McLaughlin, G. S. Buller, and J. Leach, “Long-range depth imaging using a single-photon detector array and non-local data fusion,” Scientific reports, vol. 9, no. 1, pp. 1–10, 2019.
[23] 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,” Optics Express, vol. 29, no. 8, pp. 11 917–11 937, 2021.
[24] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 7354–7363.
[25] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep

J. Wang, W. Xu, J. Cai, Q. Zhu, Y. Shi, and B. Yin, “Multi-direction

S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, “Learning

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 Transactions on Multimedia,
vol. 23, pp. 772–783, 2020.

Z. Wang, X. Ye, B. Sun, J. Yang, R. Xu, and H. Li, “Depth upsampling
based on deep edge-aware learning,” Pattern Recognition, vol. 103,
p. 107274, 2020.

D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense
two-frame stereo correspondence algorithms,” International Journal of
Computer Vision, vol. 47, no. 1, pp. 7–42, 2002.

H. Hirschmuller and D. Scharstein, “Evaluation of cost functions for
stereo matching,” in Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition, 2007, pp. 1–8.

D. Scharstein, H. Hirschmuller, Y. Kitajima, G. Krathwohl, N. Nešić,
X. Wang, and P. Westling, “High-resolution stereo datasets with
subpixel-accurate ground truth,” in Proceedings of the German
Conference on Pattern Recognition, 2014, pp. 31–42.

S. Lu, X. Ren, and F. Liu, “Depth enhancement via low-rank matrix
completion,” in Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition, 2014, pp. 3390–3397.

S. D. Cochran and G. Medioni, “3d surface description from binocular
stereo,” IEEE Transactions on Pattern Analysis & Machine Intelligence,
vol. 14, no. 10, pp. 981–994, 1992.

P. Fua, “A parallel stereo algorithm that produces dense depth maps
and preserves image features,” Machine vision and applications, vol. 6,
no. 1, pp. 35–49, 1993.

S. Birchfield and C. Tomasi, “A pixel dissimilarity measure that is
insensitive to image sampling,” IEEE Transactions on Pattern Analysis
and Machine Intelligence, vol. 20, no. 4, pp. 401–406, 1998.

D. Scharstein, “View synthesis using stereo vision,” Lecture Notes in
Computer Science, vol. 1583, 1999.

N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation
and support inference from rgb-d images,” in Proceedings of the
European Conference on Computer Vision, 2012, pp. 746–760.

A. Levin, D. Lischinski, and Y. Weiss, “Colorization using optimization,”
in Proceedings of the Special Interest Group for Computer GRAPHICS,
2004, pp. 689–694.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,”
Proceedings of the International Conference on Learning Representations,
2015.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin,
A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in
pytorch,” 2017.

K. He, J. Sun, and X. Tang, “Guided image filtering,” IEEE Transactions on
Pattern Analysis & Machine Intelligence, vol. 35, no. 6, pp. 1397–
1409, 2013.

J. Diebel and S. Thrun, “An application of markov random fields to
range sensing,” in Proceedings of the Advances in Neural Information
Processing Systems, 2005, pp. 291–298.

D. Feistl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, “Image
guided depth upsampling using anisotropic total generalized variation,”
in Proceedings of the IEEE International Conference on Computer
Vision, 2013, pp. 993–1000.

J. Park, H. Kim, Y. W. Tai, M. S. Brown, and I. Kweon, “High quality
depth map upsampling for 3d-tof cameras,” in Proceedings of IEEE
International Conference on Computer Vision, 2011, pp. 1623–1630.

B. Ham, M. Cho, and J. Ponce, “Robust image filtering using joint
static and dynamic guidance,” in Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition, 2015, pp. 4823–4831.

H. Wu, S. Zheng, J. Zhang, and K. Huang, “Fast end-to-end trainable
guided filter,” in Proceedings of the IEEE Conference on Computer Vision
and Pattern Recognition, 2018, pp. 1838–1847.

Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep joint image
filtering,” in Proceedings of the European Conference on Computer
Vision, 2016, pp. 154–169.

Y. Zhang, Y. Feng, X. Liu, D. Zhai, X. Ji, H. Wang, and Q. Dai, “Color-
guided depth image recovery with adaptive data fidelity and transferred
graph laplacian regularization,” IEEE Transactions on Circuits and
Systems for Video Technology, vol. 30, no. 2, pp. 320–333, 2019.

REFERENCES

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Kr-
ishnan, Y. Pan, G. Baldwin, and O. Beijbom, “nuscenes: A multimodal
dataset for autonomous driving,” in Proceedings of the IEEE Conference on
Computer Vision and Pattern Recognition, 2020, pp. 11 621–11 631.
X. Deng and P. L. Dragotti, “Deep convolutional neural network
Y. Li, J. B. Huang, N. Ahuja, and M. H. Yang, “Joint image filtering with
depth multi-scale guidance,” in Proceedings of the European Conference
S. Chan, A. Halimi, F. Zhu, I. Gyongy, R. K. Henderson, R. Bowman,
 depth images,” IEEE Transactions on Pattern Analysis and Machine Intelligence,
T.-W. Hui, C. C. Loy, and X. Tang, “Depth map super-resolution by
deeplabv3+,” in Proceedings of the IEEE Conference on Computer Vision
C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, “Hierarchical
X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao, “Depth restoration from
M. Liu, O. Tuzel, and Y. Taguchi, “Joint geodesic upsampling for
D. Lin and H. Huang, “Zig-zag network for semantic segmentation ofgbd images,” IEEE Transactions on Pattern Analysis and Machine Intelligence,
A. Ruget, S. McLaughlin, R. K. Henderson, I. Gyongy, A. Halimi, and
H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention gen-
dromatography assisted by monocular depth estimation,” Optical Express,
Zig-zag network for semantic segmentation ofgbd images,” IEEE Transactions on Pattern Analysis and Machine Intelligence,
T.-W. Hui, C. C. Loy, and X. Tang, “Depth map super-resolution by
deeplabv3+,” in Proceedings of the IEEE Conference on Computer Vision
H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz,
Pixel-adaptive convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1166–1175.
R. d. Lutio, S. D’aronco, J. D. Wegner, and K. Schindler, “Guided super-
res as pixel-to-pixel transformation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8829–8837.
X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li, “Punet: Progressive
manifolds,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, “Joint bilateral up-sampling,” Acm Transactions on Graphics, 26, no. 3, pp. 96.1–96.4, 2007.
M. Liu, O. Tuzel, and Y. Taguchi, “Joint geodesic upsampling of depth images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 169–176.
D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 993–1000.
X. Liu, D. Zhai, R. Chen, X. Ji, D. Zhao, and W. Gao, “Depth restoration from
rgbd data via joint adaptive regularization and thresholding on manifolds,” IEEE Transactions on Image Processing, 28, no. 3, pp. 1068–1079, 2019.
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 Transactions on Circuits and Systems for Video Technology, 30, no. 2, pp. 297–306, 2019.
T. Li, H. Lin, X. Dong, and X. Zhang, “Depth image super-resolution using correlation-controlled color guidance and multi-scale symmetric network,” Pattern Recognition, vol. 107, p. 107513, 2020.
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. 113–1, 2018.
Z. Sun, D. B. Lindell, O. Solgaard, and G. Wetzstein, “Spadnet: deep rgbd-sparse sensor fusion assisted by nonlinear depth estimation.” Optics Express, vol. 28, no. 10, pp. 14948–14962, 2020.
S. Chan, A. Halimi, F. Zhu, I. Gyongy, R. K. Henderson, R. Bowman,
S. McLaughlin, S. Buller, and J. Leach, “Long-range depth imaging using a single-photon detector array and non-local data fusion,” Scientific reports, vol. 9, no. 1, pp. 1–10, 2019.
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,” Optics Express, vol. 29, no. 8, pp. 11917–11937, 2020.
H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 7354–7363.
X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 3, pp. 511–522, 2019.
X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 2019, pp. 510–519.
Y. Mei, Y. Fan, Y. Zhang, J. Yu, Y. Zhou, D. Liu, Y. Fu, T. S. Huang, and
H. Shi, “Pyramid attention networks for image restoration,” arXiv preprint arXiv:2004.13824, 2020.
Q. Wang, B. Wu, P. Zhu, W. Z. Peihua Li, and Q. Hu, “Eca-net: Efficient
channel attention for deep convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 11531–11539.
J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, “Free-form image inpainting with gated convolution,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4470–4479.
Y. Chang, Z. Y. Liu, K. Lee, and W. Hsu, “Free-form video inpainting with 3d gated convolution and temporal patchgan,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9065–9074.
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely
connected convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261–2269.
S. Gu, Y. Li, L. V. Gool, and R. Timofte, “Self-guided network for fast
image denoising,” in Proceedings of the IEEE International Conference on Computer Vision, October 2019, pp. 2511–2520.
G. Lin, F. Liu, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for dense prediction,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1228–1242, 2019.
J. Kwak and D. Son, “Fractional residual network and solutions for real
super-resolution,” in Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 2114–2121.
K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in Proceedings of the IEEE Conference on Computer Vision, 2015, pp. 1026–1034.
W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop,
D. Rueckert, and Z. Wang, “Real-time single image and video super-
resolution using an efficient sub-pixel convolutional neural network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.
J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, “Cross-based local multi-
point filtering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2012, pp. 430–437.
G. Riegler, M. Rüther, and H. Bischof, “Atgv-net: Accurate depth super-
resolution,” in Proceedings of the European conference on computer vision. Springer, 2016, pp. 268–284.
S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, “Learning
dynamic guidance for depth image enhancement,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3769–3778.
X. Ye, X. Duan, and H. Li, “Depth super-resolution with deep edge
inference network and edge-guided depth filling,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 1398–1402.
Y. Wen, B. Sheng, P. Li, W. Lin, and D. F. Deng, “Deep color guided
coarse-to-fine convolutional network cascade for depth image super-
resolution,” IEEE Transactions on Image Processing, 28, no. 2, pp. 994–1006, 2019.
J. Wang, W. Xu, J. Cai, Q. Zhu, Y. Shi, and B. Yin, “Multi-direction
learning based depth map super-resolution with autoregressive modeling,” IEEE Transactions on Multimedia, 22, no. 6, pp. 1470–1484, 2020.
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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7792–7801.
B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep
residual networks for single image super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops, 2017, pp. 136–144.
17

[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 Transactions on Multimedia, vol. 23, pp. 772–783, 2020.

[48] Z. Wang, X. Ye, B. Sun, J. Yang, R. Xu, and H. Li, “Depth upsampling based on deep edge-aware learning,” Pattern Recognition, vol. 103, p. 107274, 2020.

[49] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” International Journal of Computer Vision, vol. 47, no. 1, pp. 7–42, 2002.

[50] H. Hirschmuller and D. Scharstein, “Evaluation of cost functions for stereo matching,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[51] 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 Proceedings of the German Conference on Pattern Recognition, 2014, pp. 31–42.

[52] S. Lu, X. Ren, and F. Liu, “Depth enhancement via low-rank matrix completion,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3390–3397.

[53] S. D. Cochran and G. Medioni, “3-d surface description from binocular stereo,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 14, no. 10, pp. 981–994, 1992.

[54] P. Fua, “A parallel stereo algorithm that produces dense depth maps and preserves image features,” Machine vision and applications, 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 Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 4, pp. 401–406, 1998.

[56] D. Scharstein, “View synthesis using stereo vision,” Lecture Notes in Computer Science, vol. 1583, 1999.

[57] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgb-d images,” in Proceedings of the European Conference on Computer Vision, 2012, pp. 746–760.

[58] A. Levin, D. Lischinski, and Y. Weiss, “Colorization using optimization,” in Proceedings of the Special Interest Group for Computer GRAPHICS, 2004, pp. 689–694.

[59] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Proceedings of the International Conference on Learning Representations, 2015.

[60] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

[61] K. He, J. Sun, and X. Tang, “Guided image filtering,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 35, no. 6, pp. 1397–1409, 2013.

[62] J. Diebel and S. Thrun, “An application of markov random fields to range sensing,” in Proceedings of the Advances in Neural Information Processing Systems, 2005, pp. 291–298.

[63] D. Ferstl, C. Reinbacher, R. Ranftl, M. Rüther, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 993–1000.

[64] J. Park, H. Kim, Y. W. Tai, M. S. Brown, and I. Kweon, “High quality depth map upsampling for 3d-tof cameras,” in Proceedings of IEEE International Conference on Computer Vision, 2011, pp. 1623–1630.

[65] B. Ham, M. Cho, and J. Ponce, “Robust image filtering using joint static and dynamic guidance,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4823–4831.

[66] H. Wu, S. Zheng, J. Zhang, and K. Huang, “Fast end-to-end trainable guided filter,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1838–1847.

[67] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Deep joint image filtering,” in Proceedings of the European Conference on Computer Vision, 2016, pp. 154–169.

[68] Y. Zhang, Y. Feng, X. Liu, D. Zhai, X. Ji, H. Wang, and Q. Dai, “Color-guided depth image recovery with adaptive data fidelity and transferred graph laplacian regularization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 320–333, 2019.