Chinese Journal of Electronics
vol. 33, no. 1, pp. 303–312, January 2024
https://doi.org/10.23919/cje.2022.00.414

RESEARCH ARTICLE

Deep Guided Attention Network for Joint Denoising and Demosaicing in Real Image

Tao ZHANG[1], Ying FU[1,2], and Jun ZHANG[2]

1. *School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
2. *Advanced Reasearch Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China*

Corresponding author: Ying FU, Email: fuying@bit.edu.cn

Manuscript Received December 5, 2022; Accepted March 16, 2023

Abstract — Denoising (DN) and demosaicing (DM) are the first crucial stages in the image signal processing pipeline. Recently, researches pay more attention to solve DN and DM in a joint manner, which is an extremely undetermined inverse problem. Existing deep learning methods learn the desired prior on synthetic dataset, which limits the generalization of learned network to the real world data. Moreover, existing methods mainly focus on the raw data property of high green information sampling rate for DM, but occasionally exploit the high intensity and signal-to-noise (SNR) of green channel. In this work, a deep guided attention network (DGAN) is presented for real image joint DN and DM (JDD), which considers both high SNR and high sampling rate of green information for DN and DM, respectively. To ease the training and fully exploit the data property of green channel, we first train DN and DM sub-networks sequentially and then learn them jointly, which can alleviate the error accumulation. Besides, in order to support the real image JDD, we collect paired raw clean RGB and noisy mosaic images to conduct a realistic dataset. The experimental results on real JDD dataset show the presented approach performs better than the state-of-the-art methods, in terms of both quantitative metrics and qualitative visualization.

Keywords — Image denoising, Image demosaicing, Joint processing, Guided attention, Paired real dataset.

Citation — Tao ZHANG, Ying FU, Jun ZHANG, “Deep Guided Attention Network for Joint Denoising and Demosaicing in Real Image,” *Chinese Journal of Electronics*, vol. 33, no. 1, pp. 303–312, 2024. doi: 10.23919/cje.2022.00.414.

I. Introduction

Most digital camera employs a single CCD/CMOS sensor to capture natural scenes. Due to 2D sensor, color filter array (CFA) is always employed to filter two-thirds of RGB information, such as the most famous Bayer pattern. Besides, due to the limitation of imaging circuit, the rest one-third of information, i.e., mosaic image, is always corrupted by various noise. Thus, recovering a high-quality color image is a highly ill-posed problem. Image denoising (DN) and demosaicing (DM) are the first crucial steps of image signal processing (ISP) pipeline in most digital camera, and their performance has vital influence on the visual appearance and downstream application of final result [1].

Due to the modular design of traditional ISP, DN and DM are independently and sequentially handled. However, it leads to error accumulation and sub-optimal recovery. Either DN needs to handle non-linear and diverse noises introduced by DM, or DM suffers from unreliable samples caused by DN. To solve this problem, researches recently focus on the joint DN and DM (JDD) image restoration and show its advantages [2], e.g., high performance and low computational complexity.

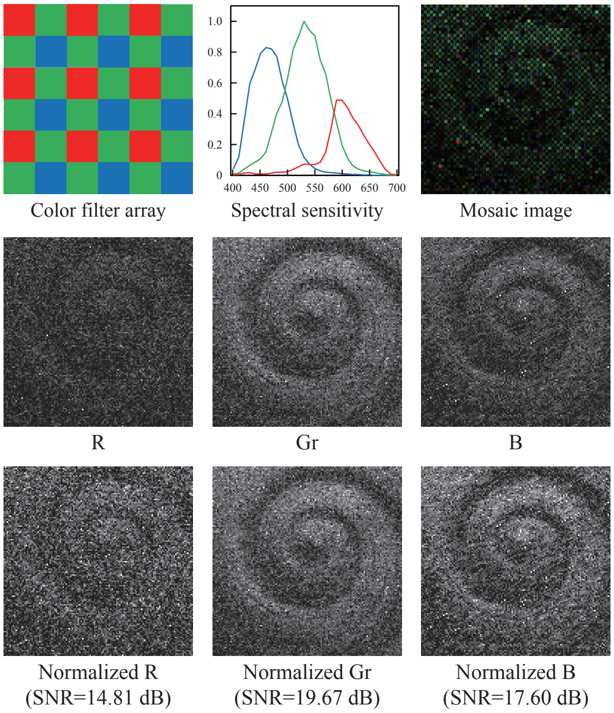
Since this restoration task is undetermined, diverse image priors are required to assist the reconstruction. Traditional methods usually solve an optimization function in an iterative manner with embedding hand-crafted priors, e.g., total variation [3], nonlocal self-similarity [4], [5]. However, the complex data in the real word cannot be sufficiently characterized by the hand-crafted priors, and there are still a number of visually disturbing artifacts appearing on some challenging high frequency regions, e.g., checkerboard and moire patterns [1].

Recently, instead of hand-crafted prior, deep learning methods [1], [6]–[8] automatically learn the desired prior with convolutional neural network (CNN). Most

Associate Editor: Prof. Chenggang YAN, Hangzhou Dianzi University.

approaches [1], [6], [7] brutally learn a mapping network between noisy image, i.e., mosaic image or decomposed four-channel RGGB image, and clean RGB image to exploit intra- and inter-channel correlation and complete the missing information. Considering the high sampling rate of green channel, Liu *et al.* [8] additionally introduced a green channel recovery branch to guide RGB image restoration. Nevertheless, all these methods are trained on synthetic data [1], [9]–[12], and CFA and Gaussian noise are utilized to synthesize the mosaic image. Due to the domain gap between synthetic and real mosaic image, these methods cannot generalize well on real raw data with complex noises.

Besides, most existing methods mainly focus on considering the raw data property for DM [8], i.e., higher sampling rate of green information, but rarely consider the raw data characteristic for DN. As human eyes can perceive green more sensitively than red and blue [13], the camera spectral sensitivity of green is designed to be larger than red and blue, which leads to higher intensity and signal-to-noise ratio (SNR) of green channel, as shown in Figure 1. These mean, due to high sampling rate and high SNR, the green channel is easy to be recovered not only for DM but also for DN.



Color filter array Spectral sensitivity Mosaic image

R Gr B

Normalized R (SNR=14.81 dB) Normalized Gr (SNR=19.67 dB) Normalized B (SNR=17.60 dB)

Figure 1 High SNR and sampling rate of green channel. Green channel has twice sampling rate than red and blue channels. Camera spectral sensitivity of green is higher than that of red and blue, which leads to higher intensity and SNR. The second line shows higher intensity and the third line shows higher SNR of green channel, respectively. Note that the Gb channel is similar to Gr.

In this work, we present a deep guided attention network (DGAN) for real image JDD, which respectively considers the high SNR and high sampling rate of green information for DN and DM, as shown in Figure 2. The network architecture of DGAN is based on UNet [14], and involves green channel guidance branch with multiple guided attention modules and decomposition and combination learning strategy. Inspired by guided filter, we design a guided attention module in local manner to adaptively generate attentive kernel weights for different spatial positions by modeling the interdependencies of more completed green channel feature in the neighborhood. To ease the learning of JDD network and fully exploit data property of green channel, we decompose JDD network into two sub-networks, where the former focuses on DN with high SNR green channel guidance and the latter takes charge of DM with high sampling rate green channel guidance. Two sub-networks are trained sequentially first, and then are combined into a whole network for jointly training to reduce the error accumulation. Besides, to support the JDD in the real world, we utilize an advanced pixelshift camera to collect a real raw dataset

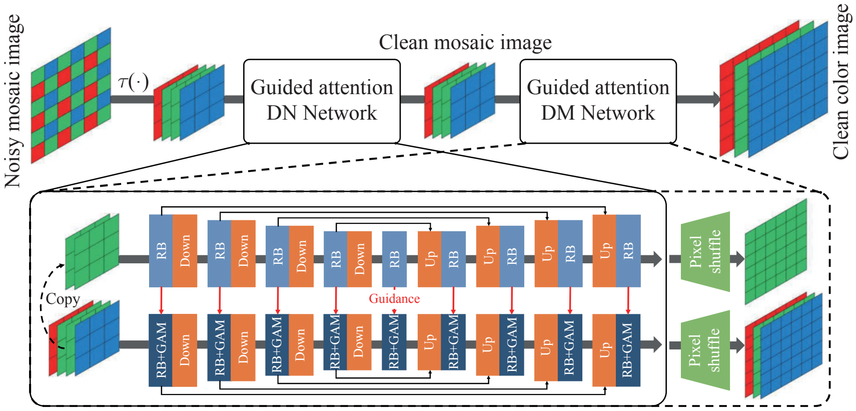


Figure 2 The architecture of proposed guided attention network, consisting of DN and DM two sub-networks. Each sub-network employs Unet as the fundamental network and residual block (RB) as basic module. We utilize green channel with high SNR and sampling rate to respectively guide the denoising and demosaicing with multiple guided attention modules (GAMs), which employ the green channel feature in the corresponding depth as guidance information. The network architecture of green channel branch is the same as the main branch, except with half feature maps. Compared with DN sub-network, DM sub-network additional utilize pixel-shuffle layers to upsample the resolution.

with paired clean full color RGB, noisy and clean mosaic images. The experimental results on real JDD dataset show that the presented approach performs better than the state-of-the-art methods, in terms of both quantitative metrics and qualitative visualization.

Our main contributions are as follows:

- We present a deep guided attention network for real image JDD, that effectively considers the green channel characteristics of high SNR and high sampling rate in raw data.
- We propose a guided attention module to adaptively guide RGB image restoration by the information in green channel recovery branch.
- We collect a real raw JDD dataset with paired noisy mosaic, clean mosaic and clean full color RGB images, and utilize a decomposition-and-combination training strategy to make the trained network more practical to the real data.

II. Related Work

The most related researches on joint image denoising and demosaicing, and guided image recovery are reviewed in this section.

1. Joint denoising and demosaicing

The aim of image DM is recovering a RGB image from a mosaic image losing two-thirds information. Various traditional methods [15]–[23] and deep learning methods [24] have been presented. Besides, due to noise commonly existing in the real world, DM methods usually need collaboration of DN methods [25], [26], and process the noisy mosaic image sequentially. Because of error accumulation, it leads to sub-optimal recovery. To solve this problem, researchers recently pay more attention on JDD image restoration and show the benefits, which can achieve higher performance and lower computational complexity [2].

Since image JDD is an extremely undetermined problem, diverse image priors are required to assist the recovery. The conventional optimization methods [3]–[5] integrate hand-crafted priors to iterative optimization algorithm, and restore the clean RGB image from noisy mosaic image. Condat *et al.* [3] integrated the total variation prior to a primal dual optimization algorithm. Heide *et al.* [4] presented an optimization method with nonlocal prior to recover color image. Tan *et al.* [5] employed alternating direction method of multipliers (ADMM) with nonlocal and total variation priors to recover color image.

Alternatively, deep learning methods [1], [6]–[8], [27] employ advanced convolutional neural networks to exploit the desired prior for JDD task. Gharbi *et al.* [1] recovered color image from noisy mosaic image with a deep convolutional neural network. Tan *et al.* [6] employed a convolutional neural network to refine the initialized color image via bilinear interpolation. Kokkinos *et al.* [7] integrated a residual DN network into unrolled majorization-minimization method for color image recovery. Xing *et al.* [27] discussed the effect of DN and DM processing order, and presented an end-to-end network for image JDD. Liu *et al.* [8] introduced additional green channel and density map guidances to design a self-guidance network for color image recovery.

The conventional methods utilize hand-crafted priors that are often limited linear characteristic and cannot sufficiently employ the image nonlinearity. The deep learning methods brutally learn the implicit mapping function from noisy mosaic to clean RGB images, but do not well consider the high SNR and high sampling rate data properties of green channel. In this work, we present a deep learning method to exploit deep prior with attentive green channel guidance for image JDD.

2. Guided image recovery

Guided image recovery utilizes auxiliary prior to assist image restoration. Guided filter [28], [29] is a well-known method that employs an additional image as guidance to generate filter weights and has been successful in many image recovery tasks, e.g., image demosaicing [30]. Recently, deep learning methods [31]–[40] have employed various auxiliary information to guided image recovery, particularly for super-resolution. Some methods [31], [32], [35] utilized RGB image as the auxiliary knowledge to guided the super-resolution of depth or hyperspectral image. Wang *et al.* [36] super-resolved the image with semantic information guidance. Zou *et al.* [34] super-resolved the image with cross-scale stereo information guidance.

Besides, self-guidance network [41] is presented for image DN, that employed the low resolution features to enhance the high resolution feature. Liu *et al.* [8] utilized the green channel property of high sampling rate and further designed a green channel sub-network to guided image JDD, where guidance information is fused with main information at the end of branches. Inspired by guided filter, we propose a guided attention module to adaptively fuse main information with guidance information. In addition, to fully exploit the guidance information in green channel, we interpolate the guided attention module into network with different depths.

III. Guided Attention Network

Firstly, we formulate the problem of joint image DN and DM, and introduce the motivation of DGAN. Then, we describe the guided attention module, that adaptively guides RGB image recovery by information in green channel recovery branch. Finally, the architecture of DGAN and the corresponding decomposition-and-combination training strategy are described. The decomposed sub-networks can effectively consider the high SNR and high sampling rate raw data properties for DN and DM, respectively.

1. Formulation and motivation

The JDD aims to handle mosaic image $Z \in \mathbb{R}^{1\times H\times W}$

corrupted with noise $n \in \mathbb{R}^{1\times H\times W}$ and recover clean RGB image $X \in \mathbb{R}^{3\times H\times W}$. H and W denote height and width of images, respectively. It can express the relationship of noisy mosaic and clean RGB images as

$$Z = Y + n = \mathcal{M}(X) + n \tag{1}$$

where Y denotes clean mosaic image and $\mathcal{M}$ is the mosaic mapping function. n includes various noise and is not limited to signal-independent Gaussian noise.

Numerous researches show that human eyes are more sensitive to green than red and blue [13]. Therefore, the CFA, e.g., Bayer pattern, in modern digital cameras are designed with higher sampling rate and spectral sensitivity of green than others, as shown in Figure 1. The higher spectral sensitivity causes higher intensity of captured green channel and the most noises in acquisition are signal-independent noise, which leads to higher SNR of green channel. Accordingly, the green channel is easier to be recovered not only for DM but also for DN.

In this paper, we first employ two networks with green channel guidance to respectively deal with DN and DM, and then fine-tune them in a joint manner. Concretely, we present a guided attention module, in which attention map is generated from guided information of green channel and is adaptive for each spatial position. We plug the guided attention module into network to recover color information with progressive green channel guidance.

2. Guided attention module

Before introducing guided attention module, we first review the guided filter [28], which has been widely used in image DN [42] and DM [30]. Guided filter is a translation-variant filter, involving a guidance image G, an input image I, and an output image O. Given the guidance image and input image, the output at the i-th pixel can be represented as

$$O_i = \sum_{j\in\mathcal{N}(i)} W_{ij}(G) I_j \tag{2}$$

where i and j denote pixel indexes and $\mathcal{N}(j)$ are the neighboring pixels of j. The filter weight W_{ij} is a transformation of the guidance image G and independent of input image I.

Inspired by the guided filter, we presented a guided attention module, in which the attention map is generated from the correlation between the features of guidance information in current position and its neighborhood, as shown in Figure 3. We first employ two 1×1 convolutional layers f_I and f_G to embed the input and guidance features with the same channels, and the embedded features are denoted as $I' = f_I(I)$ and $G' = f_G(G)$, respectively. Then, a certain element at position i in guidance feature queries the correlation with elements in its neighborhood $\mathcal{N}(i)$. Supposing the position of element in neighborhood is $j \in \mathcal{N}(i)$, the correspondence map A' can be expressed as

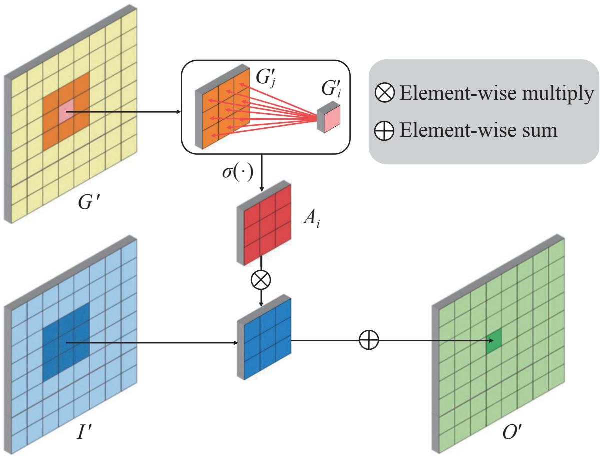


Figure 3 The guided attention module. We first employ f_G and f_I to embed the guidance and input features. Then, inspired by guided filter [28], we utilize the guidance information G' to generate local attention map A_i, which attentively filters the input information. Finally, the attended feature is mapped to input space with f_O and added with input feature.

$$A'_{ij} = {G'}_i^T G'_j \tag{3}$$

where ${G'}_i^T \in \mathbb{R}^{1\times C}$ and $G'_j \in \mathbb{R}^{C\times 1}$. The final guided attention map A_i is

$$A_i = \sigma(A'_i) \tag{4}$$

where σ is the Softmax function and forces the guided attention map sum-to-one. Specifically, the each elements calculation in A_i can be expressed as

$$A_{ij} = \frac{e^{A'_{ij}}}{\sum_{k\in\mathcal{N}(i)} e^{A'_{ik}}} \tag{5}$$

The guided attention map A is corresponding to guided filter kernel W, they are both generated by guidance information and independent of input information. Given the guided attention map A and embedded input feature I', we can obtain the attended feature in position i as

$$O'_i = \sum_{j\in\mathcal{N}(i)} A_{ij} I'_j \tag{6}$$

Finally, we employ a 1×1 convolutional layer f_O to project the attended feature to original dimension and add it with input feature. It can be represented as

$$O = f_O(O') + I \tag{7}$$

Comparing with popular self-attention [43], the guided attention map is calculated by green channel information. Due to high SNR and high sampling rate, the information of green channel is more complete than other channels. It leads to more accurate attention map calculation. Besides, taking the neighborhood with spatial size $K\times K$ as an example. Comparing conventional global attention in the vanilla transformer structure [43], the proposed guided attention reduce the memory occupation of attention map from $H \times W \times H \times W$ to $H \times W \times K \times K$, and the computational complexity from $\mathcal{O}(H^2W^2C)$ to $\mathcal{O}(HWK^2C)$. As K is much smaller than H and W, the proposed guided attention is more efficient than conventional attention.

3. Network architecture

To ease the network training, we decompose the JDD network into DN and DM two sub-networks, as shown in Figure 2.

These two sub-networks have almost the same architecture, and the main difference is that the DN sub-network employs a convolutional layer to output the denoised data, and DM sub-network additionally employs a pixelshuffle layer to upsample the spatial resolution. As numerous researches [12], [27] show that applying DN first and DM later outperforms the opposition, we first employ the DN sub-network and feed its output through DM sub-network to obtain the clean full color RGB image.

Each sub-network consists of green channel guidance branch and main branch. Both branches are based on the same representative Unet [14] architecture. The feature maps of green channel guidance branch is half of that of main branch. Each branch has 4 encoder steps and 4 corresponding decoder steps. After each encoder step, a convolution layer with 4×4 kernel size and 2 stride is employed to downsample the feature maps in $1/2\times$ scale. Before each decoder step, a deconvolution layer with 2×2 kernel size and 2 stride is utilized to upsample the feature maps in $2\times$ scale. Besides, feature maps in encoder are passed to its corresponding decoder stage through skip connections. In each encoder or decoder step, there is a residual block with two 3×3 convolution layers and an additional 1×1 convolution layer.

The existing method [8] only employs output feature of green channel branch once to guide the information recovery at the end of main branch. To fully exploit the guidance information, we utilize green channel features to guide main branch restoration with multiple times. Specifically, we interpolate the guided attention module after each residual blocks, and the main branch is guided by the features in the corresponding depth.

4. Learning strategy

As we decompose the JDD network into DN and DM two sub-networks, we present a decomposition-and-combination learning strategy to train the networks and obtain a higher recovery accuracy. Our learning strategy can be divided into three steps, including decomposed DN training, decomposed DM training and combined DN and DM training.

Firstly, we train the DN sub-network with paired clean and noisy mosaic images. Following previous works [8], we decompose the mosaic image into RGGB four channels. The L_1 error is utilized as loss function, which can be expressed as

$$\mathcal{L}_{DN}(\theta_{DN}) = ||\tau(Y) - f_{DN}(\tau(Z);\theta_{DN})||_1 \tag{8}$$

where τ, f_{DN} and θ_{DN} denote decomposition transformation, the mapping function of DN sub-network and the corresponding parameters, respectively.

Secondly, we fix the parameters θ_{DN} of DN sub-network, and train the DM sub-network. Given the pretrained DN sub-network, we can obtain a pre-denoised four channel mosaic image $\tau(\hat{Y})$. Feeding $\tau(\hat{Y})$to the DM sub-network, we want to get a full color RGB image. To this end, we train the DM sub-network with the following loss function

$$\mathcal{L}_{DM}(\theta_{DM}) = ||X - f_{DM}(\tau(\hat{Y});\theta_{DM})||_1 \tag{9}$$

where f_{DN} and θ_{DN} denote the mapping function of DM sub-network and the corresponding parameters, respectively.

Thirdly, we combine the DN and DM sub-networks, and train them in a joint manner. Given the networks trained in previous steps, we fine-tune them jointly, which can be represented as

$$\mathcal{L}_J(\theta_{DN},\theta_{DM}) = ||X - f_{DM}(f_{DN}(\tau(Z);\theta_{DN});\theta_{DM})||_1 \tag{10}$$

Apart from RGB recovery loss for the main branch, we add corresponding green channel recovery loss to equations (8)–(10) with balance parameter λ for different learning steps. The total loss of each learning step can be expressed as

$$\mathcal{L} = \mathcal{L}_M + \lambda\mathcal{L}_G \tag{11}$$

where $\mathcal{L}_M$ is the main branch loss, i.e., $\mathcal{L}_{DN}$, $\mathcal{L}_{DM}$ or $\mathcal{L}_J$, and $\mathcal{L}_G$ is the corresponding green channel branch loss, respectively.

IV. Paired Real Raw JDD Dataset

The existing deep learning JDD methods need to be learned on training datasets [1], [9]–[12]. Existing datasets for JDD have several problems. The sRGB datasets [1], [9], [10] are nonlinearly processed and demosaiced by existing DM algorithm, which mismatches the linear working space of DM approaches and introduces undesirable artifacts. The linear RGB datasets [11] are generated by raw mosaic images, however, which might alter the characteristic of signal. Recently, Qian *et al.* [12] captured linear full color RGB images with advanced pixel shift device. However, these datasets just include clean RGB image, but still synthesize the noisy mosaic image with CFA and Gaussian noise. It introduces domain gap between synthetic image and real image with complex noises, which limits the application of learned JDD algorithms to the real data.

To support this study, we utilize a camera with pixel shift technique to collect a paired real dataset, including noisy mosaic, clean mosaic and RGB images. To capture a color RGB image, pixel shift camera takes four mosaic images, as shown in Figure 4. Each mosaic image is captured with horizontal and/or vertical sensor movement. After capturing four times, the camera can fully capture the color information of each pixel. In each full color image capturing, we can obtain four pixel shifted mosaic images and a full color RGB image. After capturing clean full color RGB image, the noisy mosaic image is required to capture. According to the work in [44], we fix the imaging setting and reduce exposure time to collect noisy mosaic image. Thus, noisy/clean mosaic and noisy/clean full color RGB images can be captured in paired manner.

For dataset capturing, we utilize an advanced pixel shift camera Sony A7R4. We mount the camera on sturdy tripods and utilize a software to remotely control it. We first adjust focus, aperture, exposure time and other camera settings to improve the definition of the clean mosaic and full color RGB images. Then, the exposure time is reduced with a factor to collect noisy images. Due to multiple acquisitions of the same scene, we strictly keep the scenes in the dataset is static. After capturing,

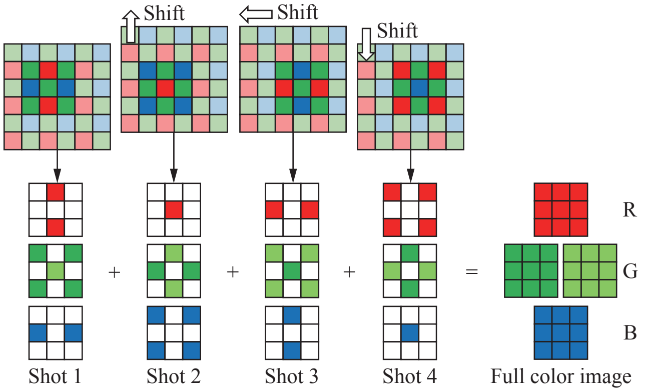


Figure 4 The working principle of pixel shift camera. The camera sensor takes four shots with physically moving in horizontal and vertical dimensions in each capturing. Then, these mosaic images are integrated to get a full color image.

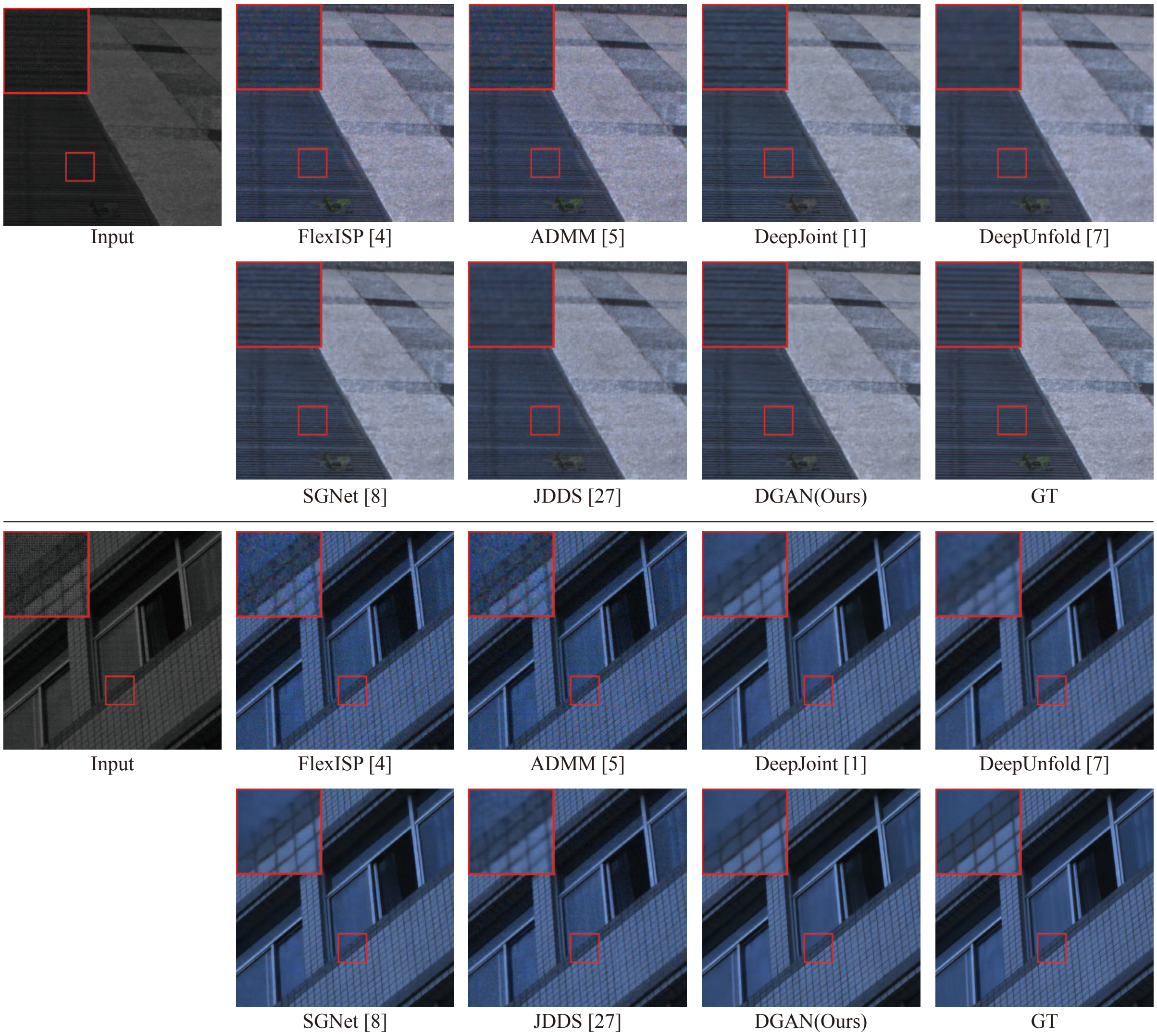


Figure 5 Visual quality comparison on two representative scenes in real JDD dataset. The input noisy image and restored results of FlexISP, ADMM, DeepJoint, DeepUnfold are shown in the first row, and the recovered results of SGNet, JDDS, DGAN and ground truth are shown in the second row.

there are 100 outdoor and indoor scenes in our dataset, whose resolution is with 9568 × 6376. We randomly select 25 scenes for testing and the rest for training.

V. Experiments

In this section, we first describe the experiment settings, such as implementation details and metrics for quantitative investigation. Besides, the proposed approach is compared with several advanced approaches on collected real raw JDD dataset. Finally, we discuss the effective of different network modules and learning strategies.

1. Settings

The window size K of guided attention module is set to 5. Following [8], the balance parameter λ is set to $30 - n_{epoch} \times 27/100$, where n_{epoch} is the number of learning epochs. During the training stage, the image in our paired real JDD dataset are randomly cropped into 256 × 256 spatial regions with overlap. We employ PyTorch for implementation. In each training step, the model is trained with Adam optimizer [45] ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 100 epochs. The initial learning rate and mini-batch size are set to 1×10^{-4} and 1, respectively.

Six state-of-the-art methods are compared with our DGAN, including two conventional methods and four deep learning methods. The conventional methods are FlexISP [4] and ADMM [5]. The deep learning methods are DeepJoint [1], DeepUnfold [7], SGNet [8] and JDDS [27]. We evaluate all methods on our paired real JDD dataset. It is worth to note that noisy map is not utilized for all deep learning methods, as real mosaic image contains various noises [46] and the noise level is difficult to be estimated.

Two evaluation metrics, *i.e.*, the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM), are utilized to investigate the performance of all algorithms. The bigger value of PSNR and SSIM means higher image quality.

2. Evaluation on real JDD dataset

Quantitative results Table 1 provides the averaged recovery results of different situations on the real JDD dataset, which quantitatively compare the performance of FlexISP, ADMM, DeepJoint, DeepUnfold, SGNet, JDDS and DGAN. We highlight the best results for each metric in bold. We can see that the proposed approach performs better than previous algorithms under both metrics. Specifically, deep learning methods exhibits remarkably higher accuracy compared with the traditional methods based on hand-crafted priors. It demonstrates the superiority of the prior modeling capability of the deep network. Compared with the deep learning methods, the proposed method fully exploits the data property of green channel, *i.e.*, high sampling rate and high SNR, and achieves better performance. It reveals the effectiveness of our deep guided attention network.

Table 1 PSNR and SSIM metrics of different algorithms on real JDD dataset

Metrics	Traditional methods		Deep learning methods				
	FlexISP [4]	ADMM [5]	DeepJoint [1]	DeepUnfold [7]	SGNet [8]	JDDS [27]	DGAN(Ours)
PSNR	30.246	30.078	41.186	41.503	42.298	42.397	**42.712**
SSIM	0.9149	0.9120	0.9832	0.9811	0.9829	0.9827	**0.9833**

Perceptual quality For visualization, we show two typical recovered scenes in Figure 5. The input noisy image and restored results of FlexISP, ADMM, DeepJoint, DeepUnfold are shown in the first row, and the recovered results of SGNet, JDDS, DGAN and ground truth are shown in the second row. The results of FlexISP and ADMM can obviously observe noise, which indicates the hand-crafted prior is inefficient for real image JDD. Our method can produce visually pleasant results with less artifact and sharper edges compared with other methods, which is consistent with quantitative results.

Computational complexity The efficiency of all deep learning methods are also quantitatively evaluated by two metrics, *i.e.*, parameters and floating-point operations (FLOPs). We show the related results in Table 2. It is worth to note that FLOPs is calculated by restoring an image with 256 × 256 resolution. We can see that our method has larger number of parameters than other methods, especially DeepJoint and DeepUnfold. It indicates that our method has more powerful capability to exploit the latent characteristic of image. Moreover, the FLOPs of DGAN is smaller than all methods, and especially compared with DeepUnfold, SGNet and JDDS in two orders of magnitude smaller. It demonstrates the efficiency of the proposed method.

Table 2 Parameters and FLOPs comparison of deep learning algorithms

Methods	Params(M)	FLOPs(G)
DeepJoint	0.56	9.39
DeepUnfold	0.38	245.60
SGNet	13.62	221.69
JDDS	6.22	399.59
DGAN(Ours)	24.86	9.18

3. Discussion

Here, we discus the effect of different guidance modules, different learning strategies, and different upsampling layers.

The effect of different guidance modules To verify the effectiveness of the green channel guidance with multiple guided attention modules, we compare it with one attention guidance, multiple concatenation guidances, one concatenation guidance and without guidance. The results are provided in Table 3, and we highlight the best results in bold. Specifically, network with green channel guidance outperforms that without guidance, which verifies the effectiveness of green channel guidance. Further, the gains of our method with multiple guidances over that with once guidance demonstrate multiple guidances can fully exploit the data property of green channel. Last but not least, our method with attention guidance is considerably better than that with concatenation guidance. It reveals the effectiveness of our guided attention module that adaptively fuses the guidance information to main branch.

Table 3 The effect of different guidance modules

Guidances	PSNR	SSIM
without guidance	41.679	0.9797
one concatenation	41.900	0.9811
multiple concatenation	42.362	0.9822
one attention	42.307	0.9821
multiple attention	**42.712**	**0.9833**

The effect of different learning strategies To evaluate the effectiveness of the decomposition-and-combination learning strategy (DN⇒DM⇒(DN→DM)), we compare it with different learning strategies, including directly end-to-end training (E2E), DN and DM separately training (DN+DM), and separately training and jointly fine-tuning ((DN+DM)⇒(DN→DM)). We provide the results in Table 4, and highlight the best results. Note that numerous researches [12], [27] show that applying DN first and DM later outperforms the opposition, so we first employ the DN sub-network and feed its output through DM sub-network to obtain the clean full color RGB image. Specifically, due to error accumulation, the DN and DM separately training performs worst. With the prior knowledge of DN and DM, (DN+DM)⇒(DN→DM) and DN⇒DM⇒(DN→DM) outperform E2E, which demonstrates the necessary of pre-training. Moreover, the gain between DN⇒DM⇒(DN→DM) and (DN+DM)⇒(DN→DM) verifies the effectiveness of our decomposition and combination learning strategies.

Table 4 The effect of different learning strategies

Strategies	PSNR	SSIM
E2E	42.542	0.9828
DN+DM	42.283	0.9821
(DN+DM)⇒(DN→DM)	42.584	0.9829
DN⇒DM⇒(DN→DM)	**42.712**	**0.9833**

The effect of different upsampling layers To upsample the spatial resolution, there mainly three operations for deep neural network, including interpolation, deconvolution and pixel-shuffle. Inspired by the advanced image super-resolution methods [47], we employ pixel-shuffle layer to upsample the spatial resolution. To evaluate the effectiveness of the pixel-shuffle upsampling layer, we compare it with interpolation and deconvolution. We provide the results in Table 5, and highlight the best results. We can see that pixel-shuffle significantly outperforms other upsampling layers, especially the interpolation layer. It verifies the superiority of pixel-shuffle layer in DM-subnetwork.

Table 5 The effect of different upsampling layers

Upsampling	PSNR	SSIM
Interpolation	41.187	0.9801
Deconvolution	42.312	0.9815
Pixel-shuffle	**42.712**	**0.9833**

VI. Conclusion

In this paper, we present a novel guided attention network for real image JDD, which considers the high SNR and high sampling rate of green information to guide the DN and DM, respectively. The designed guided attention module can adaptively guide the full color RGB image recovery, and can fully exploit the guidance of green channel by applying it multiple times in the DGAN. To ease the training, we employ a decomposition-and-combination learning strategy. Besides, we utilize pixel shift camera to collect a paired real JDD dataset containing clean RGB, clean mosaic and noisy mosaic images, making the leaned network with better generalization for the real data. The comprehensive experimental results indicate that our method have better performance than existing state-of-the-art algorithm in terms of both comprehensive quantitative metrics and visual quality. In the future, we will collect more suitable data and expand the paired real JDD dataset.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62171038, 61827901, and 62088101).

References

[1] M. Gharbi, G. Chaurasia, S. Paris, *et al.*, “Deep joint demosaicking and denoising,” *ACM Transactions on Graphics*, vol. 35, no. 6, article no. 191, 2016.

[2] K. Hirakawa and T. W. Parks, “Joint demosaicing and denoising,” *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2146–2157, 2006.

[3] L. Condat and S. Mosaddegh, “Joint demosaicking and denoising by total variation minimization,” in *Proceedings of the 19th IEEE International Conference on Image Processing*, Orlando, FL, USA, pp. 2781–2784, 2012.

[4] F. Heide, M. Steinberger, Y. T. Tsai, *et al.*, "FlexISP: A flexible camera image processing framework," *ACM Transactions on Graphics*, vol. 33, no. 6, article no. 231, 2014.

[5] H. L. Tan, X. R. Zeng, S. M. Lai, *et al.*, "Joint demosaicing and denoising of noisy Bayer images with ADMM," in *Proceedings of 2017 International Conference on Image Processing*, Beijing, China, pp. 2951–2955, 2017.

[6] D. S. Tan, W. Y. Chen, and K. L. Hua, "DeepDemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2408–2419, 2018.

[7] F. Kokkinos and S. Lefkimmiatis, "Deep image demosaicking using a cascade of convolutional residual denoising networks," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 317–333, 2018.

[8] L. Liu, X. Jia, J. Z. Liu, *et al.*, "Joint demosaicing and denoising with self guidance," in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 2240–2249, 2020.

[9] R. Timofte, S. H. Gu, J. Q. Wu, *et al.*, "NTIRE 2018 challenge on single image super-resolution: Methods and results," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, pp. 852–863, 2018.

[10] R. Timofte, E. Agustsson, L. Van Gool, *et al.*, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, HI, USA, pp. 114–125, 2017.

[11] D. Khashabi, S. Nowozin, J. Jancsary, *et al.*, "Joint demosaicing and denoising via learned nonparametric random fields," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 4968–4981, 2014.

[12] G. C. Qian, Y. H. Wang, J. J. Gu, *et al.*, "Rethinking learning-based demosaicing, denoising, and super-resolution pipeline," in *Proceedings of 2022 IEEE International Conference on Computational Photography*, Pasadena, CA, USA, pp. 1–12, 2022.

[13] J. L. Schnapf, T. W. Kraft, and D. A. Baylor, "Spectral sensitivity of human cone photoreceptors," *Nature*, vol. 325, no. 6103, pp. 439–441, 1987.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, pp. 234–241, 2015.

[15] A. Buades, B. Coll, J. M. Morel, *et al.*, "Self-similarity driven color demosaicking," *IEEE Transactions on Image Processing*, vol. 18, no. 6, pp. 1192–1202, 2009.

[16] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53–69, 2008.

[17] I. Pekkucuksen and Y. Altunbasak, "Gradient based threshold free color filter array interpolation," in *Proceedings of 2010 IEEE International Conference on Image Processing*, Hong Kong, China, pp. 137–140, 2010.

[18] G. S. Yu, G. Sapiro, and S. Mallat, "Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2481–2499, 2012.

[19] Y. Monno, D. Kiku, M. Tanaka, *et al.*, "Adaptive residual interpolation for color and multispectral image demosaicking," *Sensors*, vol. 17, no. 12, article no. 2787, 2017.

[20] L. M. Tang, Z. Fang, C. C. Xiang, *et al.*, "A variational model for staircase reduction in image denoising," *Chinese Journal of Electronics*, vol. 26, no. 2, pp. 358–366, 2017.

[21] Y. N. Xie, J. J. Huang, and Y. J. He, "One dictionary vs. two dictionaries in sparse coding based denoising," *Chinese Journal of Electronics*, vol. 26, no. 2, pp. 367–371, 2017.

[22] G. Y. Chen, G. C. Luo, L. Tian, *et al.*, "Noise reduction for images with non-uniform noise using adaptive block matching 3D filtering," *Chinese Journal of Electronics*, vol. 26, no. 6, pp. 1227–1232, 2017.

[23] R. Sadiq, M. B. Qureshi, and M. M. Khan, "De-convolution and De-noising of SAR based GPS images using hybrid particle swarm optimization," *Chinese Journal of Electronics*, vol. 32, no. 1, pp. 166–176, 2023.

[24] Z. K. Ni, K. K. Ma, H. Q. Zeng, *et al.*, "Color image demosaicing using progressive collaborative representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 4952–4964, 2020.

[25] B. Fu, Y. H. Dong, S. L. Fu, *et al.*, "Multistage supervised contrastive learning for hybrid-degraded image restoration," *Signal, Image and Video Processing*, vol. 17, no. 2, pp. 573–581, 2023.

[26] Q. Zhang, Q. Q. Yuan, M. P. Song, *et al.*, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Transactions on Image Processing*, vol. 31, pp. 6356–6368, 2022.

[27] W. Z. Xing and K. Egiazarian, "End-to-end learning for joint image demosaicing, denoising and super-resolution," in *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 3507–3516, 2021.

[28] K. M. He, J. Sun, and X. O. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[29] L. Zeng and Y. Z. Dai, "Single image dehazing based on combining dark channel prior and scene radiance constraint," *Chinese Journal of Electronics*, vol. 25, no. 6, pp. 1114–1120, 2016.

[30] Y. Monno, D. Kiku, M. Tanaka, *et al.*, "Adaptive residual interpolation for color image demosaicking," in *Proceedings of 2015 IEEE International Conference on Image Processing*, Quebec City, Canada, pp. 3861–3865, 2015.

[31] T. W. Hui, C. C. Loy, and X. O. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp. 353–369, 2016.

[32] Y. J. Li, J. B. Huang, N. Ahuja, *et al.*, "Deep joint image filtering," in *Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp. 154–169, 2016.

[33] H. T. Zheng, M. Q. Ji, H. Q. Wang, *et al.*, "CrossNet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 88–104, 2018.

[34] Y. M. Zhou, G. C. Wu, Y. Fu, *et al.*, "Cross-MPI: Cross-scale stereo for image super-resolution using multiplane images," in *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, pp. 14842–14851, 2021.

[35] Y. Fu, T. Zhang, Y. Q. Zheng, *et al.*, "Hyperspectral image super-resolution with optimized RGB guidance," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 11661–11670, 2019.

[36] X. T. Wang, K. Yu, C. Dong, *et al.*, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 606–615, 2018.

[37] S. Guo, Z. T. Liang, and L. Zhang, "Joint denoising and demosaicking with green channel prior for real-world burst images," *IEEE Transactions on Image Processing*, vol. 30, pp.

6930–6942, 2021.

[38] T. Zhang, Y. Fu, and C. Li, “Deep spatial adaptive network for real image demosaicing,” in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Palo Alto, CA, USA, pp. 3326–3334, 2022.

[39] T. Zhang, Y. Fu, and J. Zhang, “Guided hyperspectral image denoising with realistic data,” *International Journal of Computer Vision*, vol. 130, no. 11, pp. 2885–2901, 2022.

[40] Y. Fu, T. Zhang, L. Z. Wang, *et al.*, “Coded hyperspectral image reconstruction using deep external and internal learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3404–3420, 2022.

[41] S. H. Gu, Y. W. Li, L. Van Gool, *et al.*, “Self-guided network for fast image denoising,” in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, Seoul, Korea (South), pp. 2511–2520, 2019.

[42] C. L. Tsai, W. C. Tu, and S. Y. Chien, “Efficient natural color image denoising based on guided filter,” in *Proceedings of 2015 IEEE International Conference on Image Processing*, Quebec City, Canada, pp. 43–47, 2015.

[43] H. S. Zhao, J. Y. Jia, and V. Koltun, “Exploring self-attention for image recognition,” in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 10076–10085, 2020.

[44] C. Chen, Q. F. Chen, J. Xu, *et al.*, “Learning to see in the dark,” in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3291–3300, 2018.

[45] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint*, arXiv: 1412.6980, 2014.

[46] K. X. Wei, Y. Fu, J. L. Yang, *et al.*, “A physics-based noise formation model for extreme low-light raw denoising,” in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 2758–2767, 2020.

[47] Y. L. Zhang, K. P. Li, K. Li, *et al.*, “Image super-resolution using very deep residual channel attention networks,” in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 294–310, 2018.

Tao ZHANG received the B.S. degree from the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China, in 2017. He is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Beijing Institute of Technology, Beijing. His research interests include deep learning, image processing, and computational photography.
(Email: tzhang@bit.edu.cn)

Ying FU received the B.S. degree in electronic engineering from Xidian University in 2009, the M.S. degree in automation from Tsinghua University in 2012, and the Ph.D. degree in information science and technology from the University of Tokyo in 2015. She is a Professor at the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include physics-based vision, image processing, and computational photography.
(Email: fuying@bit.edu.cn)

Jun ZHANG received the B.S., M.S., and Ph.D. degrees in communications and electronic systems from Beihang University, Beijing, China, in 1987, 1991, and 2001, respectively. He was a Professor with Beihang University. He has served as the Dean for the School of Electronic and Information Engineering, and the Vice President and the Secretary for the Party Committee, Beihang University. He is currently a Professor with Beijing Institute of Technology, where he is also the Secretary. His research interests are networked and collaborative air traffic management systems, covering signal processing, integrated and heterogeneous networks, and wireless communications. He is a member of the Chinese Academy of Engineering. He has won the awards for science and technology in China many times.
(Email: buaazhangjun@vip.sina.com)