Optica Applicata, Vol. LIV, No. 4, 2024
DOI: 10.37190/oa/193937

Pseudo-panchromatic image guided transformer model for multispectral image demosaicing

Peng Chen[1], Heng Wang[2,*], Cong Wei[3], Jiangnan Yang[1], Xinyu Su[1], Jingjun Wu[3,*], Shuangli Li[1,*]

[1]School of Information Engineering, Southwest University of Science and Technology, Mianyang, 621000, Sichuan, China

[2]Department of Engineering Physics, Tsinghua University, Beijing, 100084, Beijing, China

[3]School of Electronic and Optical, Nanjing University of Science and Technology, Nanjing, 210094, Jiangsu, China

[*]Corresponding authors: slliu@swust.edu.cn (SL); wangheng21@mails.tsinghua.edu.cn (HW); jingjunwu163@163.com (JW)

The multispectral imaging system using the filter array can capture the multispectral information of the scene in one snapshot and reconstruct the complete multispectral image by demosaicing. However, the sparse sampling rate makes image captured by demosaicing a challenging problem. Although a lot of demosaicing algorithms have been developed, the existing well-performing methods have limitations in modeling non-local dependencies which lead to artifacts. To solve this problem, this paper proposes a transformer-based multispectral image demosaicing model to address the problem. The proposed model comprises a pseudo-panchromatic image generation network and a transformer-based multispectral image reconstruction network. Additionally, we designed a fusion module to combine the pseudo-panchromatic image with the raw mosaic image captured by the camera, leveraging the correlation between the band of multispectral images to improve the performance of the model. The experimental results show that the proposed method has the advantages of high reconstruction precision, strong anti-noise interference ability, and small calculation amount, which provides a better image reconstruction solution for constructing a high-quality multispectral imaging system applied to multiple scenes.

Keywords: multispectral image, demosaicing, pseudo-panchromatic image, deep learning, transformer.

1. Introduction

The multispectral image (MSI) contains the scene's two-dimensional space and the one-dimensional spectral information that can reflect the physical and chemical char-

acteristics of the target, which can better help people analyze the objects in the scene. Thus, MSI has been increasingly used in medical imaging [1], object detection [2], food safety [3], surveying and mapping [4,5] and other fields. How to efficiently acquire MSI has attracted wide attention. The traditional multispectral imaging method [6] is based on scanning, such as scanning the spectrum along the spatial dimension, which produces high hardware costs and is more time-consuming, making it difficult to exploit in complex dynamic scenes. With the widespread application of micro- and nanotechnology in fields such as terahertz [6], mid-infrared emitters [7,8], and color sensors [9], multispectral filter arrays (MSFA) featuring micro- and nanostructures have garnered significant attention from researchers. Snapshot multispectral cameras based on MSFA offer advantages such as instantaneous data acquisition, small hardware size, and low cost. The MSFA is the core component of the camera. The MSFA is a set of spectral filters integrated into a sensor. The image taken by the MSFA camera is called the raw mosaic image. Each pixel on the raw mosaic image only records a single band of sense, and the information on other bands is lost. Consequently, the missed information must be estimated from the acquired sparse spatial data. This process is called MSFA demosaicing. The MSFA demosaicing is challenging. As the number of spectral bands in the array increases, the spatial sampling rate is further reduced, the information loss in the MSI will increase and the reliability of the image will be reduced.

Many demosaicing algorithms have been developed. The researchers first considered using interpolation to solve the problem. BRAUERS *et al*. [10] proposed the weighted bilinear (WB) interpolation method. GUPTA *et al*. [11] proposed a universal multispectral image demosaicing method based on the residual method (RD). MIHOUBI *et al*. [12] applied a pseudo-panchromatic image (PPI) to the residual method for multispectral image reconstruction. The reconstruction effect is better than that of RD. Interpolation algorithms for demosaicing mainly use spectral correlation and spatial correlation. The further reduction of the sparse sampling rate will seriously weaken the spatial correlation and spectral correlation of images, making it difficult for interpolation algorithms to achieve better reconstruction results. Some researchers also use matrix solutions to solve this problem. BIAN *et al*. [13] report a generalized demosaicing method with structural and adaptive nonlocal optimization, enabling boosted reconstruction accuracy for different MSFAs. In recent years, deep learning has demonstrated exceptional capability in many areas of imaging, such as, object detection [14], image noise reduction [15], and image fusion [16]. Researchers [17-19] have applied deep learning methods to MSFA demosaicing. SHOPOVSKA *et al*. [20] uses U-net networks to leverage the cross-band dependencies to recover details in textured regions, realizing reconstructing RGB and near-infrared spectra. LIU *et al*. [21] proposes a novel end-to-end deep learning framework based on pseudo-panchromatic images, which consists of two networks, the deep PPI generation network (DPG-Net) and the deep demosaicing network (DDM-Net). FENG *et al*. [22] considered the position-coding of the filter array and proposed the MCAN network. The analysis of the aforementioned literature yields the following conclusions: 1) The introduction of PPI can effectively improve the qual-

ity of multispectral image reconstruction. 2) Deep learning algorithms based on convolutional neural networks (CNN) have greatly improved the quality of reconstruction. However, it is difficult for convolutional neural networks to capture long-range correlation and spectral similarities, leading to artifacts and other problems.

The transformer [23] is a model architecture that has garnered considerable attention, initially attaining remarkable success in natural language processing (NLP) and subsequently being adapted for visual tasks. The core concept of the transformer lies in its self-attention mechanism, which can capture dependencies between any two positions in the input data, regardless of their distance in the sequence. In the context of visual tasks, the most notable application is the vision transformer [24] (ViT), which divides an image into a series of small patches and flattens these patches into a sequence, similar to how word sequences are processed in NLP. This sequence of image patches is then fed into a standard transformer network. Through this process, the transformer learns global dependencies between patches, allowing it to capture comprehensive features across the entire image. Unlike traditional convolutional neural networks (CNNs), which rely on convolution operations to extract features, transformers utilize a global attention mechanism, offering advantages in handling long-range dependencies and complex structures. While transformers typically require more computational resources and larger datasets for training compared to CNNs, they have demonstrated strong performance across various computer vision tasks, such as image classification, object detection, and image generation, particularly on large-scale datasets.

Based on the above literature analysis, we introduce an efficient MSFA demosaicing network utilizing the transformer architecture. The algorithm comprises two main networks: the PPI generation network and the multispectral image reconstruction network. Initially, we employ the PPI generation network, leveraging the self-attention mechanism to achieve efficiency. Subsequently, the generated PPI and the raw mosaic are simultaneously sent to the multispectral image reconstruction network to reconstruct the MSI. Finally, we linearly combine the loss functions of PPI and MSI as the loss function of the entire network. Besides, we propose a new module, an enhanced version of the transformer model. This is designed to better reconstruction. This module incorporates sophisticated mechanisms to effectively utilize the intricate details provided by PPI data, thereby improving the accuracy and quality of reconstruction. Through a series of experiments, we prove the superiority of the proposed method and its applicability to real scenarios and test the stability of the model. The proposed method can be used to compose multispectral imaging systems with higher performance.

2. System model and preliminaries

2.1. MSFA camera model

The MSFA camera is an imaging device that captures information of multiple spectral bands. The MSFA camera covers the CCD with a set of tiny spectral filters, each allowing only a specific wavelength of light to pass through. These filters are arranged on the sensor in a specific array so that each pixel only receives spectral information in a cer-

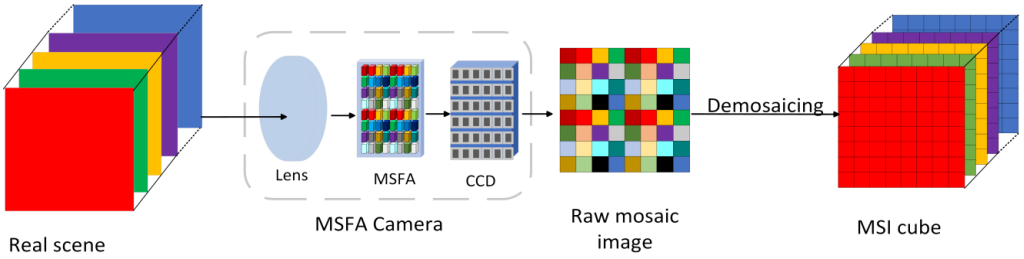

Fig. 1. Multispectral image acquisition process of MSFA-based camera.

tain band. The system includes a CCD sensor, MSFA, lens system, image processing unit, and other modules. As shown in Fig.1, the imaging process begins with light from the scene entering the camera through a lens system, carrying spectral information across different wavelengths. Upon entering the camera, the light passes through MSFA. The filtered light then reaches the CCD sensor. These signals are subsequently converted into digital data by the CCD sensor, producing a two-dimensional image array with spectral information across different bands. As each pixel in the MSFA image contains information from only one wavelength, demosaicing algorithms, such as interpolation, are typically employed to recover lost wavelength information.

Assuming ideal optics and homogeneous spectral sensitivity of the sensors, the value of the raw mosaic image at the pixel (x, y) is expressed as

$$I(x, y) = \int_{\Omega} E(\lambda) \bullet R(\lambda, x, y) \bullet T(\lambda, x, y)\, \mathrm{d}\lambda \tag{1}$$

where is the spectral wavelength range of MSI. λ is the spectral wavelength. The term is the relative spectral power distribution of the light source. $R(\lambda, x, y)$ represents the reflectivity of an object in space position (x, y) to light. $T(\lambda, x, y)$ represents the filter function of the multispectral filter array at the spatial position (x, y).

2.2. Pseudo-panchromatic image

The ground-truth (GT) PPI refers to the mean image over all bands of MSI:

$$I^{\mathrm{PPI}} = \frac{1}{C} \sum_{c=1}^{C} I^{c} \tag{2}$$

where C represents the number of bands in MSI, and I^{c} denotes spectral image of the c-th channel. MIHOUBI *et al.* [12] demonstrates that when the central wavelengths of two frequency bands are significantly separated, they correlate more with the PPI than between each other. For instance, the correlation between a spectral image centered at 455 nm and another centered at 755 nm is lower than the correlation between the 455 nm spectral image and the PPI. LIU *et al.* [21] discussed the relationship between PPI and MSI in detail and concluded that there is a strong correlation between PPI and

multispectral images. Therefore, it is of great significance to efficiently utilize the strong correlation between PPI and MSI to guide the reconstruction of MSI.

3. Methodology

This section introduces our proposed deep-learning algorithm. The comprehensive network architecture of the algorithms is initially presented, followed by a detailed description of the operation of each module.

3.1. Network framework

An overview of the proposed architecture is depicted in Fig. 2. It is a two-stage of network containing a PPI generation network (PPIGN) and a multispectral image reconstruction network (MIRN). We first apply a PPI generation network to generate PPI from the raw mosaic image.

$$I^{\mathrm{PPI}} = H_{\mathrm{PPIGN}}(I^{\mathrm{MSFA}}) \tag{3}$$

where $H_{\mathrm{PPIGN}}(\cdot)$ represents the operations of the PPI generation network.

Then, the obtained I^{PPI} and the raw mosaic image I^{MSFA} are sent to a multispectral image reconstruction network to recover MSI.

$$I^{\mathrm{recon}} = H_{\mathrm{MIRN}}(I^{\mathrm{PPI}}, I^{\mathrm{MSFA}}) \tag{4}$$

where $H_{\mathrm{MIRN}}(\cdot)$ stands for the operations of a multispectral image reconstruction network. I^{recon} stands for reconstructed multispectral image.

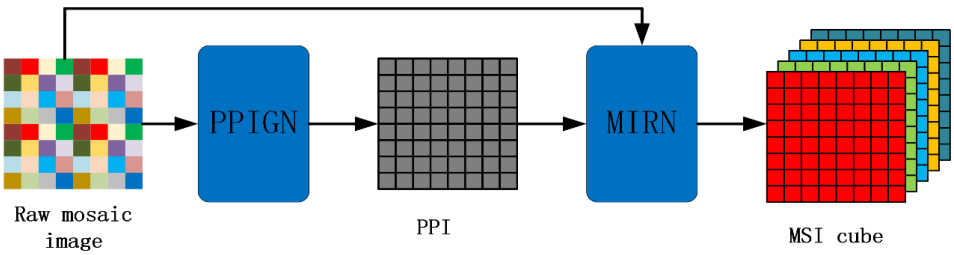

Fig. 2. The comprehensive architecture of the multispectral image reconstruction network is presented.

3.2. PPI generation network

The primary objective of the proposed PPIGN is to generate a PPI that closely approximates the real PPI. Figure 3(a) illustrates how PPIGN generates PPI from raw mosaic images. PPIGN consists of two branches: a smooth filtering branch and a high-frequency detail recovery branch. Initially, we use a smoothing filter to generate a preliminary PPI. However, there remains a significant disparity between this preliminary PPI and the actual PPI. Therefore, we introduce a network to enhance the PPI by recovering high

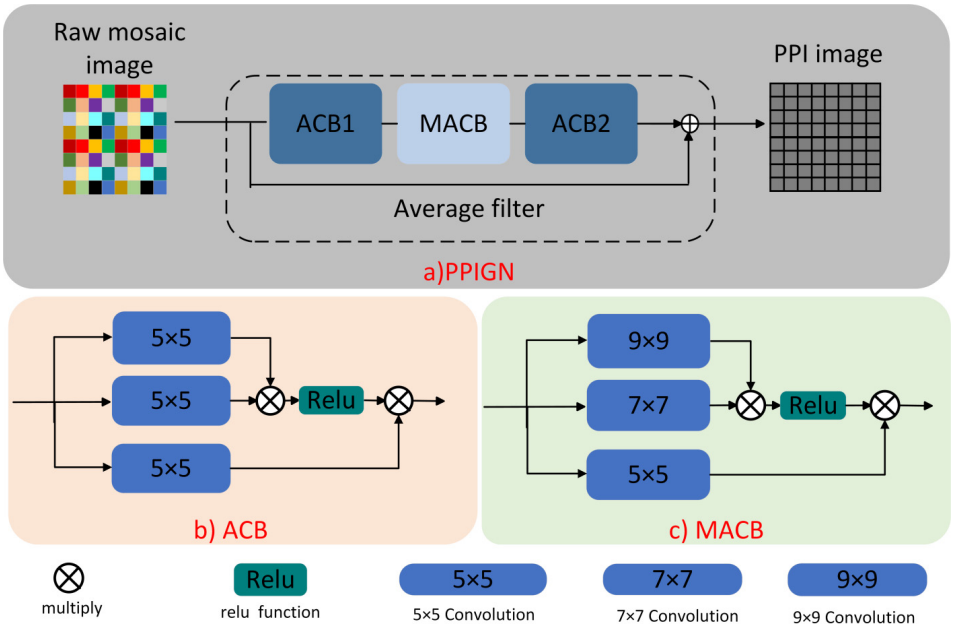

Fig. 3. The PPI generation network. (a) Overall structure of PPI generation network. (b) Structure of attention-convolution block. (c) Structure of multi-scale attention-convolution block.

-frequency information. After the high-frequency information is recovered, it is added back to the preliminary PPI to produce the final sharpened PPI.

Specifically, the high-frequency detail recovery branch comprises two attention-convolution blocks (ACB) and a multi-scale attention-convolution block (MACB). An ACB is used for shallow feature extraction. Then, a MACB is employed for the feature transformation of these extracted shallow features. Finally, an ACB is utilized to aggregate the feature information to obtain the output feature.

The structure of ACB and MACB is shown in Fig. 3(b) and (c), which are based on the transformer module. ACB consists of three 5×5 convolution kernels and one activation function. First, three 5×5 convolution kernels extract features from the input data respectively to obtain three different features. Then, two of the features are multiplied, and the activation function is applied nonlinearly to generate attention features. These attention features are then multiplied with the remaining feature to get the final output. MACB is a further improvement of ACB, mainly by transforming the two 5×5 convolution kernels for extracting attention features into 7×7 convolution kernels and 9×9 convolution kernels.

3.3. Multispectral image reconstruction network

The proposed MIRN is the cornerstone of our architecture, adopting a U-Net-like overall structure, as illustrated in Fig. 4(a). MIRN consists of an encoder, a bottleneck, a decoder, and three convolution layers, and employs up-sampling and down-sampling

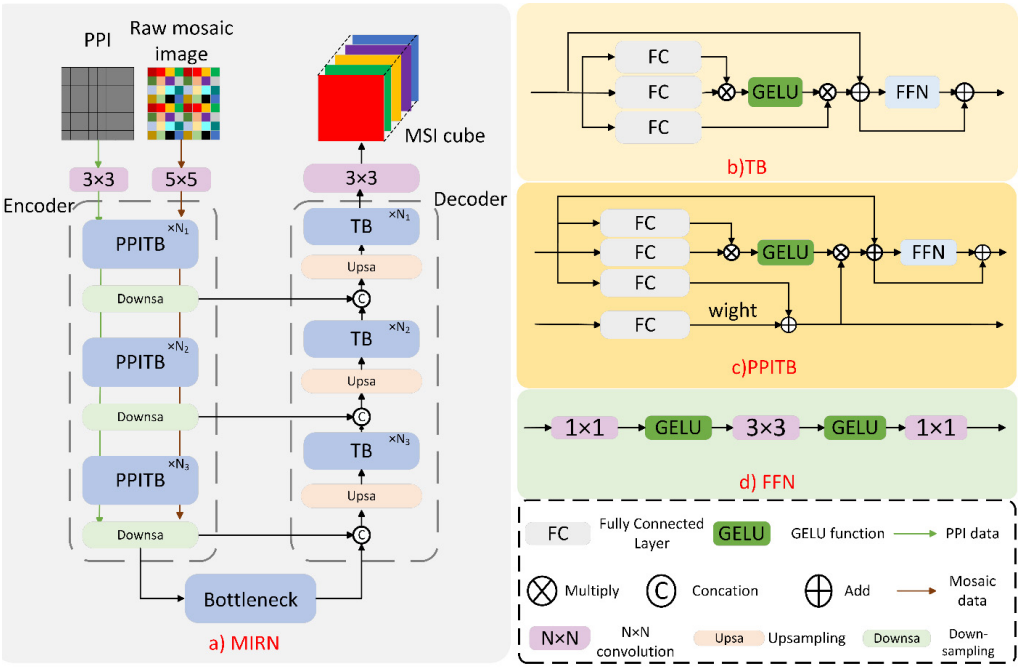

Fig. 4. Multispectral image reconstruction network. (a) Overall structure of multispectral image reconstruction network. (b) Structure of transformer block. (c) Structure of PPI-transformer block. (d) Structure of feed-forward network.

through convolution. Firstly, MIRN uses convolution kernels of different sizes to extract the initial features from the estimated PPI and the raw mosaic, respectively. Secondly, the encoder performs feature fusion and extraction of the initial features at different levels to generate deep features. Then, deep features are transformed through the bottleneck. Finally, the feature is passed through the decoder and an independent convolutional layer to generate the MSI.

The encoder is mainly composed of PPI-transformer block (PPITB) and down-sampling module. The down-sample module is a stridden 4×4 convolutional layer that down-scales the feature maps and doubles the channels. The bottleneck consists of two transformer blocks (TB). Following U-Net's structure, the symmetric decoder has the same number of layers as the encoder, but its main module is TB instead of PPITB. The whole network uses skip connections to aggregate features between encoders and decoders to reduce information loss caused by down-sampling operations.

The structure of PPITB and TB is shown in Fig. 4(b) and (c). TB consists of three fully connected layers, a GELU activation function, and a feed-forward network (FFN). The details of FFN are depicted in Fig. 4(d). To enable PPI and the image to carry out deep feature fusion, the TB is further improved to form the PPITB. Specifically, an input branch and an output branch are introduced. The input branch consists of a fully connected layer, and the features of the original branches are weighted and fused. The fused

features are divided into two parts: one part is multiplied with the other two fully connected layers as the feature output of the image, and the other part is output as further characteristics of the PPI to the PPI input at the next layer. In the encoder layer, the PPITB is stacked to achieve the feature fusion of PPI and the original mosaic at different levels.

3.4. Loss function

To train the entire framework, we utilize a combined loss function denoted as l, which aims to simultaneously minimize the reconstruction errors of PPI and demosaicing image.

$$l = l_{\mathrm{PPI}} + l_{\mathrm{HSI}} \tag{5}$$

where l_{PPI} is the loss function of PPIGN, and l_{HSI} is the loss function of MIRN. Our network is optimized by minimizing the following loss function. We define it as the L_1 norm of the difference between the estimated value and the true value. Given a training set with K training sample pairs $\{I_k^{\mathrm{MSFA}}, I_k^{\mathrm{PPI}}, I_k^{\mathrm{HSI}}\}_{k=1}^{K}$, the expressions for l_{PPI} and l_{HSI} are as follows:

$$l_{\mathrm{MSI}} = \frac{1}{K}\sum_{k=1}^{K} \left\| \hat{I}_{\mathrm{MSI}} - I_{\mathrm{MSI}} \right\|_1 \tag{6}$$

$$l_{\mathrm{PPI}} = \frac{1}{K}\sum_{k=1}^{K} \left\| \hat{I}_{\mathrm{PPI}} - I_{\mathrm{PPI}} \right\|_1 \tag{7}$$

where $\hat{I}_{\mathrm{MSI}}$ and $\hat{I}_{\mathrm{PPI}}$ what are the demosaicing results of the mosaicked training sample and the estimated PPI. I_{MSI} and I_{PPI} are the real MSI and the accurate PPI.

4. Experiments and analysis

This section involves conducting a quantitative and visual comparison with several state-of-the-art methods on a benchmark dataset and analyzing experimental results.

4.1. Dataset description

The ARAD-1K dataset [25] released by the NTIRE 2022 challenge is a high-quality publicly available multispectral reflectance dataset. The first-of-its-kind large-scale dataset for MSFA demosaicing of natural scenes contains 1000 images with 16 spectral bands covering 400–1000 nm wavelengths. Each scene includes 482×512 spatial resolution multispectral reflectance images. We perform spatial subsampling of these images to generate the raw mosaic image utilizing the 4×4 MSFA mode, without employing a dominant band. To evaluate our model quantitatively, we choose the 950

images with GT for training (900 images) and testing (50 images).

4.2. Training details

We did not use data enhancement operations during the training and implemented the proposed network using Pytorch. The whole network is optimized using Adam [26] optimizer. The training process is done for 2000 iterations with a batch size of 16. All the experiments were run on one NVIDIA GTX 3090 GPU.

4.3. Quantitative experiments metrics

We adopt the peak signal-to-noise ratio (PSNR), structural similarity [27] (SSIM), and spectral angle mapper (SAM) [28] as the quantitative evaluation metrics. PSNR is a conventional evaluation index in image processing and computer vision that measures the similarity between the demosaiced MSI and the actual MSI based on mean squared error (MSE). SSIM assesses the image degradation between the demosaiced and authentic MSI. SAM is a way of measuring the degree of similarity between multispectral images from spectral dimensions. The unit of PSNR is dB, the unit of SAM is degree, and the unit of SSIM is 1.

4.4. Demosaicing results and analysis

To illustrate the effectiveness, our proposed algorithm was compared to four traditional methods, including BETS [29], WB [10], PPID [26], SD, and two deep learning-based methods including DPD-net [21] and MCAN [22]. Which are specially designed for MSFA image demosaicing. The source codes of these hand-crafted and CNN-based methods are publicly available. We apply these algorithms to the ARAD-1K dataset for evaluation.

The comparisons between our proposed algorithm and other SOTA methods are listed in Table 1. As can be observed. our method achieved a PSNR of 48.4023, significantly higher than the other methods, indicating a pronounced advantage in minimizing reconstruction errors and improving image quality. Particularly compared to DPD-net (47.063) and MCAN (47.579), although these latter two also displayed high PSNR values, our method is closer to a lossless reconstruction. In the two evaluation indicators of SAM and SSIM, our model has a little gap with MCAN and DPD-net, but there is still a small improvement.

Figures 5 and 6 provide visual comparisons of various demosaicing methods, aim-

T a b l e 1. PSNR, SAM, and SSIM assessments of the demosaiced MSI on the ARAD-1K dataset.

Metrics	WB	BTES	SD	PPID	DPD-net	MCAN	Ours
PSNR	35.410	35.514	37.334	40.515	47.063	47.579	48.402
SAM	6.048	5.984	5.674	3.717	1.677	1.570	1.500
SSIM	0.9646	0.9651	0.9780	0.9873	0.9981	0.9983	0.9987

ing to subjectively evaluate their performance. Given the vast amount of data in MSI, this study synthesized a false color map (the three channels of the image are not the information of the RGB, but other bands) using selected bands. The two figures mainly show the details of the areas marked in red boxes in the scene. As illustrated in Fig. 5, compared to traditional interpolation algorithms, the network proposed in this paper significantly outperforms in terms of clarity and detail preservation, demonstrating the best recovery quality. Although the performance differences among our network, MCAN, and DPD-net are minimal, our network excels in detail processing. Figure 6 specifically highlights the junction between the building and the sky. Our algorithm distinctly surpasses others by producing the fewest artifacts and closely approximating the true scene, as evidenced by the seamless transitions and accurate color reproduction in the highlighted area. These results affirm the advanced capability of our method in handling complex scenarios in MSI processing, confirming its significant value in

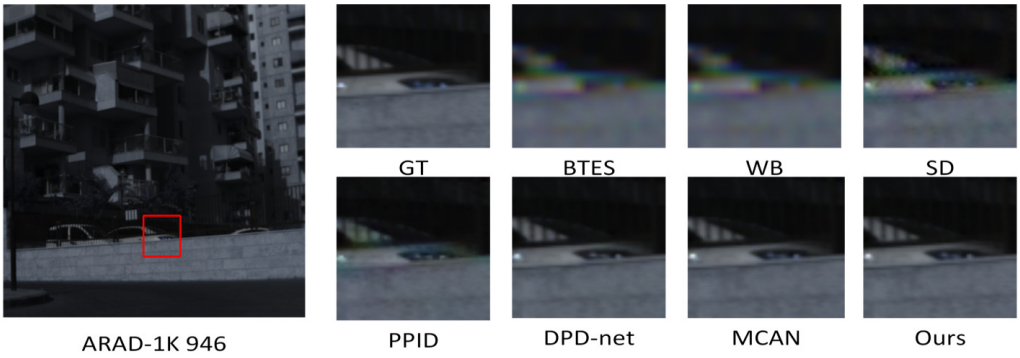

Fig. 5. Visual comparison of demosaicing methods in ARAD-1K 946 scene (false color, R:2, G:11, B:16). Our method is compared with six alternatives, BTES, WB, SD, PPID, DPD-net, MCAN. We have only presented an expanded perspective of the selected area; please zoom in for further details.

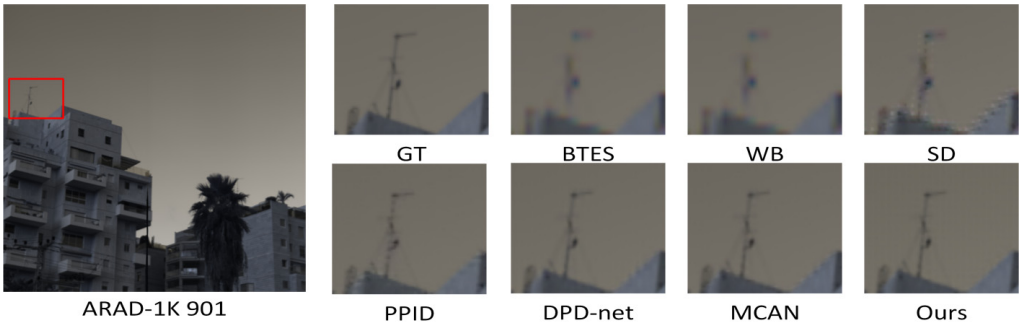

Fig. 6. Visual comparison of demosaicing methods in ARAD-1K 901 scene (false color, R:2, G:11, B:16). Our method is compared with six alternatives, BTES, WB, SD, PPID, DPD-net, MCAN. We have only presented an expanded perspective of the selected area; please zoom in for further details.

practical applications where precise color and detail reproduction are paramount.

4.5. Ablation studies and analysis

To assess the indispensability of each component within our pipeline, we executed a series of ablation studies on various configurations of our methodology. Each experimental setup is scrutinized in subsequent sections. It is important to note that apart from the dedicated hyperparameter ablation experiments, we standardized the configuration of the MIRN hyperparameters across all tests, setting two modules per layer. Within each module, the weight assigned to the PPI fusion module is consistently maintained at one.

1) The effect of different PPI: We evaluated the impact of various PPI generation methods on the quality of multispectral image reconstruction. These methods included the absence of PPI, smooth filtering, real PPI, and PPI generated via PPIGN. Our analysis, as detailed in Table 2, reveals that the incorporation of PPI significantly enhances network performance in the demosaicing process. The reconstruction results also prove that PPIGN has a better ability to generate PPI than the average filter. Notably, there exists a minor discrepancy between the outcomes derived from our model and those utilizing real PPI (only 0.13 dB lower). This discrepancy can be attributed to the inherent limitations in completely reconstructing real PPI.

T a b l e 2. PSNR of multispectral image reconstruction using PPI generated in different ways.

Way	No	Average filter	Ours	Real
PSNR	47.0755	47.4312	48.2092	48.3423

2) Effects of different weights in the PPITB module: Table 3 presents the impact of varying weights within the PPITB module on the quality of reconstructed MSI. It is observed that as the weight value changes from 0.1 to 1, the quality of the reconstructed image decreases. However, it remains relatively unchanged between weights of 1 and 10. This phenomenon can be attributed to two main reasons. First, PPI plays an auxiliary role. At weights less than 1, the features from the original mosaic image dominate, with varying weights affecting the degree of assistance and, consequently, the reconstruction quality. Second, at weights greater than 1, PPI features become dominant. However, due to the limited correlation between PPI and MSI, the reconstruction quality degrades. The minimal difference in quality between weights of 1 and 10 can be explained by the neural network's adaptive capability and optimization. Notably, architectures using a weight of zero and excluding PPI differ primarily in their loss

T a b l e 3. PSNR of reconstructed MSI with different weights in PPITB.

Weight	0.0	0.1	0.5	1.0	10.0
PSNR	48.3405	48.4023	48.3867	48.2092	48.2138

T a b l e 4. PSNR of reconstructed MSI with different numbers of modules in MIRN.

Model	N_1	N_2	N_3	PSNR
Mode 1	2	2	2	48.1933
Mode 2	2	2	4	47.7026
Mode 3	2	4	4	47.7557
Mode 4	4	2	2	48.3136
Mode 5	4	4	2	48.3854
Mode 6	4	4	4	47.7261

functions, which highlights that different loss functions can significantly enhance reconstruction performance.

3) The effect of different numbers of blocks in MIRN: Table 4 illustrates the configuration settings of MIRN layers and their corresponding experimental results. It is evident that increasing the number of modules (N1, N2) within the shallow feature extraction layers enhances the reconstruction quality of MIRN. Conversely, increasing the number of modules (N3) in the third layer significantly degrades the reconstruction quality, with an average decrease of 0.5 dB. This degradation may be attributed to the reduced correlation between PPI and MSI in deeper network layers, which introduces errors and diminishes the quality of the reconstructed images. This observation underscores the importance of optimizing layer-specific module configurations to effectively improve MIRN performance.

4.6. Running time and computational cost

The speed and computational cost of the demosaicing method are crucial factors in determining its feasibility for implementation in a real multispectral imaging system. Table 4 presents a comparison of the runtime, GFLOPs, and parameters of the most advanced demosaicking methods. The results are averaged over 50 runs using the ARAD-1K test set. All methods are implemented with PyTorch on the same machine (Intel CPU 3.6 GHz, 16 GB memory, and NVIDIA GPU GTX 3050Ti). Notably, DPD-net reconstructs MSI one band at a time, resulting in fewer parameters. As shown in the data comparison, while our network has a longer runtime compared to MCAN, it outperforms MCAN in the other three metrics. Therefore, considering all aspects, our network demonstrates superior overall performance compared to MCAN.

T a b l e 5. The demosaicing performance and complexity comparisons.

Model	PSNR	Running times [s]	Params(M)	GFLOPS
DPD-net	47.063	0.091	4.39	107.84
MCAN	47.579	0.005	13.74	71.76
Ours	48.402	0.026	9.94	25.78

4.7. Sensitivity analysis of noise

We present the sensitivity analysis of the proposed method alongside state-of-the-art methods at various noise levels in Table 5. In these experiments, Gaussian noise with zero mean and variances of 0.001, 0.01, and 0.05 was added to the ARAD-1K dataset (pixel values ranging from 0 to 1). As shown in Table 6, our method consistently achieves a high peak signal-to-noise ratio (PSNR) across varying noise levels, outperforming the current state-of-the-art methods.

T a b l e 6. Sensitivity of demosaicing methods at various noise variances on ARAD-1K dataset.

Model	0.0	0.001	0.01	0.05
DPD-net	47.06	35.99	26.08	13.96
MCAN	47.58	43.34	30.59	17.35
Ours	48.40	46.04	38.60	32.04

5. Conclusion

This paper proposes a multispectral image demosaicing model based on PPI and transformer. The method first rapidly reconstructs PPI through the attention mechanism. At the same time, the custom transformer architecture efficiently extracts the features of the generated PPI and the raw mosaic image, leading to a high-precision multispectral image reconstruction. Experiments demonstrate that our proposed PPI-transformer model outperforms existing MSFA demosaicing approaches in quantitative metrics and visual comparisons. The proposed method could be used to build high-quality MSFA-based image acquisition systems that work well in medical imaging, food quality inspection, and remote sensing applications.

Acknowledgements
This work was supported by the National Natural Science Foundation of China (Grant Nos. 62305164). Peng Chen and Heng Wang are co-first authors of the paper.

References

[1] Pu H.B., Lin L., Sun D.-W., *Principles of hyperspectral microscope imaging techniques and their applications in food quality and safety detection: A review*, Comprehensive Reviews in Food Science and Food Safety **18**(4), 2019: 853-866. https://doi.org/10.1111/1541-4337.12432

[2] Xiong F., Zhou J., Qian Y., *Material based object tracking in hyperspectral videos*, IEEE Transactions on Image Processing **29**, 2020: 3719-3733. https://doi.org/10.1109/TIP.2020.2965302

[3] Hao W.-H., Wen D.-W., *Multispectral imaging for plant food quality analysis and visualization*, Comprehensive Reviews in Food Science and Food Safety **17**(1), 2018: 220-239. https://doi.org/10.1111/1541-4337.12317

[4] Elaksher A.F., *Fusion of hyperspectral images and lidar-based dems for coastal mapping*, Optics and Lasers in Engineering **46**(7), 2008: 493-498. https://doi.org/10.1016/j.optlaseng.2008.01.012

[5] KARAGIANNIS G., *High resolution, in situ, multispectral, spectroscopic mapping imaging system applied in heritage science*, Optics and Lasers in Engineering **174**, 2024: 107971. https://doi.org/10.1016/j.optlaseng.2023.107971

[6] LI W., LIU Y., LING L., SHENG Z., CHENG S., YI Z., WU P., ZENG Q., TANG B., AHMAD S., *The tunable absorber films of grating structure of AlCuFe quasicrystal with high Q and refractive index sensitivity*, Surfaces and Interfaces **48**, 2024: 104248. https://doi.org/10.1016/j.surfin.2024.104248

[7] LIANG S., XU F., LI W., YANG W., CHENG S., YANG H., CHEN J., YI Z., JIANG P., *Tunable smart mid infrared thermal control emitter based on phase change material VO_2 thin film*, Applied Thermal Engineering **232**, 2023: 121074. https://doi.org/10.1016/j.applthermaleng.2023.121074

[8] LIANG S., CHENG S., ZHANG H., YANG W., YI Z., ZENG Q., TANG B., WU P., AHMAD S., SUN T., *Structural color tunable intelligent mid-infrared thermal control emitter*, Ceramics International **50**(13), 2024: 23611-23620. https://doi.org/10.1016/j.ceramint.2024.04.085

[9] LI W., ZHAO W., CHENG S., ZHANG H., YI Z., SUN T., WU P., ZENG Q., RAZA R., *Tunable metamaterial absorption device based on Fabry–Perot resonance as temperature and refractive index sensing*, Optics and Lasers in Engineering **181**, 2024: 108368. https://doi.org/10.1016/j.optlaseng.2024.108368

[10] BRAUERS J., AACH T., *A color filter array based multispectral camera*, [In] *12. Workshop Farbbildverarbeitung*, Ilmenau, 2006.

[11] GUPTA M., RAM M., *Weighted bilinear interpolation based generic multispectral image demosaicking method*, Journal of Graphic Era University, 2019: 108-118.

[12] MIHOUBI S., LOSSON O., MATHON B., MACAIRE L., *Multispectral demosaicing using intensity-based spectral correlation*, [In] *2015 International Conference on Image Processing Theory, Tools and Applications* (*IPTA*), IEEE, 2015: 461-466. https://doi.org/10.1109/IPTA.2015.7367188

[13] BIAN L., WANG Y., ZHANG J., *Generalized MSFA engineering with structural and adaptive nonlocal demosaicing*, IEEE Transactions on Image Processing **30**, 2021: 7867-7877. https://doi.org/10.1109/TIP.2021.3108913

[14] REDMON J., DIVVALA S., GIRSHICK R., FARHADI A., *You only look once: Unified, real-time object detection*, [In] *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016: 779-788. https://doi.org/10.1109/CVPR.2016.91

[15] MANSOUR Y., HECKEL R., *Zero-shot noise2noise: Efficient image denoising without any data*, [In] *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, BC, Canada, 2023: 14018-14027. https://doi.org/10.1109/CVPR52729.2023.01347

[16] WANG B., FENG Y., LIU H., *Multi-scale features fusion from sparse LiDAR data and single image for depth completion*, Electronics Letters **54**(24), 2018: 1375-1377. https://doi.org/10.1049/el.2018.6149

[17] ZHAO B., ZHENG J., DONG Y., SHEN N., YANG J., CAO Y.L., CAO Y.P., *PPI edge infused spatial–spectral adaptive residual network for multispectral filter array image demosaicing*, IEEE Transactions on Geoscience and Remote Sensing **61**, 2023: 5405214. https://doi.org/10.1109/TGRS.2023.3297250

[18] LI S., LIU Y., *Deep densely-connected residual learning for multispectral image demosaicing*, [In] *2023 8th International Conference on Signal and Image Processing* (*ICSIP*), IEEE, 2023: 768-772. https://doi.org/10.1109/ICSIP57908.2023.10270836

[19] 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*), IEEE, 2019: 1-5. https://doi.org/10.1109/WHISPERS.2019.8920868

[20] SHOPOVSKA I., JOVANOV L., PHILIPS W., *RGB-NIR demosaicing using deep residual U-Net*, [In] *2018 26th Telecommunications Forum* (*TELFOR*), IEEE, 2018: 1-4. https://doi.org/10.1109/TELFOR.2018.8611819

[21] 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), 2022: 622-635. https://doi.org/10.1109/JSTSP.2022.3172865

[22] 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**, 2021: 864-878. https://doi.org/10.1109/TCI.2021.3102052

[23] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., *Attention is All You Need*, arXiv:1706.03762 [cs.CL], 2017. https://doi.org/10.48550/arXiv.1706.03762

[24] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N., *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv:2010.11929 [cs.CV], 2020. https://doi.org/10.48550/arXiv.2010.11929

[25] Arad B., Timofte R., Yahel R., Morag N., Bernat A., Wu Y., Wu X., Fan Z., Xia C., Zhang F., Liu S., Li Y., Feng C., Lei L., Zhang M., Feng K., Zhang X., Yao J., Zhao Y., Ma S., He F., Dong Y., Yu S., Qiu D., Liu J., Bi M., Song B., Sun W.F., Zheng J., Zhao B., Cao Y., Yang J., Cao Y., Kong X., Yu J., Xue Y., Xie Z.., *NTIRE 2022 spectral demosaicing challenge and data set*, [In] *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), IEEE, 2022: 881-895. https://doi.org/10.1109/CVPRW56347.2022.00103

[26] Mihoubi S., Losson O., Mathon B., Macaire L., *Multispectral demosaicing using pseudo-panchromatic image*, IEEE Transactions on Computational Imaging **3**(4), 2017: 982-995. https://doi.org/10.1109/TCI.2017.2691553

[27] 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), 2004: 600-612. https://doi.org/10.1109/TIP.2003.819861

[28] Kruse F.A., Lefkoff A.B., Boardman J.W., Heidebrecht K.B., Shapiro A.T., Barloon P.J., Goetz A.F.H., *The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data*, Remote Sensing of Environment **44**(2-3), 1993: 145-163. https://doi.org/10.1016/0034-4257(93)90013-N

[29] 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), 2006: 3550-3558. https://doi.org/10.1109/TIP.2006.877476

Received August 16, 2024
in revised form September 16, 2024