Wavelength-Embedding-guided Filter-Array Transformer for Spectral Demosaicing

Haijin Zeng[1(✉)], Hiep Luong[1], and Wilfried Philips[1]

IMEC-Ghent University, Ghent 9000, Belgium
haijin.zeng@imec.be

Abstract. Spectral imaging offers the capability to unveil hidden details within the world around us. However, to fully harness this potential, it is imperative to develop effective spectral demosaicing techniques. Despite the success of learning based spectral demosaicing methods, three challenges hinder their practical use. Firstly, existing convolutional neural networks and attention-based models, struggle to capture spectral similarities and long-range dependencies. Secondly, their performance is unstable when optical characteristics, like multispectral filter array (MSFA) arrangement and wavelength distribution, change. Lastly, they lack a structured approach to incorporating imaging system physics, such as MSFA pattern. Addressing these challenges, our paper introduces the Wavelength Embedding guided Filter Array Attention Transformer (WeFAT) for effective spectral demosaicing. Specifically, inspired by the timestep embedding in denoising diffusion models, we propose a Wavelength Embedding guided Multi-head Self-Attention (We-MSA) mechanism to imbue our model with wavelength memory, facilitating adaptation to diverse cameras. This approach treats each spectral feature as a token, directly integrating wavelength information into attention calculation. Additionally, we developed a MSFA-attention Mechanism (MaM) steering We-MSA to focus on spatial regions yielding high-quality spectral data. Experimental results affirm that WeFAT exhibits strong performance consistency across diverse cameras characterized by varying spectral distributions and MSFA patterns, trained solely on ARAD dataset. It also outperforms current state-of-the-art methods in both simulated and real datasets.

Keywords: Spectral demosaicing · Wavelength embedding · Filter array attention

1 Introduction

Hyperspectral imaging captures light across a broad range of spectral bands, including those within the visible and beyond near-infrared spectrum. This provides much higher spectral resolution than the 3 spectra, leading to more accurate material characterization than is achievable through RGB imaging. This capability makes hyperspectral imaging a valuable tool in numerous fields, including medical imaging, astronomy, food quality control, remote sensing, precision agriculture and pharmaceuticals [6, 8, 46, 47].

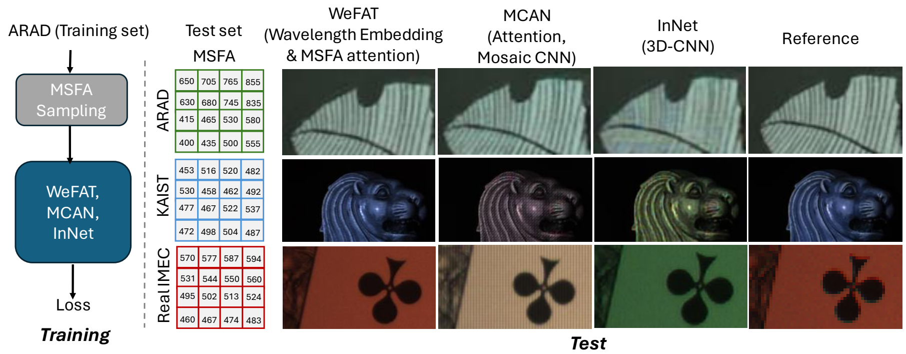

Fig. 1: Illustrating the efficacy of our wavelength embedding mechanism (False color, R: 2nd, G: 11th, B: 16th): MCAN [15] and InNet [42] exclusively trained on the ARAD dataset [2] demonstrates sub-optimal performance when applied to KAIST [10] and real IMEC camera data. In contrast, our WeFAT, which incorporates wavelength embedding and is trained solely on the ARAD dataset, exhibits superior performance on real IMEC data. For IMEC data, the reference is obtained via MSFA-based rearrangement [15], preserving the spectral bands of the ideal reference but with $\frac{1}{4}$ spatial resolution.

However, hyperspectral imaging faces limitations in computer vision due to slow acquisition times, often caused by spatial or spectral scanning. Recent advancements, including computed tomography [12, 23], light-field imaging [3, 11], and Multi-Spectral Filter Array (MSFA) cameras [24], aim to address this issue. MSFA cameras, utilizing larger Color Filter Arrays (CFAs), particularly 3×3, 4×4, or 5×5 configurations, efficiently acquire hyperspectral data, overcoming traditional limitations. Notable MSFA cameras like IMEC SNAPSHOT, XIMEA Snapshot USB3, and silios CMS series [2] are becoming more accessible. However, effective spectral demosaicing methods are crucial to fully exploit the spatial and spectral information provided by MSFA cameras. Demosaicing large MSFAs poses challenges due to their larger mosaic pattern and weaker inter-channel correlation compared to Bayer filter cameras.

Contemporary learning-based multispectral (MSI) demosaicing approaches, including convolutional neural networks (CNNs) [42, 48] and attention-based models [15], achieve high Peak Signal-to-Noise Ratio (PSNR) on specific paired training and test dataset, but often fail to effectively capture spectral similarities and long-range dependencies, particularly overlooking variations in spectral wavelengths. This limitation hampers their adaptability to cameras operating at diverse wavelengths. For example, as shown in Fig. 1, the spectral demosaicing methods MCAN [15] and InNet [42], trained on the ARAD dataset, perform well on the ARAD validation set. However, they exhibit limited capability when applied to KAIST and real imec camera data demosaicing, resulting in reconstructed images with a spectrum that deviates from the reference or contains obvious periodic artifacts. This discrepancy arises from differences in wavelength distribution and MSFA pattern across these three datasets. Furthermore, current methods inadequately utilize the potential of MSFA in modulating HSI data during spectral demosaicing. Specifically, spatial positions exhibit varying paired sampled spectral information within an MSFA, constituting a periodic pattern based on the camera's MSFA configuration.

An overlooked aspect within this field pertains to the incorporation of HSI wavelength information and MSFA pattern within the imaging process via learning-based algorithms. Consequently, existing spectral demosaicing methods exhibit restricted applicability across diverse camera datasets as shown in Fig. 1. Our study seeks to bridge this gap and rectify the deficiencies present in current CNN and attention-based methodologies. Specifically, we propose the Wavelength Embedding guided Filter-array Attention Transformer (WeFAT), a novel approach that replaces conventional convolution and attention mechanisms. A key component of our method is the Wavelength Embedding guided Multi-head Self-Attention (We-MSA) mechanism. This technique treats each spectral feature as an individual token, embedding wavelength data directly into the attention computation process as the timestep embedding in denoising diffusion [20]. Consequently, our spectral transformer retains specific wavelength information effectively, making it adaptable to cameras with diverse wavelength distributions as shown in Fig. 1. This flexibility resembles the denoising diffusion mode, which can commence from any timestep. Additionally, we propose a Multispectral Filter Arrays attention Mechanism (MaM) that directs We-MSA's focus towards spatial regions with high-fidelity spectral representations, in alignment with the MSFA's sampling pattern. In summary, our contributions are four-fold:

1. We introduce a new approach for tailoring vision transformers to perform spectral demosaicing. Our method is universally adaptable to various spectral MSFA cameras, requiring training on just one dataset.
2. Our approach features the novel Wavelength Embedding Multi-head Self-Attention (We-MSA), designed to capture inter-spectral similarities and dependencies in MSIs while incorporating wavelength information.
3. Additionally, we develop a specialized MSFA-attention Mechanism that guides We-MSA to focus on areas with accurate MSI representations, tailored to the MSFA configuration in hyperspectral imaging systems.
4. WeFAT outperforms traditional methods, as well as CNN and attention-based state-of-the-arts, across 4 MSI benchmarks and 1 real dataset with diverse MSFA patterns, e.g., 48.03 dB on ARAD with 1.71M parameters.

2 Related Work on Spectral Demosaicing

MSFA-based snapshot imaging directly converts 3-D spectral images into a 2-D raw image, with each pixel representing a single spectral band. Spectral image demosaicing is essential for reconstructing spectral images from this subsampled raw image. However, challenges arise due to low spatial correlation among neighboring pixels, leading to suboptimal performance with traditional methods [9].

The most basic spectral demosaicing technique involves interpolation algorithms [4, 18, 34, 37]. Subsequently, Miao et al. introduced the binary tree-based edge-sensing (BTES) algorithm [31] for iteratively estimating missing pixels with a periodic pattern. Mizutani et al. [22, 33] proposed the iterative spectral difference (ItSD) method, considering both spectral and spatial correlations [38, 39]. Mihoubi et al. introduced pseudopanchromatic image (PPI)-based multispectral demosaicing [28, 32, 35], leveraging spatial high-frequency information. Constrained optimization algorithms with regularizations like total variation [7], low rank [25, 29], and graph-based methods [16, 26] have been employed. Tsagkataki et al. [43] formulated spectral image demosaicing as a graph and low-rank regularized optimization problem, achieving superior performance.

Recent attention has turned to deep learning-based multispectral demosaicing methods, which use deep neural networks to map raw images to full-resolution spectral images. For instance, Feng et al. [14,15] introduced the mosaic convolution-attention network (MCAN), capturing joint spectral-spatial correlations. Similarly, 2D [19, 45] and 3D CNN [42] with a residual connection for spatial-spectral information learning based spectral demosaicing methods are also proposed. Various supervised learning-based approaches were showcased in the NRIRE 2022 spectral demosaicing challenge [2].

3 Spectral Imaging and Demosaicing

Spectral Imaging: Fabry-Pérot filters ideally enable light propagation from a specific spectral range, blocking external light. However, technical and physical constraints lead to deviations from this ideal behavior. Fig. 2 illustrates a real spectral response with notable non-linearities and additional harmonics introduced by Fabry-Pérot filters, highlighting the need for cut-off filters. As shown in Fig. 2, a Spectral Snapshot Imaging (SSI) camera captures an image of $H \times W$ pixels per exposure, where H and W denote horizontal and vertical pixel dimensions, respectively. Each pixel is associated with a distinct spectral band, resulting in a subsampled mosaic $\mathbf{Y}$ from the full-resolution tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times B}$ using the Multi-Spectral Filter Array within spectral camera. This means that in each pixel of the spectral mosaic image $\mathbf{Y}$, only one of the B bands is available, with the levels of the remaining $B-1$ bands absent.

Spectral Demosaicing: Mathematically, for a fully-defined multi-spectral image (MSI) with B bands $\{\mathbf{X}_\lambda\}_{\lambda=1}^{B} \in \mathbb{R}^{H \times W \times B}$, modulated by MSFA-based sparse band-wise binary masks $\{\mathbf{S}_\lambda\}_{\lambda=1}^{B} \in \mathbb{R}^{H \times W \times B}$, the measured spectral mosaic image is given by:

$$\mathbf{Y} = \sum_{\lambda=1}^{B} \mathbf{S}_\lambda \odot \mathbf{X}_\lambda, \tag{1}$$

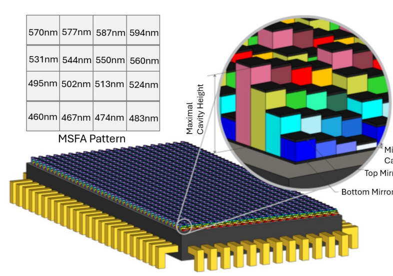

Fig. 2: A schematic representation of IMEC spectral camera [17] in a pixel-level mosaic layout and alignment of filters to tiles.

where $\odot$ denotes Hadamard product, with $\mathbf{S}_\lambda$ representing the sparse binary mask corresponding to band λ in the MSFA.

Here, we consider case of sampling operators $\mathbf{S}_\lambda$: Spectral Filter Profile Sampling: focuses on input signals modulated by spectral filter profiles, accounting for deviations from ideal responses, energy spillage into neighboring bands, and additional harmonics, but we also test our model on spectral image captured by real image captured by our imec camera in experiment section. Subsequently, our objective of spectral demosaicing is to develop a deep neural network (DNN) to learn a mapping function f that estimates a fully-defined MSI. The DNN-based MSFA demosaicing problem is thus defined as:

$$\hat{\theta} = \arg\min_{\theta} \; l(f(\mathbf{Y}; \theta), \mathbf{X}) \tag{2}$$

where $l(\cdot)$ is the loss function for the MSI cube, and θ represents the learnable network parameters.

4 Method

4.1 Network Architecture

As shown in Fig. 3, WeFAT consists of three modules: shallow feature extraction, Group Attention Transformer (GAT)-based deep feature extraction and Hyperspectral Image Reconstruction (HIR) modules. GAT employs spectral MSA with Wavelength Embedding (We-MSA) and window based MSA with MSFA Attention block (MS2A) as basic units.

Spectral Measurement Initialization. Firstly, we reverse the SSI imaging process and rearrange the spectral measurement $\mathbf{Y} \in \mathbb{R}^{H \times W}$ to obtain the initialized low resolution input $\mathbf{I}_{LR} \in \mathbb{R}^{\frac{H}{m} \times \frac{W}{m} \times B}$ as

$$\mathbf{I}_0(x, y, B_\lambda) = \mathbf{Y}(x, y) \odot \mathbf{S}_\lambda, \mathbf{I}_{LR} = f_{MR}(\mathbf{I}_0), \tag{3}$$

where f_{MR} denotes the MSFA-based Rearrangement (MR) using reverse pixel shuffle [41], e.g., given a $m \times m$ MSFA, the downscale factor of f_{MR} is m. Then, we feed $\mathbf{I}_{LR}$ into subsequent model.

Shallow and Deep Feature Extraction. For an initial low-resolution input, $\mathbf{I}_{LR}$, a 3×3 convolutional layer, $h_{sf}(\cdot)$, is employed to derive shallow features, represented as $\mathbf{F}_0 = h_{sf}(\mathbf{I}_{LR}) \in \mathbb{R}^{\frac{H}{m} \times \frac{W}{m} \times C}$, where C is the channel count of the features. This convolutional layer facilitates early visual processing, leading to enhanced stability in optimization and improved results [27]. Additionally, it efficiently maps the input image space into a higher-dimensional feature space. Subsequent extraction of deep features, $\mathbf{F}_{DF} \in \mathbb{R}^{\frac{H}{m} \times \frac{W}{m} \times C}$, from $\mathbf{F}_0$ is defined as

$$\mathbf{F}_{DF} = h_{df}(\mathbf{F}_0), \tag{4}$$

where $h_{df}(\cdot)$ is the deep feature extraction module, comprising K Group Attention Transformer (GAT) blocks with residual connections. Intermediate features $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_K$ and the final deep feature $\mathbf{F}_{DF}$ are sequentially extracted as

$$\begin{aligned} &\mathbf{F}_i = h_{GAT_i}(\mathbf{F}_{i-1}), \quad i = 1, 2, \ldots, K, \\ &\mathbf{F}_{DF} = h_{conv}(\mathbf{F}_K), \end{aligned} \tag{5}$$

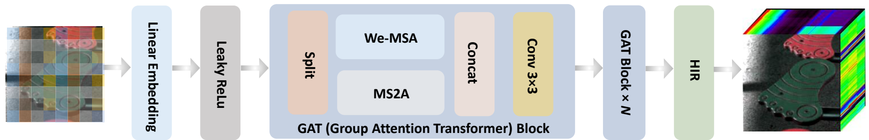

Fig. 3: Overview of the proposed **WeFAT** for MSFA imaging demosaicing, the basic backbone is GAT block with We-MSA and MS2A units shown in Fig. 4 and Fig. 5.

with $h_{GAT_i}(\cdot)$ indicating the i-th GAT and h_{conv} being the concluding convolutional layer. Incorporating a convolutional layer at this stage introduces the inductive bias of convolution operations into the Transformer-based network, enhancing the integration of shallow and deep features.

Hyperspectral Image Reconstruction. The Full-Spectral Resolution (FSR) image, $\mathbf{I}_{FSR}$, is reconstructed by amalgamating shallow and deep features as

$$\mathbf{I}_{FSR} = h_{HIR}(\mathbf{F}_0 + \mathbf{F}_{DF}) \in \mathbb{R}^{H \times W \times C}, \tag{6}$$

where $h_{HIR}(\cdot)$ symbolizes the function of the reconstruction head using sub-pixel convolution layer [41]. Shallow features predominantly encompass low-frequency data, whereas deep features are geared towards capturing high-frequency details. Incorporating an extended skip connection facilitates the direct transfer of low-frequency information to the reconstruction module. This assists the deep feature extraction component in emphasizing high-frequency details and enhances training stability.

4.2 Spectral MSA with Wavelength Embedding

Previous Wavelength Usage Scheme. Current demosaicing techniques employing wavelength primarily concentrate on utilizing wavelength-based cross-correlation to develop interpolation methodologies such as nearest-neighbor, bilinear, bicubic interpolation [40], and ItSD [33]. These methods emphasize local wavelength differences, with stronger cross-correlation for channels closer in optical wavelength. However, precise wavelength information is often overlooked in traditional and deep learning-based demosaicking for multispectral images. Considering that MSIs are organized by wavelength, we propose an embedding mechanism to encode wavelength information across spectral channels.

Our We-MSA. Leveraging non-local self-similarity is common in hyperspectral image reconstruction alongside wavelength information. CNN-based methods face challenges in effectively modeling this aspect. Given the Transformer's capability in capturing non-local dependencies and its success in visual tasks, we aim to explore its potential in MSI reconstruction. However, original Transformers are primarily designed for spatial dimensions, which may not efficiently capture spectral correlations in MSI data. Thus, we initially treat each spectral feature map as a token as [5]. Subsequently, we propose to compute self-attention along the spectral dimension incorporating wavelength embedding. Fig. 4 shows the We-MSA used in We-FAT. The input $\mathbf{X}_{in} \in \mathbb{R}^{H \times W \times C}$ ($\mathbf{X}_{in}, \mathbf{F}' = \mathrm{split}(\mathbf{F}_{i-1})$)

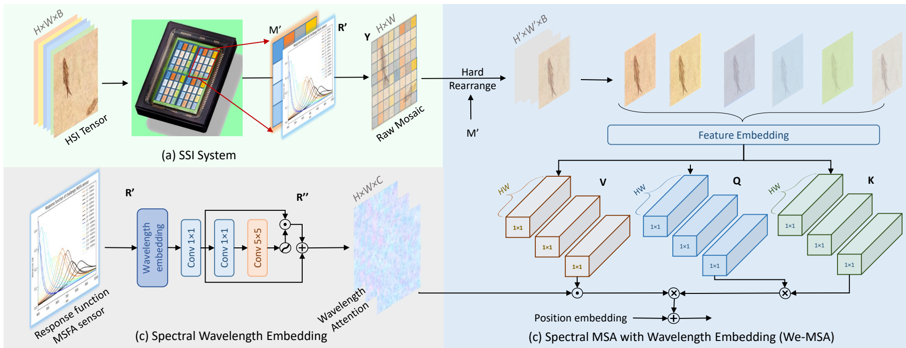

Fig. 4: Illustration of our wavelength embedding mechanism (**We-MSA**).

is reshaped into tokens $\mathbf{X} \in \mathbb{R}^{HW \times C}$. Then $\mathbf{X}$ is linearly projected into *query* $\mathbf{Q} \in \mathbb{R}^{HW \times C}$, *key* $\mathbf{K} \in \mathbb{R}^{HW \times C}$, and *value* $\mathbf{V} \in \mathbb{R}^{HW \times C}$:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}, \mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}, \mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}}, \tag{7}$$

where $\mathbf{W}^{\mathbf{Q}}$, $\mathbf{W}^{\mathbf{K}}$, and $\mathbf{W}^{\mathbf{V}} \in \mathbb{R}^{C \times C}$ are learnable parameters; *biases* are omitted for simplification. Subsequently, we respectively split $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into N *heads* along the spectral channel dimension: $\mathbf{Q} = [\mathbf{Q}_1, \ldots, \mathbf{Q}_N]$, $\mathbf{K} = [\mathbf{K}_1, \ldots, \mathbf{K}_N]$, and $\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_N]$. The dimension of each head is $d_h = \frac{C}{N}$. In attention step, firstly, the pure spectral self-attention is calculated:

$$\mathbf{A}_j = \text{softmax}(\mathbf{K}_j^{\mathrm{T}} \mathbf{Q}_j), \tag{8}$$

where $\mathbf{K}_j^{\mathrm{T}}$ denotes the transposed matrix of $\mathbf{K}_j$.

Then, in addition to the spectral attention, inspired by the timestamp embedding in denoising diffusion models [20], we embed the precise wavelength information to the transformer and make MSA has memory of different wavelength. To be specific, given the peak-wavelength vector of B bands $\mathbf{R}' \in \mathbb{R}^{B \times 1}$ shown in Fig. 4, where B is the number of wavelength (band), since the wavelength of spectral imaging system is typically between *400-1000mm*, which is largely determined by the sensitivity of the human eye and the characteristics of common materials, we firstly normalize $\mathbf{R}^{'}$ to *0-1*, :

$$\mathbf{R}_n' = \text{Normalize}(\mathbf{R}'), \tag{9}$$

Subsequently, to match the scale of the feature maps in each block, $\mathbf{R}_n'$ is expanded with shape $\frac{H}{4} \times \frac{W}{4} \times C$, and pass through an upsample operation to get $\mathbf{R}_u' \in \mathbb{R}^{H \times W \times C}$. Subsequently, $\mathbf{R}_u'$ is processed through a $conv_{1\times1}$ layer and then undergoes an identity mapping to preserve the original wavelength information. Simultaneously, a parallel branch with $conv_{1\times1}$, $conv_{1\times1}$, depth-wise $conv_{5\times5}$, sigmoid activation, and inner product layers is employed to capture both the absolute wavelength information and spectral correlations. Then we have

$$\mathbf{R}_u'' = (\mathbf{W}_1 \mathbf{R}_u') \odot (1 + f_{sa}(f_{dc}(\mathbf{W}_2 \mathbf{W}_1 \mathbf{R}_u')) \in \mathbb{R}^{H \times W \times C}, \tag{10}$$

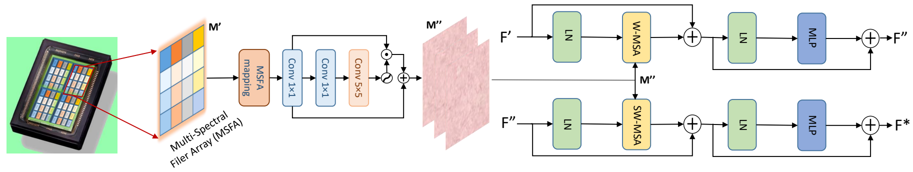

Fig. 5: **MS**FA-attention based window **M**ulti-head **S**elf-**A**ttention (**MS2A**).

where the parameters $\mathbf{W}_1$ and $\mathbf{W}_2$ represent the learnable parameters of two 1×1 convolutional layers denoted as $conv_{1\times1}$. The mapping function of the depth-wise 5×5 convolutional layer is denoted by $f_{dc}(\cdot)$. Additionally, $f_{sa}(\cdot)$ signifies the sigmoid activation. Subsequently, we reshape $\mathbf{R}''_u$ into $\mathbf{R}'' \in \mathbb{R}^{HW\times C}$ to align with the dimensions of $\mathbf{V}$. The reshaped feature map $\mathbf{R}''$ is then divided into N spectral heads denoted as $\mathbf{R}'' = [\mathbf{R}''_1, \ldots, \mathbf{R}''_N]$. Utilizing the wavelength embedding $\mathbf{R}''$ and Eq. (8), the spectral self-attention for each $head_j$ is computed as follows:

$$\mathbf{A}_j = \mathrm{softmax}(\mathbf{K}_j^{\mathrm{T}}\mathbf{Q}_j), \quad head_j = (\mathbf{R}''_j \odot \mathbf{V}_j)\mathbf{A}_j. \tag{11}$$

It is worth noting that the variation in wavelengths significantly impacts spectral density, necessitating attention weights to vary accordingly across different wavelengths. Our embedding strategy, as described in Equation (11), achieves this by adjusting the self-attention $\mathbf{A}_j$ through re-weighting $\mathbf{V}_j$ within $head_j$. Subsequently, the outputs of N *heads* are concatenated in spectral wise to undergo a linear projection and then is added with a position embedding:

$$\text{We-MSA}(\mathbf{X}) = \mathrm{reshape}\left(\left(\underset{j=1}{\overset{N}{\mathrm{Concat}}}(head_j)\right)\mathbf{W} + f_{sp}(\mathbf{V})\right) \in \mathbb{R}^{H\times W\times C}, \tag{12}$$

where the matrix $\mathbf{W}$, with dimensions $\mathbb{R}^{C\times C}$, represents learnable parameters. The function $f_{sp}(\cdot)$ is responsible for generating spectral position embeddings with respect to spectral tokens. It comprises two depth-wise convolutional layers of size 3×3, followed by a GELU activation function and reshape operations.

4.3 Shifted Window based MSFA Attention

Table 1: Ablation study on usage of MSFA, BI: Bilinear interpolation [42], MCM: Mosaic Convolution Module [15], and our MaM.

MSFA Usage	PSNR (↑)	SSIM (↑)
BI($\mathbf{Y}$)	42.88	0.981
MCM($\mathbf{Y}$)	43.22	0.986
MaM($\mathbf{X}_i$)	**47.02**	**0.993**

*$\mathbf{x}_i$ is the input feature of ith GAT block

While We-MSA effectively captures inter-spectral dependencies, it falls short in modeling spatial correlations within HSI representations. Transformers in MSI restoration lack discrimination in attention allocation to spatial regions, leading to equal treatment irrespective of the presence of sampled MSI representations. In the SSI system, a periodic MSFA pattern with dimensions of $m\times m$ samples the hyperspectral cube, ensuring position-sensitive fidelity by associating each spatial pixel with a specific wavelength. Recognizing this, we propose utilizing the MSFA pattern to guide attention towards regions with known MSI representations. Following this, we offer a summary of MSFA integration in prior spectral demosaicing techniques, followed by the introduction of the MSFA-attention Mechanism (MaM), and the resultant MS2A block.

Existing MSFA Usage Scheme. Prior methods [4, 15, 31, 32, 42, 43, 48], primarily employ interpolation or rearrangement operations on the raw mosaic to produce initialized HSIs. While this approach incorporates spatial fidelity information, it is subject to several limitations: **(i)** The operation induces significant pixel shifts, resulting in information loss and spatial discontinuity. **(ii)** It solely operates at the input level, failing to fully leverage the guidance effect of the MSFA in directing the network's attention towards regions represented by sampled HSI data. **(iii)** The absence of learnable parameters to model spatial-wise correlations restricts the effectiveness of this approach.

Our MaM. Different from previous methods, our MaM preserves all the input HSI representations and learns to direct SW-MSA [30] to pay attention to the spatial regions with sampled spectral representations. To be specific, given the MSFA $\mathbf{M}' \in \mathbb{R}^{m \times m}$ shown in Fig. 5, since MSFA has periodic layout across the sensor, we firstly repeat it periodically with shape $H \times W \times B$: $\mathbf{M}'_e = \text{expand}(\mathbf{M}')$, To match the scale of the feature maps in WeFAT, $\mathbf{M}'_e$ passes through a downsample operation. Subsequently, $\mathbf{M}'_e$ undergoes the same encoding layers as (10),

$$\mathbf{M}'' = (\mathbf{W}_1 \mathbf{M}'_e) \odot (1 + f_{sa}(f_{dc}(\mathbf{W}_2 \mathbf{W}_1 \mathbf{M}'_e))). \tag{13}$$

Then we split $\mathbf{M}''$ into N *heads* in spectral wise: $\mathbf{M}'' = [\mathbf{M}''_1, \ldots, \mathbf{M}''_N]$. For each $head_j$, MaM conducts its guidance by re-weighting $\mathbf{V}_j$ using $\mathbf{M}''_j$. Hence, when using MaM to direct SW-MSA [30], the SW-MSA module just needs to make a simple modification by re-formulating $head_j$ in Eq. (11):

$$head_j = (\mathbf{M}_j \odot \mathbf{V}_j)\mathbf{A}_j. \tag{14}$$

By incorporating MaM into the SW-MSA, we form the MS2A block, as illustrated in Fig. 5, where $\mathbf{X}_{in}, \mathbf{F}' = \text{split}(\mathbf{F}_{i-1})$. A comparative analysis between MaM and conventional MSFA utilization is presented in Table 1. By using MaM, SW-MSA can extract sampled MSI representations, enjoy the guidance of MSFA-sensitive fidelity information, and adaptively model the spatial-wise interactions.

4.4 Group Attention Transformer

The proposed MaM based MS2A and WeFAT blocks effectively represent both the MSFA pattern and wavelength information, crucial for determining the data output of different spectral cameras, within transformer blocks. To optimize computational efficiency, inspired by group convolution principles, we partition the underlying features across channels. The first half is directed to the WeFAT block, while the second half is routed to the MaM block. Subsequently, the outputs of these blocks are combined and fused using a spectral convolution layer employing pointwise convolution to capture spectral correlations between the first half and second half features, resulting in the final output. This configuration is denoted as the Group Attention Transformer (GAT) block, as shown in Fig. 3.

Table 2: Quantitative analysis of spectral demosaicing on the ARAD dataset [2].

Datasets	Method	WB [4]	BTES [31]	PPID [32]	GRMR [43]	In-Net [42]	MCAN [15]	**WeFAT-S**	**WeFAT-M**
Scene 1	PSNR (↑)	29.27	29.52	36.87	29.35	44.98	41.60	50.20	**50.35**
	SSIM (↑)	0.956	0.947	0.969	0.960	0.993	0.988	0.997	**0.997**
	SAM (↓)	0.093	0.089	0.090	0.096	0.011	0.020	0.018	**0.012**
	MRAE (↓)	-	-	-	-	0.014	0.022	0.012	**0.004**
Scene 2	PSNR ↑	30.95	31.06	39.10	30.99	41.09	36.92	45.69	**45.78**
	SSIM (↑)	0.965	0.955	0.977	0.967	0.985	0.936	0.989	**0.996**
	SAM (↓)	0.089	0.078	0.063	0.085	0.041	0.065	0.021	**0.011**
	MRAE (↓)	-	-	-	-	0.037	0.071	0.019	**0.012**
Scene 3	PSNR ↑	33.85	33.49	38.81	33.96	43.50	45.29	49.38	**50.13**
	SSIM (↑)	0.943	0.929	0.959	0.944	0.984	0.991	0.996	**0.997**
	SAM (↓)	0.113	0.134	0.079	0.113	0.033	0.030	0.017	**0.012**
	MRAE (↓)	-	-	-	-	0.043	0.038	0.020	**0.011**
50 Scenes averaged	PSNR (↑)	31.17	30.94	35.98	31.38	42.88	43.22	47.57	**48.03**
	SSIM (↑)	0.912	0.892	0.937	0.922	0.981	0.986	0.993	**0.994**
	SAM (↓)	0.158	0.176	0.121	0.150	0.034	0.034	0.020	**0.015**
	MRAE (↓)	-	-	-	-	0.043	0.044	0.023	**0.018**

5 Experiments

Datasets. We trained our model and all comparison models only on the ARAD dataset [2] from scratch. This dataset comprises 384 full spatial-spectral resolution hyperspectral image (HSI) cubes with 16 bands, each of size 320×320 pixels. Subsequently, we evaluated the trained models on the ARAD validation dataset first, followed by testing on the KAIST [10] datasets to assess their general capability. In addition, to further assess practical utility, we conduct tests on real HSI images acquired through an IMEC spectral camera with a 4×4 pattern and 1088×2048 spatial resolution, covering the spectral range of 460nm to 600nm, collected by ourselves. For these images, we down-sample the mosaic image according to the MSFA pattern to get a low-resolution (LR) reference.

Specific experiments were also performed using the CAVE [36] and the ICVL dataset [1], encompassing more than 100 scenes with diverse MSFA patterns. Details of these experiments are available in the supplementary material.

Benchmarked models. We evaluated six state-of-the-art methods to compare with WeFAT: classic interpolation-based WB [4], banary tree-based generic demosaicing algorithm (BTES) [31], pseudo-panchromatic image-directed demosaicing (PPID) [32], graph and rank regularized demosaicing (GRMR) [43], 3-D convolution-based demosaicing model InNet [42], and mosaic convolution and attention network (MCAN) [15]. All results were generated using the provided code from the respective authors with classic 4×4 mosaic pattern.

Metrics. In quantitative evaluation, we used four different indexes widely used for spectral image processing, namely (i) Peak Signal-to-Noise Ratio (PSNR), which is the classical PSNR metric averaged across bands; (ii) Structural Similarity Index Measurement (SSIM) [44]; (iii) Spectral Angle Map (SAM); (iv) Mean Relative Absolute Error (MRAE) [21].

Implementation details. We evaluated two variants of our model: **WeFAT-S**, featuring 4 GAT blocks, and **WeFAT-M**, comprising 6 GAT blocks, varying in scale. Additional details are available in the supplementary materials.

Fig. 6: Visual comparison of HSI demosaicing methods on ARAD validation dataset (False color, R: 705nm, G: 530nm, B: 555nm). The MSFA pattern is shown in Fig. 1. The image in left side is the reference image or mosaic image, the patches are reference patch and demosaiced full spatial-spectral resolution image.

5.1 MAIN RESULTS

Results on synthetic data. Firstly, we show the quantitative results of ARAD dataset which comprises 50 scenes on Table 2, In the table, results are presented for Scene 1 through Scene 3, with the superior outcomes highlighted in bold. Additionally, aggregated outcomes across all 50 scenes are provided herein, while the details for each individual scene are available in the supplementary material.

- Our wavelength embedding and MSFA attention based approach achieves state-of-the-art results in all scenes.
- All the deep learning methods consistently demonstrate high performance on the ARAD validation set, achieving PSNR values exceeding 40dB. This success can be attributed to the fact that the models undergo training on the ARAD training set. Importantly, both the training and test datasets are acquired using a camera characterized by identical wavelength distribution and a consistent MSFA layout.

Secondly, we present the quantitative results of the KAIST dataset in Table 3. The table displays the averaged PSNR, SSIM, and SAM values across all 10 scenes. In this scenario, the test data differs in wavelength and MSFA pattern compared to the training set ARAD.

- WeFAT consistently performs well across all quantitative metrics (PSNR, SSIM, and SAM). For example, while InNet and MCAN achieved average

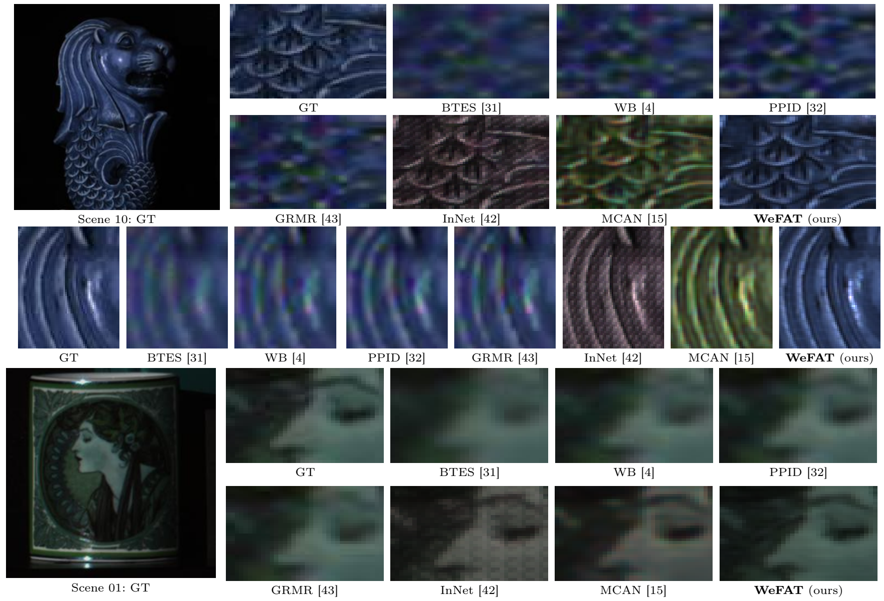

Fig. 7: Visual evaluation of HSI demosaicing techniques using the KAIST dataset. The image depicts false color rendering with wavelengths assigned as follows: Red channel (R) at 458nm, Green channel (G) at 504nm, and Blue channel (B) at 537nm. Notably, the MSFA utilized differs from the training dataset ARAD, primarily in wavelength distribution and MSFA layout, as illustrated in Fig. 1.

Table 3: Quantitative spectral demosaicing comparison using the KAIST dataset featuring 10 scenes, highlighting superior results in bold.

Dataset	Method	GRMR [43]	BTES [31]	WB [4]	PPID [32]	In-Net [42]	MCAN [15]	**WeFAT**
KAIST 10 scenes averaged	PSNR (↑)	30.209	29.218	29.933	32.198	36.04	35.23	**43.63**
	SSIM (↑)	0.909	0.874	0.898	0.926	0.946	0.892	**0.988**
	SAM (↓)	0.104	0.135	0.115	0.093	0.161	0.206	**0.052**

PSNR values of 42.88dB and 43.22dB, respectively, in the ARAD validation set, their PSNR values dropped to 36.04dB and 35.23dB, respectively, in the KAIST dataset. In contrast, WeFAT maintained a PSNR of 43.63dB. This performance advantage is due to the robust generality provided by wavelength embedding and MSFA attention mechanisms, enabling WeFAT to excel across various camera configurations.

We present visual comparisons between the ARAD dataset and the KAIST dataset in Fig. 6 and Fig. 7. Models trained exclusively on the ARAD dataset perform effectively on the ARAD validation set, showing strong spectral consistency and adequate detail. However, these models exhibit suboptimal performance when applied to the KAIST dataset. Conversely, our model, which integrates wavelength embedding and is trained solely on the ARAD dataset, demonstrates superior performance on the KAIST dataset without requiring additional external training data.

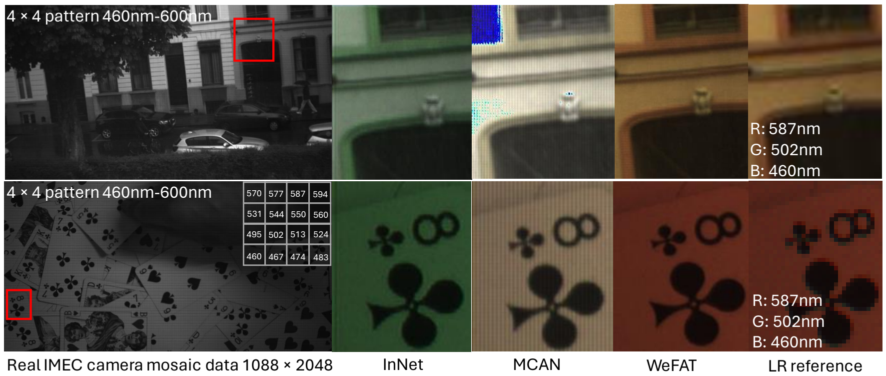

Fig. 8: Visual comparison of spectral demosaicing on the IMEC real dataset(False color, R: 587nm, G: 502nm, B: 460nm).

Results on real dataset. In addition to the simulated experiments, we also captured five real images to test the practical capability of the proposed model. The spectral camera utilized for capturing the real scene adheres to a 4×4 pattern, with a wavelength range spanning from 400nm to 660nm. Its spatial dimensions measure 1088×2048. The MSFA pattern is depicted on the left side of Fig. 8. The captured raw mosaic data is with 10bit.

We present the visual comparison of IMEC real image in Fig. 8, where the LR reference denotes the low-resolution version of the mosaic image, acquired through MSFA based hard rearrangement [15]. MCAN exhibits shortcomings in maintaining spectral consistency and has periodic artifacts. InNet manages to reconstruct the scene's structure, but its spectral fidelity is relatively low. In contrast, our WeFAT successfully preserves spectral consistency while also reconstructing meaningful details. Here, all the models are trained on ARAD dataset from scratch.

5.2 Ablation Study

Component ablation: To evaluate the efficacy of the proposed MSFA attention mechanism and wavelength embedding strategy, an ablation study was conducted comparing the WeFAT model with four component groups. The first row in Table 4 is WeFAT without MSFA attention and wavelength embedding, utilizing window-based SW-MSA and spectral transformer as the basic unit. The second and third rows correspond to WeFAT with only MSFA attention or wavelength embedding respectively, while the fourth row represents WeFAT with both mechanisms incorporated. Pure MSFA attention resulted in a 0.1dB improvement in PSNR, while solely incorporating wavelength embedding yielded a marginal 0.001 SSIM improvement but

Table 4: Ablation study on components.

MSFA Attention	Wavelength Embedding	PSNR (↑)	SSIM (↑)
✗	✗	46.92	0.992
✓	✗	47.02	0.993
✗	✓	47.87	0.994
✓	✓	**48.03**	**0.994**

Table 5: Analysis of attention complexity.

Category	GAT	Global MSA	Window MSA	S-MSA
Complexity	$O(\frac{HWC^2}{2N} + M^2HWC)$	$O(2(HW)^2C)$	$O(2M^2HWC)$	$O(\frac{2HWC^2}{N})$

with a 0.95dB PSNR increase. When wavelength embedding was combined with MSFA attention, WeFAT exhibited a significant 1.11dB PSNR improvement. This improvement can be attributed to the fact that the spectral transformer primarily focuses on spectral correlation modeling, neglecting the coupled spatial MSFA pattern information. Thus, the inclusion of MSFA attention is necessary to adequately model the absent spatial information. Here, all results are from testing on the ARAD dataset.

Complexity analysis: We firstly compare WeFAT with CNN and attention based spectral demosaicing approaches in Table 6, then analyze the computational complexity of GAT and compare it with other MSAs. We only compare the main difference, *i.e.*, the self-attention mechanism in Table 5, in which Global-MSA denotes the original global MSA [13], Window-MSA denotes the local window-based MSA [30], and M represents the window size, S-MSA is the pure spectral MSA [5]. The computational complexity of GAT and Window-MSA and S-MSA is linear to the spatial size HW. This cost is much cheaper than that of Global-MSA (quadratic to HW). Meanwhile, the We-MSA of GAT treats a whole spectral feature map as a token, while MaM employ the advance of Window-MSA with MSFA attention. Therefore, our GAT provides a comprehensive receptive field that spans both spatial and spectral dimensions. Simultaneously, it adeptly captures both global MSFA periodic information and local MSFA patterns.

Table 6: Performance-Params-FLOPs comparisons with open-source SOTA methods.

Method	Category	Params (M↓)	FLOPs (G↓)	PSNR (↑)	SSIM (↑)
InNet	CNN	0.87	1430.63	42.88	0.981
MCAN	Attention	1.37	29.90	43.22	0.986
WeFAT-S	Attention	1.33	11.60	**47.57**	**0.993**
WeFAT-M	Attention	1.71	14.03	**48.03**	**0.994**

6 Conclusion

We have investigated embedding wavelength and MSFA attention into transformer-based spectral demosaicing. Unlike prior works using attention or wavelength correlation, we do not directly utilize relative wavelength correlation to model local spectral similarity or employ MSFA-based interpolation or rearrangement operations only in the initial step. Instead, we conceptualize each spectral feature as a token, integrating wavelength information directly into the attention calculation. Furthermore, we have developed a MSFA-attention Mechanism (MaM) to steer the attention mechanism (We-MSA) towards spatial areas with sampled spectral data. This straightforward yet effective strategy demonstrates notable performance, particularly when applied to datasets with diverse camera settings, encompassing varying wavelengths and MSFA patterns.

References

1. Arad, B., Ben-Shahar, O.: Sparse recovery of hyperspectral signal from natural rgb images. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 19–34. Springer (2016)
2. Arad, B., Timofte, R., Yahel, R., Morag, N., Bernat, A., Wu, Y., Wu, X., Fan, Z., Xia, C., Zhang, F., et al.: Ntire 2022 spectral demosaicing challenge and data set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 882–896 (2022)
3. Beletkaia, E., Pozo, J.: More than meets the eye: Applications enabled by the non-stop development of hyperspectral imaging technology. PhotonicsViews **17**(1), 24–26 (2020)
4. Brauers, J., Aach, T.: A color filter array based multispectral camera. In: 12. Workshop Farbbildverarbeitung. pp. 55–64. Ilmenau (2006)
5. Cai, Y., Lin, J., Hu, X., Wang, H., Yuan, X., Zhang, Y., Timofte, R., Van Gool, L.: Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17502–17511 (2022)
6. Cao, X., Yue, T., Lin, X., Lin, S., Yuan, X., Dai, Q., Carin, L., Brady, D.J.: Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world. IEEE Signal Processing Magazine **33**(5), 95–108 (2016)
7. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imag. Vis. **20**(1), 89–97 (2004)
8. Chang, C.I.: Hyperspectral data exploitation: theory and applications. John Wiley & Sons (2007)
9. Chen, Y., Zhang, H., Wang, Y., Ying, A., Zhao, B.: Admm-dsp: A deep spectral image prior for snapshot spectral image demosaicing. IEEE Transactions on Industrial Informatics (2023)
10. Choi, I., Kim, M., Gutierrez, D., Jeon, D., Nam, G.: High-quality hyperspectral reconstruction using a spectral prior. In: Technical report (2017)
11. Cui, Q., Park, J., Smith, R.T., Gao, L.: Snapshot hyperspectral light field imaging using image mapping spectrometry. Optics letters **45**(3), 772–775 (2020)
12. Dalton, G.M., Collins, N.M., Clifford, J.M., Kemp, E.L., Limpanukorn, B., Jimenez, E.S.: Monte-carlo modeling and design of a high-resolution hyperspectral computed tomography system with multi-material patterned anodes for material identification applications. In: Developments in X-Ray Tomography XIII. vol. 11840, pp. 91–108. SPIE (2021)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Feng, K., Zeng, H., Zhao, Y., Kong, S.G., Bu, Y.: Unsupervised spectral demosaicing with lightweight spectral attention networks. IEEE Transactions on Image Processing **33**, 1655–1669 (2024)
15. Feng, K., Zhao, Y., Chan, J.C.W., Kong, S.G., Zhang, X., Wang, B.: Mosaic convolution-attention network for demosaicing multispectral filter array images. IEEE Transactions on Computational Imaging **7**, 864–878 (2021)
16. Florez-Ospina, J.F., Alrushud, A.K., Lau, D.L., Arce, G.R.: Block-based spectral image reconstruction for compressive spectral imaging using smoothness on graphs. Opt. Exp. **30**(5), 7187–7209 (2022)

17. Geelen, B., Blanch, C., Gonzalez, P., Tack, N., Lambrechts, A.: A tiny vis-nir snapshot multispectral camera. In: Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VIII. vol. 9374, pp. 194–201. SPIE (2015)
18. Gupta, M., Rathi, V., Goyal, P.: Adaptive and progressive multispectral image demosaicking. IEEE Transactions on Computational Imaging **8**, 69–80 (2022)
19. Habtegebrial, T.A., Reis, G., Stricker, D.: Deep convolutional networks for snapshot hypercpectral demosaicking. In: 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS). pp. 1–5. IEEE (2019)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
21. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. International journal of forecasting **22**(4), 679–688 (2006)
22. Jaiswal, S.P., Fang, L., Jakhetiya, V., Pang, J., Mueller, K., Au, O.C.: Adaptive multispectral demosaicking based on frequency-domain analysis of spectral correlation. IEEE Transactions on Image Processing **26**(2), 953–968 (2016)
23. Koundinyan, S., Thompson, K.R., Suknot, A.: Material identification and classification using machine learning techniques with hyperspectral computed tomography. Tech. rep., Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (2018)
24. Lapray, P.J., Wang, X., Thomas, J.B., Gouton, P.: Multispectral filter arrays: Recent advances and practical implementation. Sensors **14**(11), 21626–21659 (2014)
25. Li, Z., Nie, F., Chang, X., Nie, L., Zhang, H., Yang, Y.: Rank-constrained spectral clustering with flexible embedding. IEEE Trans. Neural Netw. Learn. Syst. **29**(12), 6073–6082 (Dec 2018)
26. Li, Z., Nie, F., Chang, X., Yang, Y., Zhang, C., Sebe, N.: Dynamic affinity graph construction for spectral clustering using multiple features. IEEE Trans. Neural Netw. Learn. Syst. **29**(12), 6323–6332 (Dec 2018)
27. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
28. Liu, S., Zhang, Y., Chen, J., Lim, K.P., Rahardja, S.: A deep joint network for multispectral demosaicking based on pseudo-panchromatic images. IEEE Journal of Selected Topics in Signal Processing **16**(4), 622–635 (2022)
29. Liu, Y., Yuan, X., Suo, J., Brady, D., Dai, Q.: Rank minimization for snapshot compressive imaging. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 2990–3006 (Dec 2019)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
31. Miao, L., Qi, H., Ramanath, R., Snyder, W.E.: Binary tree-based generic demosaicking algorithm for multispectral filter arrays. IEEE Transactions on Image Processing **15**(11), 3550–3558 (2006)
32. Mihoubi, S., Losson, O., Mathon, B., Macaire, L.: Multispectral demosaicing using pseudo-panchromatic image. IEEE Transactions on Computational Imaging **3**(4), 982–995 (2017)
33. Mizutani, J., Ogawa, S., Shinoda, K., Hasegawa, M., Kato, S.: Multispectral demosaicking algorithm based on inter-channel correlation. In: 2014 IEEE Visual Communications and Image Processing Conference. pp. 474–477. IEEE (2014)

34. Monno, Y., Kikuchi, S., Tanaka, M., Okutomi, M.: A practical one-shot multispectral imaging system using a single image sensor. IEEE Transactions on Image Processing **24**(10), 3048–3059 (2015)
35. Pan, Z., Li, B., Bao, Y., Cheng, H.: Deep panchromatic image guided residual interpolation for multispectral image demosaicking. In: 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS). pp. 1–5. IEEE (2019)
36. Park, J.I., Lee, M.H., Grossberg, M.D., Nayar, S.K.: Multispectral imaging using multiplexed illumination. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
37. Rathi, V., Goyal, P.: Convolution filter based efficient multispectral image demosaicking for compact msfas. In: VISIGRAPP (4: VISAPP). pp. 112–121 (2021)
38. Rathi, V., Goyal, P.: Multispectral image demosaicking based on novel spectrally localized average images. IEEE Signal Processing Letters **29**, 449–453 (2021)
39. Rathi, V., Goyal, P.: Generic multispectral demosaicking using spectral correlation between spectral bands and pseudo-panchromatic image. Signal Processing: Image Communication **110**, 116893 (2023)
40. Rifman, S.S.: Digital rectification of erts multispectral imagery. In: NASA. Goddard Space Flight Center Symp. on Significant Results obtained from the ERTS-1, Vol. 1, Sect. A and B. No. PAPER-I6 (1973)
41. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
42. Shinoda, K., Yoshiba, S., Hasegawa, M.: Deep demosaicking for multispectral filter arrays. arXiv preprint arXiv:1808.08021 (2018)
43. Tsagkatakis, G., Bloemen, M., Geelen, B., Jayapala, M., Tsakalides, P.: Graph and rank regularized matrix recovery for snapshot spectral image demosaicing. IEEE Transactions on Computational Imaging **5**(2), 301–316 (2018)
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
45. Wisotzky, E.L., Daudkane, C., Hilsmann, A., Eisert, P.: Hyperspectral demosaicing of snapshot camera images using deep learning. In: DAGM German Conference on Pattern Recognition. pp. 198–212. Springer (2022)
46. Zhang, T., Fu, Y., Li, C.: Hyperspectral image denoising with realistic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2248–2257 (October 2021)
47. Zhang, T., Fu, Y., Wang, L., Huang, H.: Hyperspectral image reconstruction using deep external and internal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8559–8568 (2019)
48. Zhao, B., Zheng, J., Dong, Y., Shen, N., Yang, J., Cao, Y., Cao, Y.: Ppi edge infused spatial-spectral adaptive residual network for multispectral filter array image demosaicing. IEEE Transactions on Geoscience and Remote Sensing (2023)