Deep Attentional Guided Image Filtering

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

***Abstract*— Guided filter is a fundamental tool in computer vision and computer graphics, which aims to transfer structure information from the guide image to the target image. Most existing methods construct filter kernels from the guidance itself without considering the mutual dependency between the guidance and the target. However, since there typically exist significantly different edges in two images, simply transferring all structural information from the guide 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 and then adaptively combine them by modeling the pixelwise dependency between the two images. Meanwhile, we propose a multiscale guided image filtering module to progressively generate the filtering result with the constructed kernels in a coarse-to-fine manner. Correspondingly, a multiscale 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 (SR), cross-modality restoration, and semantic segmentation. Moreover, our scheme achieved the first place in the real depth map SR challenge held in ACM ICMR 2021. The codes can be found at https://github.com/zhwzhong/DAGF.**

***Index Terms*— Attentional kernel learning (AKL), cross-modality restoration, dual regression, guided filter (GF), guided super-resolution (SR).**

I. Introduction

GUIDED 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 to handle visual signals in various domains and modalities, where one modal signal serves as guidance to improve the quality of the other [1], [7]. It has been a useful tool for many computer vision tasks, such as depth map super-resolution (SR) [8], [9], [10], [11], [12], [13], [14], [15], depth map completion [16], [17], cross-modality image restoration [18], [19], [20], [21], structure–texture separation [22], [23], and scene understanding [24], [25].

In the literature, GF has been extensively studied, ranging from bilateral filter [26] to emerging deep learning-based ones. The pioneer bilateral filter [26] constructs spatially varying kernels, where local image structures of the guidance image are explicitly involved in the filtering process through the photometric similarity. The guided image filtering scheme proposed in [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. 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 propose to use an optimization-based manner to find mutual structures for propagation while suppressing inconsistent ones. For example, Yang et al. [27] proposed a low-rank model with offset learning to reduce the distortions caused by the noise and outliers of the input images. However, it is challenging to select reference structures and propagate them properly with handcrafted objective functions. In recent years, learning-based approaches for GF design have become 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 [28], a dynamic filter network (DFN) is proposed where pixelwise filters are generated dynamically using a separate subnetwork conditioned on the guidance. Unlike DFN, Su et al. [5] adapted a standard spatially invariant kernel to each pixel by multiplying it with a spatially varying filter. Although with increased flexibility due to their adaptive nature, they still suffer from the same drawback as in [1] and [26] that only guidance information is considered in a filter design.

Some methods attempt to exploit the target and the guidance information together. For instance, Li et al. [4] proposed to use two subnetworks to extract features from the target and the guidance, and then, the extracted features are concatenated and sent to the fusion network for structure transferring. Zhao et al. [17] proposed a recurrent distance transform pooling module to enlarge the receptive field for sparse data. Moreover, they propose an error correction mechanism to prevent the spread of wrong input values. Zhang et al. [29] proposed a multitask GAN for depth completion and semantic segmentation. In their mode, the reconstructed depth and generated semantic image serve as guidance for each other to improve the overall performance. Yan et al. [30] proposed a joint learning framework where they use depth completion as an auxiliary

Manuscript received 17 February 2022; revised 26 November 2022 and 7 February 2023; accepted 1 March 2023. Date of publication 31 March 2023; date of current version 4 September 2024. This work was supported by the National Natural Science Foundation of China under Grant 92270116 and Grant 62071155. *(Corresponding author: Xianming Liu.)*

Zhiwei Zhong, Xianming Liu, Junjun Jiang, and Debin Zhao are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: zhwzhong@hit.edu.cn; csxm@hit.edu.cn; jiangjunjun@hit.edu.cn; dbzhao@hit.edu.cn).

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

Digital Object Identifier 10.1109/TNNLS.2023.3253472

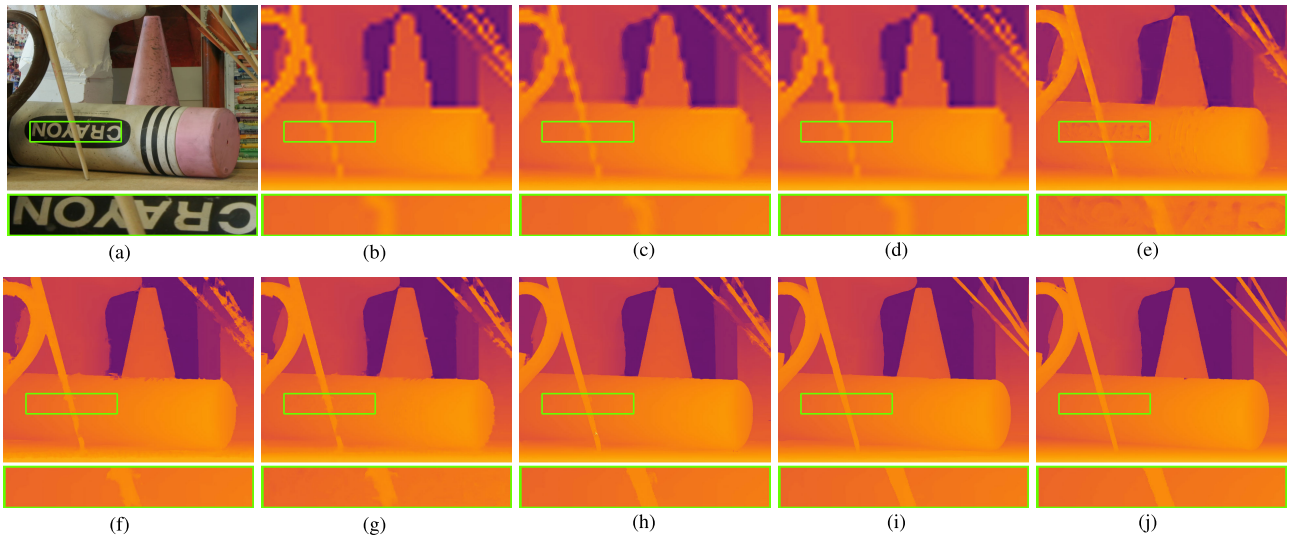


Fig. 1. Guided image filtering on an RGB/D image pair for 16× guided SR. Results of (c) and (d) suffer from edge blurring artifacts and results of (e) and (h) suffer from texture-copying artifacts. Our result produces much sharper edges. Please enlarge the PDF for more details. (a) Guidance image. (b)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

task to guide the depth SR process. Wang et al. [31] proposed a two-stage network that first learns the edge information of the target and then uses the edge map and the guidance to reconstruct the target image. Ye et al. [32] proposed a multibranch aggregation network to progressively recover the degraded depth. Dong et al. [33] assumed that the target can be linearly represented by the guidance and proposed to learn the spatially variant linear representation coefficients. Instead of regressing the filtering results directly from the network, Kim et al. [6] proposed to use spatially variant weighted averages, where the set of neighbors and the corresponding kernel weights are learned end-to-end. However, in the designed networks of these methods, simple concatenation or elementwise multiplication is exploited to combine multimodal information, which is not that effective. There is no mechanism to distinguish the contributions of the guidance and target to the final filtering result, and this may 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 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 [3] and guided SR [34], multimodal data have significantly different characteristics due to the difference of sensing principle, making the guidance not always reliable. In view of this, we argue that the purpose of GF should be twofold: 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 in which the guidance is no longer reliable, so we have to rely on the target itself for reconstruction. Most existing GF methods only concern structure transfer 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], and [26], 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 elementwise multiplication [4], [6]. "Dual regression" and "smart combination" bring the main motivations of our method.

Accordingly, in this article, we propose an effective deep attentional guided image filtering (DAGF) scheme, which constructs filter kernels by fully considering information from both guidance and target images. Specifically, an attentional kernel learning (AKL) module is proposed to generate dual sets of filter kernels from the guidance and target. Moreover, pixelwise contributions of the guidance and 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 SR methods. It can be found that our proposed method is capable of producing a high-resolution (HR) depth image with clear boundaries as well as avoiding texture-copying artifacts. The main contributions of the proposed method are summarized as follows.

1) We propose a general DAGF framework, in which the detailed network structures can be flexibly equipped according to the user's computational environment.
2) We propose an 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 pixelwise dependencies between the two images in a learning manner. Compared to existing kernel generation approaches, our method is more robust when there are inconsistent structures between the guidance and target.
3) We propose a multiscale guided filtering module, which generates the filtering result in a coarse-to-fine manner. Correspondingly, we propose a multiscale 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 GF framework that learns the multiscale kernels to filter the target image at different scales in the embedding space.

4) We evaluate the performance of our method on various computational photography and computer vision tasks, such as guided image SR (GSR), cross-modality image restoration, and semantic segmentation. The quantitative and qualitative results demonstrate the effectiveness and universality of our method.
5) 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 a fair comparison. The codes can be found at https://github.com/zhwzhong/DAGF.

The remainder of this article is organized as follows. Section II gives a brief introduction of the related works. Section III introduces the proposed method. Section IV provides experimental comparisons. Ablation experiments are presented in Section V to analyze the network hyperparameters and verify the advantage of each component proposed in our model. We conclude this article in Section VI.

II. GFs Revisiting

A. Classical GFs

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

$$\boldsymbol{f}_u = \sum_v \boldsymbol{W}_{u,v}(\boldsymbol{g}, \boldsymbol{t})\boldsymbol{t}_v \tag{1}$$

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

In the classical bilateral filter and the guided image filter, $\boldsymbol{W}_{u,v}$ depends only on the guidance $\boldsymbol{g}$. Specifically, the filter weight in the bilateral filter is defined as

$$\boldsymbol{W}^{\mathrm{BF}}_{u,v} = \frac{1}{C_u}\exp\left(-\frac{\|u-v\|}{\sigma_s}\right)\exp\left(-\frac{\|\boldsymbol{g}_u - \boldsymbol{g}_v\|}{\sigma_r}\right) \tag{2}$$

where C_u is the normalization parameter and σ_s and σ_r are parameters for geometric and photometric similarity, respectively. In the guided image filter [1], the filter kernel weight is defined as

$$\boldsymbol{W}^{\mathrm{GIF}}_{u,v} = \frac{1}{|\boldsymbol{N}_k|^2}\sum_{k:(u,v)\in\boldsymbol{N}_k}\left(1 + \frac{(\boldsymbol{g}_u - \mu_k)(\boldsymbol{g}_v - \mu_k)}{\sigma_k^2 + \epsilon}\right) \tag{3}$$

where $|\boldsymbol{N}_k|$ is the number of pixels in a window $\boldsymbol{N}_k$ and μ_k and σ_k^2 are the mean and variance of $\boldsymbol{g}$ in $\boldsymbol{N}_k$, respectively.

B. Learning-Based GFs

Among deep learning-based approaches for GF design, the DFN [28] first defines a filter-generating network (FGN) that takes the guidance $\boldsymbol{g}$ as input to obtain location-specific dynamic filters $\boldsymbol{F}_\theta = \mathrm{FGN}(\boldsymbol{g}, \boldsymbol{\theta})$, which are then applied to the target image $\boldsymbol{t}$ to yield the output $\boldsymbol{f} = \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{t})$. Pixel-adaptive convolution [5] defines the filter kernel by multiplying a spatially varying filter on the standard spatially invariant kernel

$$\boldsymbol{f}_u = \sum_{v\in\boldsymbol{N}_u}\boldsymbol{K}(\boldsymbol{g}_u, \boldsymbol{g}_v)\boldsymbol{W}[p_u - p_v]\boldsymbol{t}_v + b \tag{4}$$

where $\boldsymbol{W}$ is the spatially invariant kernel, $\boldsymbol{K}(\cdot,\cdot)$ is a varying filter kernel function that has a fixed form such as Gaussian, and $[p_u - p_v]$ denotes the index offset of kernel weights. From the above formulation, it can be found that, similar to the bilateral filter and guided image filter, DFN and pixel-adaptive convolution also only depend on the guidance $\boldsymbol{g}$ 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 [4] alleviates this drawback by jointly leveraging features of both the guidance and the target. It designs two-branch subnetworks to extract features from the guidance and the target, which are passed through a fusion subnetwork to output the filtering result. The joint filter $\boldsymbol{\Phi}$ is learned in an end-to-end manner by the following optimization:

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

where $\boldsymbol{f}^{gt}$ is the ground truth of the output. In contrast to the implicit filter learning approach of DJF, deformable kernel networks (DKNs) [6] explicitly learn the kernel weights $\boldsymbol{K}$ and offsets $\boldsymbol{s}$ using two-branch subnetworks from the two images. Concretely, the filtering is performed by

$$\boldsymbol{f}_u = \sum_{v\in\boldsymbol{N}_u}\boldsymbol{W}_{u,\boldsymbol{s}(v)}(\boldsymbol{g}, \boldsymbol{t})\boldsymbol{t}_{\boldsymbol{s}(v)} \tag{6}$$

with

$$\boldsymbol{W}(\boldsymbol{g}, \boldsymbol{t}) = \boldsymbol{K}(\boldsymbol{g}) \odot \boldsymbol{K}(\boldsymbol{t}) \tag{7}$$

where $\boldsymbol{K}(\boldsymbol{g})$ and $\boldsymbol{K}(\boldsymbol{t})$ are kernel weights learned from the guidance and the target, respectively, and $\odot$ denotes the elementwise multiplication. Although DJF and DKN achieve better performance than previous methods, they treat the guidance and target 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, they fuse multimodal features though elementwise 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 DAGF scheme to more effectively leverage multimodal 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

$$\boldsymbol{f}_u = \sum_{v\in\boldsymbol{N}_u}\boldsymbol{A}_{u,v}\boldsymbol{W}^g_{u,v}\boldsymbol{t}_v + \sum_{v\in\boldsymbol{N}_u}(1 - \boldsymbol{A}_{u,v})\boldsymbol{W}^t_{u,v}\boldsymbol{t}_v \tag{8}$$

where $\boldsymbol{W}^g_{u,v}$ and $\boldsymbol{W}^t_{u,v}$ are filter kernels computed from the guidance and the target, respectively; and $\boldsymbol{A}_{u,v}$ denotes the pixelwise 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 unreliable structure propagation.

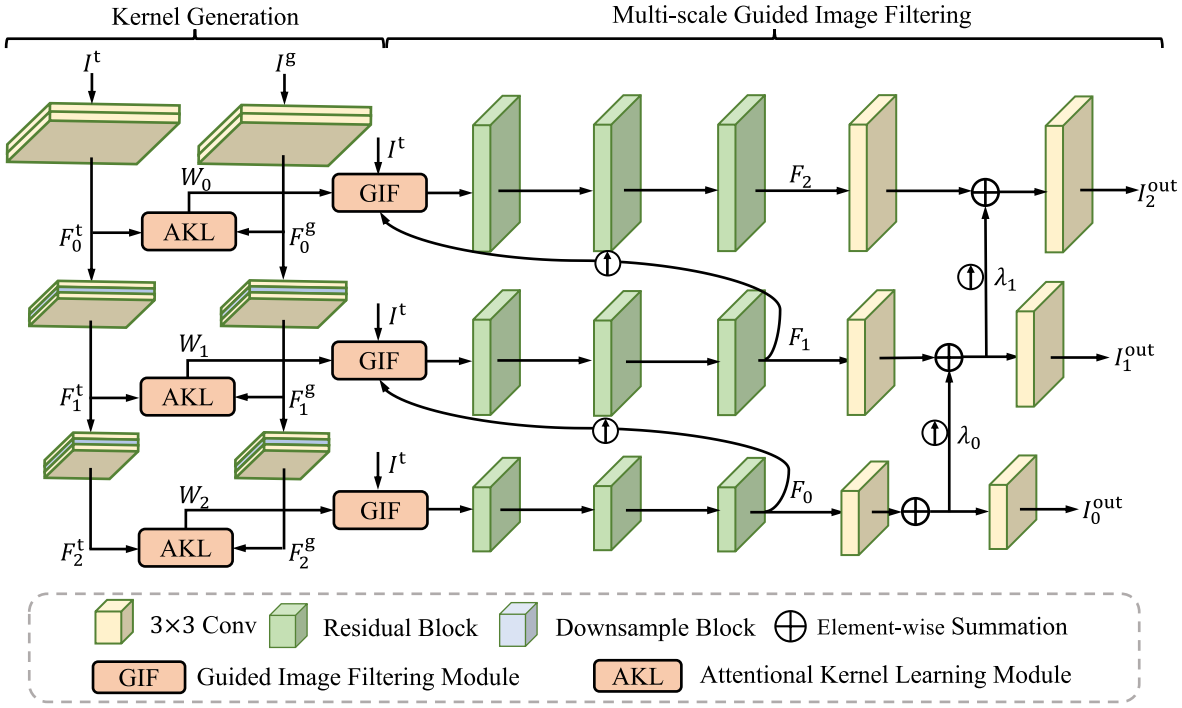


Fig. 2. Network architecture of the proposed DAGF with the number of pyramid level $m = 3$. DAGF consists of a kernel generation network for constructing filter kernels and a multiscale guided image filtering network with the purpose of filtering target image by using the generated kernels.

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 attentioanl 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 $\boldsymbol{I}^t \in \mathbb{R}^{H\times W\times C^t}$ (e.g., low-resolution (LR) depth) and a guidance image $\boldsymbol{I}^g \in \mathbb{R}^{H\times W\times C^g}$ (e.g., HR color image) as inputs and generates a reconstructed image $\boldsymbol{I}^{\text{out}} \in \mathbb{R}^{H\times W\times C^t}$ as output, where H, W, and C denote the height, width, and channels respectively.

Fig. 2 shows the overall architecture of the proposed network, which is composed of kernel generation subnetwork and multiscale guided filtering subnetwork. Instead of directly predicting kernels in the image domain and enlarging its receptive field by using the deformable sampling strategy as in [6], 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.

1) In the filter kernel generation subnetwork, the multiscale features of $\boldsymbol{I}^t$ and $\boldsymbol{I}^g$ are fed into the AKL module to generate filter kernels $\{W_i\}$. The network architecture of AKL is shown 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.
2) In the guided filtering subnetwork, with the pixelwise filter kernels, features of the target image are processed in a coarse-to-fine manner to get the upsampled features.

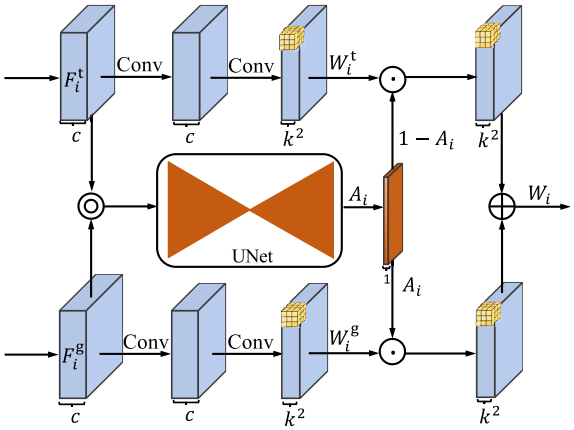


Fig. 3. Network architecture of the proposed AKL module, where $\odot$ denotes the elementwise multiplication and $\circledcirc$ denotes the concatenation operation.

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

B. Filter Kernel Generation

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

$$\boldsymbol{F}_0^t = \text{Conv}\big(\text{Conv}\big(\boldsymbol{I}^t\big)\big) \tag{9}$$

$$\boldsymbol{F}_i^t = \text{Down}\big(\boldsymbol{F}_{i-1}^t\big), 0 < i < m \tag{10}$$

where m denotes the level of pyramid network and Conv$(\cdot)$ is the convolution operator; Down$(\cdot)$ represents the downsample block with a scale factor of 2, which is implemented by two convolution layers and an inverse pixel-shuffle operation [35]. 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, instead of only relying on the guidance as done in most existing methods. To this end, as shown in Fig. 3, we propose an 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

$$\boldsymbol{W}_i^t = \text{Conv}\big(\text{Conv}\big(\boldsymbol{F}_i^t\big)\big), 0 \le i < m \tag{11}$$

$$\boldsymbol{W}_i^g = \text{Conv}\big(\text{Conv}\big(\boldsymbol{F}_i^g\big)\big), 0 \le i < m \tag{12}$$

where $\boldsymbol{W}_i^t$ and $\boldsymbol{W}_i^g$ are the ith constructed filter kernels from the target and guidance features, respectively. The spatial resolution of ith kernels is the same as one of its corresponding input features, while the number of channels is 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 an adaptive kernel combination module based on a lightweight UNet architecture, which takes both guidance and target features as inputs and models the pixelwise dependencies among them in a learning manner. This process is formulated as

$$\boldsymbol{A}_i = \text{UNet}\big(\big[\boldsymbol{F}_i^t, \boldsymbol{F}_i^g\big]\big), 0 \le i < m \tag{13}$$

where UNet is a five-layer U-like [36] network; $[\cdot,\cdot]$ denotes the concatenation operation; and $\boldsymbol{A}_i$ 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 GF kernels can be derived as

$$\boldsymbol{W}_i = \boldsymbol{A}_i \odot \boldsymbol{W}_i^g + (\boldsymbol{1} - \boldsymbol{A}_i) \odot \boldsymbol{W}_i^t, 0 \le i < m \tag{14}$$

where $\boldsymbol{W}_i$ is the generated ith filter kernel, $\boldsymbol{1}$ denotes the all-1 matrix, and $\odot$ means the elementwise multiplication.

C. Multiscale Guided Filtering

After generating the GF kernels, the following step is to perform filtering on the target image, which is done by the guided filtering subnetwork. As shown in the right of Fig. 2, it takes the target image $\boldsymbol{I}^t$ as the input and progressively filters the input target image by using the learned filter kernels $\{\boldsymbol{W}_0, \ldots, \boldsymbol{W}_{m-1}\}$ in a coarse-to-fine manner.

Specifically, given $\boldsymbol{I}^t$, we first utilize Bicubic to resize it to the same resolution as its corresponding filter kernels

$$\hat{\boldsymbol{I}}^t = \text{Bicubic}\big(\boldsymbol{I}^t\big). \tag{15}$$

Then, the filtering process can be formulated as

$$\boldsymbol{F}_0 = \text{ResNet}\big(\text{GIF}\big(\text{Conv}\big(\hat{\boldsymbol{I}}^t\big)\big), \boldsymbol{W}_{m-1}\big) \tag{16}$$

$$\boldsymbol{F}_i = \text{ResNet}\Big(\text{GIF}\Big(\Big[\boldsymbol{F}_{i-1}^{\uparrow}, \text{Conv}\big(\hat{\boldsymbol{I}}^t\big)\Big], \boldsymbol{W}_{m-1-i}\Big)\Big), 0 < i < m \tag{17}$$

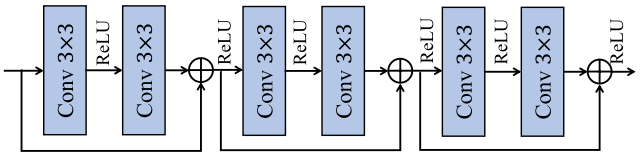


Fig. 4. Network architecture of the ResNet used in our model, where $\oplus$ means the elementwise summation.

where $[\cdot,\cdot]$ means the concatenation operation. ResNet$(\cdot)$ consists of three cascade residual blocks [37], as shown in Fig. 4, and we use it to increase the nonlinearity of our model. It is worth noting that, since the main purpose of this article is to demonstrate the power of the proposed DAGF framework, we turn to use the very simple residual block as the basic block, and other advanced blocks, such as the dense block and transformer module, can also be used here to further improve the performance of our model. $\boldsymbol{F}_i$ is the ith filtered target feature and $\uparrow$ is an upsampling operation. GIF$(\cdot)$ is a filtering operation that conducts the filtering operation on the corresponding target features. Specifically, we first reshape the third dimension of the filter from k^2 to $k \times k$, and then, the filtering process for a pixel $\{(u,v)|0 \le u < H, 0 \le v < W\}$ can be defined as follows:

$$\boldsymbol{F}(u,v) = \sum_{x=-\sigma}^{\sigma} \sum_{y=-\sigma}^{\sigma} \boldsymbol{W}_{u,v}(x,y) \cdot \tilde{\boldsymbol{F}}(u-x, v-y) \tag{18}$$

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

Based on $\{\boldsymbol{F}_i\}_{i=0}^{m-2}$, we can obtain the filter results of DAGF by using the proposed the multiscale fusion strategy

$$\hat{\boldsymbol{F}}_0 = \text{Conv}(\boldsymbol{F}_0) \tag{19}$$

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

$$\boldsymbol{I}_i^{\text{out}} = \text{Conv}\big(\hat{\boldsymbol{F}}_i\big) + \boldsymbol{I}^t, 0 \le i < m-1 \tag{21}$$

where λ_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. $\{\boldsymbol{I}_i^{\text{out}}\}_{i=0}^{m-2}$ are the intermediate multiscale results and $\boldsymbol{I}_{m-1}^{\text{out}}$ is the final filtering result of the proposed scheme.

D. Loss Function

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

1) *L_1 Loss:* $\mathcal{L}_1$ measures the pixelwise errors between the output image $\boldsymbol{I}_{m-1}^{\text{out}}$ and the ground-truth image

$$\mathcal{L}_1 = ||\boldsymbol{I}^h - \boldsymbol{I}_{m-1}^{\text{out}}||_1. \tag{22}$$

2) *Multistage Loss:* To stabilize the network training process and promote the multistage guided filtering module to learn more effective parameters, we propose a multistage loss to enforce all intermediate results to be close

to the ground-truth residual image

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

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

3) *Boundary-Aware Loss:* Optimizing the pixelwise loss 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 high-frequency parts, we propose a boundary-aware loss to promote our model to generate sharper boundaries. Specifically, we first employ the Sobel operator ∇ to detect the boundary information of the ground truth and the network output and obtain the boundary mask $\boldsymbol{M}$

$$\boldsymbol{M} = \left(\nabla_x \boldsymbol{I}^h - \nabla_x \boldsymbol{I}_{m-1}^{\text{out}}\right) \odot \left(\nabla_y \boldsymbol{I}^h - \nabla_y \boldsymbol{I}_{m-1}^{\text{out}}\right) \tag{24}$$

and then, the boundary-aware loss is defined as

$$\mathcal{L}_{\text{ba}} = ||\boldsymbol{M} \odot \boldsymbol{I}^h - \boldsymbol{M} \odot \boldsymbol{I}_{m-1}^{\text{out}}||_1 \tag{25}$$

where $\odot$ denotes the elementwise multiplication.

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

$$\mathcal{L} = \omega_1 \cdot \mathcal{L}_1 + \omega_2 \cdot \mathcal{L}_{\text{ba}} + \omega_3 \cdot \mathcal{L}_{\text{ms}} \tag{26}$$

where ω_1, ω_2, and ω_3 are hyperparameters to balance these loss functions. We set $\omega_3 = 1$ to stabilize the training procedure at an early stage and then progressively decay to zero with the training progress to boost the performance of final output. We set $\omega_1 = 1$ and $\omega_2 = 10$.

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 3×3. The ablation study presented in the following will verify the effectiveness of our configuration. The hyperparameters of our model are $\omega_1 = 1$, $\omega_2 = 10$, and $\omega_3 = 1$. All the convolution layers within the proposed methods are sized of 3×3 and the channels of intermediate features are 32. We use PReLU [38] as the default activation function. We utilize PixelShuffle [39] and InvPixelShuffle [35] as the upsampling and downsampling operators to resize the features in our model.

In the training phase, the batch size is 32 and we randomly crop 256×256 image patches from the target and guidance images as inputs. We augment the training data with random flipping and rotation. Adam [40] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is employed as an optimizer. The initial learning rate is set as 1×10^{-4} and we halve it every 80 epochs and stop the training after 100 epochs. Our model is implemented by Pytorch [41] and trained on one RTX 1080ti GPU. Training the proposed method roughly takes two days for the NYU v2 [42] dataset.

Our network takes three-channel guidance and one-channel target images as inputs. For the multichannel target images (e.g., flash/nonflash image denoising), 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 a three-channel 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 GSR (e.g. depth image SR and saliency map SR, see Section IV-A), cross-modality image restoration (e.g. joint depth image SR and denoising, and flash/nonflash image denoising, see Section IV-B), and image semantic segmentation (Section IV-C).

For a 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 SR

It is a classic computer vision task, which aims to reconstruct an HR image from an LR one with the help of an HR image from another modality. For example, we can obtain an HR depth by GSR using an LR depth and an HR RGB image as inputs, where the HR RGB image serves as the guidance.

Following the experimental settings in [4] and [6], we train our model on the task of depth image SR and then evaluate the performance of the model on tasks of depth image SR and saliency map SR, and the latter one is used to verify the generalization ability of our model.

B. RGB-Guided Depth SR

For this task, we use the first 1000 RGB-D image pairs from the NYU v2 dataset [42] as the training set. In order to make a fair comparison with existing methods, we exploit the nearest neighbor downsampling as the standard downsampling operator to generate the LR target image. To show the effectiveness of the proposed method, we further conduct experiments on Bicubic downsampling as done in [6]. The performance of the proposed method is evaluated on the following four standard benchmark datasets.

1) *Sintel Dataset [50]:* This dataset consists of 1064 image pairs that are obtained by an animated 3-D movie.
2) *NYU v2 Dataset [42]:* 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.
3) *Lu Dataset [51]:* It contains six image pairs captured by the ASUS Xtion Pro camera.
4) *Middlebury Dataset [52], [53]:* 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. [51].

We compare our method with 13 state-of-the-art methods and adopt root-mean-square error (RMSE) as the evaluation metric. Lower RMSE values mean higher recovery quality. Table I lists 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 Table I, our method achieves the best results among all the compared methods on both synthetic and real datasets. The superior performance benefits from the more precise filter kernels learned and the multiscale 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

TABLE I

QUANTITATIVE COMPARISON FOR DEPTH IMAGE SR ON FOUR STANDARD RGB/D DATASETS IN TERMS OF AVERAGE RMSE VALUES. FOLLOWING THE EXPERIMENTAL SETTING IN [6] AND [43], WE CALCULATE THE AVERAGE RMSE VALUES IN CENTIMETER FOR THE NYU V2 [42] 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 IS HIGHLIGHTED IN **BOLDFACE**, WHILE THE SECOND BEST ONES ARE UNDERSCORED. THE LOWER RMSE MEANS 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 [44]	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 [45]	3.39	5.41	12.03	4.48	7.58	17.46	6.98	11.23	28.13	8.30	13.05	19.96
Experiment results for depth map super-resolution (Nearest-neighbour down-sampling).												
DGF [46]	3.92	6.04	10.02	2.73	5.98	11.73	4.50	8.98	16.77	7.53	11.53	17.50
DMSG [47]	1.79	3.39	5.87	2.48	4.74	7.51	3.48	6.07	10.27	6.80	9.09	11.81
DJFR [4]	1.98	3.61	6.07	2.22	4.54	7.48	3.38	5.86	10.11	7.05	9.12	12.61
DSRN [48]	2.08	3.26	5.78	2.57	4.46	6.45	3.49	5.70	9.76	7.29	9.43	11.62
CUNet [3]	2.11	3.23	5.74	2.51	4.42	6.45	3.10	5.17	9.09	6.88	9.01	11.18
PAC [5]	1.91	3.20	5.60	2.48	4.37	6.60	2.82	5.01	8.64	6.79	8.36	11.02
SVLRM [33]	1.86	2.94	5.30	2.13	4.25	6.97	2.37	4.84	8.68	6.92	8.54	11.65
DKN [6]	1.93	3.17	5.49	2.35	4.16	6.33	2.46	4.76	8.50	6.84	8.61	11.21
DAGF(Ours)	**1.78**	**2.73**	**4.75**	**1.96**	**3.81**	**6.16**	**2.35**	**4.62**	**7.81**	**6.72**	**8.35**	**10.64**
Experiment results for depth map super-resolution (Bicubic down-sampling).												
DGF [46]	1.94	3.36	5.81	2.45	4.42	7.26	3.21	5.92	10.45	5.91	8.02	11.17
DMSG [47]	1.88	3.45	6.28	2.30	4.17	7.22	3.02	5.38	9.17	4.73	6.26	8.36
DJFR [4]	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 [48]	1.77	3.05	4.96	1.77	3.10	5.11	3.00	5.16	8.41	4.49	6.53	9.28
CUNet [3]	1.45	2.78	4.96	1.51	2.49	5.23	2.02	4.43	8.01	4.59	6.47	9.11
PAC [5]	1.32	2.62	4.58	1.20	2.33	5.19	1.89	3.33	6.78	4.42	6.13	8.42
SVLRM [33]	**1.11**	2.13	4.34	0.93	2.19	5.44	1.51	3.21	6.98	4.05	5.83	8.60
DKN [6]	1.23	2.12	4.24	0.96	2.16	5.11	1.62	3.26	6.51	4.38	5.89	8.40
DAGF (Ours)	1.15	**1.80**	**3.70**	**0.83**	**1.93**	**4.80**	**1.36**	**2.87**	**6.06**	**3.84**	**5.59**	**7.44**

method, we present the visual results for 16× depth image SR in Fig. 5, from which we can see that our method can generate results with clear boundaries and less artifacts.

C. RGB-Guided Saliency Map SR

To further demonstrate the generalization ability of the proposed method, we apply the model that is trained on the NYU v2 dataset directly to the task of saliency map SR without any fine-tuning step. Similar to DKN [6], we use 5168 image pairs from the DUT-OMRON dataset [49] to evaluate the SR performance. We use Bicubic interpolation (8×) to generate the LR saliency maps and then super-resolve them with the corresponding HR color image as guidance. The quantitative results in terms of F-measure are listed in Table II. As can be seen from Table II, 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 randomly select two images and visualize the recovered HR saliency map obtained by different methods in Fig. 6. It can be observed that the results of Bicubic are oversmoothed, in which the structure details are severely damaged. DMSG [47] and DJFR [4] struggle to generate clear boundaries. The results of DKN [6] 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.

D. Cross-Modality Image Restoration

For the task of cross-modality image restoration, we first conduct experiments on joint depth image SR 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/nonflash and RGB/nearing infrared (NIR) image pairs. Finally, we conduct experiments on the ToFMark dataset [45]. It contains three real-world depth images acquired by time of flight (ToF) camera, which has complicated multimodality degradation.

E. Joint Depth Image SR and Denoising

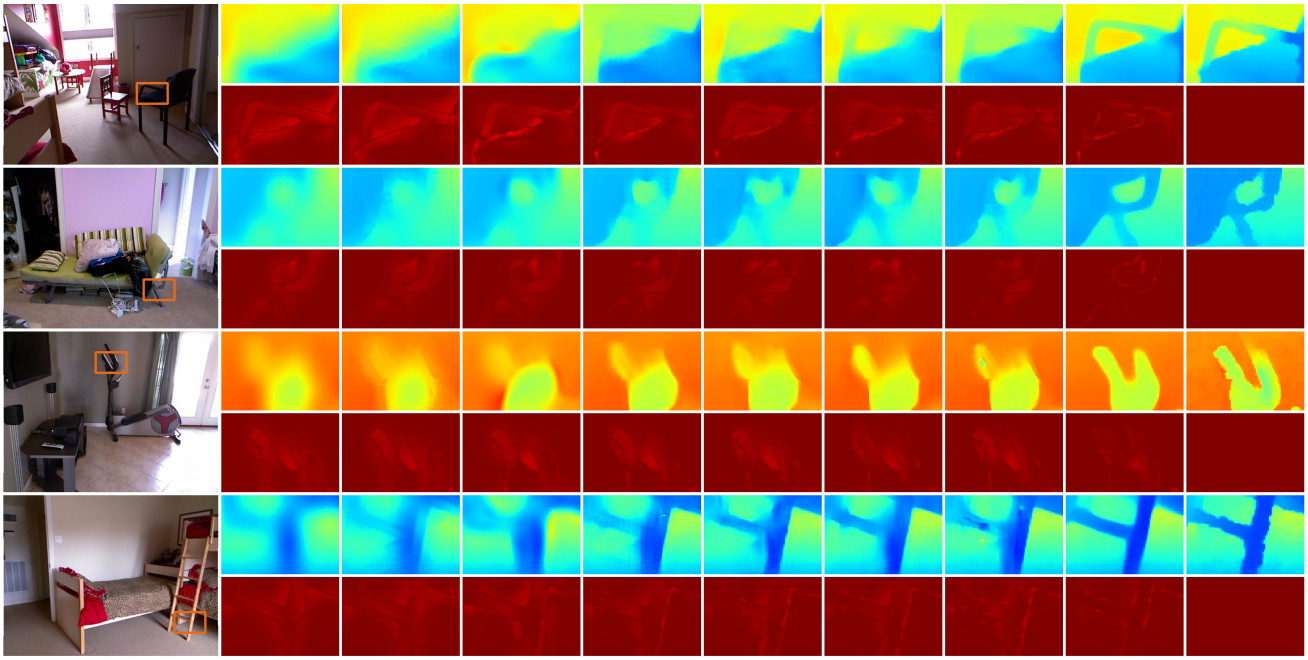
Depth images acquired by ranging sensors are typically noisy. In order to simulate the data acquisition process of the depth sensor, we add Gaussian noise with variance as 25 to the LR target depth images. We use the same experimental settings as the task of GSR in Section IV-A to train our model. Since most of the existing methods do not provide experimental results for this task, we retrain all 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× and 16× cases, which are more difficult to recover. This is mainly because: 1) we employ a pyramid architecture to extract multimodality features for guided kernel generation, and thus, the multiscale complementary information can be obtained; 2) for guided image filtering, we leverage

TABLE II

QUANTITATIVE COMPARISON OF 8× SALIENCY MAP SR ON THE DUT-OMRON DATASET [49]. 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 [47]	DJFR [4]	PAC [5]	SVLRM [33]	DKN [6]	DAGF (Ours)
Fscore	0.853	0.821	0.910	0.901	0.908	0.923	<u>0.926</u>	**0.932**



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 5. Qualitative comparison for recovered depth maps (16×). (a) Guidance image. (b) Target image. (c) GF [1]. (d) DMSG [47]. (e) DJFR [2]. (f) PAC [3]. (g) SVLRM [33]. (h) DKN [6]. (i) DAGF. (j) Ground-truth depth map. Please enlarge the PDF for more details.

TABLE III

QUANTITATIVE COMPARISON FOR JOINT DEPTH IMAGE SR AND DENOISING ON FOUR STANDARD RGB/D DATASETS IN TERMS OF AVERAGE RMSE VALUES. FOLLOWING THE EXPERIMENTAL SETTING IN [6] AND [43], WE CALCULATE THE AVERAGE RMSE VALUES IN CENTIMETER FOR THE NYU V2 [42] 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 IS HIGHLIGHTED IN **BOLDFACE**, WHILE THE SECOND BEST ONES ARE UNDERSCORED

Datasets	Middlebury			Lu			NYU v2			Sintel		
Method	4×	8×	16×	4×	8×	16×	4×	8×	16×	4×	8×	16×
DGF [46]	2.70	4.44	6.38	4.06	5.85	8.39	6.52	9.23	13.00	6.94	9.03	12.05
DMSG [47]	1.80	2.70	4.75	<u>1.97</u>	3.42	6.14	3.60	5.31	<u>8.07</u>	5.08	<u>6.64</u>	<u>8.86</u>
DJFR [4]	1.87	3.16	5.35	2.20	4.04	6.84	4.01	6.21	9.83	5.52	7.58	10.31
DSRN [48]	1.99	3.12	5.17	2.26	3.85	6.69	4.35	6.31	9.64	5.85	7.59	10.14
CUNet [3]	1.77	2.87	4.58	2.02	3.60	5.90	3.84	6.00	9.23	5.28	7.13	9.55
PAC [5]	1.91	2.94	5.09	2.57	3.44	6.18	4.22	6.24	9.60	<u>4.92</u>	7.32	9.89
SVLRM [33]	1.80	<u>2.65</u>	4.91	2.04	3.61	6.55	3.47	<u>5.27</u>	8.82	4.97	6.77	9.44
DKN [6]	<u>1.76</u>	2.68	<u>4.55</u>	2.02	<u>3.34</u>	<u>5.97</u>	<u>3.45</u>	5.29	8.55	4.97	<u>6.64</u>	9.18
DAGF (Ours)	**1.67**	**2.61**	**4.25**	**1.79**	**3.20**	**5.71**	**3.25**	**4.96**	**7.76**	**4.76**	**6.47**	**8.66**

the coarse-to-fine strategy to filter the LR target image, and thus, the structure details can be progressively recovered; and 3) compared to single loss at the end of network, the proposed multiscale loss can bring stronger supervision to our model.

Fig. 7 further shows the visual superiority of the proposed method for joint depth image SR and denoising (16× Bicubic downsampling and Gaussian noise). The results of GF [1], MUF [2], and SDF [54] still contain much 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 pixelwise linear representations. The methods of PAC [5] and DJFR [4] 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 [6]. In contrast, our method is

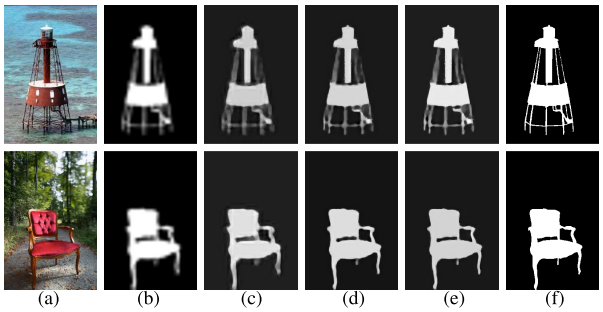


Fig. 6. Visual comparison of 8× saliency map SR on the DUT-OMRON dataset [49]. (a) Guidance (RGB). (b) LR image (Bicubic). (c) DJFR [4]. (d) DKN [6]. (e) DAGF. (f) Ground truth. Please enlarge the PDF for more details.

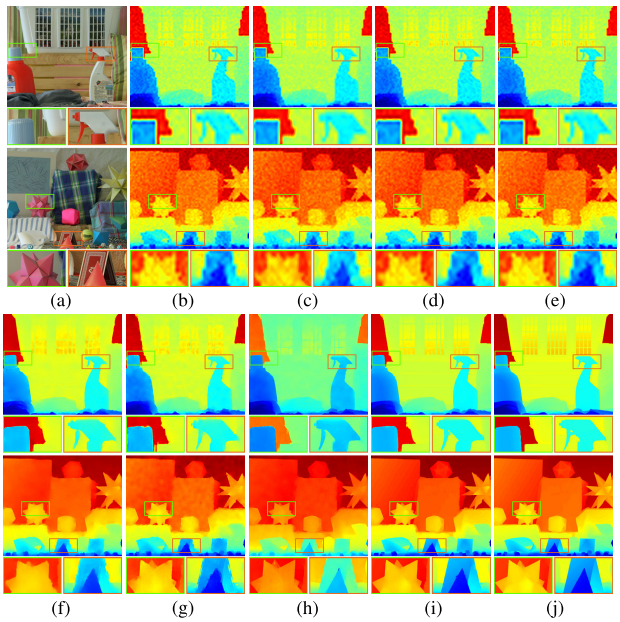


Fig. 7. Qualitative comparison of joint depth map SR and denoising. Please enlarge the PDF for more details. (a) Guidance. (b) Target. (c) GF [1]. (d) MUF [2]. (e) SDF [54]. (f) PAC [5]. (g) DJFR [4]. (h) DKN [6]. (i) DAGF. (j) GT.

able to remove the noise effectively and produces the clearest and sharpest boundaries.

F. 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 nonflash image denoising and NIR-guided color image restoration. Fig. 8 shows the visual comparison between existing state-of-the-art methods and ours. All the deep learning-based methods (e.g. DJFR [4] and DKN [6]) are tested with the same setting as ours. Among the compared methods, SDF [54] and RTV [22] are specially designed for this task. As can be seen from Fig. 8, DJFR [4] cannot remove noise, and the results of DKN [6] suffer from halo artifacts. On the contrary, the proposed DAGF can produce more convincing results with fewer artifacts. The method of RTV [22], which is specially designed for this task, obtains the best performance.

G. Realistic Depth Image SR

To further evaluate the robustness of the proposed method, we conduct experiments on the ToFMark dataset [45], which include real ToF sensor data and thus have complicated multimodality degradation. Following the experimental protocol of DGDIE [55], we first perform image completion on the acquired depth images and then send them to our model (4× SR and denoising) trained on the NYU v2 dataset [42] to obtain the final results. We compare our method with a recently proposed deep learning-based method (e.g. DKN [6]) and some traditional methods (e.g. TGV [45], SDF [54], and DGDIE [55]). As shown in Table IV, our method constantly obtains the best objective results for the three test images. Fig. 9 presents the visual comparison results for two images (*books* and *devil*). From these figures, it is easy to observe that the results of SDF [54] suffer from texture-copying artifacts. The results of DKN [6] are smooth and blurred since DKN generates filter kernels without considering the inconsistence between color and depth image. The results of DGDIE [55] 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.

TABLE IV

QUANTITATIVE COMPARISON FOR REALISTIC DEPTH IMAGE SR IN TERMS OF RMSE VALUES ON THE TOFMARK [45] 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 [7]	16.03	18.79	27.57
GF [1]	15.74	18.21	27.04
TGV [45]	12.36	15.29	14.68
SDF [54]	12.66	14.33	10.68
Yang [8]	12.25	14.71	13.83
DGDIE [55]	12.32	14.06	9.66
DKN [6]	<u>11.81</u>	<u>13.54</u>	<u>9.11</u>
DAGF (Ours)	**11.80**	**13.47**	**9.07**

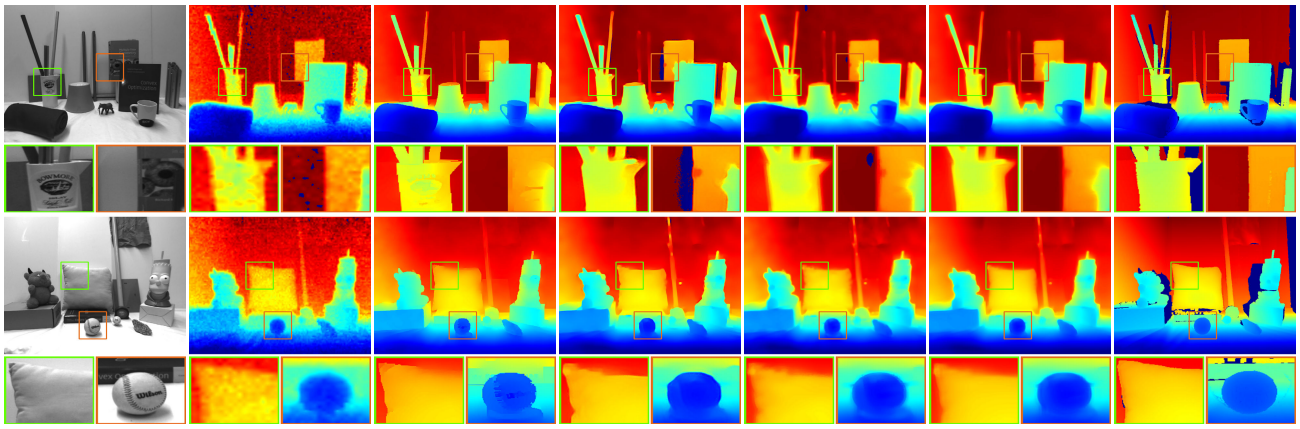
H. Semantic Segmentation

Semantic segmentation is a fundamental computer vision task, which aims at assigning predefined labels to each pixel of an image. In DGF, Wu et al. [46] proposed to use guided image filtering as a layer to replace the time-consuming fully connected conditional random field (CRF) [57] for semantic segmentation. We demonstrate that the proposed DAGF can be applied to this problem. Following DGF [46], we plug the proposed model into DeepLab-v2 [56] and train the whole network in an end-to-end manner, and thus, the offline postprocessing of CRFs can be avoided. We utilize the Pascal VOC 2012 dataset [58] in our experiment, which contains 1264, 1229, and 1456 images for training, validation, and testing, respectively. Similar to DGF [46], we augment the training set with the annotations provided in [59], resulting in 10 582 images. The 1449 images in the validation set are employed to evaluate the proposed method.

We use the mean intersection-over-union (IoU) score as an evaluation metric and report the quantitative results for the validation set of the Pascal VOC dataset [58] in Table V. The baseline denotes DeepLab-v2 [56] without CRF. As can be seen from Table V, our method achieves slightly better performance than the compared methods in terms of mIoU values. Moreover, we visualize the segmentation results among

(a) (b) (c) (d) (e) (f) (g)

Fig. 8. Visual comparison of cross-modality image restoration. Top: flash-guided nonflash image denoising. Bottom: NIR-guided color image denoising. Please enlarge the PDF for more details. (a) Guidance. (b) Target. (c) SDF [54]. (d) RTV [22]. (e) DJFR [4]. (f) DKN [6]. (g) DAGF.



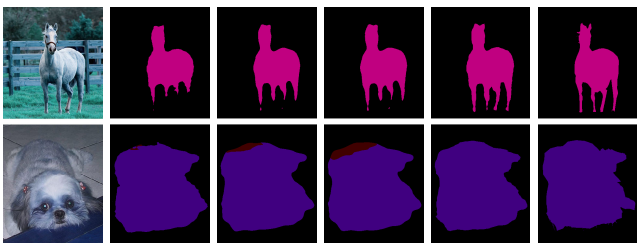
(a) (b) (c) (d) (e) (f) (g)

Fig. 9. Visual comparison of realistic depth map SR on ToFMark [45] dataset. Please enlarge the PDF for more details. (a) Guidance. (b) Target. (c) SDF [54]. (d) DGDIE [55]. (e) DKN [6]. (f) DAGF. (g) GT.

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 IS HIGHLIGHTED IN **BOLDFACE**, WHILE THE SECOND ONES ARE UNDERSCORED

Methods	Mean IoU
Deeplab-V2 [56]	70.69
DenseCRF [57]	71.98
DGF [46]	72.96
DJFR [4]	73.30
FDKN [6]	<u>73.60</u>
DAGF (Ours)	**73.76**



(a) (b) (c) (d) (e) (f)

Fig. 10. Visual comparison of semantic segmentation on the validation set of Pascal VOC 2012 dataset [58]. (a) Input image (RGB). (b) DeepLab-v2 [56]. (c) DGF [46]. (d) FDKN [6]. (e) Ours (DAGF). (f) Ground truth. Please enlarge the PDF for more details.

our 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 hyperparameters setting in our model and then conduct a series of ablation experiments to investigate the effectiveness of our main contributions, e.g. AKL module (mentioned in Section III-B), multiscale fusion (mentioned in Section III-C) with deep supervision (mentioned in Section III-D), and boundary-aware loss (mentioned in Section III-D). In this study, we train different variants of our model on the commonly used NYU v2 dataset [42] with 16× nearest neighbor downsampling and evaluate the performance of them on four benchmark datasets. The experimental settings are the same as in Section IV-A.

A. Hyperparameters Setting

For the hyperparameters setting, we first investigate the influence of the size $k \times k$ of the learned kernels (i.e., W_0, W_1, and W_2 in Fig. 2) and the number of pyramid level m. Enlarging k or m can increase the receptive field of our model but at the expense of higher computational complexity. To seek an appropriate tradeoff between complexity and performance, we conduct experiments on the task of depth map SR 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 size of kernels $k \times k$. The possible reason for this phenomenon is that the receptive field is sufficient for this task when $m = 3$ and $k = 3$ and larger m or k will burden the network optimization process. Therefore, we set $m = 3$ and $k = 3$ in our experiments. Then, we conduct

TABLE VI
ABLATION STUDY. QUANTITATIVE COMPARISON OF DIFFERENT SIZES OF KERNEL ($k \times k$) AND THE NUMBER OF PYRAMID LEVEL (m) ON THE NYU V2 DATASET [42]

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

experiments to analyze the effect of ω_1, ω_2, and ω_3 in (26). ω_1 controls the weight of the $\mathcal{L}_1$ loss, which is used to maintain fidelity, and increasing ω_1 is equivalent to decreasing ω_2 and ω_3. Thus, we keep ω_1 as 1 to analyze ω_2 and ω_3. The results are shown in Tables VII–IX, from which we can see that when $\omega_2 = 10$ and $\omega_3 = 1$, the model can achieve the best performance. Hence, we set $\omega_2 = 10$ and $\omega_3 = 1$ in our model.

B. Ablation Experiments

As shown in Fig. 2, our model consists of two parts: kernel generation subnetwork and multiscale guided image filtering subnetwork. For the kernel generation subnetwork, 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 the guided image filtering subnetwork, we progressively filter the target image with the learned multiscale kernels. In order to fully integrate the intermediate filtered results, we propose a multiscale feature fusion strategy and a multistage 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 hybrid loss functions, e.g., pixelwise loss $\mathcal{L}_1$, multiscale loss $\mathcal{L}_{\text{ms}}$, and boundary-aware loss $\mathcal{L}_{\text{ba}}$. To analyze the contribution of each component of our model, we implement seven variants of our model.

1) *Model1:* It takes (target, target) as inputs for kernel generation and is trained with the $\mathcal{L}_1$ loss.
2) *Model2:* It takes (guidance, guidance) as inputs for kernel generation and is trained with the $\mathcal{L}_1$ loss.
3) *Model3:* It takes (target, guidance) as inputs, uses elementwise multiplication to combine the generated two sets of kernels, and is trained with the $\mathcal{L}_1$ loss.
4) *Model4:* It takes (target, guidance) as inputs, uses elementwise summation to combine the generated two sets of kernels, and is trained with the $\mathcal{L}_1$ loss.
5) *Model5:* It takes (target, guidance) as inputs, uses the learned weight map to adaptively combine the generated two sets of kernels, and is trained with the $\mathcal{L}_1$ loss.
6) *Model6:* Model5 trained with the $\mathcal{L}_1$ loss and $\mathcal{L}_{\text{ms}}$ loss.
7) *Model7:* This model learns the filter kernels directly from both the guidance and target images, and the training losses include the $\mathcal{L}_1$ loss, the $\mathcal{L}_{\text{ms}}$ loss, and the $\mathcal{L}_{\text{ba}}$ loss.
8) *DAGF:* Model5 trained with the $\mathcal{L}_1$ loss, $\mathcal{L}_{\text{ms}}$ loss, and $\mathcal{L}_{\text{ba}}$ loss. This is our full model.
9) *Model8:* It means the DAGF variant that takes (target, target) as inputs for kernel generation.
10) *Model9:* It denotes the DAGF variant that takes (guidance, guidance) as inputs for kernel generation.

It is noteworthy that we adjust the number of convolutional layers in multiscale guided image filtering subnetwork to guarantee that each variant could have roughly the same number of parameters as our final model. The quantitative results are shown in Table VII, from which we can see that the full model (DAGF) achieves the best reconstruction performance in four test datasets compared to the ablated models, and each component proposed in our model can significantly improve the performance of the network. Moreover, we build a new model (Model7) to verify which is the best to generate filter kernels. Model7 learns the filter kernels directly from both the guidance and target images, and it can be obtained by modifying our model as follows: 1) delete the guidance subnetwork; 2) change the first convolution layer of the target subnetwork so that it can take the concatenated target and guidance images as input; and 3) enlarge the number of channels to make the parameters of Model7 roughly equal with our model. As shown in Table VII, Model7 is worse than our method. Because of the modality gap between the target and the guidance images, directly feeding the concatenated result may let the network overlook incompatible problems between modalities. In the following, we will provide a detailed analysis of each component of our method.

C. Effectiveness of AKL

In this article, 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 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 the kernels from the target image only, and thus, the reconstruction accuracy is relatively low. With the assistance of the 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), which indicates that constructing kernels from both target and guidance images enjoys some advantages over using only the guidance. Nevertheless, using elementwise 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 (14). As shown 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 $\boldsymbol{I}^h - \boldsymbol{I}^{\text{out}}$. As shown in Fig. 11, the result of

TABLE VII

ABLATION STUDY. QUANTITATIVE COMPARISON OF DIFFERENT COMPONENTS FOR 16× DEPTH IMAGE SR. WE CHOSE RMSE AS THE EVALUATION METRIC, AND THE LOWER VALUES INDICATE BETTER PERFORMANCE

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	✓	✓				✓	✓	5.01	6.47	8.31	11.09	7.72
DAGF	✓	✓			✓	✓	✓	**4.75**	**6.16**	**7.81**	**10.64**	**7.34**

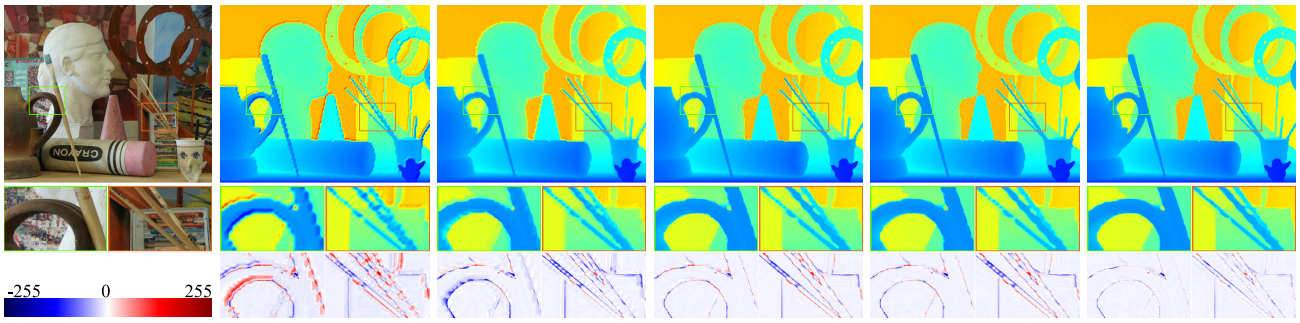


Fig. 11. Ablation Study. Visual comparison of an example without and with the proposed AKL module for depth image SR. The first row is the super-resolved depth images and the last row is the error map ($\boldsymbol{I}^h - \boldsymbol{I}^{\text{out}}$). Please enlarge the PDF for more details. (a) Guidance. (b) Model1. (c) Model2. (d) Model3. (e) Model4. (f) Model5.

TABLE VIII

ABLATION STUDY. QUANTITATIVE COMPARISONS OF DIFFERENT VALUES OF w_2. w_3 IS SET AS 0

	$\omega_2 = 0$	$\omega_2 = 1$	$\omega_2 = 10$	$\omega_2 = 20$
RMSE	8.49	8.42	**8.36**	8.38

TABLE IX

ABLATION STUDY. QUANTITATIVE COMPARISONS OF DIFFERENT VALUES OF w_3. w_2 IS SET AS 0

	$\omega_3 = 0$	$\omega_3 = 1$	$\omega_3 = 10$	$\omega_3 = 20$
RMSE	8.49	**7.92**	8.04	8.14

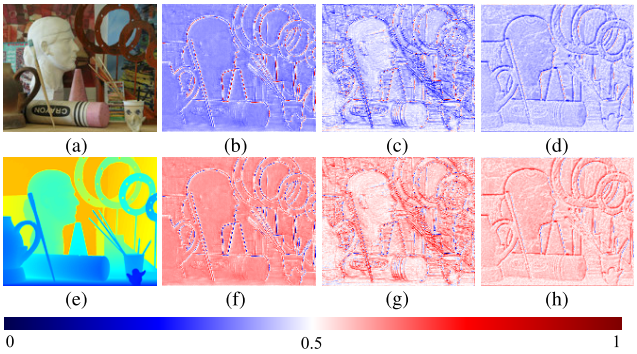


Fig. 12. Ablation study. Visualization of the learned multiscale attention maps ($\boldsymbol{A}_i$ represents the ith learned attention map [see (14)]. We resize the attention map to the same size for better visualization. (a) Guidance. (b) $\boldsymbol{A}_0$. (c) $\boldsymbol{A}_1$. (d) $\boldsymbol{A}_2$. (e) Target. (f) $\mathbf{1} - \boldsymbol{A}_0$. (g) $\mathbf{1} - \boldsymbol{A}_1$. (h) $\mathbf{1} - \boldsymbol{A}_2$.

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 that 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. Due to the proposed AKL 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. 12 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. 12, the structure regions are lighter than the texture regions, and this indicates that our model can adaptively select relevant information from the guidance image while avoiding texture overtransfer issues.

D. 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 ($\mathcal{L}_{\text{ba}}$). Table VII shows that $\mathcal{L}_{\text{ba}}$ loss is helpful in improving the reconstruction accuracy. Fig. 13 presents an example visual comparison with and without the $\mathcal{L}_{\text{ba}}$ loss. Obviously, $\mathcal{L}_{\text{ba}}$ further improves visual quality,

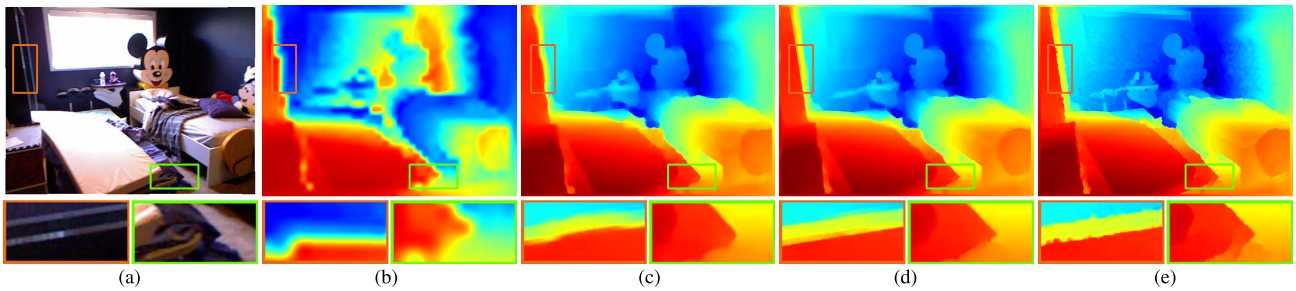


Fig. 13. Ablation study. Visual comparison of an example without and with the proposed boundary-aware loss for depth image SR. (a) Guidance. (b) Target. (c) w/o $\mathcal{L}_{\text{ba}}$. (d) w/ $\mathcal{L}_{\text{ba}}$. (e) GT.

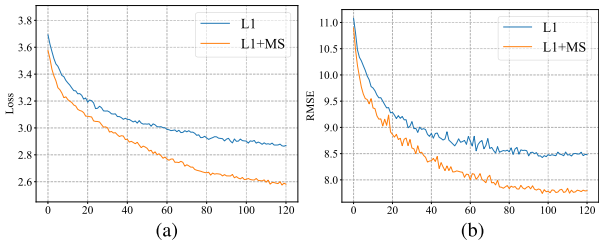


Fig. 14. Ablation study. (a) Training loss and (b) validation RMSE values on the NYU v2 dataset [42] for 16× depth image SR. MS denotes the proposed multistage loss $\mathcal{L}_{\text{ms}}$.

yielding more precise edges. The boundaries on the door frame and corner of the mattress are sharper and clearer, which verifies the effectiveness of the proposed boundary-aware loss.

E. Effectiveness of Multiscale Fusion and Deep Supervision

In this article, we propose a multiscale framework for guided image filtering. Specifically, in order to obtain both high-level structure information and low-level details, we propose to fuse multilevel filtered outputs. Moreover, a multistage loss is introduced to enforce the intermediate results to be close to the ground-truth target image. The quantitative results are shown in Table VII. As expected, Model6 trained with a hybrid loss of $\mathcal{L}_1$ and $\mathcal{L}_{\text{ms}}$ further improves the reconstruction accuracy. Fig. 14 further shows the train (left) and test (right) RMSE plots. We observe that the multistage loss ($\mathcal{L}_{\text{ms}}$) is able to accelerate convergence velocity and produce results with lower RMSE values.

F. Effectiveness of Guidance Branch

The general principle of guide image filtering is that we can transfer the valuable structures contained in the guidance image to the target image. Recently, various approaches have been proposed for guided image filtering. Nevertheless, most of them focus on designing an 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 GSR and noisy depth SR to evaluate the role of the guidance image. The RMSE comparisons are shown in Fig. 15. Model8 means the DAGF variant that takes (target, target) as inputs for kernel generation, and Model9 denotes the DAGF variant that takes (guidance, guidance) as inputs for kernel generation. The results show that the guidance image can provide significant assistance for the 8× and 16× cases, and the vanilla DAGF achieves the best performance. However, for the 4× case, the guidance information has a negligible effect. The main reason is that the target image is not severely

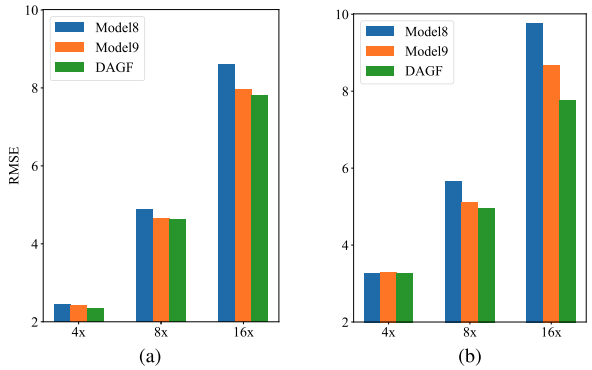


Fig. 15. Ablation study. Average RMSE values for (a) depth image SR and (b) joint depth image SR and denoising.

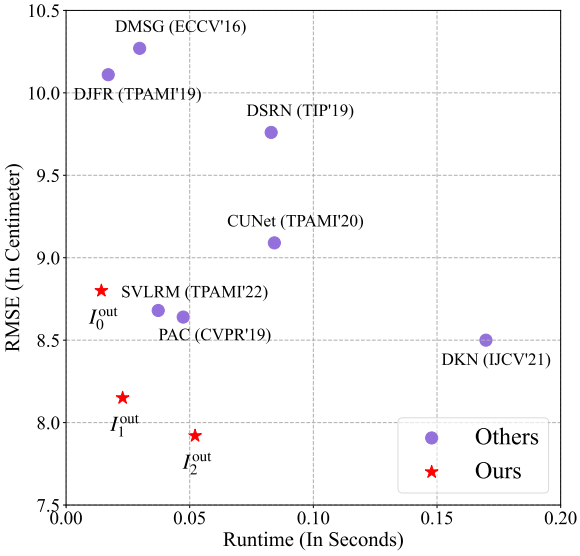


Fig. 16. Average runtime (in seconds) and RMSE comparison for 16× depth image SR on the NYU v2 dataset [42]. All the runtimes are evaluated on the same NVIDIA 1080Ti GPU with a depth image size of 480 × 640.

damaged by downsampling degradation; therefore, the target image can be easily recovered by Model8. For the more difficult cases (8× and 16×), the target image is badly polluted, and the guidance image would play an important role in the reconstruction process.

G. Performance Versus Complexity Analysis

In Fig. 16, we compare the running time among our method and other comparison methods on NYU v2 [42] for 16× depth image SR. For a 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^{out}, I_1^{out}, and I_2^{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 shown in Fig. 16, the final result I_2^{out} achieves the best RMSE result than DKN [6] and CUNet [3] but needs less time. The time cost for I_0^{out} is the least, and the performance of I_0^{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 and, otherwise, reduce the level of pyramid. Overall, our method can achieve a better tradeoff between the reconstruction performance and computational complexity.

VI. Conclusion

In this article, 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 AKL module is proposed to generate dual sets of filter kernels from the guidance and target images and then adaptively combine the learned kernels in a learning manner. Furthermore, a multiscale 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 multiscale fusion with deep supervision to regularize and combine multiple filtering results. Finally, the 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.

References

[1] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[2] X. Shen, C. Zhou, L. Xu, and J. Jia, "Mutual-structure for joint filtering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3406–3414.

[3] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2021.

[4] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Joint image filtering with deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1909–1923, Aug. 2019.

[5] H. Su, V. Jampani, D. Sun, O. Gallo, E. Learned-Miller, and J. Kautz, "Pixel-adaptive convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11166–11175.

[6] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *Int. J. Comput. Vis.*, vol. 129, pp. 579–600, Oct. 2021.

[7] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.

[8] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3443–3458, Aug. 2014.

[9] M. M. Ibrahim, Q. Liu, R. Khan, J. Yang, E. Adeli, and Y. Yang, "Depth map artefacts reduction: A review," *IET Image Process.*, vol. 14, no. 12, pp. 2630–2644, Oct. 2020.

[10] X. Ye et al., "Depth super-resolution via deep controllable slicing network," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1809–1818.

[11] J. Yang, W. Xu, X. Ye, P. Frossard, and K. Li, "Graph based non-uniform sampling and reconstruction of depth maps," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2324–2328.

[12] J. Yang, Z. Jiang, X. Ye, and K. Li, "Depth super-resolution with color guidance: A review," in *RGB-D Image Analysis and Processing*, P. L. Rosin, Y.-K. Lai, L. Shao, and Y. Liu, Eds. Cham, Switzerland: Springer, 2019, pp. 51–65, doi: 10.1007/978-3-030-28603-3_3.

[13] J. Yang, X. Ye, K. Li, and C. Hou, "Depth recovery using an adaptive color-guided auto-regressive model," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 158–171.

[14] Z. Jiang, H. Yue, Y.-K. Lai, J. Yang, Y. Hou, and C. Hou, "Deep edge map guided depth super resolution," *Signal Process., Image Commun.*, vol. 90, Jan. 2021, Art. no. 116040.

[15] J. Yang, X. Ye, and P. Frossard, "Global auto-regressive depth recovery via iterative non-local filtering," *IEEE Trans. Broadcast.*, vol. 65, no. 1, pp. 123–137, Mar. 2018.

[16] Z. Xie, X. Yu, X. Gao, K. Li, and S. Shen, "Recent advances in conventional and deep learning-based depth completion: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 2, 2022, doi: 10.1109/TNNLS.2022.3201534.

[17] Y. Zhao, M. Elhousni, Z. Zhang, and X. Huang, "Distance transform pooling neural network for LiDAR depth completion," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 13, 2021, doi: 10.1109/TNNLS.2021.3129801.

[18] H. Yoshino, C. Dong, Y. Washizawa, and Y. Yamashita, "Kernel Wiener filter and its application to pattern recognition," *IEEE Trans. Neural Netw.*, vol. 21, no. 11, pp. 1719–1730, Nov. 2010.

[19] 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, Germany: Springer, 2019, pp. 110–115.

[20] X. Deng and P. Dragotti, "Deep coupled ISTA network for multi-modal image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 1683–1698, 2020.

[21] P. Song, X. Deng, J. F. C. Mota, N. Deligiannis, P. L. Dragotti, and M. R. D. Rodrigues, "Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 57–72, 2020.

[22] L. Xu, Q. Yan, Y. Xia, and J. Jia, "Structure extraction from texture via relative total variation," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, 2012.

[23] L. Karacan, E. Erdem, and A. Erdem, "Structure-preserving image smoothing via region covariances," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–11, Nov. 2013.

[24] D. Fan, Z. Lin, Z. Zhang, M. Zhu, and M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, Jun. 2021.

[25] J. Jiang, J. Liu, J. Fu, X. Zhu, Z. Li, and H. Lu, "Global-guided selective context network for scene parsing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1752–1764, Apr. 2022.

[26] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.

[27] S. Yang, K. Zhang, and M. Wang, "Learning low-rank decomposition for pan-sharpening with spatial–spectral offsets," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3647–3657, Aug. 2018.

[28] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 667–675.

[29] C. Zhang, Y. Tang, C. Zhao, Q. Sun, Z. Ye, and J. Kurths, "Multitask GANs for semantic segmentation and depth completion with cycle consistency," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 12, pp. 5404–5415, Dec. 2021.

[30] Z. Yan et al., "Learning complementary correlations for depth super-resolution with incomplete data in real world," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 26, 2022, doi: 10.1109/TNNLS.2022.3208330.

[31] Z. Wang, X. Ye, B. Sun, J. Yang, R. Xu, and H. Li, "Depth upsampling based on deep edge-aware learning," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107274.

[32] X. Ye et al., "PMBANet: Progressive multi-branch aggregation network for scene depth super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 7427–7442, 2020.

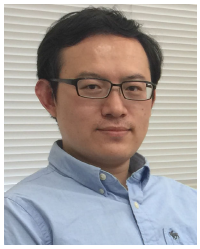
[33] J. Dong, J. Pan, J. S. Ren, L. Lin, J. Tang, and M.-H. Yang, "Learning spatially variant linear representation models for joint filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8355–8370, Nov. 2022.

[34] R. D. Lutio, S. D'aronco, J. D. Wegner, and K. Schindler, "Guided super-resolution as pixel-to-pixel transformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8828–8836.

[35] J. Kwak and D. Son, "Fractal residual network and solutions for real super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2114–2121.
[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[39] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[41] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. NIPS-W*, 2017, pp. 1–4.
[42] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.
[43] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.
[44] J. Diebel and S. Thrun, "An application of Markov random fields to range sensing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 291–298.
[45] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 993–1000.
[46] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1838–1847.
[47] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 353–369.
[48] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, May 2019.
[49] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.
[50] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.
[51] S. Lu, X. Ren, and F. Liu, "Depth enhancement via low-rank matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3390–3397.
[52] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[53] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[54] B. Ham, M. Cho, and J. Ponce, "Robust guided image filtering using nonconvex potentials," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 192–207, Jan. 2018.
[55] S. Gu, W. Zuo, S. Guo, Y. Chen, C. Chen, and L. Zhang, "Learning dynamic guidance for depth image enhancement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3769–3778.
[56] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[57] P. Krhenbühl and V. Koltun, *Efficient Inference in Fully Connected CRFs With Gaussian Edge Potentials*. Red Hook, NY, USA: Curran Associates, 2012.
[58] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
[59] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.

Zhiwei Zhong received the B.S. degree in computer science from Heilongjiang University, Harbin, China, in 2017. He is currently pursuing the Ph.D. degree in computer science with the Harbin Institute of Technology (HIT), Harbin.

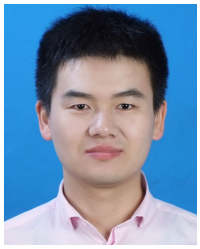
His research interests include image processing, computer vision, and deep learning.



Xianming Liu (Member, IEEE) 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, Hamilton, ON, 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 IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, and IEEE TRANSACTIONS ON MULTIMEDIA, and top conferences, such as International Conference on Machine Learning (ICML), Computer Vision and Pattern Recognition Conference (CVPR), International Joint Conferences on Artificial Intelligence (IJCAI).

Dr. Liu was a recipient of the IEEE International Conference on Multimedia and Expo (ICME) 2016 Best Student Paper Award.



Junjun Jiang (Member, IEEE) 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 the 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. His research interests include image processing and computer vision.

Dr. Jiang won the Finalist of the World's FIRST 10K Best Paper Award at ICME 2017 and the Best Student Paper Runner-Up Award at Magnetism and Magnetic Materials (MMM) 2015. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and the 2015 ACM Wuhan Doctoral Dissertation Award.

Debin Zhao (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1985, 1988, and 1998, respectively.

He is currently a Professor with the Department of Computer Science, Harbin Institute of Technology. He has published over 200 technical papers in refereed journals and conference proceedings in the areas of image and video coding, video processing, video streaming and transmission, and pattern recognition.

Xiangyang Ji (Member, IEEE) 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, in 2008.

He joined Tsinghua University, Beijing, in 2008, where he is currently a Professor with the Department of Automation, School of Information Science and Technology.