

# Deep Attentional Guided Image Filtering

Zhiwei Zhong, Xianming Liu, *Member, IEEE*, Junjun Jiang, *Member, IEEE*, Debin Zhao, *Member, IEEE*,  
Xiangyang Ji, *Member, IEEE*

arXiv:2112.06401v2 [cs.CV] 28 Feb 2022

**Abstract**—Guided filter is a fundamental tool in computer vision and computer graphics which aims to transfer structure information from guidance image to target image. Most existing methods construct filter kernels from the guidance itself without considering the mutual dependency between the guidance and target. However, since there typically exist significantly different edges in two images, simply transferring all structural information of the guidance to the target would result in various artifacts. To cope with this problem, we propose an effective framework named deep attentional guided image filtering, the filtering process of which can fully integrate the complementary information contained in both images. Specifically, we propose an attentional kernel learning module to generate dual sets of filter kernels from the guidance and the target, respectively, and then adaptively combine them by modeling the pixel-wise dependency between the two images. Meanwhile, we propose a multi-scale guided image filtering module to progressively generate the filtering result with the constructed kernels in a coarse-to-fine manner. Correspondingly, a multi-scale fusion strategy is introduced to reuse the intermediate results in the coarse-to-fine process. Extensive experiments show that the proposed framework compares favorably with the state-of-the-art methods in a wide range of guided image filtering applications, such as guided super-resolution, cross-modality restoration, texture removal, and semantic segmentation. Moreover, our scheme achieved the first place in real depth map super-resolution challenge held in ACM ICMR’2021<sup>1</sup>.

**Index Terms**—Guided filter, dual regression, attentional kernel learning, guided super-resolution, cross-modality restoration.

## I. INTRODUCTION

**G**UIDED filter (GF), also named joint filter, is tailored to transfer structural information from a guidance image to a target one. The popularity of GF can be attributed to its ability in handling visual signals in various domains and modalities, where one modal signal serves as the guidance to improve the quality of the other one. It has been a useful tool for many image processing and computer vision tasks, such as depth map super-resolution [2], [5], scale-space filtering [8], [9], cross-modality image restoration [1], [3], [10], structure-texture separation [11], [12], [13], image semantic segmentation [7], [14] and so on.

In the literature, GF has been extensively studied, ranging from the classical bilateral filter to the emerging deep learning-based ones. The pioneer bilateral filter [15] constructs spatially-varying kernels, where local image structures of the

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: {zhwzhong,csxm,jiangjunjun}@hit.edu.cn.

X. Ji is with the Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: xyji@tsinghua.edu.cn.

guidance image are explicitly involved into filtering process through the photometric similarity. The guided image filtering scheme proposed by He et al. [1] takes a more rigorous manner to exploit the structure information of the guidance, which computes a locally linear model over the guidance image for filtering. These filters consider only the information contained in the guidance image in filtering. However, since there typically exist significantly different edges in the two images, simply transferring all patterns of the guidance to the target would introduce various artifacts. Some works [3], [8] propose to utilize the optimization-based manner to find mutual structures for propagation while suppressing inconsistent ones. However, it is challenging to select reference structures and propagate them properly by hand-crafted objective functions. In addition, the computational complexity of these methods is usually high.

In recent years, learning-based approaches for GF design are becoming increasingly popular, which derive GF in a purely data-driven manner. They allow the networks to learn how to adaptively select structures to transfer, and thus have the ability to handle more complicated scenarios. For instance, in [16], a dynamic filter network (DFN) is proposed where pixel-wise filters are generated dynamically using a separate sub-network conditioned on the guidance. Unlike DFN, Su et al. [6] adapts a standard spatially invariant kernel at each pixel by multiplying it with a spatially varying filter. Although with increased flexibility thanks to their adaptive nature, [16] and [6] still suffer from the same drawback as [1], [15] that only the guidance information is considered in filters design. Some recent methods attempt to exploit the target and guidance information jointly. For instance, Li et al. [5] propose to leverage two sub-networks to extract informative features from both the target and guidance images, which are then concatenated as inputs for the fusion network to selectively transfer salient structures from the guidance to the target. Instead of regressing the filtering results directly from the network, Kim et al. [7] proposes to use spatially variant weighted averages, where the set of neighbors and the corresponding kernel weights are learned in an end-to-end manner. However, in the designed networks of these methods, the simple concatenation or element-wise multiplication is exploited to combine multimodal information, which is not that effective. There is no mechanism to distinguish the contributions of the guidance and the target to the final filtering result, and thus would also lead to erroneous structure propagation. In addition, the guidance and target images are treated as independent information since existing methods typically utilize two separate networks for feature extraction, thus the complementary information contained in the two images cannot be fully exploited.

By reviewing existing GF methods, it can be found that most

<sup>1</sup><https://icmr21-realsr-challenge.github.io/#Leaderboard>

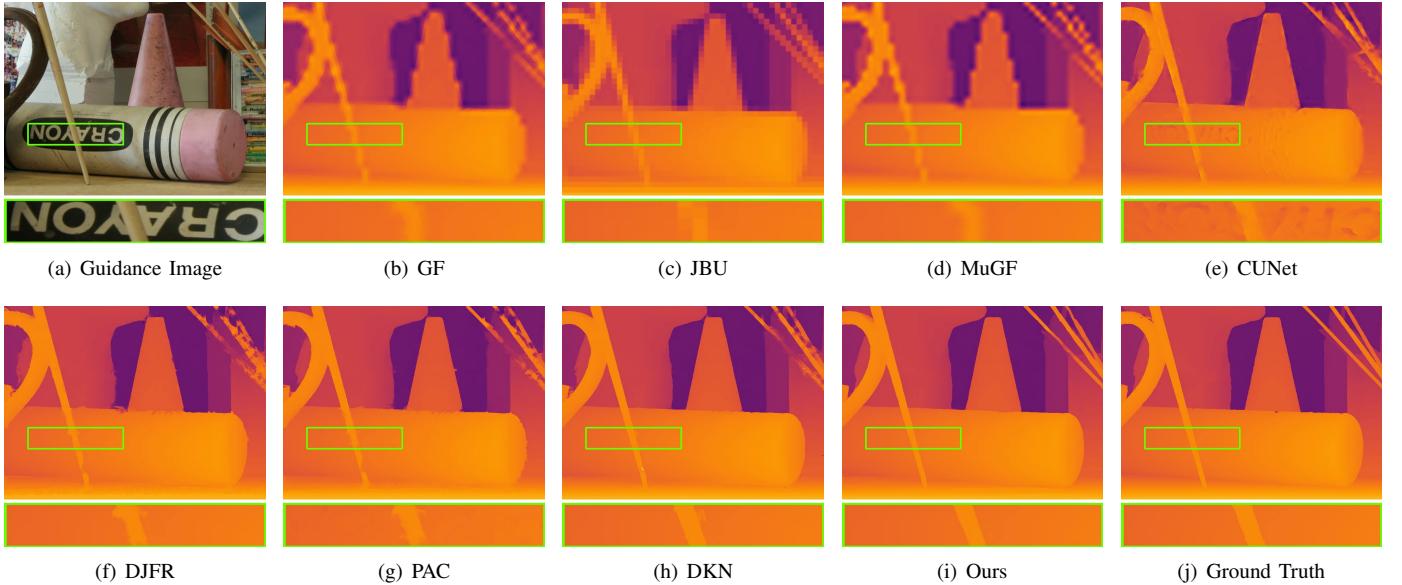


Fig. 1. Guided image filtering on a RGB/D image pair for  $16 \times$  guided super-resolution: (a) Guidance image, (b) GF [1], (c) JBU [2], (d) MuGF [3], (e) CUNet [4], (f) DJFR [5], (g) PAC [6], (h) DKN [7], (i) Ours, (j) Ground truth. The results of (b)-(c) suffer from edge blurring artifact and the results of (d)-(h) suffer from texture-copying artifacts. Our result produces much sharper edges. Please enlarge the PDF for more details.

of them concentrate their efforts on how to transfer structural information from the guidance to the target. However, for some scenarios, such as cross-modality image restoration [4] and guided super-resolution [17], multi-modal data has significantly different characteristics due to the difference of sensing principle, making the guidance not always trustworthy. In view of this, we argue that the purpose of GF should be two-folds: 1) apply the guidance as a prior for reconstruction of regions in the target where there are structure-consistent contents; and 2) derive a plausible prediction for regions in the target with inconsistent contents of the guidance. The latter represents the case that the guidance is no longer reliable, so we have to rely on the target itself for reconstruction. Most existing GF methods only concern structure transferring from the guidance, but neglect structure prediction from the target, leading to erroneous or extraneous artifacts in the output. It implies that instead of performing regression on guidance only, as done in [1], [15], we should perform *dual regression* on both the guidance and the target, and combine them adaptively in a smarter manner instead of simple concatenation or element-wise multiplication, as done in [5], [7]. “Dual regression” and “smart combination” bring the main motivations of our proposed method.

Accordingly, in this paper, we propose an effective deep attentional guided image filtering scheme, which constructs filter kernels by fully considering information from both guidance and target images. Specifically, an attentional kernel learning module is proposed to generate dual sets of filter kernels from the guidance and the target, respectively. Moreover, pixel-wise contributions of the guidance and the target to the final filtering result are automatically learned. In this way, we can adaptively apply the guidance as a prior for reconstruction of target regions where there are structure-consistent contents with the guidance; and derive a prediction for target regions

with inconsistent contents by regression on the target itself. We show an illustrated example in Fig. 1, which presents the visual filtering results comparison of our scheme with the state-of-the-art guided depth super-resolution methods. It can be found that our proposed method is capable of producing high-resolution depth image with clear boundaries as well as avoiding texture-copying artifacts.

The main contributions of the proposed method are summarized as follows:

- We propose an attentional kernel learning (AKL) module for guided image filtering, which generates dual sets of filter kernels from both guidance and target, and then adaptively combines these kernels by modeling the pixel-wise dependencies between the two images in a learning manner. Compared with existing kernel generation approaches, the proposed method is more robust when there are inconsistent structures between the guidance and the target.
- We propose a multi-scale guided filtering module, which generates the filtering result in a coarse-to-fine manner. Correspondingly, we propose a multi-scale fusion strategy with deep supervision to fully explore the intermediate results in the coarse-to-fine process. To the best of our knowledge, this is the first guided filter framework that learns the multi-scale kernels to filter the target image at different scales in the embedding space.
- We evaluate the performance of the proposed method on various computational photography and computer vision tasks, such as guided image super-resolution, cross-modality image restoration, texture removal, and semantic segmentation. The quantitative and qualitative results demonstrate the effectiveness and universality of the proposed method.
- Considering that there is no standard protocol to train

and evaluate the performance of guided image filtering algorithms, we reimplement eight recently proposed state-of-the-art deep learning-based guided filtering models and unify their settings to facilitate fair comparison. All of the codes and trained models are publicly available<sup>2</sup> to encourage reproducible research.

The remainder of this paper is organized as follows. Sect. II gives a brief introduction to the relevant works of guided filter. Sect. III introduces the proposed method for guided image filtering. Sect. IV provides experimental comparisons with existing state-of-the-art methods for a varied range of guided filtering tasks. Ablation experiments are presented in Sect. V to analyze the network hyper-parameters and verify the advantage of each components proposed in our model. We conclude the paper in Sect. VI.

## II. GUIDED FILTERS REVISITING

In this section, we start with a revisiting of formal definitions of popular variants of guided filters in the literature, and then explain our generalization of them to derive the proposed deep attentional guided image filter.

### A. Classical Guided Filters

Define the guidance image as  $\mathbf{g}$  and the target image as  $\mathbf{t}$ , the output  $\mathbf{f}$  of guided filtering can be represented as:

$$\mathbf{f}_i = \sum_j \mathbf{W}_{i,j}(\mathbf{g}, \mathbf{t}) \mathbf{t}_j, \quad (1)$$

where  $i$  and  $j$  are pixel coordinates;  $\mathbf{W}_{i,j}$  is the filter kernel weight, whose parameters  $(\mathbf{g}, \mathbf{t})$  mean that it can be derived from either  $\mathbf{g}$  or  $\mathbf{t}$ , or both.

In the classical bilateral filter and guided image filter,  $\mathbf{W}_{i,j}$  is only dependent on the guidance  $\mathbf{g}$ . Specifically, the filter weight in bilateral filter is defined as:

$$\mathbf{W}_{i,j}^{BF} = \frac{1}{C_i} \exp\left(-\frac{\|i-j\|}{\sigma_s}\right) \exp\left(-\frac{\|\mathbf{g}_i - \mathbf{g}_j\|}{\sigma_r}\right), \quad (2)$$

where  $C_i$  is the normalization parameter;  $\sigma_s$  and  $\sigma_r$  are parameters for geometric and photometric similarity, respectively. In guided image filter (He et al., [1]), the filter kernel weight is defined as:

$$\mathbf{W}_{i,j}^{GIF} = \frac{1}{|\mathcal{N}_k|^2} \sum_{k:(i,j) \in \mathcal{N}_k} \left( 1 + \frac{(\mathbf{g}_i - \mu_k)(\mathbf{g}_j - \mu_k)}{\sigma_k^2 + \epsilon} \right), \quad (3)$$

where  $|\mathcal{N}_k|$  is the number of pixels in a window  $\mathcal{N}_k$ ;  $\mu_k$  and  $\sigma_k^2$  are the mean and variance of  $\mathbf{g}$  in  $\mathcal{N}_k$ .

### B. Learning-based Guided Filters

Among deep learning based approaches for guided filter design, dynamic filter network [16] first defines a filter-generating network (FGN) that takes the guidance  $\mathbf{g}$  as input to obtain location-specific dynamic filters  $\mathbf{F}_\theta = \text{FGN}(\mathbf{g}, \theta)$ , which are then applied to the target image  $\mathbf{t}$  to yield the output  $\mathbf{f} = \mathbf{F}_\theta(\mathbf{t})$ . Pixel-adaptive convolution [6] defines the filter

kernel by multiplying a spatially varying filter on standard spatially invariant kernel:

$$\mathbf{f}_i = \sum_{j \in \mathcal{N}_i} \mathbf{K}(\mathbf{g}_i, \mathbf{g}_j) \mathbf{W}[p_i - p_j] \mathbf{t}_j + b, \quad (4)$$

where  $\mathbf{W}$  is the spatially invariant kernel;  $\mathbf{K}(\cdot, \cdot)$  is a varying filter kernel function that has a fixed form such as Gaussian,  $[p_i - p_j]$  denotes the index offset of kernel weights. From the above formulation, it can be found that, similar to bilateral filter and guided image filter, dynamic filter network and pixel-adaptive convolution also only depend on the guidance  $\mathbf{g}$  in defining the filter kernels. When there are inconsistent structures in the guidance and the target, this approach would generate annoying artifacts in the output.

The recent deep joint filtering (DJF) method [5] alleviates this drawback by jointly leveraging features of both the guidance and the target. It designs two-branch sub-networks to extract features from the guidance and the target respectively, which are passed through a fusion sub-network to output the filtering result. The joint filter  $\Phi$  is learned in an end-to-end manner by the following optimization:

$$\Phi^* = \arg \min_{\Phi} \|\mathbf{f}^{gt} - \Phi(\mathbf{g}, \mathbf{t})\|^2, \quad (5)$$

where  $\mathbf{f}^{gt}$  is the ground truth of the output. In contrast to the implicit filter learning approach of DJF, deformable kernel networks (DKN) [7] explicitly learns the kernel weights  $\mathbf{K}$  and offsets  $s$  using two-branch sub-networks from the two images. Concretely, the filtering is performed by

$$\mathbf{f}_i = \sum_{j \in \mathcal{N}_i} \mathbf{W}_{i,s(j)}(\mathbf{g}, \mathbf{t}) \mathbf{t}_{s(j)}, \quad (6)$$

with

$$\mathbf{W}(\mathbf{g}, \mathbf{t}) = \mathbf{K}(\mathbf{g}) \odot \mathbf{K}(\mathbf{t}), \quad (7)$$

where  $\mathbf{K}(\mathbf{g})$  and  $\mathbf{K}(\mathbf{t})$  are kernel weights learned from the guidance and the target, respectively,  $\odot$  denotes element-wise multiplication. Although DJF and DKN achieve better performance than previous methods, they treat the guidance and target images as independent information and utilize separate networks for kernel learning, thus the complementary information contained in the two images cannot be fully exploited. In addition, the fusion approach of multi-modal weights through element-wise multiplication is not effective, in which the guidance and the target contribute equally to the final filtering results.

### C. Our Strategy

Considering the drawbacks of existing methods, we propose a deep attentional guided image filtering scheme to more effectively leverage multi-modal information. Our method performs dual regression on both guidance and target, and combines them adaptively using an attention mechanism. Mathematically, our filtering process can be generally formulated as

$$\mathbf{f}_i = \sum_{j \in \mathcal{N}_i} \mathbf{A}_{i,j} \mathbf{W}_{i,j}^g \mathbf{t}_j + \sum_{j \in \mathcal{N}_i} (1 - \mathbf{A}_{i,j}) \mathbf{W}_{i,j}^t \mathbf{t}_j, \quad (8)$$

<sup>2</sup><https://github.com/zhwzhong/DAGF>

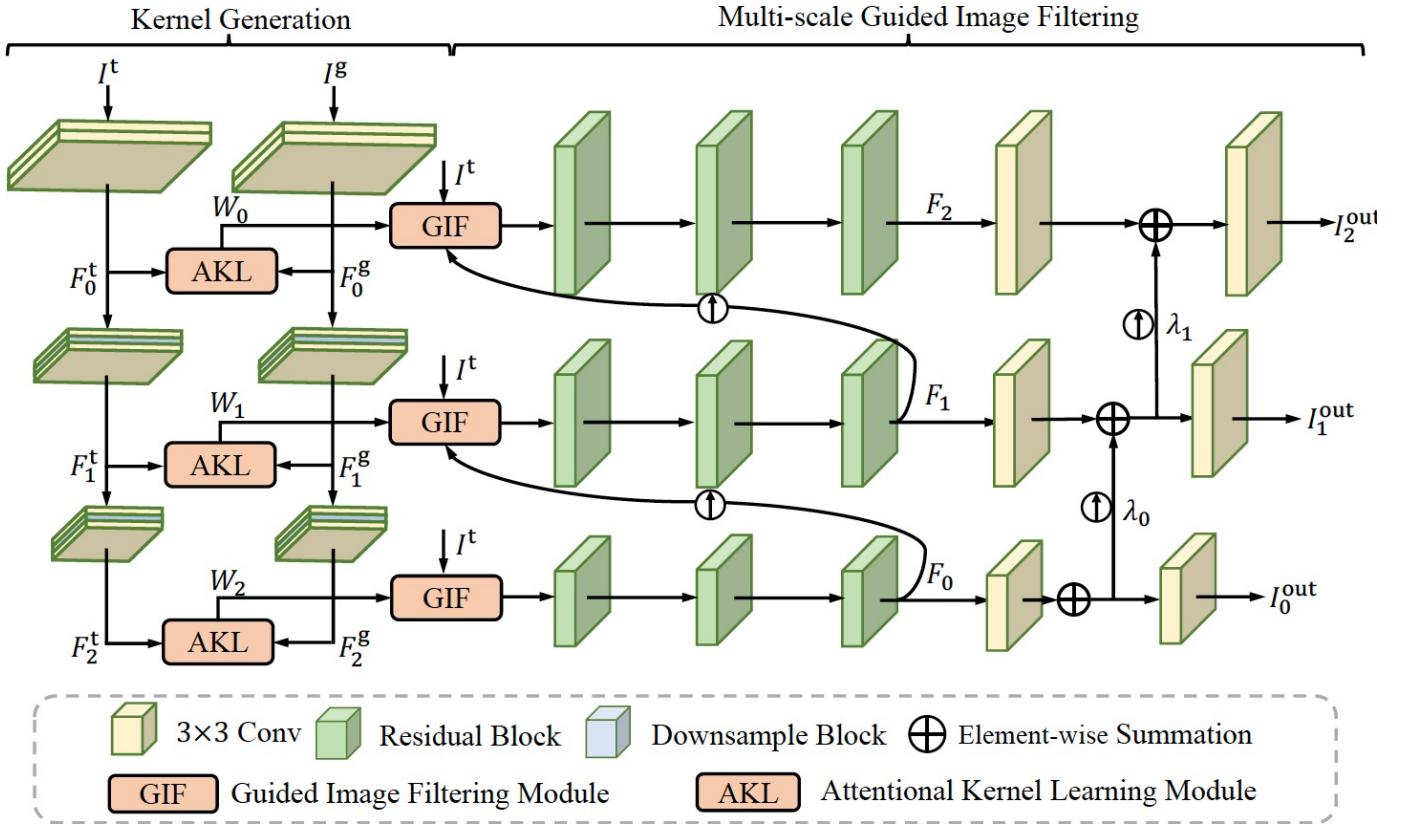


Fig. 2. The network architecture of the proposed deep attentional guided image filtering (DAGF) with the number of pyramid level  $m = 3$ . DAGF consists of a kernel generation network for constructing filter kernels and a multi-scale guided image filtering network with the purpose of filtering target image by using the generated kernels.

where  $W_{i,j}^g$  and  $W_{i,j}^t$  are filter kernels computed from the guidance and the target respectively;  $A_{i,j}$  denotes the pixel-wise reliability weight of the guidance image, which is determined automatically by considering both guidance and target information. The above formulation means that, when the guidance information is not trustworthy, we should turn to use the target information itself for regression, so as to prevent the unreliable structure propagation.

### III. PROPOSED METHOD

An effective guided image filtering scheme should be able to identify the consistent structures contained in the guidance as well as avoid transferring extraneous or erroneous contents to the target. In this section, we introduce in detail the proposed deep attentional guided image filtering (DAGF) framework for this purpose, where the complementary information contained in the two images can be fully explored in both kernel generation and image filtering process.

#### A. Network Architecture

The DAGF takes a target image  $I^t \in \mathbb{R}^{H \times W \times C^t}$  (e.g., low-resolution depth) and a guidance image  $I^g \in \mathbb{R}^{H \times W \times C^g}$  (e.g., high-resolution color image) as inputs, and generates a reconstructed image  $I^{\text{out}} \in \mathbb{R}^{H \times W \times C}$  as output, where  $H, W$  and  $C$  denote the height, width and the number of channels respectively.

Fig. 2 illustrates the overall architecture of the proposed network, which is composed of *kernel generation sub-network* and *multi-scale guided filtering sub-network*. Instead of directly predicting kernels in image domain and enlarging its receptive field by using the deformable sampling strategy as [7], we employ a pyramid architecture to achieve a large receptive field, and conduct filter learning in the feature domain since deep features are more robust with respect to appearance difference of the target and the guidance.

- In the filter kernel generation sub-network, the multi-scale features of  $I^t$  and  $I^g$  are fed into the attentional kernel learning (AKL) module to generate filter kernels  $\{W_i\}$ . The network architecture of AKL is illustrated in Fig. 3, where an attentional contribution module based on U-Net architecture is designed to adaptively fuse the filter kernels generated by the guidance and the target.
- In the guided filtering sub-network, with the derived pixel-wise filter kernels, features of the target image are processed in a coarse-to-fine manner to get the upsampled features.

The above process is repeated until arriving the final scale. In the following, we will elaborate these two sub-networks and the loss function design for network training.

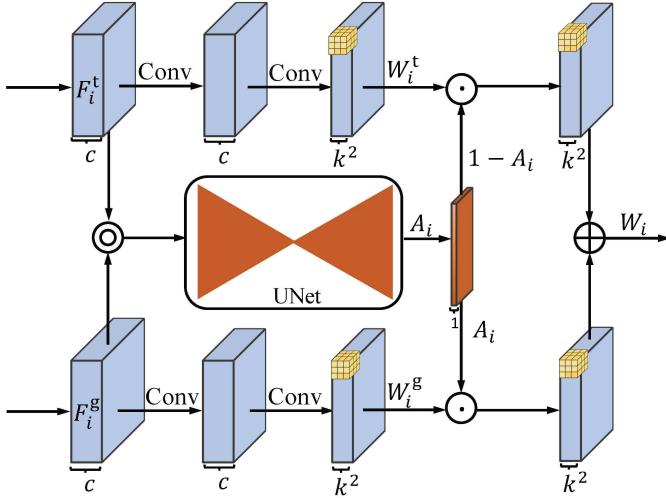


Fig. 3. The network architecture of the proposed attentional kernel learning (AKL) module, where  $\odot$  denotes the element-wise multiplication and  $\oplus$  is concatenation operation.

### B. Filter Kernel Generation

The filter kernel generation sub-network is tailored to generate spatial-variant kernels by considering the mutual dependency between the target and the guidance. As illustrated in the left part of Fig. 2, given  $I^t$  and  $I^g$  as inputs, we first employ two-branch pyramid network to extract multi-scale features  $\{F_i^t, 0 \leq i < m\}$  and  $\{F_i^g, 0 \leq i < m\}$  from the target and the guidance, respectively. We take the target branch as an example, which is done by:

$$F_0^t = \text{Conv}(\text{Conv}(I^t)), \quad (9)$$

$$F_i^t = \text{Down}(F_{i-1}^t), 0 < i < m, \quad (10)$$

where  $m$  denotes the level of pyramid network and  $\text{Conv}(\cdot)$  is the convolution operator;  $\text{Down}(\cdot)$  represents the down-sample block with scale factor 2, which is implemented by two convolution layers and a inverse pixel-shuffle [18] operation.

For guided image filtering, the prior information for reconstruction is either from the guidance image if there are consistent structures between the guidance and target images, or from the target image itself if there is no reliable guidance information. This inspires us to design dual regression over the guidance and the target, respectively, instead of only relying on the guidance as done in most existing methods. To this end, as shown in Fig. 3, we propose an attentional kernel learning (AKL) module. It takes the extracted guidance and target features as inputs and consists of two steps: dual kernels generation and adaptive kernels combination.

The first step is the dual kernels generation, which is formulated as:

$$W_i^t = \text{Conv}(\text{Conv}(F_i^t)), 0 \leq i < m, \quad (11)$$

$$W_i^g = \text{Conv}(\text{Conv}(F_i^g)), 0 \leq i < m, \quad (12)$$

where  $W_i^t$  and  $W_i^g$  are the  $i$ -th constructed filter kernels from the target and guidance features respectively. The spatial-resolution of  $i$ -th kernels is the same as the one of its corresponding input features while the number of channels

are  $k^2$  where  $k$  is the desired filter kernel size. However, these kernels generated by the target or guidance information alone cannot explore the dependencies among them, making the filtering outputs suffer from blurring or texture copying artifacts. To alleviate this problem, we introduce adaptive kernels combination module based on a light-weight UNet architecture, which takes both guidance and target features as inputs and models the pixel-wise dependencies among them in a learning manner. This process is formulated as:

$$A_i = \text{UNet}([F_i^t, F_i^g]), 0 \leq i < m, \quad (13)$$

where UNet is a five-layer U-like [19] network,  $[\cdot, \cdot]$  denotes concatenation operation;  $A_i$  is the output of this module, which can be considered as an attention map to adaptively combine kernels constructed from guidance and target features. The final guided filter kernels can be derived as:

$$W_i = A_i \odot W_i^g + (1 - A_i) \odot W_i^t, 0 \leq i < m, \quad (14)$$

where  $W_i$  is the generated  $i$ -th filter kernel;  $1$  denotes the all-1 matrix;  $\odot$  means element-wise multiplication.

### C. Multi-scale Guided Filtering

After generating the guided filter kernels, the following step is to perform filtering on the target image, which is done by the guided filtering sub-network. As shown in the right part of Fig. 2, it takes the target image  $I^t$  as the input, and progressively filters the input target image by using the learned filter kernels  $\{W_0, \dots, W_{m-1}\}$  in a coarse-to-fine manner.

Specifically, given  $I^t$  as input, we first utilize Bicubic to resize it to the same resolution as its corresponding filter kernels:

$$\hat{I}^t = \text{Bicubic}(I^t). \quad (15)$$

Then the filtering process can be formulated as:

$$F_0 = \text{ResNet}(\text{GIF}(\text{Conv}(\hat{I}^t)), W_0), \quad (16)$$

$$F_i = \text{ResNet}(\text{GIF}([F_{i-1}^\uparrow, \text{Conv}(\hat{I}^t)], W_i)), 0 < i < m, \quad (17)$$

where  $[\cdot, \cdot]$  means concatenation operation and  $\text{ResNet}(\cdot)$  is the function including three residual blocks (He et al., [20]);  $F_i$  is  $i$ -th filtered target feature;  $\uparrow$  is upsampling operation.  $\text{GIF}(\cdot)$  is a filtering operation that conducts filtering operation on the corresponding target features. Specifically, we first reshape the third dimension of the filter from  $k^2$  to  $k \times k$ , then the filtering process for a pixel  $\{(u, v) | 0 \leq u < H, 0 \leq v < W\}$  can be defined as following:

$$F(u, v) = \sum_{x=-\sigma}^{\sigma} \sum_{y=-\sigma}^{\sigma} W_{u,v}(x, y) \cdot \tilde{F}(u - x, v - y), \quad (18)$$

where  $\sigma = \lfloor k/2 \rfloor$ ;  $\tilde{F}$  is the output of the GIF module.

Based on  $\{F_i\}_{i=0}^{m-2}$ , we can obtain the filter results of DAGF by using the proposed multi-scale fusion strategy:

$$\hat{F}_0 = \text{Conv}(F_0), \quad (19)$$

$$\hat{F}_i = \text{Conv}(F_i) + \lambda_{i-1} \cdot \hat{F}_{i-1}^\uparrow, 0 < i < m, \quad (20)$$

$$I_i^{\text{out}} = \text{Conv}(\hat{F}_i) + I^t, 0 \leq i < m-1, \quad (21)$$

where  $\lambda_i$  is a learnable parameter that is initialized as 0. The parameter enables the output layer first to rely on features of the current layer and then gradually learn to combine high-level features from previous layers. Therefore, the output of the last layer can enjoy the merit of preserving both high-level contextual details and low-level spatial information.  $\{\mathbf{I}_i^{\text{out}}\}_{i=0}^{m-2}$  are the intermediate multi-scale results and  $\mathbf{I}_{m-1}^{\text{out}}$  is the final filtering result of the proposed scheme.

#### D. Loss function

We adopt the residual learning strategy to train the proposed method. Let  $\mathbf{I}^g$  and  $\mathbf{I}^t$  be the input guidance and target image,  $\mathbf{I}^h$  be the corresponding ground-truth image. The proposed DAGF network aims to learn the residual between  $\mathbf{I}^h$  and  $\mathbf{I}^t$ . The overall all loss function is composed of three terms: a  $L_1$  loss  $\mathcal{L}_1$ , a multi-stage loss  $\mathcal{L}_{ms}$  and a boundary-aware loss  $\mathcal{L}_b$ :

- **L<sub>1</sub> loss.**  $\mathcal{L}_1$  measures the pixel-wise errors between the output image  $\mathbf{I}_{m-1}^{\text{out}}$  and its corresponding residual image  $\mathbf{I}^r$ :

$$\mathcal{L}_1 = \|\mathbf{I}^h - \mathbf{I}_{m-1}^{\text{out}}\|_1. \quad (22)$$

- **Multi-stage loss.** To stabilize the network training process and promote the multi-stage guided filtering module to learn more effective parameters, we propose a multi-stage loss to enforce all intermediate results to be close to the ground truth residual image:

$$\mathcal{L}_{ms} = \frac{1}{m-1} \sum_{i=0}^{m-2} \|\mathbf{I}^h - \text{Bicubic}(\mathbf{I}_i^{\text{out}})\|_1, \quad (23)$$

where  $m$  is the number of pyramid levels. We use Bicubic interpolation to resize the output image  $\mathbf{I}_i^{\text{out}}$  to the same resolution as the ground truth target image  $\mathbf{I}^h$ .

- **Boundary-aware loss.** Optimizing the pixel-wise loss (e.g.,  $\mathcal{L}_1$  and  $\mathcal{L}_2$ ) typically cannot preserve high-frequency structure information well, and tends to produce blurry images as all pixels are treated equally. To mitigate this problem and encourage the network to give more emphasis on the high-frequency parts, we propose a boundary-aware loss to promote our model to generate sharper boundaries. Specifically, we first employ Sobel operator  $\nabla$  to detect the boundary information of the ground truth and the network output, and obtain the boundary mask  $\mathbf{M}$ :

$$\mathbf{M} = (\nabla_x \mathbf{I}^h - \nabla_x \mathbf{I}_{m-1}^{\text{out}}) \odot (\nabla_y \mathbf{I}^h - \nabla_y \mathbf{I}_{m-1}^{\text{out}}), \quad (24)$$

then the boundary-aware loss is defined as:

$$\mathcal{L}_{ba} = \|\mathbf{M} \odot \mathbf{I}^h - \mathbf{M} \odot \mathbf{I}_{m-1}^{\text{out}}\|_1, \quad (25)$$

where  $\odot$  denotes element-wise multiplication.

With these three losses, the total loss is then formulated as:

$$\mathcal{L} = \omega_1 \cdot \mathcal{L}_1 + \omega_2 \cdot \mathcal{L}_{ba} + \omega_3 \cdot \mathcal{L}_{ms}, \quad (26)$$

where  $\omega_1, \omega_2$  and  $\omega_3$  are hyper-parameters to balance these loss functions. We set  $\omega_3 = 1$  to stabilize the training procedure at early stage and then progressively decay to zero with the training progress to boost the performance of final output. We set  $\omega_1 = 1, \omega_2 = 10$ , respectively.

#### E. Implementation Details

In our model, we set the number of pyramid levels as  $m = 3$  and the size of generated kernel in AKL modules as  $3 \times 3$ . The ablation study presented below will verify the effectiveness of our configuration. The hyper-parameters of our model are  $\omega_1 = 1, \omega_2 = 0.001$  and  $\omega_3 = 1$ . All the convolution layers within the proposed methods are sized of  $3 \times 3$  and the channels of intermediate features are 32. We use PReLU [21] as the default activation function. We utilize PixelShuffle [22] and InvPixelShuffle [18] as the up-sampling and down-sampling operators to resize the features in our model.

In the training phase, the batch size is set as 32 and we random crop  $256 \times 256$  image patches from the target and guidance images as inputs. We augment the training data with random flipping and rotation. Adam [23] with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  is employed as optimizer. The initial learning rate is set as  $1 \times 10^{-4}$  and we halve it every 80 epochs, stop the training after 100 epochs. Our model is implemented by Pytorch [24] and trained on one RTX 1080ti GPU. Training the proposed method roughly takes 2 day for NYU v2 [25] datasets.

Our network takes three channels guidance and one channel target images as inputs. For the multi-channels target images, we apply the trained model separately for each channel and the outputs are combined to obtain the final result. For the single-channel guidance image, we copy the single-channel three times to generate three-channels guidance image.

## IV. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the performance of the proposed method on a wide range of guided image filtering tasks, including guided image super-resolution (e.g. depth image super-resolution and saliency map super-resolution, Sect. IV-A), cross-modality image restoration (e.g. joint depth image super-resolution and denoising, and flash/non-flash image denoising, Sect. IV-B), texture removal (Sect. IV-C) and image semantic segmentation (Sect. IV-D).

For fair comparison, the results for the compared methods are generated by using the source codes released by their authors with the default parameter settings, and all learning based methods are trained and tested on the same datasets.

### A. Guided Image Super-resolution

Guided image super-resolution (GSR) is a classic computer vision task which aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) one with the help of a HR image from another modality. For example, we can obtain a HR depth by GSR using a LR depth and a HR RGB image as inputs, where the HR RGB image serves as the guidance.

Following the experimental settings of [5], [7], we train our model on the task of depth image super-resolution, and then evaluate the performance of the model on tasks of depth image super-resolution and saliency map super-resolution, the latter one is used to verify the generalization ability of our model.

**RGB-guided Depth Super-resolution.** For this task, we use the first 1000 RGB-D image pairs from NYU v2 dataset [25] as the training set. In order to make fair comparison with

TABLE I

QUANTITATIVE COMPARISON FOR DEPTH IMAGE SUPER-RESOLUTION ON FOUR STANDARD RGB/D DATASETS IN TERMS OF AVERAGE RMSE VALUES. FOLLOWING THE EXPERIMENTAL SETTING OF [7], [26], WE CALCULATE THE AVERAGE RMSE VALUES IN CENRIMETER FOR NYU v2 [25] DATASET. FOR OTHER DATASETS, WE COMPUTE THE RMSE VALUES BY SCALING THE DEPTH VALUE TO THE RANGE [0, 255]. THE BEST PERFORMANCE FOR EACH CASE ARE HIGHLIGHTED IN **BOLDFACE** WHILE THE SECOND BEST ONES ARE UNDERSCORED. FOR RMSE METRIC, THE LOWER VALUES MEAN THE BETTER PERFORMANCE.

Datasets	Middlebury			Lu			NYU v2			Sintel		
Method	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	10.11	14.51	19.95
MRF [27]	4.26	7.43	11.80	4.90	9.03	14.19	7.84	13.98	22.20	9.87	13.45	18.19
GF [1])	4.01	7.22	11.70	4.87	8.85	14.09	7.32	13.62	22.03	8.83	12.60	18.78
TGV [28])	3.39	5.41	12.03	4.48	7.58	17.46	6.98	11.23	28.13	8.30	13.05	19.96
SDF [8])	3.14	5.03	8.83	4.65	7.53	11.52	5.27	12.31	19.24	9.20	13.63	19.36
FBS [14])	2.58	4.19	7.30	3.03	5.77	8.48	4.29	8.94	14.59	8.29	10.31	16.18
JBU [2])	2.44	3.81	6.13	2.99	5.06	7.51	4.07	8.29	13.35	8.25	11.74	16.02
Experiment results for depth map super-resolution (Nearest-neighbour down-sampling).												
DGF [29])	3.92	6.04	10.02	2.73	5.98	11.73	4.50	8.98	16.77	7.53	11.53	17.50
DJF [26])	2.14	3.77	6.12	2.54	4.71	7.66	3.54	6.20	10.21	7.09	9.12	12.36
DMSG [30])	<u>1.79</u>	3.39	5.87	2.48	4.74	7.51	3.48	6.07	10.27	6.80	9.09	11.81
DJFR [5])	1.98	3.61	6.07	<u>2.22</u>	4.54	7.48	3.38	5.86	10.11	7.05	9.12	12.61
DSRN [31])	2.08	3.26	5.78	2.57	4.46	6.45	3.49	5.70	9.76	7.29	9.43	11.62
PAC [6])	1.91	3.20	5.60	2.48	4.37	6.60	2.82	5.01	8.64	6.79	<u>8.36</u>	<u>11.02</u>
DKN [7])	1.93	<u>3.17</u>	<u>5.49</u>	2.35	<u>4.16</u>	<u>6.33</u>	<u>2.46</u>	<u>4.76</u>	<u>8.50</u>	<u>6.84</u>	8.61	11.21
DAGF(Ours)	<b>1.78</b>	<b>2.73</b>	<b>4.75</b>	<b>1.96</b>	<b>3.81</b>	<b>6.16</b>	<b>2.35</b>	<b>4.62</b>	<b>7.81</b>	<b>6.72</b>	<b>8.35</b>	<b>10.64</b>
Experiment results for depth map super-resolution (Bicubic down-sampling).												
DGF [29])	1.94	3.36	5.81	2.45	4.42	7.26	3.21	5.92	10.45	5.91	8.02	11.17
DJF [26])	1.68	3.24	5.62	1.65	3.96	6.75	2.80	5.33	9.46	5.30	7.53	10.41
DMSG [30])	1.88	3.45	6.28	2.30	4.17	7.22	3.02	5.38	9.17	4.73	6.26	<u>8.36</u>
DJFR [5])	1.32	3.19	5.57	1.15	3.57	6.77	2.38	4.94	9.18	4.90	7.39	10.33
DSRN [31])	1.77	3.05	4.96	1.77	3.10	<u>5.11</u>	3.00	5.16	8.41	4.49	6.53	9.28
PAC [6])	1.32	2.62	4.58	1.20	2.33	5.19	1.89	3.33	6.78	4.42	6.13	8.42
DKN [7])	<u>1.23</u>	<u>2.12</u>	4.24	<u>0.96</u>	<u>2.16</u>	<u>5.11</u>	<u>1.62</u>	<u>3.26</u>	<u>6.51</u>	4.38	<u>5.89</u>	8.40
DAGF(Ours)	<b>1.15</b>	<b>1.80</b>	<b>3.70</b>	<b>0.83</b>	<b>1.93</b>	<b>4.80</b>	<b>1.36</b>	<b>2.87</b>	<b>6.06</b>	<b>3.84</b>	<b>5.59</b>	<b>7.44</b>

existing methods, we exploit the nearest-neighbour down-sampling as the standard downsampling operator to generate LR target image from the ground-truth. Three scales are considered, including 4×, 8×, 16×. To show the effectiveness of the proposed method, we further conduct experiments on Bicubic downsampling as done in [7]. The performance of the proposed method is evaluated on the following four standard benchmark datasets:

- Sintel dataset [32]: this dataset consists of 1064 image pairs which are obtained by an animated 3D movie.
- NYU v2 dataset [25]: this dataset contains 1449 image pairs acquired by Microsoft Kinect. We use the last of 449 image pairs to evaluate the performance of our method.
- Lu dataset [33]: it contains 6 image pairs captured by ASUS Xtion Pro camera.
- Middlebury dataset [34], [36]: this dataset is captured by structure light, and we utilize the 30 image pairs from 2001-2006 datasets with the missing depth values generated by Lu et al. [33].

We compare our method with 13 state-of-the-art methods, including two local filtering-based methods: GF [1] and JBU [2]; four global optimization-based methods: MRF [27], TGV [28], SDF [8] and FBS [14]; seven deep learning-based methods: DGF [29], DJF [26], DMSG [30], DJFR [5],

DSRN [31], PAC [6] and DKN [7]. We adopt Root Mean Square Error (RMSE) as the evaluation metric. Lower RMSE values mean higher recovery quality.

Table I summarizes the quantitative comparison results between ours and other state-of-the-art methods. The best performance is highlighted in bold. As can be seen from this table, our method achieves the best results among all the compared methods on both synthetic and real datasets (e.g. the Sintel and NYU v2 dataset) and on three scales. The superior performance benefits from the more precise filter kernels learned and the multi-scale filtering process. Compared with the second best results (underlined), our results obtain the gains of 0.12(4×), 0.24(8×) and 0.39(16×) with respect to average RMSE values.

To further analyze the performance of the proposed method, we present the visual results for 8× depth image super-resolution in Fig. 4. It can be observed that the results of JBU [2] suffer from jaggy artifacts. The results of GF [1] are over-smoothed, which indicates that the local filter is not effective at large scale factors (e.g. 8×). Compared to GF [1] and JBU [2], the learning-based methods are capable of generating results with clearer boundaries. However, for finer details, e.g., the arm in the second image and the rope in the last image, the compared learning-based methods exhibit

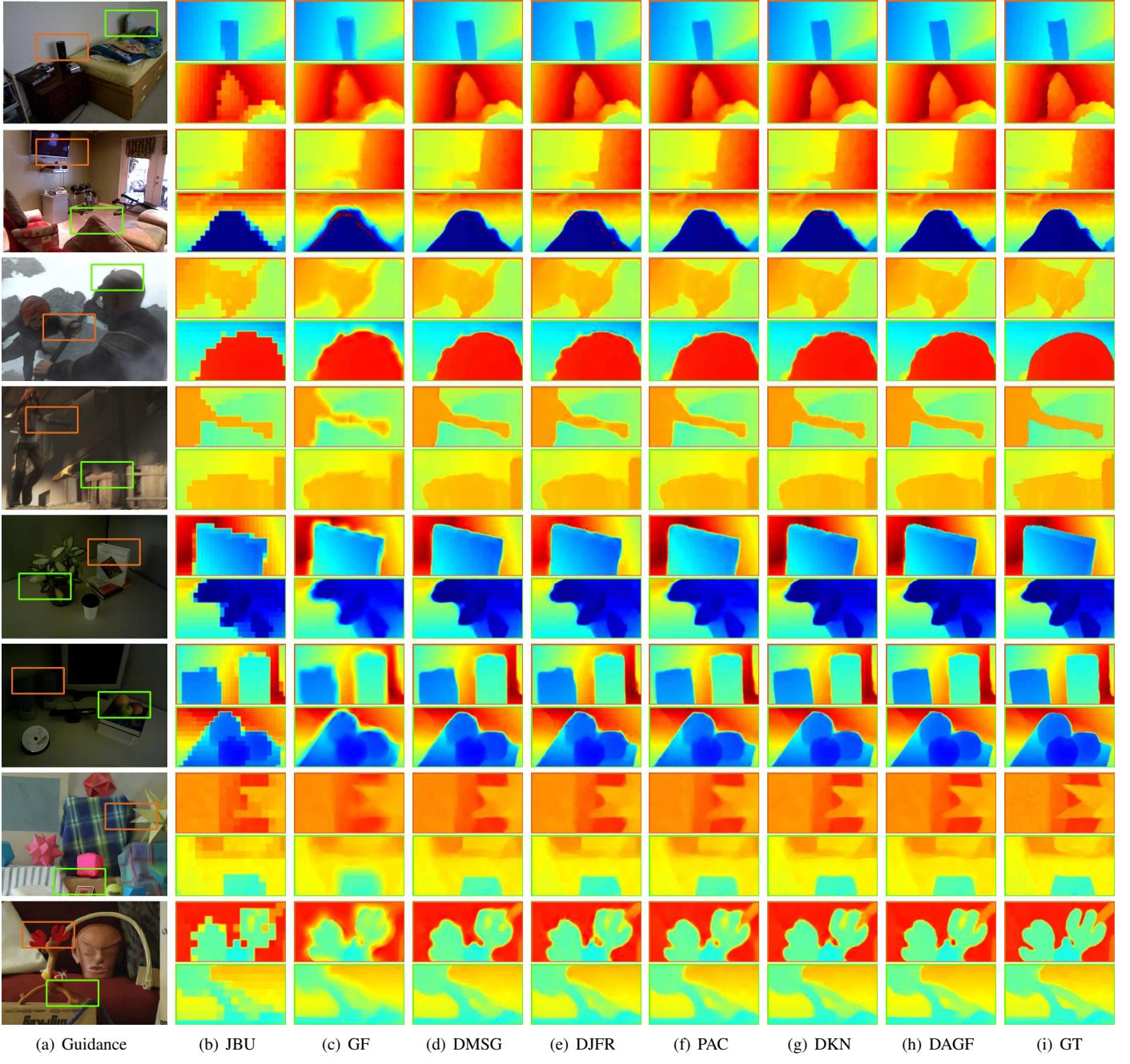


Fig. 4. Qualitative comparison for recovered depth maps (8 $\times$ ). (a) Guidance image, (b): JBU [2], (c): GF [1], (d): DMSG [30], (e): DJFR [5], (f): PAC [6], (g): DKN [7], (h): DAGF, and (i) Ground truth depth map. Top to bottom: Each two rows present recovered depth maps on the NYU v2 [25], Sintel [32], Lu [33] and Middlebury [34] datasets respectively. Please enlarge the PDF for more details.

TABLE II

QUANTITATIVE COMPARISON OF 8 $\times$  SALIENCY MAP SUPER-RESOLUTION ON THE DUT-OMRON DATASET [35]. FOLLOWING DJFR [5], WE USE F-MEASURE TO CALCULATE THE DIFFERENCE BETWEEN THE PREDICTED SALIENCY MAP AND THE CORRESPONDING GROUND-TRUTH. THE BEST PERFORMANCE FOR EACH CASE IS HIGHLIGHTED IN **BOLDFACE** WHILE THE SECOND ONE IS UNDERSCORED FOR F-MEASURE, THE HIGHER VALUES MEAN THE BETTER PERFORMANCE.

Methods	Bicubic	GF [1]	DMSG [30]	DJFR [5]	PAC [6]	FDKN [7]	DKN [7]	DAGF (Ours)
Fscore	0.853	0.821	0.910	0.901	0.922	0.921	<u>0.926</u>	<b>0.932</b>

obvious artifacts such as blurring on the arm and wrong estimation on the rope, which implies that the downsampling

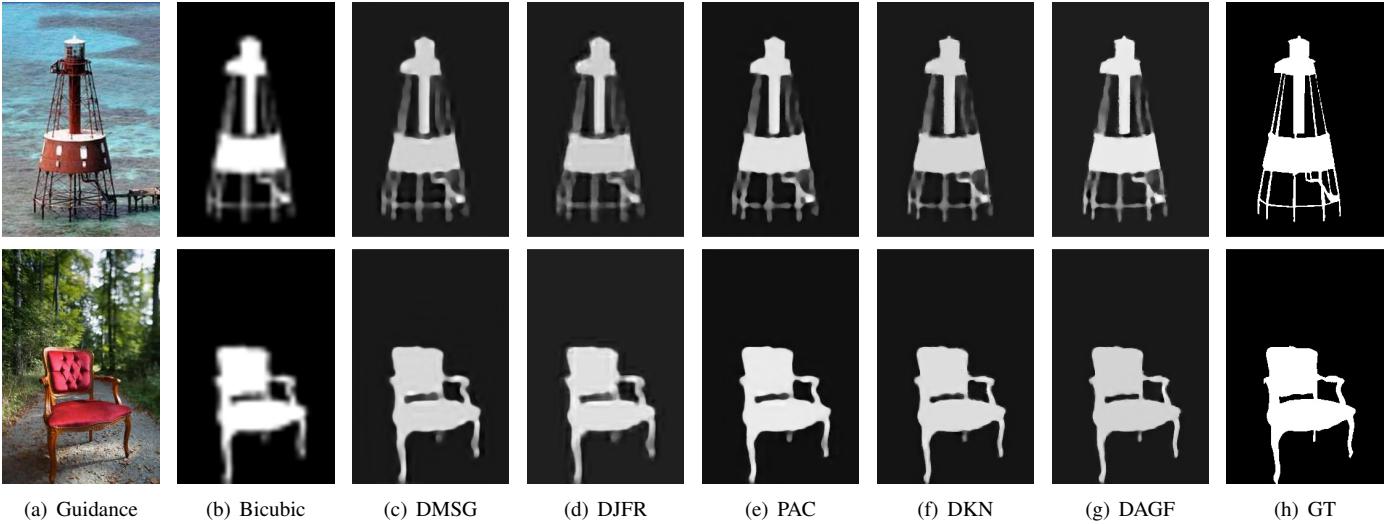


Fig. 5. Visual comparison of  $8\times$  saliency map super-resolution on the DUT-OMRON dataset [35]: (a): Guidance, (b): low-resolution image, (c): DMSG [30], (d): DJFR [5], (e): PAC [6], (f): DKN [7], (g): DAGF, (h): Ground truth. Please enlarge the PDF for more details.

degradation brings significantly damage on the small objects and therefore makes those regions harder to recover. On the contrary, the results obtained by the proposed method are clearer, sharper, and more faithful to the ground truth image.

**RGB-guided Saliency Map Super-resolution.** To further demonstrate the generalization ability of the proposed method, we apply the model that is trained on NYU v2 dataset directly to the task of saliency map super-resolution without any fine-tuning step. Similar to DKN [7], we use 5168 image pairs from DUT-OMRON dataset [35] to evaluate the SR performance. We use bicubic interpolation ( $8\times$ ) to generate the low-resolution saliency maps and then super-resolve them with the corresponding high-resolution color image as the guidance. The quantitative results in terms of F-measure are listed in Table II. As can be seen from this table, our DAGF achieves the best result among all the compared methods and outperforms the second best method by a large margin, which demonstrates the generalization ability of the proposed method. In addition, we random select two images and visualize the recovered high-resolution saliency map obtained by different methods in Fig. 5. It can be observed that the results of Bicubic are over-smoothed, in which the structure details are severely damaged. DMSG [30] and DJFR [5] struggle to generate clear boundaries. The results of DKN [7] have certain artifacts around the edge area. In contrast, our method is able to generate high-quality saliency maps as well as keep the sharpest boundaries, which indicates that the proposed method can fully take advantage of the guidance image and effectively transfer meaningful structure information.

### B. Cross-modality Image Restoration

For the task of cross-modality image restoration, we first conduct experiments on joint depth image super-resolution and denoising to show the superiority of the proposed method. Moreover, to verify the ability of the proposed method on dealing with various visual domains, we apply the trained

models on two noise reduction tasks using flash/non-flash and RGB/NIR image pairs. Finally, we conduct experiments on ToF Mark dataset [39]. It contains three real world depth images acquired by Time of Flight (ToF) camera, which have complicated multi-modality degradation

### Joint Depth Image Super-resolution and Denoising.

Depth images acquired by ranging sensors are typically noisy. In order to simulate the data acquisition process of the depth sensor, we add Gaussian noise with variance as 25 to the low-resolution target depth images. We use the same experimental settings as the task of GSR in Sect IV-A to train our model. We compare our method with ten state-of-the-art methods, including GF [1], MUF [3] and SDF [8], which are traditional model-based methods; and DGF [29], DJF [26], DMSG [30], DJFR [5], DSRN [31], PAC [6], DKN [7], which are deep learning-based methods. Since most of the existing methods do not provide experimental results for this task, we retrain all the deep learning-based methods with the same training and test dataset as ours.

The quantitative results in terms of RMSE values for four benchmark datasets are reported in Table III, from which we can see that the proposed method can obtain consistently better results than the existing state-of-the-art methods, especially for the  $8\times$  and  $16\times$  cases which are more difficult to recover. This is mainly because that: 1) we employ a pyramid architecture to extract multi-modality features for guided kernel generation, thus the multi-scale complementary information can be obtained; 2) for guided image filtering, we leverage the coarse-to-fine strategy to filter the low-resolution target image and thus the structure details can be progressively recovered; 3) compared to single loss at the end of network, the proposed multi-scale loss can bring stronger supervision to our model.

Fig. 6 further demonstrates the visual superiority of the proposed method for joint depth image super-resolution and denoising ( $16\times$  Bicubic downsampling and Gaussian noise). The results of GF [1], MUF [3] and SDF [8] still contain much

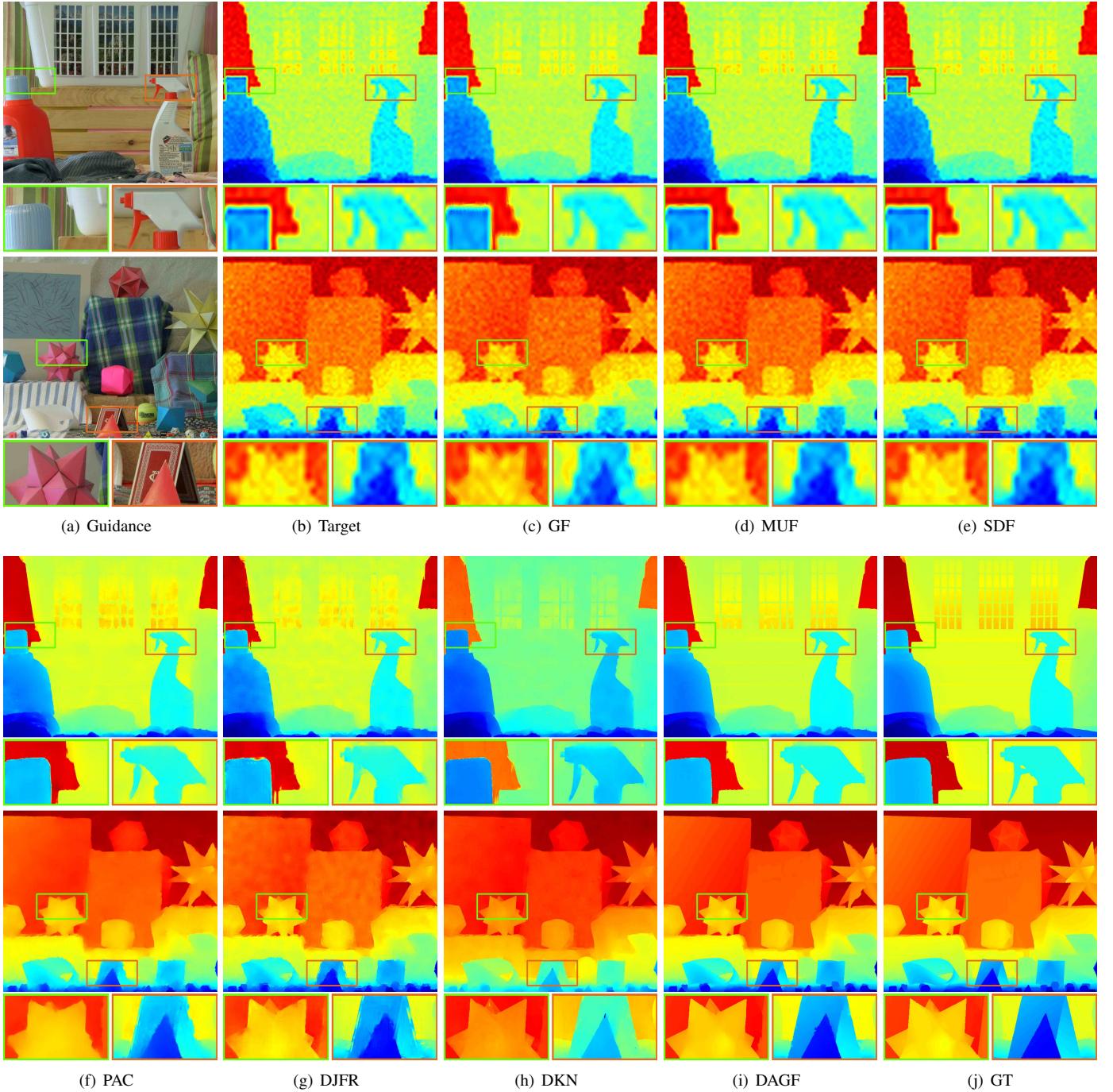


Fig. 6. Qualitative comparison of joint depth map super-resolution and denoising. Please enlarge the PDF for more details. (a): Guidance Image, (b): Target image, (c): GF [1], (d): MUF [3], (e): SDF [8], (f): PAC [6], (g): DJFR [5], (h): DKN [7], (i): DAGF and (j): Ground-truth image. ease enlarge the PDF for more details.

noise, and the visual quality of the whole image is poor. This is due to that these methods are based on the locally linear assumption and they employ the mean filter to calculate the coefficients for pixel-wise linear representations. The methods of PAC [6] and DJFR [5] can remove noise well, while they cannot preserve the sharp edge and introduce ringing artifacts. The results of DKN are clearer and sharper than previous methods. However, they suffer from color distortion, which attributes to the batch normalization used in DKN [7]. In

contrast, our method is able to remove the noise effectively and produces the clearest and sharpest boundaries.

**Cross-modality Image Restoration.** We further demonstrate that our model trained for depth image denoising can be generalized to address other cross-modality image restoration tasks, such as flash guided non-flash image denoising and NIR guided color image restoration. Fig. 8 shows the visual comparison among existing state-of-the-art methods and ours. All of the deep learning-based methods (e.g. DJFR [5] and

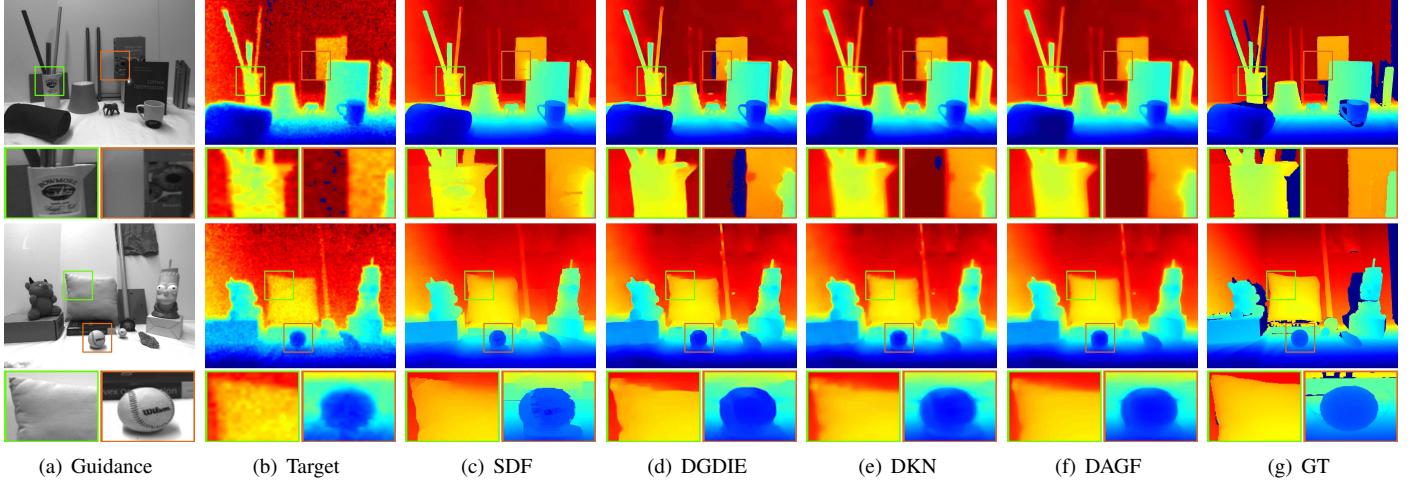


Fig. 7. Visual comparison of realistic depth map super-resolution on two examples (*books* and *devil*) from ToFMark [28] dataset: (a): Guidance image, (b): Target image, (c): SDF [8], (d): DGDIE [37], (e): DKN [7], (f): DAGF, (e): Ground truth. Please enlarge the PDF for more details.

TABLE III

QUANTITATIVE COMPARISON FOR JOINT DEPTH IMAGE SUPER-RESOLUTION AND DENOISING ON FOUR STANDARD RGB/D DATASETS IN TERMS OF AVERAGE RMSE VALUES. FOLLOWING THE EXPERIMENTAL SETTING OF [7], [26]), WE CALCULATE THE AVERAGE RMSE VALUES IN CENTIMETER FOR NYU v2 [25] DATASET. FOR OTHER DATASETS, WE COMPUTE THE RMSE VALUES BY SCALING THE DEPTH VALUE TO THE RANGE [0, 255]. THE BEST PERFORMANCE FOR EACH CASE ARE HIGHLIGHTED IN **BOLDFACE** WHILE THE SECOND BEST ONES ARE UNDERSCORED. FOR RMSE METRIC, THE LOWER VALUES MEAN THE BETTER PERFORMANCE.

Datasets	Middlebury			Lu			NYU v2			Sintel		
	Method	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×
DGF [29]	2.70	4.13	6.38	4.06	5.85	8.39	6.52	9.23	13.00	6.94	9.03	12.05
DJF [26]	1.80	2.99	5.16	1.85	3.13	5.39	3.74	5.95	9.61	4.88	6.93	10.05
DMSG [30]	1.79	2.69	4.75	1.88	<u>2.79</u>	4.84	3.60	5.31	<u>8.07</u>	4.74	6.36	<u>8.72</u>
DJFR [5]	1.86	3.07	5.27	1.91	3.21	5.51	4.01	6.21	9.90	5.10	7.12	10.23
DSRN [31]	1.84	2.99	4.70	1.97	2.98	5.94	4.36	6.31	9.75	5.49	7.21	9.80
PAC [6]	1.81	2.94	5.08	1.93	3.44	6.18	4.23	6.24	9.54	5.40	7.32	9.89
DKN [7]	<u>1.76</u>	<u>2.68</u>	<u>4.55</u>	<u>1.81</u>	2.82	<u>4.81</u>	<u>3.39</u>	<u>5.24</u>	8.41	<u>4.51</u>	<u>6.25</u>	9.20
DAGF (Ours)	<b>1.72</b>	<b>2.61</b>	<b>4.24</b>	<b>1.74</b>	<b>2.72</b>	<b>4.51</b>	<b>3.25</b>	<b>5.01</b>	<b>7.54</b>	<b>4.42</b>	<b>6.09</b>	<b>8.25</b>

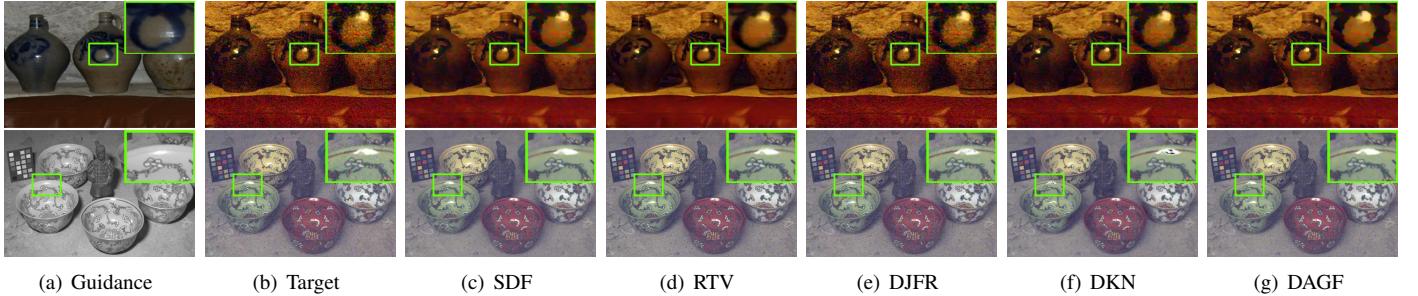


Fig. 8. Visual comparison of cross-modality image restoration. Top: flash guided non-flash image denoising. Bottom: NIR guided color image denoising. (a): Guidance image, (b): Target image, (c): SDF [8], (d): RTV [13], (e): DJFR [5], (f): DKN [7], (g): DAGF. Please enlarge the PDF for more details.

DKN [7]) are tested with the same setting as ours. Among the compared methods, SDF [8] and RTV [13] are specially designed for this task. As can be seen from Fig. 8, DJFR [5] cannot remove noise, and the results of DKN [7] suffer from halo artifacts. On the contrary, the proposed DAGF can produce more convincing results with less artifact. The method of RTV [13] which is specially designed for this task, obtains

the best performance.

**Realistic Depth Image Super-resolution.** To further evaluate the robustness of the proposed method, we conduct experiments on ToFmark dataset [39], which include real ToF sensor data and thus have complicated multi-modality degradation. Following the experimental protocol of DGDIE [37], we first perform image completion on the acquired depth images

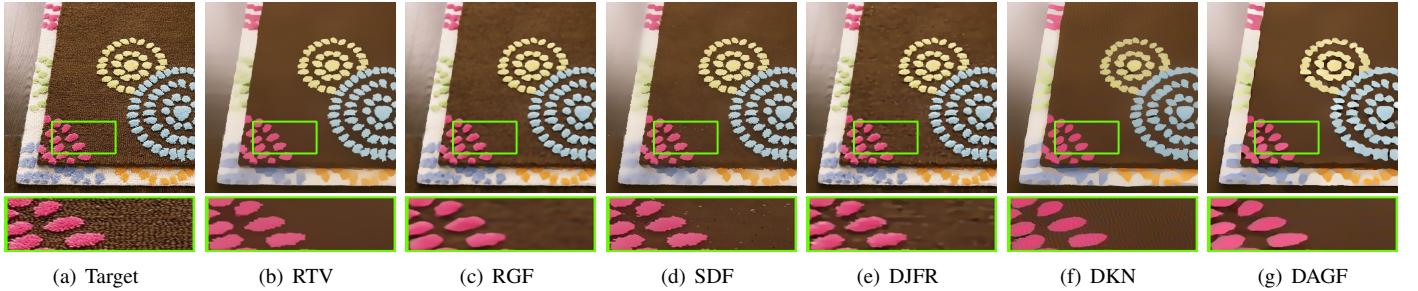


Fig. 9. Visual comparisons of texture remove results. (a): Target image, (b): RTV [13], (c): RGF [9], (d): SDF [8], (e): DJFR [5], (f): DKN [7], (g): DAGF. Please enlarge the PDF for more details.

TABLE IV  
QUANTITATIVE COMPARISON FOR REALISTIC DEPTH IMAGE  
SUPER-RESOLUTION IN TERMS OF RMSE VALUES ON THE TOFMARK [28]  
DATASET. THE BEST PERFORMANCE FOR EACH CASE ARE HIGHLIGHTED  
IN **BOLDFACE** WHILE THE SECOND ONES ARE UNDERSCORED.

Methods	Books	Devil	Shark
Bilinear	17.10	20.17	18.66
JBu [2]	16.03	18.79	27.57
GF [1]	15.74	18.21	27.04
TGV [28]	12.36	15.29	14.68
SDF [8]	12.66	14.33	10.68
Yang [38]	12.25	14.71	13.83
DGDIE [37]	12.32	14.06	9.66
DKN [7]	<u>11.81</u>	<u>13.54</u>	<u>9.11</u>
DAGF (Ours)	<b>11.80</b>	<b>13.47</b>	<b>9.07</b>

TABLE V  
QUANTITATIVE COMPARISON FOR SEMANTIC SEGMENTATION IN TERMS  
OF AVERAGE IOU ON THE VALIDATION SET OF PASCAL VOC 2012. THE  
BEST PERFORMANCE FOR EACH CASE ARE HIGHLIGHTED IN **BOLDFACE**  
WHILE THE SECOND ONES ARE UNDERSCORED.

Methods	Mean IoU
DeepLab-V2 [40]	70.69
DenseCRF [41]	71.98
DGF [29]	72.96
DJFR [5]	73.30
FDKN [7]	<u>73.60</u>
DAGF (Ours)	<b>73.76</b>

and then send them to our model ( $4 \times$  super-resolution and denoising) trained on NYU v2 dataset [25] to obtain the final results. We compare our method with a recently proposed deep learning-based method (e.g. DKN [7]) and some traditional methods (e.g. TGV [28], SDF [8], DGDIE [37]). As shown in Table IV, our method constantly obtains the best objective results for the three test images. Fig. 7 presents visual comparison results for two images (*books* and *devil*). From these figures, it is easy to observe that the results of SDF [8] suffer from texture-copying artifacts. The results of DKN [7] are smooth and blurred, since DKN generates filter kernels without considering the inconsistency between color and depth image. The results of DGDIE [37] are clear but

they deviate from the ground truth. By comparison, the results of the proposed method are sharper and much closer to the ground truth, especially at the boundary regions.

### C. Texture Removal

Texture removal is the task of extracting semantically meaningful structures from textured surfaces. For this task, we use the textured image itself as the guidance, and apply our model trained for depth image denoising iteratively to remove small-scale textures. We compare our method with RTV [13], RGF [9], SDF [8], DJFR [5] and DKN [7]. For deep learning-based methods, we follow DKN [7], set the number of iterations as 4, and for other methods we carefully fine-tune the parameters to provide the best results. The visual comparison are presented in Fig. 9. Obviously, our method outperforms other compared methods, and it can painlessly remove small-scale textures as well as preserve the global color variation and main edges.

### D. Semantic Segmentation

Semantic segmentation is a fundamental computer vision task, which aims at assigning pre-defined labels to each pixel of an image. In DGF [29], the author proposed to use guided image filtering as a layer to replace the time-consuming fully connected conditional random filed (CFR) [41] for semantic segmentation. We demonstrate that the proposed DAGF can be applied to this problem. Following DGF [29], we plug the proposed model into DeepLab-v2 [40] and train the whole network in an end-to-end manner, and thus the offline post-processing of CRFs can be avoided. We utilize the Pascal VOC 2012 dataset [42] in our experiment, which contains 1264, 1229 and 1456 images for training, validation and testing, respectively. Similar to DGF [29], we augment the training set with the annotations provided by [43], resulting in 10582 images. The 1449 images in the validation set are employed to evaluated the proposed method.

We use the mean intersection-over-union (IoU) score as evaluation metric and report the quantitative results for the validation set of Pacal VOC dataset [42] in Table V. The baseline denotes DeepLab-v2 [40] without CRF. As can be seen from this table, our method outperforms the second best model DKN [7] by 0.16% mIoU and other models by a large margin. We visualize the segmentation results among our

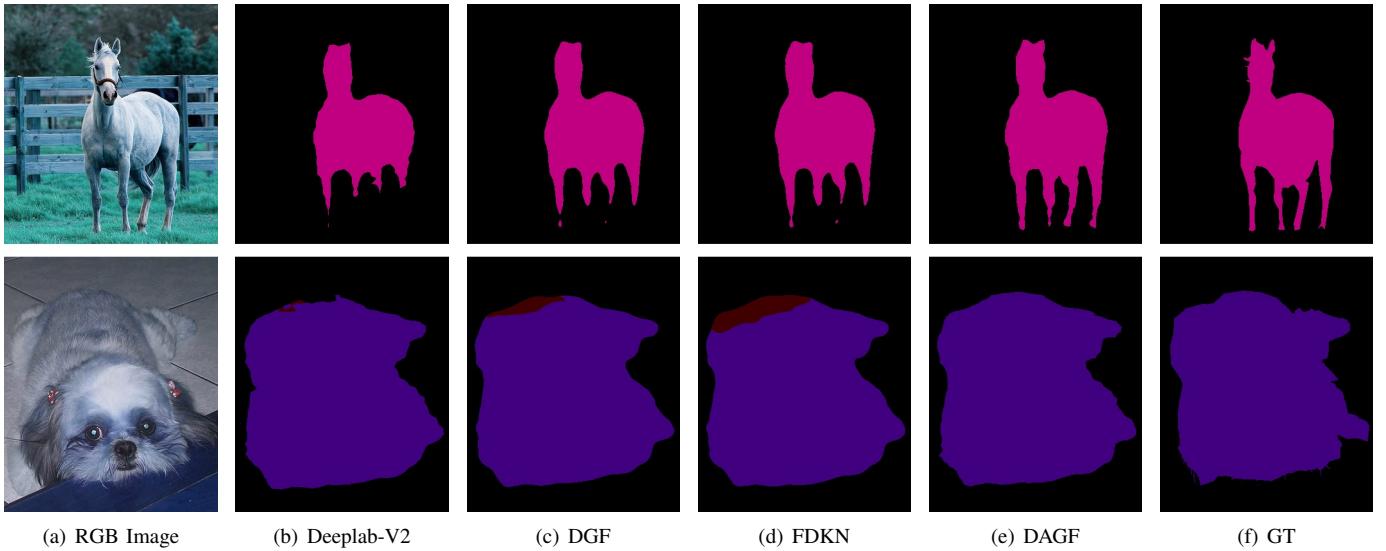


Fig. 10. Visual comparison of semantic segmentation on the validation set of Pascal VOC 2012 dataset [42]. (a): RGB image, (b): Deeplab-V2 [40], (c): DGF [29], (d): FDKN [7], (e): DAGF, (f): ground truth image. Please enlarge the PDF for more details.

TABLE VI

**ABLATION STUDY. QUANTITATIVE COMPARISON OF DIFFERENT SIZE OF KERNEL ( $k \times k$ ) AND THE NUMBER OF PYRAMID LEVEL ( $m$ ).**

Kernel Size	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$1 \times 1$	12.8	11.72	11.41	11.26
$3 \times 3$	8.97	8.12	7.81	7.75
$5 \times 5$	8.60	8.02	7.73	7.98
$7 \times 7$	8.67	7.94	7.78	7.99

We report average RMSE values of the last 449 image pairs in NYU v2 dataset [25].

method and other compared methods in Fig 10, from which we can see that our method is capable of generating results with accurate and complete object boundaries.

## V. ABLATION STUDY

In this section, we first present the hyper-parameters setting in our model, and then conduct a series of ablation experiments to investigate the effectiveness of our main contributions, e.g., attentional kernel learning module (mentioned in Sect. III-B), multi-scale fusion (mentioned in Sect. III-C) with deep supervision (mentioned in Sect. III-D) and boundary-aware loss (mentioned in Sect. III-D). In this study, we train different variants of our model on the commonly used NYU v2 dataset (Silberman et al., [25]) with  $16 \times$  nearest-neighbour downsampling and evaluate the performance of them on four benchmark datasets. The experimental settings are the same as Sect. IV-A.

### A. Hyper-parameters setting

For network hyper-parameters setting, we investigate the influence the size  $k \times k$  of learned filter kernels in our kernel generation sub-network (e.g.,  $W_0, W_1, W_2$  in Fig. 2) and the number of pyramid level  $m$  in our model for multi-modality feature extraction to the final performance. Enlarging

$k$  or  $m$  can increase the receptive field of our model but at the expense of higher computational complexity. To seek an appropriate trade-off between complexity and performance, we conduct experiments on the task of depth map super-resolution with different  $k$  and  $m$ , and the results are summarized in Table VI. From this table, we can see that the reconstruction performance is significantly improved when the number of pyramid levels  $m$  increased from 1 to 3. However, when  $m$  is too large, e.g.,  $m = 4$ , the improvements are small or even worse. We can draw the same conclusion for the size of filter kernels  $k \times k$ . The possible reason for this phenomenon is that the receptive filed is enough for this task when  $m = 3, k = 3$  and larger  $m$  or  $k$  will burden the optimization process of network. Therefore, we set  $m = 3, k = 3$  in our experiments.

### B. Ablation Experiments

As shown in Fig. 2, our model consists of two part: kernel generation sub-network and multi-scale guided image filtering sub-network. For kernel generation sub-network, we propose to generate dual sets of kernels from the guidance and target images, and employ a tiny network to learn a weight map to adaptively combine the two sets of kernels. For guided image filtering sub-network, we progressively filter the target image with the learned multi-scale kernels. In order to fully integrate the intermediate filtered results, we propose a multi-scale feature fusion strategy and a multi-stage loss. To encourage our model to give more emphasis to the high-frequency and to generate visual pleasing results, we propose to train our model with hybird loss functions, e.g., pixel-wise loss  $\mathcal{L}_1$ , multi-scale loss  $\mathcal{L}_{ms}$ , and boundary-aware loss  $\mathcal{L}_{ba}$ . To analyze the contribution of each component of our model, we implement seven variants of our model:

- **Model1**, which takes (target, target) as inputs for kernel generation, and is trained with  $\mathcal{L}_1$  loss.

TABLE VII

**ABLATION STUDY.** QUANTITATIVE COMPARISON OF DIFFERENT COMPONENTS FOR  $16 \times$  DEPTH IMAGE SUPER-RESOLUTION. WE CHOSE RMSE AS THE EVALUATION METRIC, AND THE LOWER VALUES INDICATE BETTER PERFORMANCE. MODEL7 IS OUR FINAL MODEL (DAGF).

Model	Kernel Generation		Kernel Combination			$\mathcal{L}_{ms}$	$\mathcal{L}_{ba}$	Middlebury	Lu	NYU v2	Sintel	Average
	Target	Guidance	MUL	SUM	AKL							
Model1	✓							7.08	7.87	11.99	13.67	10.15
Model2		✓						5.68	7.19	9.09	11.82	8.45
Model3	✓	✓	✓					5.47	6.84	9.07	11.74	8.28
Model4	✓	✓		✓				5.36	6.90	8.99	11.65	8.23
Model5	✓	✓			✓			5.06	6.57	8.49	11.18	7.82
Model6	✓	✓			✓	✓		4.88	6.19	7.92	10.89	7.47
Model7	✓	✓			✓	✓	✓	<b>4.75</b>	<b>6.16</b>	<b>7.81</b>	<b>10.64</b>	<b>7.34</b>

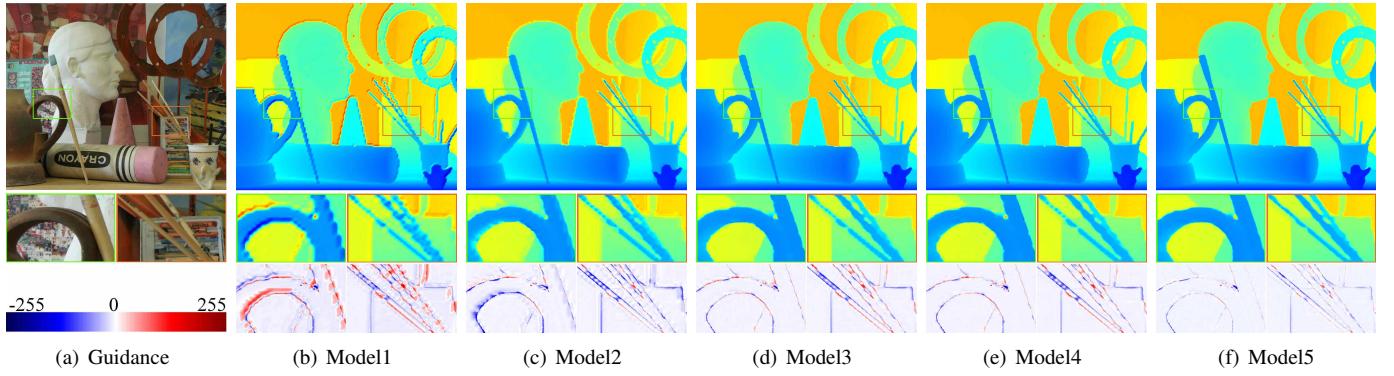


Fig. 11. **Ablation Study.** Visual comparison of an example without and with the proposed attentional kernel learning module (AKL) for depth image super-resolution. The first row shows the super-resolved depth images and the last row shows the error map ( $I^h - I^{out}$ ). Please enlarge the PDF for more details.

- **Model2**, which takes (guidance, guidance) as inputs for kernel generation, and is trained with  $\mathcal{L}_1$  loss.
- **Model3**, which takes (target, guidance) as inputs for kernel generation, and uses element-wise multiplication to combine the generated two sets of kernels, and is trained with  $\mathcal{L}_1$  loss.
- **Model4**, which takes (target, guidance) as inputs for kernel generation, and uses element-wise summation to combine the generated two sets of kernels, and is trained with  $\mathcal{L}_1$  loss.
- **Model5**, which takes (target, guidance) as inputs for kernel generation, and uses the learned weight map to adaptively combine the generated two sets of kernels, and is trained with  $\mathcal{L}_1$  loss.
- **Model6**, which is Model5 but trained with  $\mathcal{L}_1$  loss and  $\mathcal{L}_{ms}$  loss.
- **Model7**: which is Model5 but trained with  $\mathcal{L}_1$  loss,  $\mathcal{L}_{ms}$  loss and  $\mathcal{L}_{ba}$  loss. This is our full model.

It is noteworthy that we adjust the number of convolutional layers in multi-scale guided image filtering sub-network to guarantee that each variant could have roughly the same number of parameters with our final model. The quantitative results are shown in Table V, from which we can see that the full model (Model7) achieves the best reconstruction performance across four testing datasets when compared with the ablated models, and every component proposed in our model can

boost the network performance significantly. In the following, we will give a detailed analysis of each component in our method. **Effectiveness of Attentional Kernel Learning (AKL):** In this paper, we propose to use AKL to generate filter kernels for guided image filtering. Specifically, it first generates dual sets of kernels by using the extracted guidance and target features respectively, and then adaptively combines the generated kernels by the learned attention maps. To demonstrate the effectiveness of AKL, we implement several variants (e.g., different inputs for kernel construction and different kernel fusion strategies) of the proposed method, including Model1–Model5. The quantitative results on the four testing datasets are reported in Table VII. As can be seen from this table, Model1 generates kernels from target image only, thus the reconstruction accuracy is relatively low. With the assistance of guidance image, Model2 obtains a significant improvement compared with Model1, which implies that the guidance information is helpful for filter kernel generation. However, the guidance images are not always reliable, such as color images captured in bad weather or low-light conditions. In view of this, Model3 and Model4 generate dual sets of kernels from the guidance and target images, respectively, and the difference between the two models is the strategy of kernel combination. As shown in Table VII, Model3 and Model4 can further improve the accuracy over Model2 (The average RMSE is dropped from 8.45 to 8.28 and 8.23),

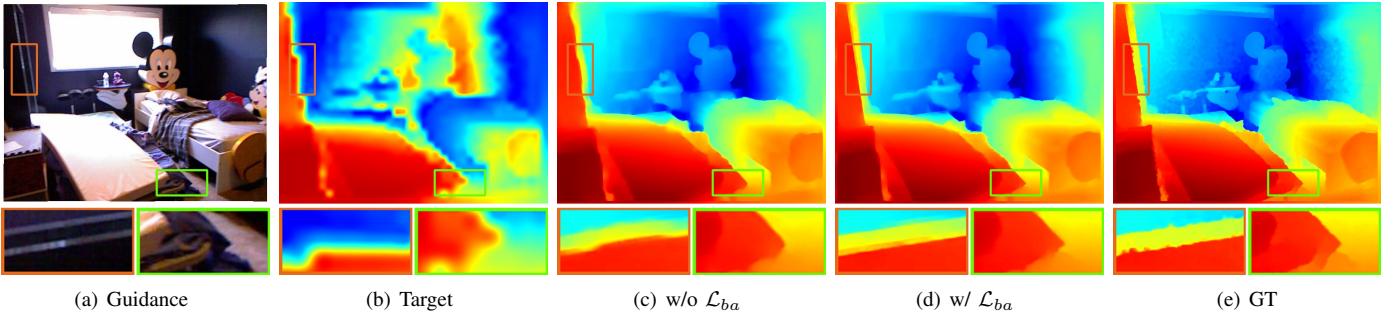


Fig. 12. **Ablation Study.** Visual comparison of an example without and with the proposed boundary-aware loss for depth image super-resolution. Please enlarge the PDF for more details.

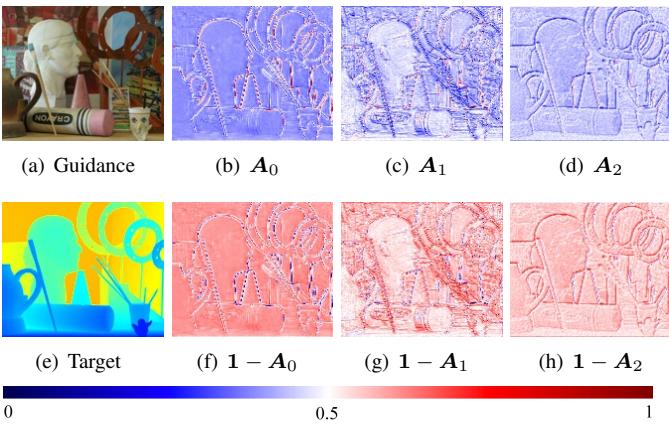


Fig. 13. **Ablation Study.** Visualization of the learned multi-scale attention maps for kernel combination. We resize the attention map to the same size for better visualization. Please enlarge the PDF for more details.

which indicates that constructing kernels from both target and guidance images enjoys some benefits over using only the guidance. Nevertheless, using element-wise multiplication or summation to combine the generated kernels would limit the capacity of the network, since they ignore the inconsistency between guidance and target images. To solve this problem, we first learn an attention map, and then utilize the attention map to selectively combine the dual kernels as in Eq. 14. As depicted in Table VII, equipped with AKL, compared with Model3, Model5 reduces the average RMSE from 8.32 to 7.82.

To visually show the effect of AKL, we present in Fig. 11 the super-resolved depth images (first row) and error maps (last row) with different configurations. The error map is obtained by  $I^h - I^{out}$ . As shown in Fig. 11, the result of Model1 is blur and lack of high-frequency details. For the error map of Model1, most of the values at the image boundaries are positive, which means that the boundaries generated by Model1 are weaker than the ones of ground truth. The reason is that the kernels generated from the target image only cannot produce the high-frequency details which are lost by the image degradation process. On the contrary, most values in the error map of Model2 are negative, although the depth boundaries are enhanced, the texture-copying artifacts seriously influence the super-resolved depth maps. Thanks to the proposed attentional

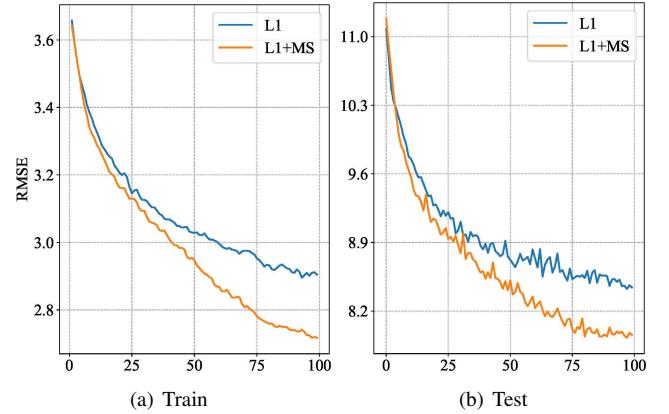


Fig. 14. **Ablation Study.** Training and testing RMSE values on NYU v2 dataset (Silberman et al., [25]) for  $16 \times$  depth image super-resolution. MS denotes the proposed multi-stage loss  $\mathcal{L}_{ms}$ .

kernel learning theme that constructs kernels by fully integrating complementary information contained in both guidance and target images, the visual effect and reconstruction accuracy of Model5 are substantially improved.

Moreover, we visualize the attention map in Fig. 13 to further validate the capability of the proposed AKL, from which we can see that the kernels generated from the target and guidance images are both important for the task of guided filtering as most of the pixel values in the attention maps are in the range of [0.4, 0.6]. In addition, as shown in the first row of Fig. 13, the structure regions are lighter than texture regions, and this indicates that our model can adaptively select relevant information from the guidance image while avoiding texture over-transfer issues.

**Effectiveness of Multi-scale Fusion and Deep Supervision:** In this paper, we propose a multi-scale framework for guided image filtering. Specifically, in order to obtain both high-level structure information and low-level details, we propose to fuse multi-level filtered outputs. Moreover, a multi-stage loss is introduced to enforce the intermediate results to be close to the ground-truth target image. The quantitative results are illustrated in Table VII. As expected, Model6 trained with a hybrid loss of  $\mathcal{L}_1$  and  $\mathcal{L}_{ms}$  further improves the reconstruction accuracy. Fig. 14 further shows the train (left) and test (right) RMSE plots. We observe that the multi-stage loss ( $\mathcal{L}_{ms}$ ) is

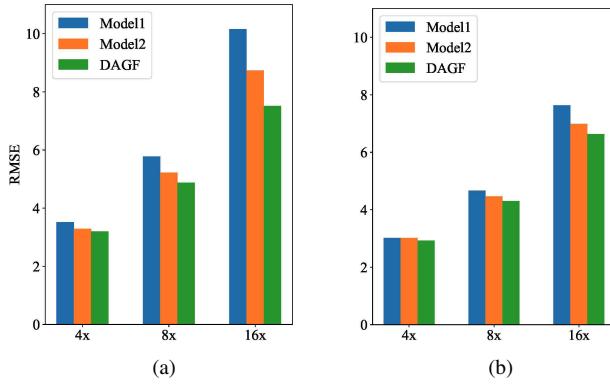


Fig. 15. **Ablation Study.** Average RMSE values for depth image super-resolution. The low-resolution depth image are obtained by (a): Nearest-neighbour downsampling, (b): Bicubic downsampling and Gaussian noise.

able to accelerate convergence velocity and produce results with lower RMSE values.

**Effectiveness of Boundary-aware Loss:** To encourage the network to pay more attention to high-frequency information, we propose to train our model with boundary-aware loss ( $\mathcal{L}_{ba}$ ). Table VII demonstrates that  $\mathcal{L}_{ba}$  loss is helpful in improving the reconstruction accuracy (Model7). Fig. 12 presents an example visual comparison with and without the  $\mathcal{L}_{ba}$  loss. Obviously,  $\mathcal{L}_{ba}$  improves the visual quality further, yielding more precise edges. The boundaries on the doorframe and corner of the mattress are sharper and clearer, which verifies the effectiveness of the proposed boundary-aware loss.

**Effectiveness of Guidance Branch:** The general principle of guide image filtering is that we can transfer the valuable structure information contained in guidance image to the target image. Recently, various approaches have been proposed for guided image filtering. Nevertheless, most of them focus on designing advanced algorithm for efficiently transferring structures from the guidance to the target image, and the contributions of guidance images under different conditions are rarely explored. Here, we conduct experiments on several applications of guided image super-resolution, e.g., depth image super-resolution (nearest-neighbour downsampling) and noisy depth super-resolution (Bicubic downsampling and Gaussian noise). As shown in Fig. 15, we evaluate the role of guidance image and compute the average RMSE value for each upsampling factor. Model1 takes the target image as input for kernel generation, while Model2 takes the guidance images as input for kernel generation. The results show that the guidance image can provide significant assistance for the  $8\times$  and  $16\times$  cases, and the model (DAGF) equipped with the proposed AKL can further improve the performance. However, for the  $4\times$  case, which is easy to recover, the guidance information has a negligible effect. The main reason is that the target image is not severely damaged by downsampling degradation, therefore, the target image can be easily recovered by Model1. For the more difficult cases ( $8\times$  and  $16\times$ ), the target image is badly polluted, the guidance image would play an important role in the reconstruction process.

**Performance vs. Complexity Analysis:** In Fig. 16, we

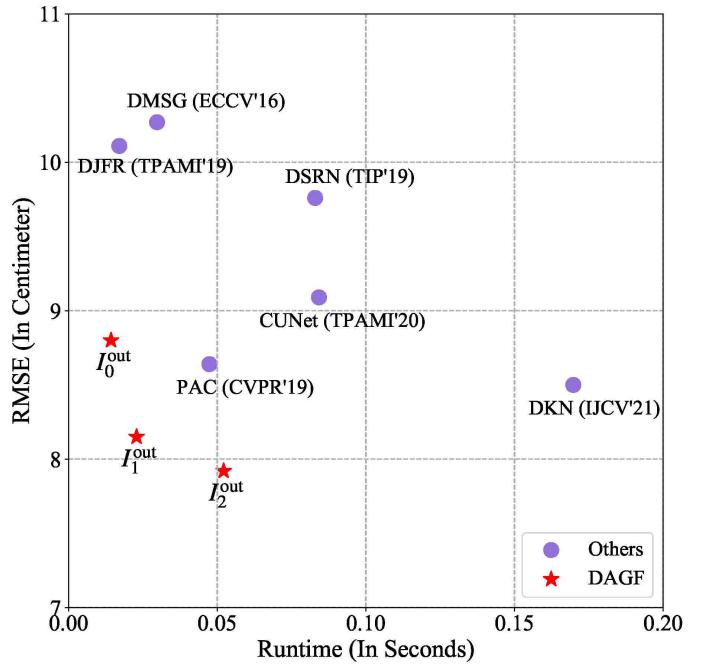


Fig. 16. Average runtime (in seconds) and root mean square error (RMSE) comparison for  $16\times$  depth image super-resolution on NYU v2 dataset [25]. All the runtimes are evaluated on the same NVIDIA 1080Ti GPU with depth image size  $480 \times 640$ .

compare the running time among our method and other comparison methods on NYU v2 (Silberman et al., [25]) for  $16\times$  depth image super-resolution. For fair comparison, all the running times are obtained on the same machine by one NVIDIA 1080Ti GPU. As shown in Fig. 2, our method produces multiple results  $I_0^{\text{out}}$ ,  $I_1^{\text{out}}$ ,  $I_2^{\text{out}}$ , and we first resize them to the same resolution as the ground truth target image by a simple bilinear interpolation method, and then calculate the RMSE values. As illustrated in Fig. 16, the final result  $I_2^{\text{out}}$  achieves the best RMSE result than DKN (Kim et al., [7]) and DSRN (Guo et al., [31]) but needs less time. The time cost for  $I_0^{\text{out}}$  is the least, and the performance of  $I_0^{\text{out}}$  is comparable to other methods. If the purpose is to achieve the performance as best as possible, we can increase the level of pyramid, otherwise reduce the level of pyramid. Overall, our method can achieve a better trade-off between the reconstruction performance and computational complexity.

## VI. CONCLUSION

In this paper, we present an effective network architecture for guided image filtering, which can automatically select and transfer important structures from the guidance to the target image. Specifically, an attentional kernel learning module (AKL) is proposed to generate dual sets of filter kernels from the guidance and target images, respectively, and then adaptively combine the learned kernels in a learning manner. Furthermore, a multi-scale guided image filtering framework is introduced, which takes the generated kernels and target image as inputs and progressively filters the target image in a coarse-to-fine manner. Moreover, to fully explore the intermediate results in the coarse-to-fine process, we propose

a multi-scale fusion with deep supervision to regularize and combine multiple filtering results. Finally, boundary-aware loss is introduced to enhance the high-frequency details of guided filtering. Experimental results on various guided image filtering applications show the superiority and flexibility of the proposed model and the ablation experiments demonstrate the effectiveness of each component in our method.

## ACKNOWLEDGMENT

## REFERENCES

- [1] 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.
- [2] 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.
- [3] X. Shen, C. Zhou, L. Xu, and J. Jia, “Mutual-structure for joint filtering,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3406–3414.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] B. Kim, J. Ponce, and B. Ham, “Deformable kernel networks for joint image filtering,” *International Journal of Computer Vision*, pp. 1–22, 2020.
- [8] B. Ham, M. Cho, and J. Ponce, “Robust guided image filtering using nonconvex potentials,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 192–207, 2018.
- [9] Z. Qi, X. Shen, X. Li, and J. Jia, “Rolling guidance filter,” in *Proceedings of the European conference on computer vision*, 2014, pp. 815–830.
- [10] B. Stimpel, C. Syben, F. Schirrmacher, P. Hoelter, A. Dörfler, and A. Maier, “Multi-modal super-resolution with deep guided filtering,” in *Bildverarbeitung für die Medizin 2019*, Wiesbaden, 2019, pp. 110–115.
- [11] L. Xu, Q. Yan, Y. Xia, and J. Jia, “Structure extraction from texture via relative total variation,” *ACM transactions on graphics*, vol. 31, no. 6, pp. 1–10, 2012.
- [12] L. Karacan, E. Erdem, and A. Erdem, “Structure-preserving image smoothing via region covariances,” *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 1–11, 2013.
- [13] L. Xu, Q. Yan, Y. Xia, and J. Jia, “Structure extraction from texture via relative total variation,” *ACM Trans. Graph.*, vol. 31, no. 6, 2012.
- [14] J. T. Barron and B. Poole, “The fast bilateral solver,” in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 617–632.
- [15] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *Proceedings of the Sixth International Conference on Computer Vision*, 1998, pp. 839–846.
- [16] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, “Dynamic filter networks,” in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 667–675.
- [17] R. D. Lutio, S. D’aronco, J. D. Wegner, and K. Schindler, “Guided super-resolution as pixel-to-pixel transformation,” in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8828–8836.
- [18] J. Kwak and D. Son, “Fractal 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.
- [19] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
- [20] 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.
- [21] 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 International Conference on Computer Vision*, 2015, pp. 1026–1034.
- [22] 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.
- [23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv: Learning*, 2014.
- [24] 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,” in *NIPS-W*, 2017.
- [25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *Proceedings of the 12th European conference on Computer Vision - Volume Part V*, 2012, pp. 746–760.
- [26] 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.
- [27] J. Diebel and S. Thrun, “An application of markov random fields to range sensing.” *Advances in Neural Information Processing Systems*, pp. 291–298, 2005.
- [28] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruechter, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 993–1000.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 611–625.
- [33] 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.
- [34] D. Scharstein and C. Pal, “Learning conditional random fields for stereo,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [35] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, “Saliency detection via graph-based manifold ranking,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.
- [36] H. Hirschmüller and D. Scharstein, “Evaluation of cost functions for stereo matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [37] 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.
- [38] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, “Color-guided depth recovery from rgbd data using an adaptive autoregressive model,” *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3443–3458, 2015.
- [39] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruechter, and H. Bischof, “Image guided depth upsampling using anisotropic total generalized variation,” in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 2013.
- [40] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [41] P. Krhenbühl and V. Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” *Curran Associates Inc.*, 2012.
- [42] M. Everingham, S. Eslami, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [43] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 991–998.

## REFERENCES

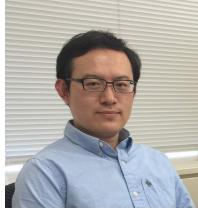
- [1] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.
- [2] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *Acm Transactions on Graphics*, vol. 26, no. 3, pp. 96.1–96.4, 2007.
- [3] X. Shen, C. Zhou, L. Xu, and J. Jia, "Mutual-structure for joint filtering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3406–3414.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, pp. 1–22, 2020.
- [8] B. Ham, M. Cho, and J. Ponce, "Robust guided image filtering using nonconvex potentials," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 192–207, 2018.
- [9] Z. Qi, X. Shen, X. Li, and J. Jia, "Rolling guidance filter," in *Proceedings of the European conference on computer vision*, 2014, pp. 815–830.
- [10] B. Stimpel, C. Syben, F. Schirrmacher, P. Hoelter, A. Dörfler, and A. Maier, "Multi-modal super-resolution with deep guided filtering," in *Bildverarbeitung für die Medizin 2019*, Wiesbaden, 2019, pp. 110–115.
- [11] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM transactions on graphics*, vol. 31, no. 6, pp. 1–10, 2012.
- [12] L. Karacan, E. Erdem, and A. Erdem, "Structure-preserving image smoothing via region covariances," *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 1–11, 2013.
- [13] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Trans. Graph.*, vol. 31, no. 6, 2012.
- [14] J. T. Barron and B. Poole, "The fast bilateral solver," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 617–632.
- [15] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the Sixth International Conference on Computer Vision*, 1998, pp. 839–846.
- [16] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 667–675.
- [17] R. D. Lutio, S. D'armonco, J. D. Wegner, and K. Schindler, "Guided super-resolution as pixel-to-pixel transformation," in *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8828–8836.
- [18] J. Kwak and D. Son, "Fractal 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.
- [19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
- [20] 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.
- [21] 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 International Conference on Computer Vision*, 2015, pp. 1026–1034.
- [22] 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.
- [23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv: Learning*, 2014.
- [24] 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," in *NIPS-W*, 2017.
- [25] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proceedings of the 12th European conference on Computer Vision - Volume Part V*, 2012, pp. 746–760.
- [26] 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.
- [27] J. Diebel and S. Thrun, "An application of markov random fields to range sensing," *Advances in Neural Information Processing Systems*, pp. 291–298, 2005.
- [28] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 993–1000.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 611–625.
- [33] 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.
- [34] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [35] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.
- [36] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [37] 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.
- [38] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from rgbd data using an adaptive autoregressive model," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3443–3458, 2015.
- [39] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 2013.
- [40] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [41] P. Krhenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Curran Associates Inc.*, 2012.
- [42] M. Everingham, S. Eslami, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [43] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 991–998.



**Zhiwei Zhong** received the B.S. degree in computer science from the Heilongjiang University, Harbin, China, in 2017. He is currently pursing the Ph.D. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, China. His research interests include image processing, computer vision and deep learning.



**Xiangyang Ji** received the B.S. degree in materials science and the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He joined Tsinghua University, Beijing, in 2008, where he is currently a Professor with the Department of Automation, School of Information Science and Technology. He has authored over 100 referred conference and journal papers. His current research interests include signal processing, image/video compressing, and intelligent imaging.



**Xianming Liu** received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2006, 2008, and 2012, respectively. In 2011, he spent half a year at the Department of Electrical and Computer Engineering, McMaster University, Canada, as a Visiting Student, where he was a Post-Doctoral Fellow from 2012 to 2013. He was a Project Researcher with the National Institute of Informatics (NII), Tokyo, Japan, from 2014 to 2017. He is currently a Professor with the School of Computer Science and Technology, HIT. He has published over 50 international conference and journal publications, including top IEEE journals, such as T-IP, T-CSVT, T-IFS, and T-MM, and top conferences, such as ICML, CVPR, IJCAI. He was a receipt of the IEEE ICME 2016 Best Student Paper Award.



**Junjun Jiang** received the B.S. degree from the Department of Mathematics, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer, Wuhan University, Wuhan, China, in 2014. From 2015 to 2018, he was an Associate Professor at China University of Geosciences, Wuhan. Since 2016, he has been a Project Researcher with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. He won the Finalist of the World's FIRST 10K Best Paper Award at ICME 2017, and the Best Student Paper Runner-up Award at MMM 2015. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and 2015 ACM Wuhan Doctoral Dissertation Award. His research interests include image processing and computer vision.



**Debin Zhao** received the B.S., M.S., and Ph.D. degrees in computer science from Harbin Institute of Technology, China in 1985, 1988, and 1998, respectively. He is now a professor in the Department of Computer Science, Harbin Institute of Technology. He has published over 200 technical articles in refereed journals and conference proceedings in the areas of image and video coding, video processing, video streaming and transmission, and pattern recognition.