Mosaic Convolution-Attention Network for Demosaicing Multispectral Filter Array Images

Kai Feng, Yongqiang Zhao, Jonathan Cheung-Wai Chan, Seong G. Kong, Xun Zhang, and Binglu Wang

***Abstract*—This paper presents a mosaic convolution-attention network (MCAN) for demosaicing spectral mosaic images captured using multispectral filter array (MSFA) imaging sensors. MSFA-based multispectral imaging systems acquire multispectral information of a scene in a single snap-shot operation. A complete multispectral image is reconstructed by demosaicing an MSFA-based spectral mosaic image. To avoid aliasing and artifacts in demosaicing, we utilize joint spatial-spectral correlation in a raw mosaic image. The proposed MCAN includes a mosaic convolution module (MCM) and a mosaic attention module (MAM). The MCM extracts features via a learning approach with a margin between splitting the periodic spectral mosaic and keeping the underlying spatial information of the raw image. Based on the strategy of position-sensitive weight sharing, MCM assigns the same weight to pixels with the same relative position in an MSFA. The MAM uses a position-sensitive feature aggregation strategy to describe the loading of mosaic patterns within the feature maps, which gradually reduces mosaic distortion through the attention mechanism. The experimental results on synthetic as well as real-world data show that the proposed scheme outperforms state-of-the-art methods in terms of spatial details and spectral fidelity.**

***Index Terms*—Multispectral imaging, multispectral image demosaicing, multispectral filter array, Convolution-attention network, deep learning.**

I. Introduction

MULTISPECTRAL images contain rich spatial and spectral information, which is useful in identifying the material characteristics of an object of interest. It is more effective than color images in tasks such as food safety inspection [1], [2], land cover classification [3], [4], and object tracking [5]–[7].

Manuscript received December 15, 2020; revised April 12, 2021, June 8, 2021, and July 7, 2021; accepted July 26, 2021. Date of publication August 4, 2021; date of current version August 18, 2021. This work was supported in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grants JCYJ20170815162956949 and JCYJ20180306171146740, in part by the National Natural Science Foundation of China (NSFC) under Grant 61771391, in part by Key R & D plan of Shaanxi Province under Grant 2020ZDLGY07-11, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JM6056, in part by Korea National Research Foundation under Grant NRF-2016R1D1A1B01008522, and in part by the Yulin smart energy big data application joint Key Laboratory. *(Corresponding author: Yongqiang Zhao.)*

Kai Feng, Yongqiang Zhao, Xun Zhang, and Binglu Wang are with the Research and Development Institute, Northwestern Polytechnical University at Shenzhen, Shenzhen 518057, China (e-mail: 2018100620@mail.nwpu.edu.cn; zhaoyq@nwpu.edu.cn; xunzhang.zx@gmail.com; wbl921129@gmail.com).

Jonathan Cheung-Wai Chan is with the Department of Electronics and Informatics, Vrije Universiteit Brussel, 1050 Brussels, Belgium (e-mail: jcheungw@etrovub.be).

Seong G. Kong is with the Department of Computer Engineering, Sejong University, Seoul 05006, Korea (e-mail: skong@sejong.edu).

Digital Object Identifier 10.1109/TCI.2021.3102052

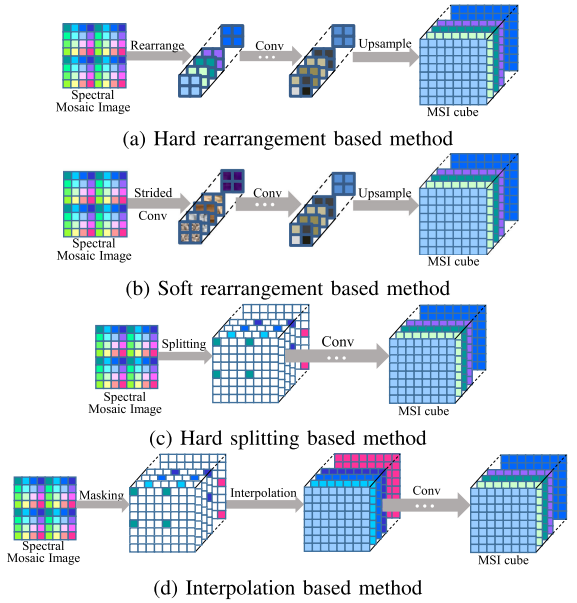

Fig. 1. Existing deep learning-based demosaicing pipeline to produce a multispectral image (MSI) cube from a spectral mosaic image.

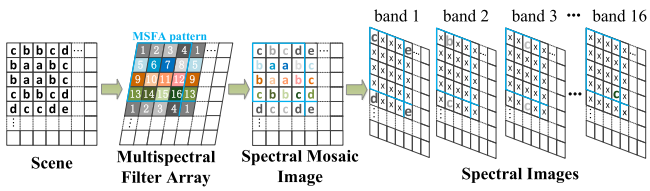

Fig. 2. The observation model of the multispectral imaging system based on MSFA.

Conventional multispectral imaging techniques acquire spectral information at multiple spectral bands and spatial information sequentially. Due to the sensor structure, conventional multispectral imagers require multiple exposures or scanning which can be unsuitable for capturing spectral-spatial information from dynamic scenes of moving objects. Inspired by the color filter array (CFA), multispectral filter array (MSFA) imaging techniques acquire spectral and spatial information simultaneously from a single image shot. Its ability to take snapshots and compact the sensor size offers great potential in a wide range of applications.

An MSFA comprises various spectral filters in the basic pattern (Fig. 2). With the pixelated MSFA, spectral information at one of the predetermined bands is captured at each pixel. Similar to color imaging sensors with CFA, a spectral mosaic

image is obtained by using an MSFA-based imager. A simple rearrangement causes aliasing distortion in the spatial and spectral domains. Multispectral demosaicing is the key step in MSFA-based imaging techniques and references the procedure of obtaining a fully-defined and high-quality (with no spatial and spectral distortions) multispectral image (MSI) cube [8]. After demosaicing, advanced work regarding multispectral data analysis [9]–[11] can be implemented on the obtained MSI cube.

Traditional multispectral demosaicing methods have been proposed with spatial and/or spectral correlation [12]–[15]. Without incorporating spatio-spectral correlation, their demosaicing results are often spectrally distorted, causing blurry edges in the spatial domain.

Convolutional neural network (CNN)-based multispectral demosaicing methods [16]–[18] implicitly explore the spatio-spectral correlation of the dataset. Before extracting deeper features, various approaches including soft rearrangement [16], hard splitting [17], or interpolation [18] have been adopted to initially handle the periodic spectral mosaic of the input (Fig. 1). The soft-rearrangement in Fig. 1(b) uses standard strided convolutions in learning to rearrange the raw spectral mosaic image, which outperforms the hard rearrangement as shown in Fig. 1(a). The hard rearrangement directly rearranges the raw spectral mosaic image into a multiband image, which is a common technique in color demosaicing. Rearrangement-based initial processing weakens the natural underlying full-resolution spatial information in the raw spectral mosaic image, thus increasing the difficulty of subsequent network reconstruction. Hard splitting in Fig. 1(c) avoids the abovementioned problem, which directly splits the different bands from the raw spectral image and leaves the missing spectral information blank. However, sparse input for standard convolution causes checkerboard artifacts in the results and causes some problems to the convergence of the network. Interpolating sparse cubes as in Fig. 1(d) is a compromise but causes the spatio-spectral aliasing. Because standard convolution explores spatial correlation for ordinary images whose pixels belong to the same spectral band, it cannot be directly applied to handle periodic spectral mosaic images. It is desirable to design a special convolution implementation for multispectral demosaicing to softly split the periodic spectral mosaic and keep the underlying full spatial information in the raw image.

This paper proposes a mosaic convolution module (MCM) to softly split the periodic spectral mosaic in the raw image during learning. Unlike the global weight sharing strategy with standard convolutions, a new position-sensitive strategy is used in MCM, which assigns the same weight to pixels that belong to the same spectral band. This strategy considers periodic spectral mosaic changes in the raw image. MCM softly splits the spectral bands into full spatial resolution feature maps called spectral feature maps, avoiding the spatial damage caused by hard or soft rearrangement.

After the soft splitting of the MCM, the periodic spectral mosaic is not completely split, and a similar periodic mosaic distortion appears in each channel of the spectral feature maps. CNNs need a certain depth to achieve a sufficiently large receptive field [19]. However, due to the high training cost of the MCM, a stacking MCM is unrealistic. Therefore, we also propose a mosaic attention module (MAM) to reduce periodic mosaics in deeper spectral feature maps. A new position-sensitive feature aggregation strategy called mosaic pooling is used in MAM. Mosaic pooling aggregates the features in the same relative position of the MSFA pattern. It enables the MAM to focus on loading of mosaic patterns within each channel of the spectral feature maps. Stacking the mosaic residual attention blocks (MRABs), which are composed of MAM and standard convolution, the proposed demosaicing scheme gradually reduces the mosaic distortion of deeper features at a low cost. Using these modules compatible with the characteristics of spectral mosaic images, this paper presents a mosaic convolution-attention network (MCAN) for multispectral demosaicing.

Our proposed demosaicing technique makes the following contributions:

1) Presents an end-to-end network scheme for multispectral image demosaicing according to the natural joint spatio-spectral correlation in mosaic images.
2) Proposes a position-sensitive weight sharing strategy used in the mosaic convolution module to softly split the periodic spectral mosaic and keeps the underlying full spatial information of a raw image.
3) Proposes a position-sensitive feature aggregation strategy used in the mosaic attention module to reduce the mosaic distortion within each spectral feature map.

Extensive experiments on synthetic data as well as real-world scenes demonstrate the merits of our MCAN over state-of-the-art demosaicing techniques in terms of fidelity in the spatial and spectral domains.

The remainder of this paper is organized as follows. Section II describes related work. Section III presents the observation model of the multispectral demosaicing. The details of our proposed multispectral demosaicing method are presented in Section IV. Section V provides an ablation study and other experimental comparisons with state-of-the-art multispectral demosaicing techniques.

II. Related Work

A. Filter Array-Based Snapshot Spectral Imaging Systems

Existing filter array-based snapshot spectral imaging systems can be divided into CFA-based imagers and MSFA-based imagers according to the number of filters in the array pattern. To collect rich spectral information, the CFA-based imager requires the design of broad filters [20], [21] or combines compressing spectral imaging techniques [22], [23]. These imagers have a high spectral resolution but rely on reconstruction methods or sacrifice the compactness of the system. For an MSFA-based imager, its spectral resolution is linked to the number of filters in the array, but it reduces the challenge to the reconstruction method and keeps the sensor compact. The number of bands and the choice of filters can be adapted to specific applications. The pattern of MSFAs is arranged to ensure spectral consistency and spatial uniformity of the acquired images [24]. In recent years, there have been quite a few proposals for the design of MSFA mosaic patterns [25]–[27]. However, manufacturing

difficulties and high costs have made many practical industrial implementations of various solutions impossible. We consider a 16-band typical MSFA-based system [28] that is available on the market.

B. Traditional Multispectral Demosaicing Methods

Traditional demosaicing methods herein refer to non-deep-learning based methods. To solve the demosaicing problem and to reconstruct the fully-defined MSI cube from MSFAs, many researchers have attempted to utilize the spatial and/or spectral correlation. For spatial correlation, general assumptions from color demosaicing can be readily adopted, such as weighted bilinear (WB) interpolation [29] and edge-sensing [24]. For spectral correlation, typical approaches assume that the spectral band values of the same position are correlated [13], [29]. Considering that the spectral distance of different bands will affect the spectral correlation, a pseudo-panchromatic image (PPI) that is strongly correlated with all bands is estimated as an intermediary [13]. In [15], a low-rank and graph regularized method (GRMR) was proposed for multispectral demosaicing, and more realistic spectral sensitive functions (SSFs) were considered in the GRMR to model the spectral correlation. The main drawback of such methods is that spatial and spectral correlations are utilized separately and prespecified manually.

C. Deep Learning-Based Demosaicing Methods

In recent years, researchers have investigated a variety of deep learning-based approaches to CFA demosaicing [30]–[36] but have mainly focused on the Bayer pattern with a dominant green band. The four main types of methods used are shown in Fig. 1 Inspired by these methods, three different approaches have attempted to account for applying CNNs to multispectral demosaicing [16]–[18]. Since CNNs are not specifically designed for spectral mosaics as in arrays of MSFAs, the spectral mosaic pattern must be processed at the beginning of the network. The downsampling-based method (DsNet) [16] uses strided convolutions to softly rearrange the spectral mosaic pattern, as shown in Fig. 1(b). While it reduces a significant amount of computation, there are checkerboard artifacts in the results due to deconvolution layers in the reconstruction stage. The splitting-based method (SpNet) in Fig. 1(c) uses the split sparse band image as the input of the network, which keeps the full spatial information of the raw image [17]. However, performing standard convolution on sparse input will cause checkerboard artifacts and certain difficulties in the convergence of the network. In [18], a deep network for MSFA demosaicing (InNet) is applied where the input is a bilinear interpolated MSI cube, as shown in Fig. 1(d). InNet tends to produce spatial artifacts because of the initial interpolation, and spatio-spectral aliasing is an obstacle for convolution networks.

A large and growing body of literature has investigated the spectral image resolution enhancement [37]–[40], which has many similarities with multispectral demosaicing. Compared to spectral super-resolution, spectral demosaicing does not change the spectral resolution and reconstructs only the missed pixels of each band image, which is similar to spatial super-resolution. One of the most crucial differences between spatial super-resolution and multispectral demosaicing is that multiband low-resolution images rearranged from raw spectral images have a field of view error [41], [42].

However, the network structures of the existing rearrangement-based and interpolation-based multispectral demosaicing methods are almost the same as super-resolution networks. This work aims to prove that it is necessary to design special networks for multispectral demosaicing.

D. Attention Mechanism

The attention mechanism originates from machine translation [43] and has been proven to be effective in CNN-based computer vision tasks, such as image classification [44], [45], object detection [46], [47] and semantic segmentation [48], [49]. The earliest and most commonly used mechanism is channel attention (CA) [50], which explicitly adjusts the channel of the feature maps (the output of the convolutional layer) globally. The entire module is lightweight and helps the network convergence. On the basis of channel attention, there are additional studies combining spatial attention with channel attention, such as CBAM [51], scSE [52], and SSC [53].

In recent years, some work has introduced the attention mechanism into demosaicing [33], [54], but they are all about the CFA pattern and essentially use channel attention.

III. Observation Model

This section describes the observation model for MSFA-based snapshot imaging systems. The incident light is first projected into the multispectral filter array through the objective lens, and the filter array represents a set of preselected spectral wavebands. The filtered incident light is then captured by a detector and the mosaic image is generated. As shown in Fig. 2, at each pixel of a spectral mosaic image Y, only one out of the B bands is available and the levels of the $B-1$ others are missing. Mathematically, considering that a fully-defined MSI with B bands $\{X_\lambda\}_1^B \in \mathbb{R}^{M\times N}$ is modulated by a group of sparse band-wise binary masks $\{S_\lambda\}_1^B \in \mathbb{R}^{M\times N}$, the measured spectral mosaic image is formulated as

$$Y = \sum_{\lambda=1}^{B} S_\lambda \cdot X_\lambda \tag{1}$$

where ' $\cdot$ ' means the element-wise product. S_λ is a sparse band-wise binary mask, which only has values at the positions that correspond to band λ on the MSFA. The observation model in (1) can be rewritten as

$$Y = T * X \tag{2}$$

where T denotes the masking matrix of MSFA and is determined by S_λ.

Our goal is to use an end-to-end CNN to learn a mapping function F to estimate a fully-defined MSI counterpart. More formally, CNN-based MSFA demosaicing requires solving the following problem:

$$\hat{\theta} = \arg\min_{\theta} l^{MSI}(F(Y;\theta), X) \tag{3}$$

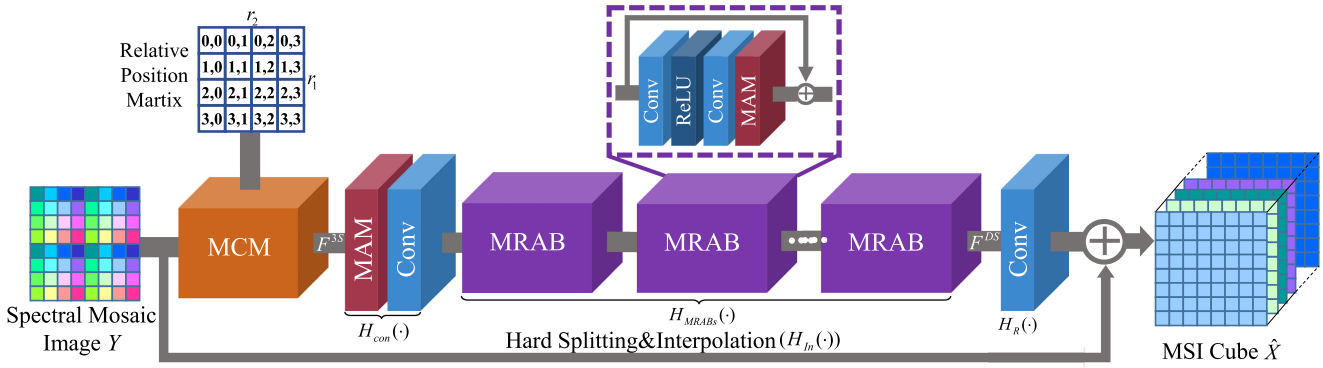

Fig. 3. A schematic diagram of the proposed mosaic convolution-attention network (MCAN), which consists of a mosaic convolution module (MCM), mosaic attention module (MAM), and mosaic residual attention block (MRAB).

where $l^{MSI}(\cdot)$ is the loss function for the fully-defined MSI cube, and θ are the parameters in the networks.

IV. Mosaic Convolution-Attention Network (MCAN)

This section describes the proposed MCAN in detail. We first introduce the overall network framework of the MCAN and then present the architectures of the mosaic convolution module (MCM) and mosaic attention module (MAM).

A. Network Framework

Fig. 3 shows that MCAN consists of three parts: soft spectral splitting, deeper spectral feature extraction and reconstruction. We are given Y and $\hat{X}$ as the input and output of MCAN. We apply one mosaic convolution module to initially and softly split the periodic spectral information from the raw mosaic image

$$F^{3S} = H_{MCM}(Y) \tag{4}$$

where $H_{MCM}(\cdot)$ stands for the operation of the mosaic convolution module. Then the soft split spectral feature F^{3S} is used for MRABs to extract the deeper spectral feature

$$F^{DS} = H_{MRABs}(H_{con}(F^{3S})) \tag{5}$$

where $H_{con}(\cdot)$ denotes the first mosaic attention module and convolution layer, which are mainly for converting the number of channels in F^{3S} to facilitate subsequent MRABs, and $H_{MRABs}(\cdot)$ represents several stacked MRABs. Each MRAB has two standard convolution layers, followed by one mosaic attention module (MAM) with skip connections to exploit feature interdependencies (see the dotted box in purple in Fig. 3). The receptive field of our proposed MRABs is large and assures effective feature extraction.

To reconstruct the fully-defined multispectral image (MSI) cube, we use a convolution layer to map F_{DS} to the residual image cube. Taking the skip connection into consideration, the final demosaicing MSI cube $\hat{X}$ can be obtained by adding the bilinear interpolated cube with the reconstructed residual image cube

$$\hat{X} = H_{In}(Y) + H_R(F^{DS}) = H_{MCAN}(Y) \tag{6}$$

where $H_{In}(\cdot)$ denotes the operation of hard splitting and interpolation to obtain the bilinear interpolated cube, $H_R(\cdot)$ is the reconstruction layer, and $H_{MCAN}(\cdot)$ is the function of the whole MCAN.

The MCAN will be optimized with a certain loss function. Popular loss functions include L_2 [55], L_1 [56], [57], and perceptual losses [58]. For better convergence in demosaicing [59], we adopt L_1 as our loss function. Given a training set with N spectral mosaic images and their fully-defined MSI counterparts denoted by $\{Y^i, X^i\}_{i=1}^N$, the goal of training the MCAN is to optimize the L_1 loss function:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left\| H_{MCAN}(Y^i) - X^i \right\|_1 \tag{7}$$

where θ denotes the parameter set of the MCAN. The loss function is optimized by Adam [60].

B. Mosaic Convolution Module

To maintain the full spatial information of the raw spectral mosaic image, the spatial size of the extracted feature maps should be consistent with the spectral mosaic image, which means that the stride of convolution should be 1. When the convolution kernel slides on the input spectral mosaic image with a stride of 1, the spectral distribution in the receptive field changes periodically. Therefore, it is unreasonable to continue to use the standard convolution with global weight sharing. We propose the mosaic convolution module (MCM) with a position-sensitive weight sharing strategy, which periodically change the weights of the convolution kernel when the convolution kernel is sliding. Increasing the channels of the feature maps can be regarded as softly splitting the periodic spectral information from the raw mosaic image. The obtained feature maps are called soft split spectral feature maps F^{3S}. Fig. 4 presents a block diagram of the MCM that outputs a single spectral feature map. The formulation of MCM is as follows.

Let F_c^{3S} denote the c-th soft split spectral feature map extracted by the MCM. The size of F_c^{3S} is consistent with the size of the raw image Y (both are $M \times N$), which keeps the underlying full spatial information. Suppose that the size of the MSFA pattern is $r_1 \times r_2$ (the pattern size in all diagrams in this

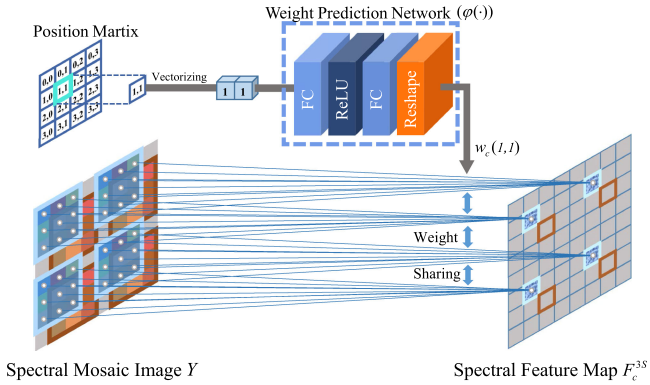

Fig. 4. Mosaic convolution module (MCM) showing a one-channel case for simplicity.

paper is 4×4); then, the change period of the weights of the convolution kernel should also be $r_1 \times r_2$. The value of position (m, n) on the spectral feature map F_c^{3S} is decided by the patch of the spectral mosaic image centered at (m, n) and the weights of the kernel. The weights are specified by the relative position of the pixel in the MSFA pattern. Thus, the convolution module is formulated as

$$F_c^{3S}(m,n) = \varphi(Y(m,n), w_c(m,' n')) \tag{8}$$

where $F_c^{3S}(m,n)$ denotes the spectral feature at (m,n) of F_c^{3S}. $Y(m,n)$ denotes the local mosaic patch centered at pixel (m,n) in MSFA. $(m,' n') = (m \bmod r1, n \bmod r2)$ represents the relative position of this pixel in the MSFA pattern. $w_c(m,' n')$ is the weight of the c-th kernel for the relative position $(m,' n')$. $\varphi(\cdot)$ is the feature mapping function to calculate the feature values. The matrix product is used as the feature mapping function in this work.

For the pixels that belong to the same relative position in the MSFA pattern, the spectral distribution of their neighborhoods is similar, in such a way that the same weight is assigned. We call this a position-sensitive weight sharing strategy. Regarding the generation of weights, different from the standard convolution, we use a network to predict the weights of the kernels based on the position information as suggested in [61]. The weight prediction is formulated as

$$W(m,' n') = \phi(v_{m,'n'}; \theta') \tag{9}$$

where $W(m,' n')$ is the concatenation of $\{w_c(m,' n')\}_{c=1}^{C}$ specified by the relative position $(m,' n')$ of the pixel in MSFA. Here, $v_{m,'n'}$ is a vector related to $(m,' n')$. $\phi(\cdot)$ is the weight prediction network, which takes $v_{m,'n'}$ as input. θ' is the parameter of the weight prediction network.

For $\phi(\cdot)$, which generates the position-sensitive weights for a certain MSFA pattern, the proper input is the normalization of $(m,' n')$, and the input can be formulated as

$$v_{m,'n'} = \left(\frac{m'+1}{r_1}, \frac{n'+1}{r_2} \right) \tag{10}$$

Algorithm 1 describes the detailed steps of the mosaic convolution module.

Algorithm 1: Mosaic Convolution Module (MCM).

Input: the input mosaic image Y, the size of spectral mosaic image (M, N), the weight prediction function W, the size of MSFA pattern (r_1, r_2)

Output: the feature maps

```
for m = 0 : 1 : M do
  for n = 0 : 1 : N do
    (m,' n') = (m mod r1, n mod r2)
    v(m,' n') = ((m'+1)/r1, (n'+1)/r2)
    the loacal mosaic patch on (m,n):Y(m,n)
    weights predicted by φ:W(m,' n')
    fv = Y(m,n) * W(m,' n')
    the feature value on (m,n) is fv
  end for
end for
```

C. Mosaic Attention Module

In the forward propagation of the entire model, the spatial size of the spectral feature maps is consistent with the raw image, and thus, each feature point is dominated by the raw pixel with the same spatial position [19]. After the soft splitting of the MCM, the periodic spectral mosaic will not be completely split, and a similar periodic mosaic distortion will appear in each channel of the spectral feature maps. At this point, we propose a mosaic attention module (MAM) that uses a new position-sensitive feature aggregation strategy. Combining MAM and standard convolutions to form MRAB helps to reduce periodic mosaic distortion within the spectral feature maps. The following describes the details of MAM.

A new feature aggregation strategy called mosaic pooling is used in MAM. As shown in Fig. 5, given a spectral feature map F with the size $M \times N \times C'$, mosaic pooling aggregates the feature points with the same relative position in the MSFA pattern for each channel of F. Mosaic pooling forms a spectral mosaic descriptor $z \in \mathbb{R}^{r_1 \times r_2 \times C'}$, which can describe the loading of periodic mosaic patterns in the spectral feature maps. The (m,n,c')-th value of z is computed as

$$z(m,n,c') = \frac{1}{M/r_1 \times N/r_2} \sum_{i=0}^{\frac{M}{r_1}-1} \sum_{j=0}^{\frac{N}{r_2}-1} F(m+i\times r_1, n+j\times r_2, c') \tag{11}$$

where $r_1 \times r_2$ is the size of the MSFA pattern. To fully exploit the spectral correlation from the spectral mosaic descriptor, we implement two standard convolution layer and a simple gating mechanism to obtain the basic mosaic attention map s with a size of $r_1 \times r_2 \times C'$

$$s = H_{Rs}(\sigma(W_E(\delta(W_S(z))))) \tag{12}$$

where $\sigma(\cdot)$ is the sigmoid function, $\delta(\cdot)$ is the ReLU function, W_S and W_E represent convolution layers with weights $W_S \in \mathbb{R}^{r_1 \times r_2 \times \frac{C}{d}}$ and $W_E \in \mathbb{R}^{1\times 1 \times r_1 r_2 C}$, d is the scale factor to squeeze the information of z, and $H_{Rs}(\cdot)$ represents a reshaping operation to obtain the output size of $r_1 \times r_2 \times C'$.

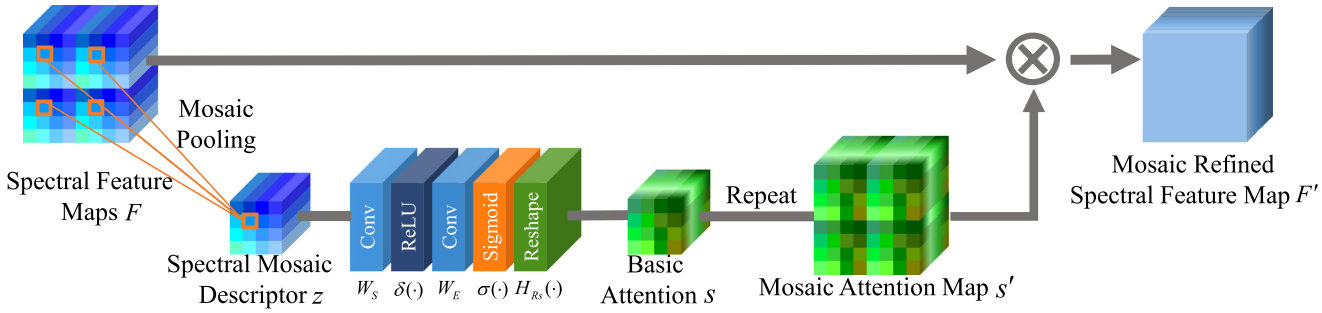

Fig. 5. A schematic diagram of our proposed mosaic attention module.

Algorithm 2: Mosaic Attention Module (MAM).

Input: the input spectral feature map F, the function of mosaic pooling H_{MP}, two function of standard convolution layer W_S, W_E

Output: the refined feature maps F'

1: obtain spectral mosaic descriptor z by mosaic pooling using equation (11):

$$z = H_{MP}(F)$$

2: obtain the basic mosaic attention map: s:

$$s = H_{Rs}(\sigma(W_E(\delta(W_S(z)))))$$

3: repeat the s along spatial dimension to obtain the full mosaic attention map s':

$$s'(m, n, c) = s(m \bmod r_1, n \bmod r_2, c)$$

4: obtain the mosaic refined spectral feature map:

$$F' = s' \cdot F$$

In the feature aggregation, we treat the feature points that belong to the same relative position in the MSFA pattern as a group. Correspondingly, in order to obtain the full mosaic attention map $s' \in R^{M \times N \times C'}$, we directly repeat s along the spatial dimension

$$s'(m, n, c') = s(m \bmod r_1, n \bmod r_2, c') \tag{13}$$

s' can be used to reduce the mosaic distortion within each channel and to emphasize informative channels. Finally, we refine the input spectral feature map via s'

$$F' = s' \cdot F \tag{14}$$

where '$\cdot$' stands for the element-wise product. F' is a mosaic refined spectral feature maps.

Algorithm 2 describes the detailed steps of the mosaic attention module.

V. Experiments

A. Implementations Details

For the mosaic convolution module (MCM), the weight prediction network consists of two fully connected layers and one activation layer. Each position vector input will output one group of weights with the shape $(inC, outC, k_1, k_2)$. Here inC is the number of raw mosaic images, and $inC = 1$ in this paper because we discuss single frame demosaicing. The $outC$ is the number of channels of the extracted feature maps, and the $outC = r_1 r_2$ to be consistent with the number of spectral bands. The k_1 and k_2 represent the size of the convolution kernel. Since the output size $(k_1 \times k_2 \times inC \times outC)$ is very large compared with the input vector size (2), we set the number of hidden neurons to 128. In addition, the activation function is ReLU [62]. For the kernel size, we set k_1 and k_2 to be the smallest odd number larger than or equal to r_1 and r_2 to make the receptive field of the convolution kernel contain each spectral band.

For the standard convolution layer $H_{con}(\cdot)$ at the head of the MARBs, the size and number of convolution kernels are set to 3×3 and 64, respectively. We set the MARB number to 2 in the stage of deeper spectral feature extraction, the kernel size of the standard convolution kernel used in MARBs is fixed to 3, and the channels of extracted features are fixed to 64. The scale factor d of all MAMs is set to 0.25. The number of filters in W_S and W_E is 1.

For the reconstruction layer, the size of the convolution kernel is set to 3×3, and the number of kernels is set to $r_1 r_2$ in the nonredundant case to match the number of bands of the camera.

This paper mainly conducts experiments on a 4×4 MSFA pattern, so $r_1 = r_2 = 4$. The stride size of the mosaic convolution used in soft spectral splitting and standard convolutions used in deeper spectral feature extraction are all 1, and there is suitable padding to ensure that the spatial dimensions remain unchanged after convolution. Each convolution is followed by a ReLU layer, except for the last reconstruction layer.

B. Datasets and Metrics

The CAVE laboratory at Columbia University released a high-quality multispectral image dataset in [63]. There are 32 scenes in the CAVE database, including various items used in real-life. We divide it into 26 scenes for training and 6 scenes for validation and testing. The data format of multispectral images is the reflectance, which is beneficial for us in synthesizing the radiance data. To test generalization performance, we use an additional 30 reflectance scenes in the TT-31 [26] database and 52 radiance scenes in the ICVL [64] database. The demosaicing results are evaluated with PSNR, SSIM [65], SAM [66], and ERGAS [67].

TABLE I
DEMOSAICING RESULTS (PSNR/SSIM/SAM/ERGAS) AT THREE TYPICAL SCENES AND THE AVERAGED RESULTS OVER ALL TEST SCENES FOR DIFFERENT METHODS IN THE CAVE DATASET UNDER THE D65 ILLUMINANT

Methods	Beads				Fake and real peppers				Photo and face				Average of all test			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
WB	26.352	0.9075	9.938	42.344	35.853	0.9880	7.207	25.219	34.285	0.9822	10.890	33.161	31.594	0.9675	9.430	34.106
PPID	35.391	0.9893	7.725	15.039	43.994	0.9982	4.936	9.991	42.810	0.9976	7.488	12.446	40.140	0.9961	6.358	12.132
GRMR	25.739	0.8980	12.609	45.504	34.018	0.9792	9.607	32.093	33.686	0.9794	11.063	35.802	30.370	0.9585	10.883	39.719
DsNet	34.091	0.9855	8.499	17.487	46.897	0.9991	6.060	7.067	45.459	0.9987	8.647	9.218	41.001	0.9955	7.471	12.258
SpNet	36.565	0.9916	6.198	13.252	48.102	0.9991	6.624	6.162	46.683	0.9987	9.733	8.114	43.082	0.9974	6.933	9.143
InNet	36.547	0.9917	6.371	13.238	46.028	0.9988	5.899	7.829	45.066	0.9985	7.903	9.698	42.415	0.9973	6.344	9.882
Ours	**39.461**	**0.9958**	**4.956**	**9.380**	**50.307**	**0.9995**	**4.246**	**4.813**	**48.879**	**0.9993**	**6.127**	**6.339**	**46.140**	**0.9987**	**4.816**	**6.561**

TABLE II
DEMOSAICING RESULTS (PSNR/SSIM/SAM/ERGAS) AT THREE TYPICAL SCENES AND THE AVERAGED RESULTS OVER ALL TEST SCENES FOR DIFFERENT METHODS IN THE CAVE DATASET UNDER THE HA ILLUMINANT

Methods	Beads				Fake and real peppers				Photo and face				Average of all test			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
WB	25.757	0.9454	11.224	34.152	35.857	0.9932	8.379	17.328	34.104	0.9833	12.233	31.590	31.435	0.9798	10.517	27.150
PPID	31.865	0.9865	9.157	17.525	42.025	0.9984	**6.933**	8.522	41.205	0.9968	9.387	13.984	38.098	0.9956	8.303	12.393
GRMR	24.314	0.9215	15.579	42.626	32.138	0.9712	13.160	32.943	32.532	0.9703	15.201	39.608	28.917	0.9583	14.326	40.062
DsNet	32.404	0.9883	10.158	16.249	43.715	0.9988	9.312	7.212	43.273	0.9980	11.197	11.075	39.441	0.9964	9.880	11.163
SpNet	33.329	0.9902	7.878	14.796	45.054	0.9991	9.387	6.108	44.907	0.9984	10.430	9.373	41.040	0.9972	8.986	9.334
InNet	35.226	0.9940	7.590	11.598	45.453	0.9992	6.995	5.796	44.149	0.9984	8.830	10.020	41.931	0.9981	7.669	8.218
Ours	**37.380**	**0.9963**	**6.592**	**9.169**	**48.499**	**0.9996**	7.027	**4.075**	**47.304**	**0.9992**	**8.350**	**7.048**	**44.678**	**0.9989**	**7.063**	**6.140**

For the degradation methods to generate the radiance mosaic images, following [13], we use the SSFs of the IMEC camera that we purchased and the available illuminants to synthesize the radiance label images and the mosaic images. For the ICVL radiance database, we do not need to additionally simulate illuminants.

C. Training Details

We randomly augment the radiance multispectral image cube by using various CIE standard illuminants (D65, F12) or two real illuminants (HA, LD) released in [13], flipping horizontally or vertically and rotating 90°, 180°, and 270°. During the training of the network, we randomly extract 32 mosaic images with a size of 128×128 as a batch input. The optimizer is Adam [60]. The learning rate is initialized to 2×10^{-3} for all of the layers and is multiplied by 0.5 for every 1000 epochs. The training process will stop after 5000 epochs. All of the experiments were run on one NVIDA GTX 1080Ti GPU.

D. Comparisons With Existing Demosaicking Methods

We compare the proposed demosaicing method with six existing approaches: weighted bilinear interpolation (WB) [29], multispectral demosaicing using pseudo-panchromatic image (PPID) [13], graph and rank regularized matrix recovery (GRMR) [15], downsampling-based demosaicing network (DsNet) [16], hard splitting-based demosaicing network (SpNet) [17], and interpolation-based demosaicing network (InNet) [18].

Tables 1 and 2 show the quantitative results for the three typical examples and the average over all 6 test images for the various methods in the CAVE dataset under D65 and HA illuminant. The best results are highlighted in bold. These tables indicate that the proposed demosaiking method outperforms the existing methods in the spatial and spectral domains. It is noted that the results of the GRMR are not very good because it is more suitable for multiframe input.

Since the PSNR/SSIM metrics do not always faithfully react to the visual quality of the images, we have also included the subjective quality comparison the results on six test scenes of the CAVE dataset in Fig. 6. The error maps are the absolute errors between the ground truth and the demosaicing results. Note that all images are 8 bits with values from 0 to 255. Our method offers much better MSI image demosaicing results, which demonstrates that our method can provide higher spatial accuracy. To further analyze the spectral performance of the proposed method, the absolute error between the ground truth and the demosaicing results of the scenes in Fig. 6 along the spectral for all methods are shown in Fig. 7. Our method are much closer to the ground truth, showing higher spectral fidelity.

We compare CNN-based methods. Fig. 8 shows local details of the pseudocolor ground truth and demosaicing results of the three CNN-based methods and our method for two CAVE scenes. The pseudocolor image can better visualize the performance of the spectral restoration. The results of DsNet and SpNet have some checkerboard distortion caused by the upsampling operator of DsNet and the input sparsity issue of SpNet. The results of InNet show no checkerboard distortion but some streak artifacts on the two green beads on the left. The reflective point of the first pink bead above produces color distortion. The outcome reflects that the spatio-spectral aliasing caused by the input is difficult to eliminate through subsequent operations.

E. Ablation Studies

To evaluate the generalization ability on the distribution variation between the training and testing data, we test all of the methods on the TT31 and ICVL datasets, and all of the

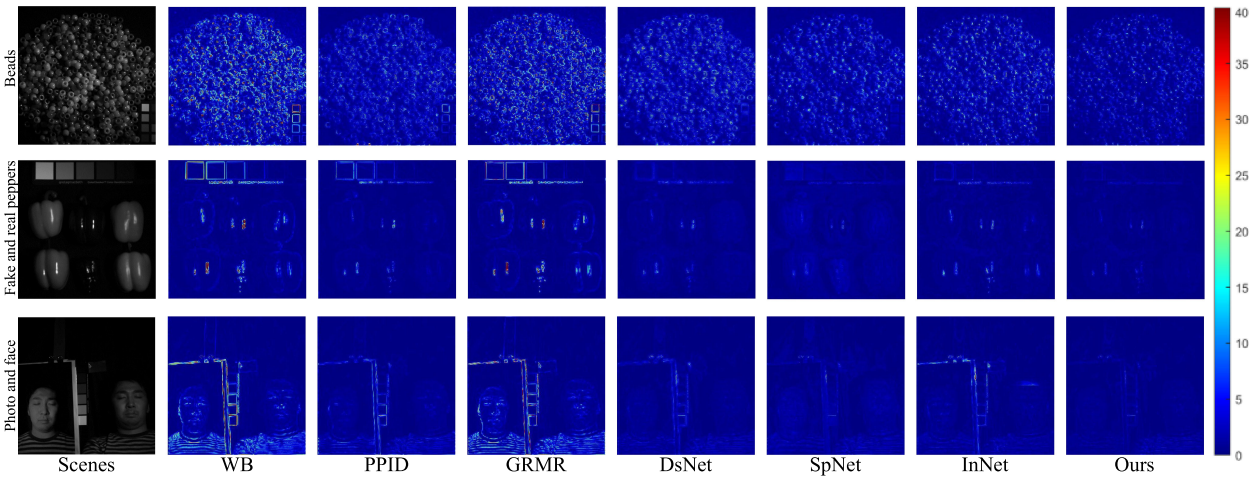

Fig. 6. Visual quality comparison of representative scenes of the CAVE dataset under the D65 illuminant at 480~nm. The error maps for WB/PPID/GRMR/DsNet/SpNet/InNet/our demosaicing results and the scenes are shown from left to right.

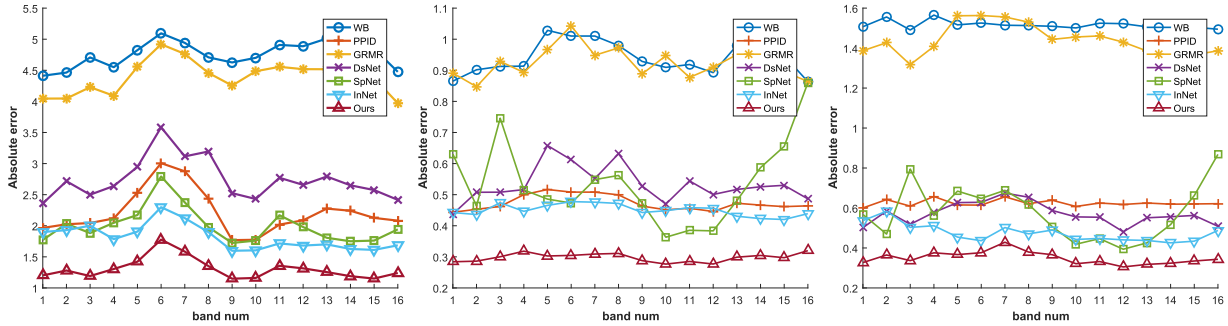

Fig. 7. The absolute error between the ground truth and the demosaicing results of the scenes in Fig. 6 along the spectral for all methods. Better view with zooming in.

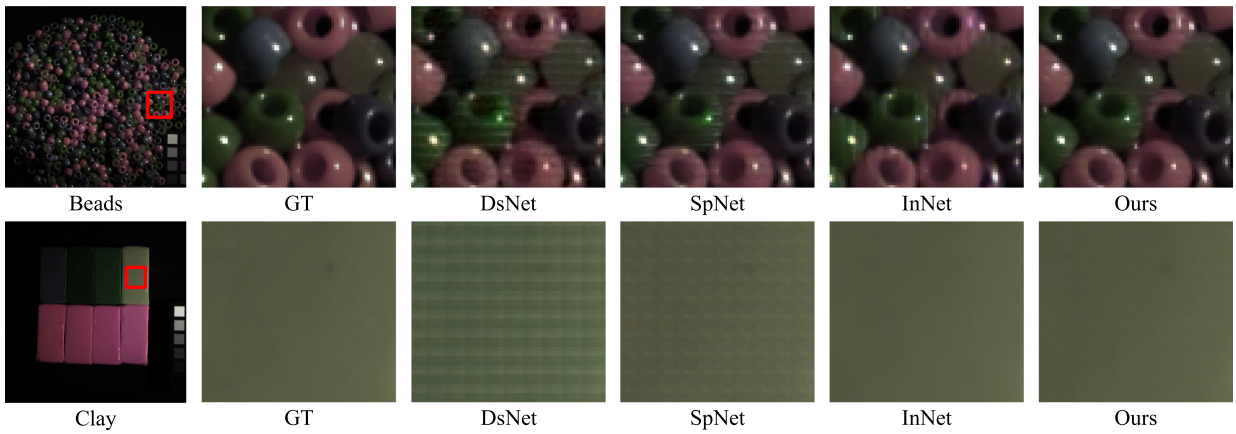

Fig. 8. The local details of the pseudocolor groud truth and demosaicing results of the three CNN-based methods and our method for two CAVE scenes.

CNN-based methods are trained on the CAVE dataset. According to Tables 3 and 4, our method still outperforms existing methods quantitatively. We further show subjective quality comparison results for six representative scenes of the TT31 and ICVL datasets in Figs. 9 and 11. The corrresponding spectral quality comparisons are shown in Fig. 10 and Fig. 12. Our method has good generalization ability on the data distribution variation quantitatively and qualitatively.

To demonstrate the effect of the proposed mosaic convolution module (MCM), we use the plain ResNet without any attention as our base model, and then we study the networks with different initializations in Table 5:

1) Directly rearrange the raw spectral mosaic image to form the low-resolution multiband image. Then, the pixelshuffle technique [68] is applied to reconstruct the MSI cube, labelled as a hard rearrangement;

TABLE III
EVALUATION GENERALIZATION ABILITY OF THE DISTRIBUTION VARIATION BETWEEN THE TRAINING AND TESTING DATA. ALL LEARNING-BASED MODELS ARE TRAINED ON THE CAVE DATASET AND TESTED ON THE TT31 DATASET UNDER THE D65 ILLUMINANT

Methods	Butterfly				Character				Fan				Average of all test			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
WB	30.006	0.9848	2.098	9.101	24.521	0.9821	5.923	18.713	26.343	0.9629	4.974	15.650	29.195	0.9767	4.070	15.141
PPID	35.843	0.9961	1.210	4.604	32.303	0.9971	3.469	7.647	33.588	0.9933	2.392	6.804	35.553	0.9953	2.427	7.123
GRMR	30.510	0.9866	3.759	8.968	24.020	0.9801	7.832	19.965	25.804	0.9597	6.602	16.689	29.583	0.9784	6.355	15.422
DsNet	42.283	0.9992	1.105	2.205	39.169	0.9994	3.196	3.497	37.204	0.9970	1.863	4.541	39.264	0.9974	2.725	5.189
SpNet	43.923	0.9994	0.940	1.843	42.226	0.9997	2.027	2.455	37.929	0.9975	1.542	4.204	40.186	0.9974	2.396	4.991
InNet	42.959	0.9993	1.008	2.044	40.562	0.9996	2.406	2.986	35.737	0.9959	2.094	5.343	39.920	0.9977	2.151	4.637
Ours	**47.701**	**0.9997**	**0.670**	**1.176**	**45.320**	**0.9999**	**1.632**	**1.713**	**39.234**	**0.9981**	**1.325**	**3.622**	**43.464**	**0.9989**	**1.731**	**3.245**

TABLE IV
EVALUATION GENERALIZATION ABILITY OF THE DISTRIBUTION VARIATION BETWEEN THE TRAINING AND TESTING DATA. ALL LEARNING-BASED MODELS ARE TRAINED ON THE CAVE DATASET AND TESTED ON THE ICVL DATASET

Methods	Grf_0328-0949				Objects_0924-1641				Omer_0331-1118				Average of all test			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
WB	34.400	0.9776	1.473	8.126	33.433	0.9737	0.721	4.932	37.517	0.9872	1.653	7.347	34.634	0.9814	1.461	7.640
PPID	41.985	0.9963	0.786	3.351	40.601	0.9955	0.383	2.023	44.834	0.9977	0.933	3.116	41.576	0.9964	0.833	3.383
GRMR	34.754	0.9793	1.550	7.818	33.518	0.9743	0.758	4.880	37.748	0.9879	1.740	7.156	34.786	0.9821	1.543	7.491
DsNet	45.834	0.9986	0.805	2.153	47.348	0.9993	0.418	0.963	48.381	0.9991	0.928	2.073	45.683	0.9988	0.818	2.146
SpNet	46.264	0.9987	0.843	2.064	47.294	0.9994	0.488	0.976	48.856	0.9992	0.994	1.985	45.961	0.9987	0.896	2.190
InNet	46.283	0.9986	0.683	2.046	47.961	0.9992	0.343	0.866	48.942	0.9991	0.887	1.944	46.225	0.9986	0.787	2.151
Ours	**49.099**	**0.9993**	**0.546**	**1.482**	**51.063**	**0.9996**	**0.276**	**0.608**	**51.850**	**0.9996**	**0.690**	**1.388**	**49.086**	**0.9994**	**0.602**	**1.488**

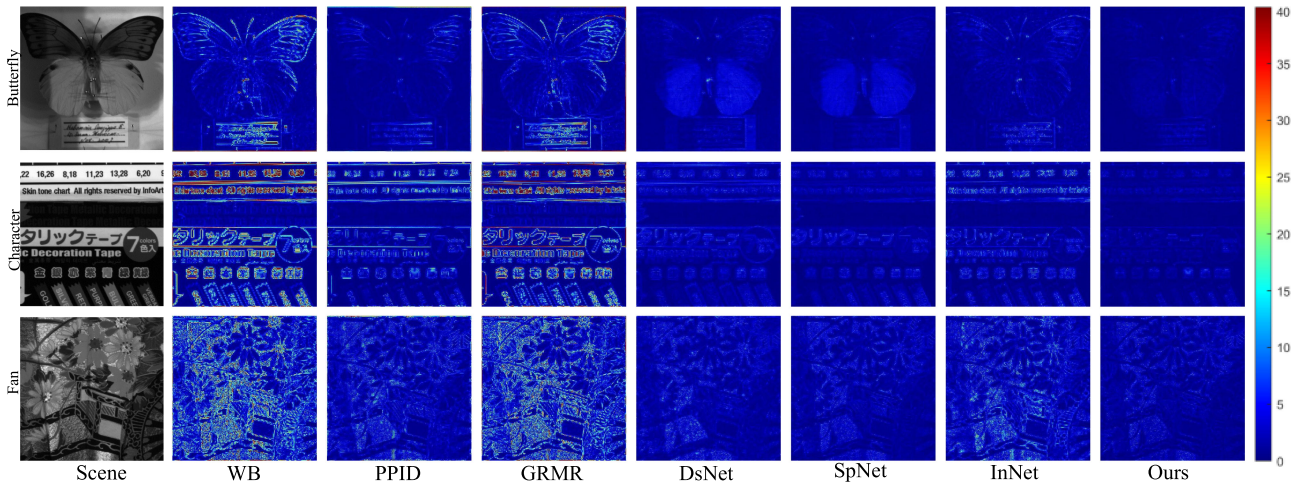

Fig. 9. Visual quality comparison on three representative scenes of the TT31 dataset under the D65 illuminant at $480^{\sim}nm$. The error maps for WB/PPID/GRMR/DsNet/SpNet/InNet/our demosaicing results and the scenes are shown from left to right.

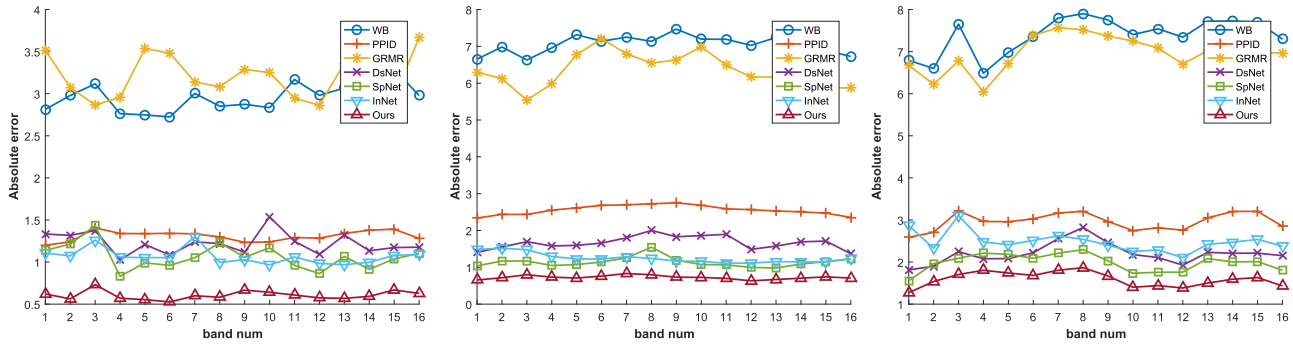

Fig. 10. The absolute error between the ground truth and the demosaicing results of the scenes in Fig. 9 along the spectral for all methods. Better view with zooming in.

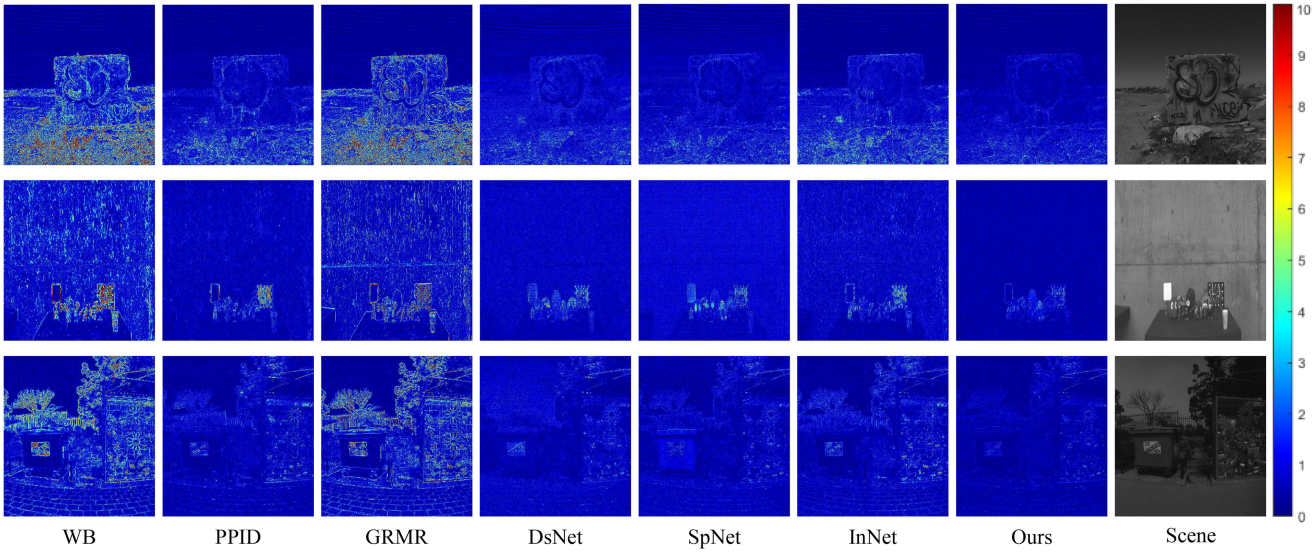

Fig. 11. Visual quality comparison on three representative scenes of the ICVL dataset at 480 nm. The error maps for WB/PPID/GRMR/DsNet/SpNet/InNet/our demosaicing results and the scenes are shown from left to right.

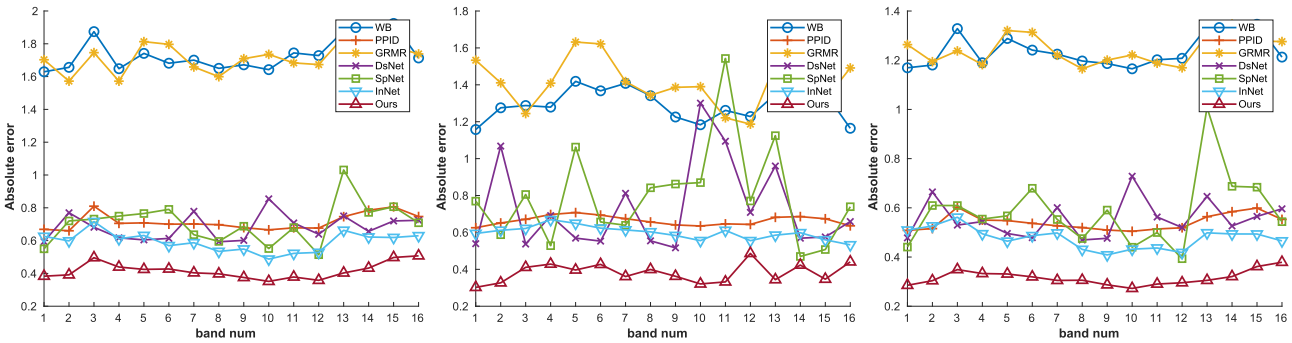

Fig. 12. The absolute error between the ground truth and the demosaicing results of the scenes in Fig. 11 along the spectral for all methods. Better view with zooming in.

TABLE V
THE IMPACT OF OUR PROPOSED MOSIAC CONVOLUTION ON THE DEMOSAICING PERFORMANCE AND COMPUTIONAL COST ON THE CAVE DATASET UNDER THE D65 ILLUMINANT

Initial Processing	PSNR	SAM	GFLOPs
Hard Rearrangment	34.94	9.576	**4.1**
Soft Rearrangment	40.89	7.534	8.3
Standard Convolution	38.53	8.568	63.4
Interpolation	40.62	7.454	64.9
Hard Splitting	42.26	7.191	64.9
Mosaic Convolution	**43.28**	**6.676**	63.2

2) Use the strided standard convolution to softly rearrange the raw image, labelled as a soft rearrangement, which is used in DsNet;
3) Directly use the standard convolution with a stride of 1 on the raw image, labelled as a standard convolution;
4) Split the raw spectral mosaic image directly, labelled as hard splitting, which is used in SpNet;
5) Interpolate the split sparse multiband image, labelled as Interpolation, which is used in InNet;
6) Use our proposed mosaic convolution to softly split the spectral mosaic image, labelled a mosaic convolution.

All of these networks have the same width and depth, and the training settings are the same as mentioned in part A of Section V, except for the learning rate. Since the hard splitting approach could cause problems with network convergence, we uniformly adjust the initial learning rate to 2×10^{-4} for a fair comparison. The synthetic CAVE images under D65 illuminant are used for testing. The rearrangement case is the worst, which shows that this conventional approach in color demosaicing is unsuitable for multispectral demosaicing. The hard splitting and mosaic convolution (soft splitting) approaches are the best, which shows that keeping the full spatial resolution of the raw spectral mosaic image unchanged at the beginning helps the multispectral demosaicing networks. However, compared to our

TABLE VI
THE IMPACT OF OUR PROPOSED MOSIAC ATTENTION ON THE DEMOSAICING PERFORMANCE AND COMPUTIONAL COST ON THE CAVE DATASET UNDER THE D65 ILLUMINANT

Scale Factor		PSNR	SAM	GFLOPs
MC-NA		44.58	6.039	**63.25**
MC-CA		45.12	5.583	63.39
MC-CBAM		45.20	5.482	63.33
MC-MA (Ours)	2	45.95	5.035	63.2908
	1	45.97	5.055	63.2898
	0.5	46.05	4.872	63.2892
	0.25	46.14	4.816	63.2889
	0.125	**46.28**	**4.796**	63.2888

soft splitting using mosaic convolution, the sparse multiband input caused by hard splitting is challenging for deep networks.

To demonstrate the effect of the proposed mosaic attention module (MAM), we use ResNet with mosaic convolution as our base model and then study the networks with different attention in Table 6:

1) the base model with no attention, labelled as MC-NA;
2) the base model with channel attention [50], labelled as MC-CA;
3) the base model with spatial-channel attention [51], labelled as MC-CBAM;
4) the base model with mosaic attention, which is our full model, labelled as MC-MA.

All of these networks are trained under the same default setting mentioned in part A of Section V for a fair comparison. The synthetic CAVE images under the D65 illuminant are used for testing. Table 6 shows that the MAMs with different scale factors are all better in terms of spatial accuracy and spectral fidelity and require a smaller amount of calculation than the other attention modules.

The limitation of the MSFA-based imagers is the spectral resolution linked to the number of bands in the MSFA pattern. To address this concern, we analyze how the number of bands in the array will affect our method. We fixed the spectral sampling range to be 400∼700 nm, and then uniformly selected 4, 9, 16 and 25 bands to compose four different MSFA patterns. We re-trained our models for these synthetic patterns, evaluated WB and our methods on all test images of the CAVE dataset under the D65 illuminant and report the PSNR in Fig. 13. Compared with WB, the PSNR of our method is better as the number of bands increases. Our method has greater advantages at higher spectral resolution, which is important, because our focus is the hyperspectral-oriented filters.

F. Performance on Real-World Data

Finally, we tested our proposed method on real-world data acquired by a real 4×4 IMEC snapshot camera. Since no ground truth is available for this case, Fig. 14 and 15 provide four illustrative scenes. For each example scene, the enlarged raw

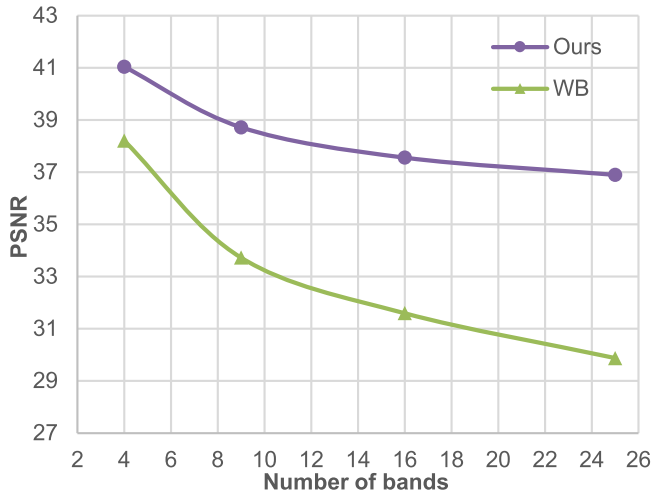

Fig. 13. Comparsions of different spectral resolutions. We train our MCAN using four ideal MSFA patterns with different spectral resolutions and evaluate on the CAVE test data.

TABLE VII
THE DEMOSAICING PERFORMANCE ON THE CAVE DATASET UNDER THE D65 ILLUMINANT AND COMPLEXITY COMPARISONS

Methods	PSNR	SAM	Running times(ms)	GFLOPs
DsNet	41.001	7.471	1.27	2.7
SpNet	43.082	6.933	3.23	324.3
InNet	42.415	6.344	2.91	945.2
Ours	46.140	4.816	2.74	63.3

mosaic patch and the corresponding pseudocolor demosaicing results of each method are provided. The indoor scenes are illuminated by LED lights, and the outdoor scenes are shot at noon. The demosaicing results generated by WB/GRMR/PPID/InNet tend to have severe artifacts. The results of DsNet/SpNet/Ours are better than those of the other methods. There still exists some degree of checkerboard distortion on the results of DsNet/SpNet. In the results of the "color chart", the existence of this distortion is more obvious. Our method performs well at both high and low frequencies and is also robust to illumination.

G. Running Time and Computational Cost

The speed and computational cost of the demosaicing method are important in determining whether it can be implemented on a real multispectral imaging system. Table 7 compares the running times and GFLOPs of several state-of-the-art CNN-based multispectral demosaicing methods, along with their average demosaicing performance on all test images of the CAVE dataset under the D65 illuminant. The running time is averaged over twenty times to reduce the impact of the machine startup. The GFLOPs are calculated in a single forward pass for a 512×512 pixel input spectral mosaic image. All of the methods are implemented with PyTorch using the same machine (Intel CPU 3.6 GHz, 16 GB memory, and NVIDA GPU GTX 1080Ti). Due to the efficiency of MCM and MAM, our method does not need to stack many convolutional layers violently, and thus, algorithm runs at competitive speed and computational cost.

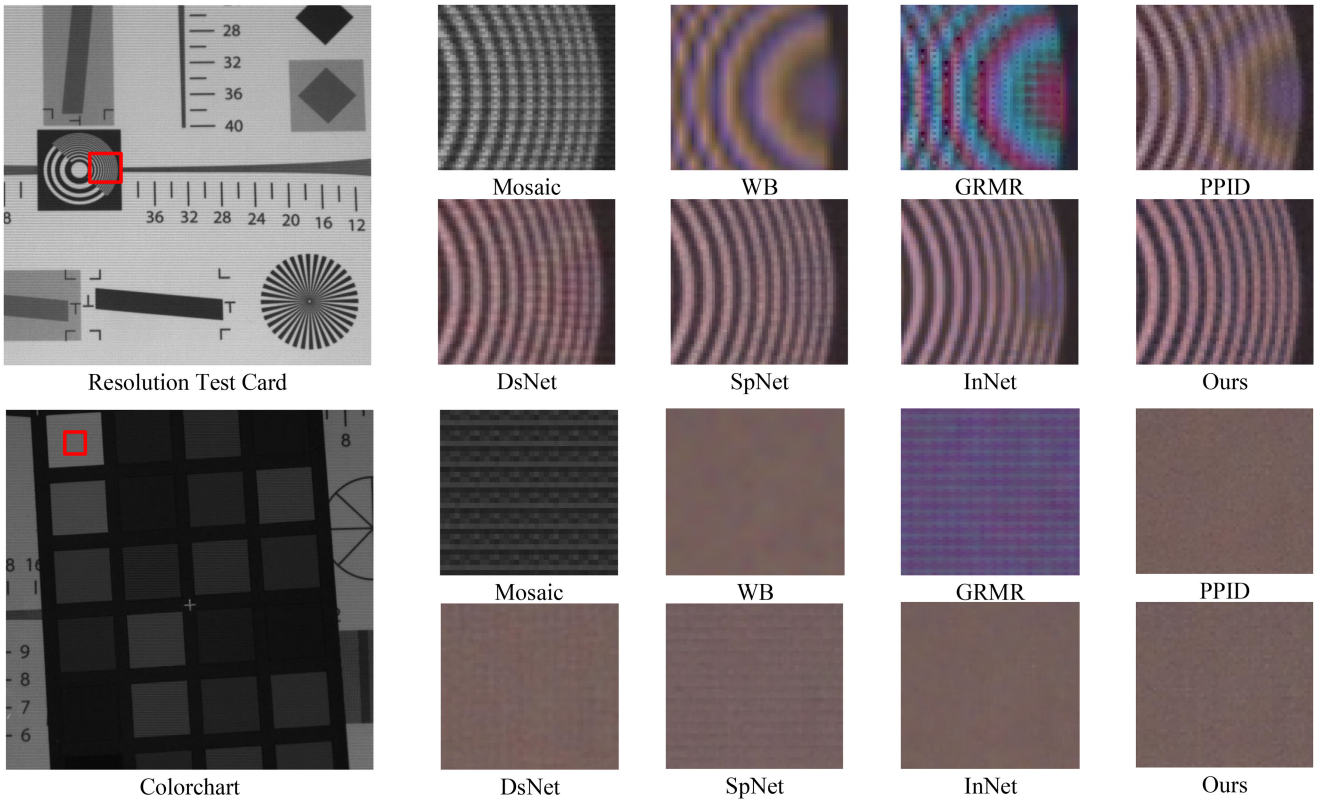

Fig. 14. The demosaicing pseudocolor results from captured spectral mosaic images of real-world indoor scenes.

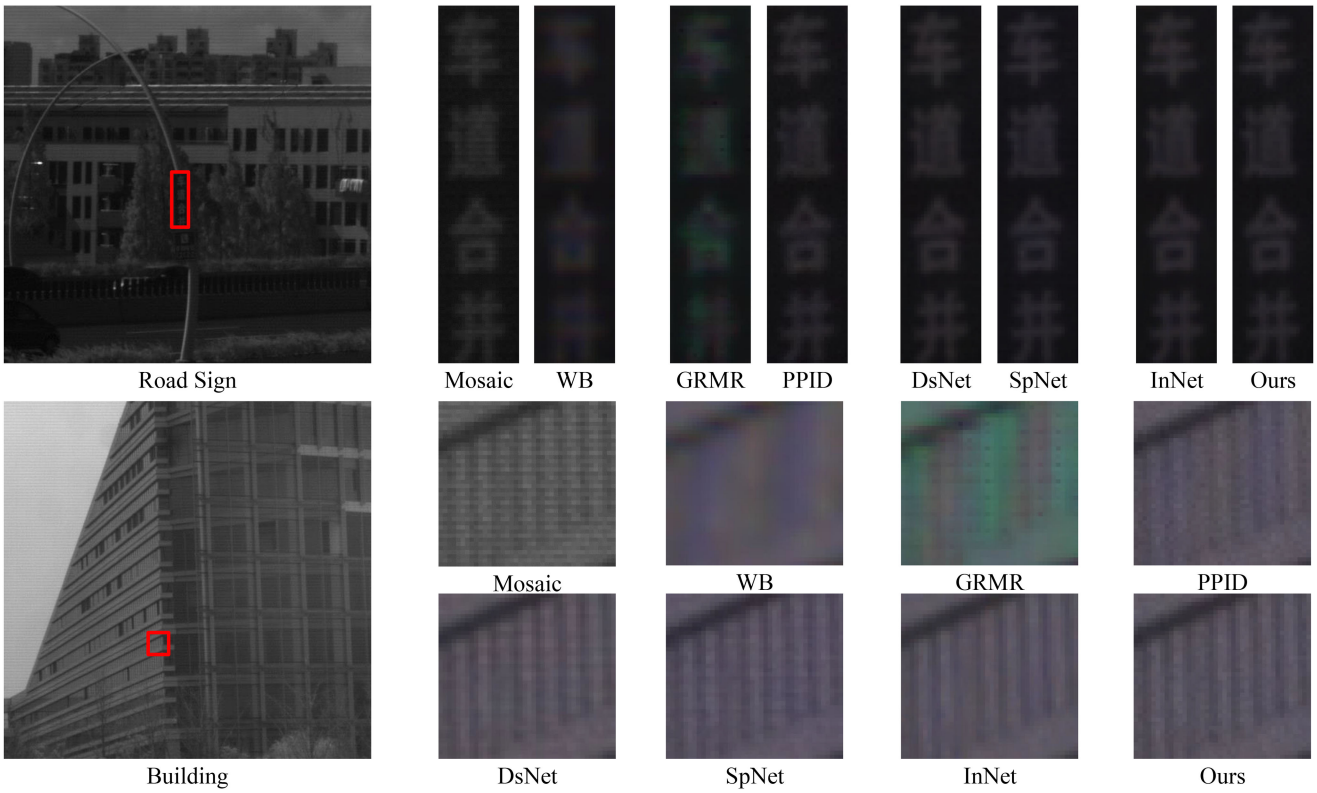

Fig. 15. The demosaicing pseudocolor results from captured spectral mosaic images of real-world outdoor scenes.

TABLE VIII
SENSITIVITY OF DEMOSAICING METHODS AT VARIOUS NOISE VARIANCES ON CAVE DATASET UNDER D65 ILLUMINANT

σ	0	0.5	1	1.5	2
WB	31.594	31.549	31.502	31.427	31.329
DsNet	41.001	40.542	39.977	39.257	38.504
SpNet	43.008	42.854	42.275	41.483	40.605
InNet	42.415	42.493	41.913	41.109	40.196
Ours	**46.140**	**45.563**	**44.051**	**42.400**	**40.864**

TABLE IX
SENSITIVITY OF DEMOSAICING METHODS AT VARIOUS NOISE VARIANCES ON CAVE DATASET UNDER HA ILLUMINANT

σ	0	0.5	1	1.5	2
WB	31.435	31.356	31.281	31.172	31.024
DsNet	39.441	38.991	38.446	37.731	36.970
SpNet	41.040	40.639	40.155	39.452	38.651
InNet	41.931	41.360	40.658	39.760	38.755
Ours	**44.678**	**43.556**	**42.227**	**40.670**	**39.176**

H. Sensitivity Analysis of Noise

We present the sensitivity analysis of the proposed method as well as the state-of-the-art methods at various noise levels in Tables 8 and 9. In the experiments, different Gaussian noise variances with zero means are added to the CAVE dataset under the D65 and HA illuminants. The noise variances σ are uniformly selected within the range of [0, 2]. This noise variance range is determined by the working characteristics we summarized when actually using MSFA-based imager. Tables 8 and 9 show that our method maintains a high PSNR when the noise level varies, compared to the state-of-the-art methods.

VI. CONCLUSION

This paper presents a mosaic convolution-attention network for demosaicing spectral mosaic images captured using multispectral filter array imaging sensors. Its mosaic convolution module and mosaic attention module directly learn the deep prior from the raw mosaic image dataset. The proposed method takes advantage of joint spectral-spatial correlation in the raw spectral mosaic image to avoid aliasing and artifacts in demosaicing. Our experimental results verified that MCAN accurately reconstructed spatial structures and high-fidelity spectral information on different datasets without re-training or fine-tuning. It is also a general multispectral demosaicing method, which has been verified on different spectral resolution MSFA patterns. Due to the efficient and well-designed modules, our method does not need to stack many convolutional layers violently, and thus, the algorithm runs at competitive speed and computational cost. Finally, we evaluated our MCAN on real-world mosaic images with a 4×4 pattern size, which supports its effectiveness in network design. Furthermore, the proposed method can be applied to other MSFA-based MSI reconstructions, e.g., raw mosaic denoising [69], video/multiple snapshot imaging systems [70], joint demosaicing and super-resolution enhancement [33].

REFERENCES

[1] H. Pu, L. Lin, and D.-W. Sun, "Principles of hyperspectral microscope imaging techniques and their applications in food quality and safety detection: A review," *Comprehensive Rev. Food Sci. Food Saf.*, vol. 18, no. 4, pp. 853–866, 2019.

[2] A. Falkovskaya and A. Gowen, "Literature review: Spectral imaging applied to poultry products," *Poultry Sci.*, vol. 99, no. 7, pp. 3709–3722, 2020.

[3] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.

[4] W. Liu, X. Shen, B. Du, I. W. Tsang, W. Zhang, and X. Lin, "Hyperspectral imagery classification via stochastic HHSVMs," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 577–588, Feb. 2019.

[5] F. Xiong, J. Zhou, and Y. Qian, "Material based object tracking in hyperspectral videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3719–3733, 2020, doi: 10.1109/TIP.2020.2965302.

[6] B. Uzkent, A. Rangnekar, and M. J. Hoffman, "Tracking in aerial hyperspectral videos using deep kernelized correlation filters," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 449–461, Jan. 2019.

[7] L. Chen *et al.*, "Object tracking in hyperspectral-oriented video with fast spatial-spectral features," *Remote Sens.*, vol. 13, no. 10, p. 1922, 2021.

[8] Y. Monno, M. Tanaka, and M. Okutomi, "Multispectral demosaicking using guided filter," in *Digital Photogr. VIII*, vol. 8299. *Int. Society Optics Photonics*, SPIE, pp. 204–210, 2012. [Online]. Avilable: https://doi.org/10.1117/12.906168

[9] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[10] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[11] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[12] J. Mizutani, S. Ogawa, K. Shinoda, M. Hasegawa, and S. Kato, "Multispectral demosaicking algorithm based on inter-channel correlation," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, 2014, pp. 474–477.

[13] S. Mihoubi, O. Losson, B. Mathon, and L. Macaire, "Multispectral demosaicing using pseudo-panchromatic image," *IEEE Trans. Comput. Imag.*, vol. 3, no. 4, pp. 982–995, Dec. 2017.

[14] S. Ogawa *et al.*, "Demosaicking method for multispectral images based on spatial gradient and inter-channel correlation," in *Proc. Int. Conf. Image Signal Process.*. Springer, 2016, pp. 157–166.

[15] G. Tsagkatakis, M. Bloemen, B. Geelen, M. Jayapala, and P. Tsakalides, "Graph and rank regularized matrix recovery for snapshot spectral image demosaicing," *IEEE Trans. Comput. Imag.*, vol. 5, no. 2, pp. 301–316, Jun. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8584100/

[16] K. Dijkstra, J. van de Loosdrecht, L. R. B. Schomaker, and M. A. Wiering, "Hyperspectral demosaicking and crosstalk correction using deep learning," *Mach. Vis. Appl.*, vol. 30, no. 1, pp. 1–21, Feb. 2019. [Online]. Available: http://link.springer.com/10.1007/s00138-018-0965-4

[17] T. A. Habtegebrial, G. Reis, and D. Stricker, "Deep convolutional networks for snapshot hypercpectral demosaicking," in *Proc. 10th Workshop Hyperspectral Imag. Signal Process.: Evol. Remote Sens.*, 2019, pp. 1–5, iSSN: 2158–6276.

[18] K. Shinoda, S. Yoshiba, and M. Hasegawa, "Deep demosaicking for multispectral filter arrays," 2018, *arXiv:1808.08021*.

[19] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," in *Proc. 30th Int. Conf. Neural Informat. Process. Syst.*, 2016, pp. 4905–4913.

[20] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Joint camera spectral response selection and hyperspectral image recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, doi: 10.1109/TPAMI.2020.3009999.

[21] H. Song *et al.*, "Deep-learned broadband encoding stochastic filters for computational spectroscopic instruments," *Adv. Theory Simulations*, vol. 4, no. 3, 2021, Art. no. 2000299.

[22] C. V. Correa, H. Arguello, and G. R. Arce, "Snapshot colored compressive spectral imager," *JOSA A*, vol. 32, no. 10, pp. 1754–1763, 2015.

[23] H. Rueda, H. Arguello, and G. R. Arce, "Dmd-based implementation of patterned optical filter arrays for compressive spectral imaging," *JOSA A*, vol. 32, no. 1, pp. 80–89, 2015.

[24] L. Miao, H. Qi, R. Ramanath, and W. E. Snyder, "Binary tree-based generic demosaicking algorithm for multispectral filter arrays," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3550–3558, Nov. 2006.

[25] R. Shrestha, J. Y. Hardeberg, and R. Khan, "Spatial arrangement of color filter array for multispectral image acquisition," in *Sensors, Cameras, Systems Industrial, Scientific, Consumer Applications XII*, editor., Ralf Widenhorn and Valérie Nguyen, vol. 7875. International Society for Optics and Photonics, SPIE, pp. 20–28, 2011. [Online]. Avilable: https://doi.org/10.1117/12.872253

[26] Y. Monno, S. Kikuchi, M. Tanaka, and M. Okutomi, "A practical one-shot multispectral imaging system using a single image sensor," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3048–3059, Oct. 2015.

[27] M. Kawase, K. Shinoda, and M. Hasegawa, "Demosaicking using a spatial reference image for an anti-aliasing multispectral filter array," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4984–4996, Oct. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8692717/

[28] B. Geelen, N. Tack, and A. Lambrechts, "A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic," in *Advanced Fabrication Technol. Micro/Nano Optics Photonics VII*, vol. 8974. Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VII, Editor., Georg von Freymann and Winston V. Schoenfeld and Raymond C. Rumpf, International Society for Optics and Photonics, SPIE, pp. 80–87, 2014. [Online]. Avilable: https://doi.org/10.1117/12.2037607

[29] J. Brauers and T. Aach, "A color filter array based multispectral camera," in 12. *Workshop Farbbildverarbeitung. Ilmenau*, Editor., German Color Group, Ilmenau: Oct. 2006.

[30] R. Tan, K. Zhang, W. Zuo, and L. Zhang, "Color image demosaicking via deep residual learning," *IEEE Int. Conf. Multimedia Expo (ICME)*, vol. 2, no. 4, p. 6, Jul. 2017.

[31] L. Liu, X. Jia, J. Liu, and Q. Tian, "Joint demosaicing and denoising with self guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2240–2249.

[32] S. W. Zamir *et al.*, "Cycleisp: Real image restoration via improved data synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2696–2705.

[33] X. Xu, Y. Ye, and X. Li, "Joint demosaicing and super-resolution (JDSR): Network design and perceptual optimization," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 968–980, 2020, doi: 10.1109/TCI.2020.2999819.

[34] B. Henz, E. S. Gastal, and M. M. Oliveira, "Deep Joint Design of Color Filter Arrays and Demosaicing," in *Comput. Graph. Forum*, vol. 37, no. 2, pp. 389–399, 2018.

[35] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3291–3300.

[36] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Trans. Graph. (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.

[37] L. Gao, D. Hong, J. Yao, B. Zhang, P. Gamba, and J. Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2269–2280, Mar. 2021.

[38] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vis.*. Springer, 2020, pp. 208–224.

[39] Y. Bu *et al.*, "Hyperspectral and multispectral image fusion via graph laplacian-guided coupled tensor decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 648–662, Jan. 2021.

[40] J. Xue, Y.-Q. Zhao, Y. Bu, W. Liao, J. C.-W. Chan, and W. Philips, "Spatial-spectral structured sparse low-rank representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 3084–3097, 2021.

[41] Y. Xie, C. Liu, S. Liu, W. Song, and X. Fan, "Snapshot imaging spectrometer based on pixel-level filter array (pfa)," *Sensors*, vol. 21, no. 7, 2021, Art. no. 2289.

[42] K. Monakhova, K. Yanny, N. Aggarwal, and L. Waller, "Spectral diffusercam: Lensless snapshot hyperspectral imaging with a spectral filter array," *Optica*, vol. 7, no. 10, pp. 1298–1307, 2020.

[43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.

[44] C. Liu, H. Xie, Z.-J. Zha, L. Yu, Z. Chen, and Y. Zhang, "Bidirectional attention-recognition model for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1785–1795, Jul. 2020.

[45] J. Ma, L. Zhang, and Y. Sun, "Roi extraction based on multiview learning and attention mechanism for unbalanced remote sensing data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6210–6223, Sep. 2020.

[46] B. Wang, L. Yang, and Y. Zhao, "Polo: Learning explicit cross-modality fusion for temporal action localization," *IEEE Signal Process. Lett.*, vol. 28, pp. 503–507, 2021, doi: 10.1109/LSP.2021.3061289.

[47] W. Zhang, B. Wang, S. Ma, Y. Zhang, and Y. Zhao, "I2net: Mining intra-video and inter-video attention for temporal action localization," *Neurocomputing*, vol. 444, pp. 16–29, 2021, [Online]. Avilable: https://doi.org/10.1016/j.neucom.2021.02.085.

[48] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[49] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, "Matnet: Motion-attentive transition network for zero-shot video object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 8326–8338, 2020, doi: 10.1109/TIP.2020.3013162.

[50] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020, doi: 10.1109/TPAMI.2019.2913372.

[51] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[52] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Aug. 2018.

[53] J. Ma, H. Zhang, P. Yi, and Z. Wang, "Scscn: A separated channel-spatial convolution net with attention for single-view reconstruction," *IEEE Trans. Ind. Electron.*, vol. 67, no. 10, pp. 8649–8658, Oct. 2020.

[54] Q. Kang, Y. Fu, and H. Huang, "Deep color image demosaicking with feature pyramid channel attention," in *Proc. IEEE Int. Conf. Multimedia & Expo Workshops*, 2019, pp. 246–251.

[55] W. Wei, J. Nie, Y. Li, L. Zhang, and Y. Zhang, "Deep recursive network for hyperspectral image super-resolution," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1233–1244, 2020, doi: 10.1109/TCI.2020.3014451.

[56] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2017, pp. 624–632.

[57] W. Wei, Y. Sun, L. Zhang, J. Nie, and Y. Zhang, "Boosting one-shot spectral super-resolution using transfer learning," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1459–1470, 2020, doi: 10.1109/TCI.2020.3031070.

[58] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 694–711.

[59] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980* .

[61] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-sr: A magnification-arbitrary network for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1575–1584.

[62] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[63] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

[64] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in *Proc. Eur. Conf. Comput. Vis.*. Springer, 2016, pp. 19–34.

[65] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[66] F. A. Kruse *et al.*, "The Spectral Image Processing System (sips)-Interactive Visualization and Analysis of Imaging Spectrometer Data," in *Proc. AIP Conf. Proc.*, American Institute of Physics, 1993, vol. 283, no. 1, pp. 192–201.

[67] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" *Fusion Earth Data: Merging Point Measurements, Raster Maps Remotely Sensed Images*, SEE/URISCA, pp. 99–103, 2000.

[68] 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.*, 2016, pp. 1874–1883.
[69] Z. Pan, B. Li, H. Cheng, and Y. Bao, "Deep residual network for MSFA raw image denoising," in *Proc. ICASSP 2020-2020 IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2020, pp. 2413–2417, iSSN: 2379–190X.
[70] P.-J. Lapray, J.-B. Thomas, and P. Gouton, "High dynamic range spectral imaging pipeline for multispectral filter array cameras," *Sensors*, vol. 17, no. 6, p. 1281, Jun. 2017. [Online]. Available: http://www.mdpi.com/1424-8220/17/6/1281

Kai Feng received the B.S. degree in automation from Chang'an University, Xi'an, China, in 2018. Since 2018, he has been working toward the Ph.D. degree in control science and engineering with the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2018. His research interests include computational imaging, hyperspectral imaging, deep learning, and image processing.

Yongqiang Zhao received the B.S., M.S., and Ph.D. degrees in control science and engineering from the Northwestern Polytechnic University, Xian, China, in 1998, 2001, and 2004, respectively. From 2007 to 2009, he was a Postdoctoral Researcher with McMaster University, Hamilton, ON, Canada, and Temple University, Philadelphia, PA, USA. He is currently a Professor with the Northwestern Polytechnical University. His research interests include polarization vision, hyperspectral imaging, compressive sensing, and pattern recognition.

Jonathan Cheung-Wai Chan received the Ph.D. degree from The University of Hong Kong, Hong Kong, in 1999. From 1998 to 2001, he was a Research Scientist with the Department of Geography, University of Maryland, College Park, MD, USA. From 2001 to 2005, he was with the Interuniversity Micro-Electronics Centre, Leuven, Belgium. From 2005 to 2011, he was with Geography Department, Vrije Universiteit Brussel (VUB), Brussel, Belgium. From 2013 to 2014, he was a Marie Curie Fellow with Fondazione Edmund Mach, San Michele all'Adige, Italy. He is currently a Senior Researcher and a Guest Professor with the Department of Electronics and Informatics, VUB. His research interests include land-cover classification with machinelearning algorithms, detailed mapping using hyperspectral data, and spatial and spectral enhancement of satellite hyperspectral images.

Seong G. Kong received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1982 and 1987, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 1991. He was an Associate Professor of electrical and computer engineering with the University of Tennessee, Knoxville, TN, USA, and with Temple University, Philadelphia, PA, USA. He was also the Chair of the Department of Electrical Engineering, Soongsil University, Seoul, South Korea, and the Graduate Program Director with the Electrical and Computer Engineering Department, Temple University, Philadelphia, PA, USA. He is currently a Professor of computer engineering and the Director of strategic planning with Sejong University, Seoul, South Korea. His research interests include image processing, computer vision, hyperspectral imaging, and intelligent systems. He was the recipient of the Honorable Mention Paper Award from the American Society of Agricultural and Biological Engineers in 2005 and the Most Cited Paper Award from the Computer Vision and Image Understanding journal in 2007 and 2008. He was an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS and the Guest Editor of the *Journal of Sensors*.

Xun Zhang received the B.S. degree in automation from Chang'an University, Xi'an, China, in 2019. Since 2019, he has been working toward the Ph.D. degree in control science and engineering with the School of Automation, Northwestern Polytechnical University, Xi'an, China. His research interests include pattern recognition, computer vision, machine learning, and image processing.

Binglu Wang received the master's degree in robotics from University West, Trollhattan, Sweden, in 2016. Since 2017, he has been working toward the Ph.D. degree in control science and engineering with the School of Automation, Northwestern Polytechnic University, Xi'an, China. His research interests include computer vision, robotics science, and deep learning.