Pansharpening via Multiscale Embedding and Dual Attention Transformers

Wensheng Fan, Fan Liu, *Member, IEEE*, and Jingzhi Li

***Abstract*—Pansharpening is a fundamental and crucial image processing task for many remote sensing applications, which generates a high-resolution multispectral image by fusing a low-resolution multispectral image and a high-resolution panchromatic image. Recently, vision transformers have been introduced into the pansharpening task for utilizing global contextual information. However, long-range and local dependencies modeling and multiscale feature learning are all essential to the pansharpening task. Learning and exploiting these various information raises a big challenge and limits the performance and efficiency of existing pansharpening methods. To solve this issue, we propose a pansharpening network based on multiscale embedding and dual attention transformers (MDPNet). Specifically, a multiscale embedding block is proposed to embed multiscale information of the images into vectors. Thus, transformers only need to process a multispectral embedding sequence and a panchromatic embedding sequence to efficiently use multiscale information. Furthermore, an additive hybrid attention transformer is proposed to fuse the embedding sequences in an additive injection manner. Finally, a channel self-attention transformer is proposed to utilize channel correlations for high-quality detail generation. Experiments over QuickBird and WorldView-3 datasets demonstrate the proposed MDPNet outperforms state-of-the-art methods visually and quantitatively with low running time. Ablation studies further verify the effectiveness of the proposed multiscale embedding and transformers in pansharpening.**

***Index Terms*—Attention mechanism, image fusion, multiscale embedding, pansharpening, remote sensing, vision transformer (ViT).**

I. Introduction

MULTISPECTRAL (MS) images are widely used for various remote sensing applications such as land cover classification [1], environmental change detection [2], [3], and agriculture monitoring [4]. Due to physical constraints, there is a tradeoff between spatial and spectral resolutions during satellite imaging. The satellite can only provide low-spatial-resolution (LR) MS images and corresponding high-spatial-resolution (HR) PAN images [5]. To obtain HRMS images, image processing is needed. Image processing applies procedures to an image to enhance it or derive valuable information from it [6]. As a remote sensing image processing technique, pansharpening sharpens LRMS images using their corresponding PAN images to produce HRMS images. Therefore, pansharpening can improve the performance of remote sensing applications such as land-use classification [7].

In the past several decades, many pansharpening algorithms have been developed. They can be roughly grouped into four main categories: component substitution (CS), multiresolution analysis (MRA), variational optimization (VO), and deep learning (DL) [8], [9], [10]. The first three classes are traditional algorithms that emerged decades ago. The DL-based methods have arisen recently and achieved promising outcomes.

CS-based algorithms usually transform the up-sampled MS image into another space to separate out its spatial component, and then, replace it with the PAN image to enrich spatial details. Well-known CS algorithms include those exploiting intensity-hue-saturation (IHS) transform [11], Gram–Schmidt (GS) transform [12], and band-dependent spatial detail (BDSD) [13]. MRA-based methods typically use multiscale decomposition or high-pass filtering to extract spatial details from the PAN image and obtain the HRMS image via detail injection. Representative MRA approaches include additive wavelet luminance (AWL) [14], smoothing filter-based intensity modulation (SFIM) [15], and generalized Laplacian pyramids with modulation transfer function (MTF-GLP) [16]. VO-based approaches build a model with suitable regularization terms based on certain priors or assumptions and utilize an effective algorithm to optimize the model. Typical VO-based methods include Bayesian-based fusion methods [17], [18], sparse representation-based detail injection [19], and total variation (TV) [20]. There are also hybrid methods that combine different kinds of traditional approaches and even combine them with DL techniques to complement each other [21], [22], [23].

DL techniques are also widely applied to the remote sensing field and have shown great potential continuously no matter in specific tasks such as cross-city semantic segmentation [24] or in universal foundation model development [25]. As for the pansharpening task, inspired by the image super-resolution method based on convolutional neural network (CNN) [26], Masi et al. [27] proposed an efficient three-layer CNN for pansharpening (PNN), which produced promising outcomes in the pansharpening task. Introducing more domain knowledge, Yang et al. [28] proposed PanNet, which learns the spatial details to be

Manuscript received 12 May 2023; revised 23 July 2023, 24 November 2023, and 12 December 2023; accepted 14 December 2023. Date of publication 18 December 2023; date of current version 10 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61703299 and in part by the Basic Research Project Foundation of Shanxi Province under Grant 202203021221094. *(Corresponding author: Fan Liu.)*

Wensheng Fan is with the College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan 030024, China (e-mail: fanwensheng9603@163.com).

Fan Liu and Jingzhi Li are with the College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Jinzhong 030600, China (e-mail: liufan@tyut.edu.cn; 916381552@qq.com).

Digital Object Identifier 10.1109/JSTARS.2023.3344215

injected into the MS image in the high-pass filtering domain via residual learning [29], which successfully protects both spatial and spectral information in the deep network. Based on the observation that the scale of features varies among different ground objects, Yuan et al. [30] proposed a multiscale and multidepth CNN (MSDCNN) and achieved superior pansharpening performance. The exploration and utilization of multiscale features have since become a key concern in developing DL-based methods. To enhance the fidelity of pan-sharpened images, generative adversarial network techniques are applied to the pansharpening task [31], [32], [33]. These methods typically design one or several discriminators to judge the fidelity of pansharpening outcomes. On the other hand, to model the long-range dependencies in the images, the transformer techniques [34], [35] are recently applied to the pansharpening task [36]. Zhou et al. [37] designed a customized transformer for pansharpening, which enhances the spatial details of the pan-sharpened image via both hard-attention and soft-attention mechanisms. DR-NET [38] inserts Swin transformer [39] blocks into a hierarchical u-shaped architecture to combine the transformer with a typical multiscale pansharpening scheme, which reduces the detail loss in the down-sampling process. To mutually reinforce spatial and spectral features, Zhang et al. [40] propose a cross-interaction kernel attention network for the improvement of dynamic convolution-based pansharpening.

The aforementioned methods are designed by following observations. On the one hand, remote sensing images contain a large number of repetitive ground objects. Ground objects with similar spatial and spectral information can be close or far apart. On the other hand, ground objects in remote sensing images are in many sizes. Thus, there are several potential drawbacks in transformer-based methods regarding the following aspects.

1) Long-range and local dependencies modeling and multiscale feature learning are all essential to pansharpening. It is difficult for transformer-based method to fully and efficiently exploit these various information, which may cause performance limitation and high computational complexity.
2) The traditional self-attention transformers can only play the role of long-range feature extraction. They have little to do with feature fusion. This prevents transformer from modeling the dependencies between the original LRMS image and the fusion product and is unfavorable to generating details complement to the LRMS image.
3) The dependencies along spatial and channel dimensions are not fully utilized in existing transformer-based methods. Correlations among both dimensions are crucial to the spatial and spectral quality of pansharpening outcomes.

To solve these problems, we propose a pansharpening network based on multiscale embedding and dual attention transformers (MDPNet) is proposed in this article. Specifically, a multiscale embedding block is proposed to embed the multiscale information of the LRMS and PAN images into two sequences of vectors. Thus, only processing this pair of multiscale embedding sequences is enough to efficiently realize the utilization of multiscale features. To fuse the two sequences, a feature fusion module based on the additive hybrid attention transformer (AHAT) is proposed considering the long-range dependencies in the spatial dimension. Finally, a detail generation module based on channel self-attention transformer (CSAT) is proposed to generate details for detail injection considering the correlations among feature channels. Experiments over datasets collected by QuickBird (QB) and WorldView-3 (WV3) satellites demonstrate that our MDPNet outperforms state-of-the-art methods, and has lower running time than other transformer-based approaches. The main contributions of this article are as follows.

1) We propose a multiscale embedding block to embed the multiscale information of the LRMS and PAN images into two sequences of embedding vectors. Then, the transformer only needs to process these two sequences without any down-sampling operations and separate treatments for efficiency.
2) A feature fusion module based on the AHAT is proposed to fuse the LRMS and PAN embedding sequences. Considering domain-specific knowledge, we use additive spatial information injection in the AHAT to transfer the texture and structure features from the PAN embedding sequence to the LRMS one.
3) A detail generation module based on the CSAT is proposed to generate details considering the correlations among feature channels. The module enhances the fused feature maps via interaction along the channel dimension, and thus improves the quality of resulting details.

The rest of this article is organized as follows. Section II reviews related works. Section III elaborates the proposed method. Section IV analyses the experimental results. Finally, Section V concludes this article.

II. Related Work

A. Additive Detail Injection-Based Methods

Additive detail injection is a unified framework for traditional CS-based and MRA-based pansharpening methods [41]. CS approaches typically use linear transformations and only substitute the spatial component. Thus, the transformation and substitution process can be recast into an additive detail injection model as follows [42]:

$$\mathbf{H}_k = \widetilde{\mathbf{L}}_k + \mathbf{G}_k \cdot (\mathbf{P} - \mathbf{I}_{\mathrm{L}}) \tag{1}$$

where $\mathbf{H}_k$ denotes the kth band of the desired HRMS image. $\widetilde{\mathbf{L}} \in \mathbb{R}^{H\times W\times B}$ is the interpolated LRMS image at the PAN scale, where W and H are the width and height of the PAN image. B is the number of MS bands. $\mathbf{G}_k$ denotes the injection gain matrix. $\mathbf{P} \in \mathbb{R}^{H\times W\times 1}$ is the PAN image. $\mathbf{I}_{\mathrm{L}}$ is the intensity component of $\widetilde{\mathbf{L}}$. The calculation approaches of $\mathbf{I}_{\mathrm{L}}$ and $\mathbf{G}_k$ distinguish different CS-based methods. The GSA [43] algorithm determines the optimal weights to obtain $\mathbf{I}_{\mathrm{L}}$ via multivariate regression at the reduced resolution. The BDSD [13] calculates $\mathbf{I}_{\mathrm{L}}$ for each MS band separately with different weights.

MRA-based methods typically rely on an iterative decomposition process to obtain the low-pass PAN image $\mathbf{P}_{\mathrm{L}}$ [9]. And the difference between $\mathbf{P}$ and $\mathbf{P}_{\mathrm{L}}$ is the detail to be added to $\widetilde{\mathbf{L}}$. The general additive detail injection model for MRA approaches

can be formulated as

$$\mathbf{H}_k = \widetilde{\mathbf{L}}_k + \mathbf{G}_k \cdot (\mathbf{P} - \mathbf{P}_{\mathrm{L}}). \tag{2}$$

The ways to compute $\mathbf{P}_{\mathrm{L}}$ and $\mathbf{G}_k$ are the main differences among MRA-based methods. AWL [14] method uses shift-invariant "à trous" wavelet decomposition to extract the PAN details and adds them to the intensity component of $\widetilde{\mathbf{L}}$. MTF-GLP [16] uses Gaussian filters that match the MS sensor's MTF to extract $\mathbf{P}_{\mathrm{L}}$, which makes the obtained detail just complement $\widetilde{\mathbf{L}}$.

The additive detail injection model has also inspired many DL-based methods and gives them an explicit physical interpretation. Detail injection-based CNN (DiCNN) [44] learns details in an end-to-end manner, and thus, they can be directly added to $\widetilde{\mathbf{L}}$ to avoid separately predicting $\mathbf{G}_k$ and $\mathbf{P}_{\mathrm{L}}$. FusionNet [45] uses the difference between the duplicated PAN image and $\widetilde{\mathbf{L}}$ as the network input to avoid the calculation of $\mathbf{I}_{\mathrm{L}}$ or $\mathbf{P}_{\mathrm{L}}$, and thus, overcomes the limitations of traditional CS and MRA schemes. In remote sensing, addition is also a popular strategy to fuse distinctive image representations extracted by different networks. For example, Hong et al. [46] fuse the features extracted by miniGCNs and CNNs via additive strategy to improve hyperspectral image classification performance.

B. Multiscale Deep Neural Networks for Pansharpening

Multiscale information contained in the LRMS and PAN images has proven to be quite useful in the pansharpening task. MSDCNN [30] extracts such information using parallel convolutional layers of different kernel sizes, which complements a fundamental serial CNN and significantly improves its performance. In addition to using convolutional kernels of different sizes, image pyramids can also be used to introduce the multiscale property into DL methods. Laplacian pyramid pansharpening petwork (LPPN) [47] applies pyramidal decomposition to both PAN and LRMS images. Then, multiple subnetworks are utilized to process the pyramid layers separately, which facilitates the full use of multiscale information.

C. Vision Transformers (ViTs) for Pansharpening

The ViT [35] partitions an image into patches and embeds them into a sequence of vectors. Thus, the long-range dependencies among patches can be modeled by transformer blocks via attention mechanisms. With the success of the ViT in image recognition tasks, many attempts have been made to apply ViT to high-resolution image processing [39], [48]. In remote sensing, a group-wise spectral embedding approach is proposed in SpectralFormer [49] to focus on the spectral characteristics for accurate hyperspectral image classification. This demonstrates that a flexible embedding strategy can deeply affect the function of transformers.

In the pansharpening field, ViTs have also achieved promising outcomes. Zhou et al. [37] proposed a customized transformer for pansharpening and put it into a detail injection-based framework to extract long-range features. HyperTransformer [50] captures multiscale long-range features by using transformer blocks at different scales of a backbone network. DR-NET [38] incorporates transformer blocks in the encoder of a Unet-like CNN [51] to reduce the loss of details during down sampling.

In this article, the proposed method avoids down sampling to improve spatial detail preservation and efficiency. The utilization of multiscale information is carried out by the embedding process. Furthermore, similarities in both spatial and channel dimensions are captured by the proposed AHAT and CSAT blocks to generate high-quality details for detail injection.

III. Methodology

A. Overall Network Architecture

The overall architecture of the proposed MDPNet is depicted in Fig. 1, which consists of two multiscale embedding blocks, a feature fusion module based on the AHAT and a detail generation module based on the CSAT. $\widetilde{\mathbf{L}}$ and $\mathbf{P}$ are embedded into multiscale embedding sequences $\mathbf{E}_{\mathrm{L}}^0$ and $\mathbf{E}_{\mathrm{P}}$ through two multiscale embedding blocks, respectively. The embedding sequences are fused via two stacked AHAT blocks, one with a regular window partitioning strategy (W-AHAT) and the other with a shifted window partitioning strategy (SW-AHAT) [39]. Long-range features along the spatial dimension are captured through these two blocks. The two stacked AHAT blocks output fused vector sequence $\mathbf{E}_{\mathrm{L}}^2$ as follows:

$$\mathbf{E}_{\mathrm{L}}^2 = \text{SW-AHAT}\left(\text{W-AHAT}\left(\mathbf{E}_{\mathrm{L}}^0, \mathbf{E}_{\mathrm{P}}\right), \mathbf{E}_{\mathrm{P}}\right). \tag{3}$$

Subsequently, the fused sequence $\mathbf{E}_{\mathrm{L}}^2$ is reshaped back to feature maps $\mathbf{F}$ and fed into the CSAT-based detail generation module to produce residual details $\mathbf{D} \in \mathbb{R}^{H\times W\times B}$. The HRMS image $\mathbf{H} \in \mathbb{R}^{H\times W\times B}$ is obtained by adding $\mathbf{D}$ to $\widetilde{\mathbf{L}}$ following the widely used detail injection framework. The detail generation and injection process can be summarized as follows:

$$\mathbf{H} = \widetilde{\mathbf{L}} + \text{CSAT}(\mathbf{F}). \tag{4}$$

The specific structures of the multiscale embedding block, the AHAT-based feature fusion module, and the CSAT-based detail generation module will be elaborated in the following.

B. Multiscale Embedding Block

As shown in Fig. 2, for each pixel in the input image, the multiscale embedding module split out s different sizes of patches centered on the pixel and flattens them into vectors. It is noteworthy that the 1×1 pixel itself is also retained as a vector to preserve details at the finest scale. All the vectors are projected to l-dimensional embedding vectors via corresponding linear layers, respectively. Then, these embedding vectors are concatenated to form an sl-dimensional multiscale embedding vector, which represents the multiscale information around the pixel. Finally, The multiscale embedding vectors for all the pixels in $\widetilde{\mathbf{L}}$ and $\mathbf{P}$ comprise the multiscale embedding sequences $\mathbf{E}_{\mathrm{L}}^0 \in \mathbb{R}^{HW\times sl}$ and $\mathbf{E}_{\mathrm{P}} \in \mathbb{R}^{HW\times sl}$, respectively. Both the high-resolution property and multiscale information of $\widetilde{\mathbf{L}}$ and $\mathbf{P}$ are maintained in the two sequences for long-range feature extracting and merging.

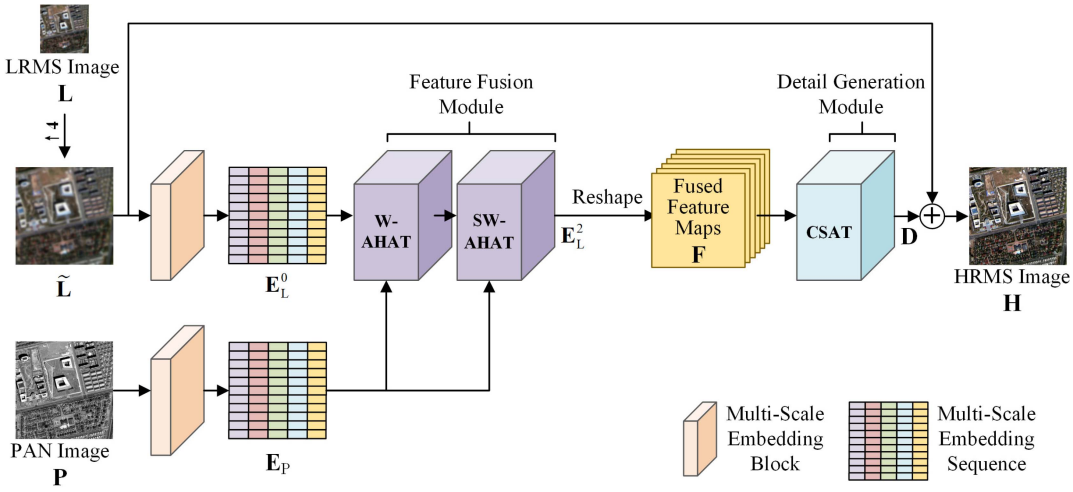


Fig. 1. Architecture of the proposed MDPNet. ↑ 4 denotes up sampling the LRMS image by a factor of 4 using bicubic interpolation.

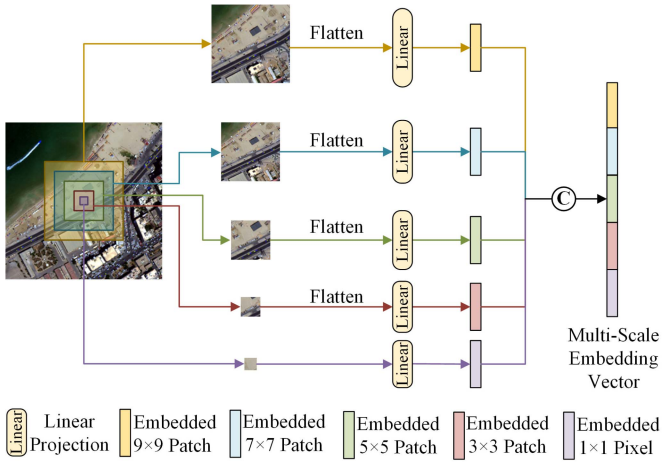


Fig. 2. Schematic diagram of the multiscale embedding block.

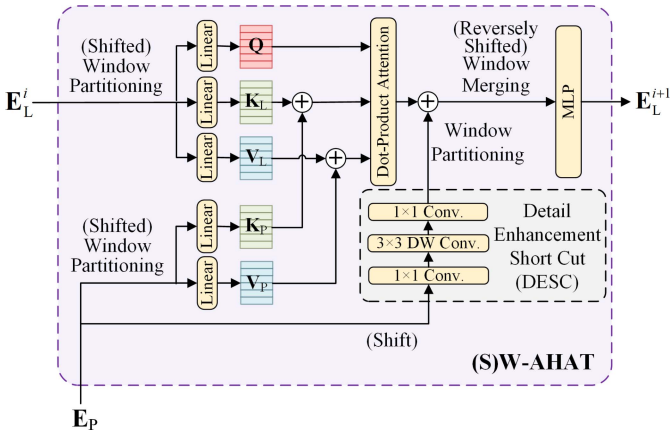


Fig. 3. Structure of the (S)W-AHAT blocks. The W-AHAT block uses regular window partitioning and merging. The SW-AHAT block uses shifted window partitioning and merging [39].

C. AHAT-Based Feature Fusion Module

The vanilla transformer block [34] consists of a self-attention mechanism and a feed-forward network. The self-attention mechanism can capture self-similarity and extract long-range features from an embedding sequence. However, in the pan-sharpening task, there are LRMS and PAN embedding sequences that need long-range feature extraction and fusion. Besides, in the additive detail injection framework, both spectral and spatial information in $\widetilde{\mathbf{L}}$ and $\mathbf{P}$ are useful for detail generation since the details have a spectral dimension, which has already been proven in [45]. This inspires us to design a hybrid attention mechanism that can extract useful spectral and spatial features related to the up-sampled LRMS image $\widetilde{\mathbf{L}}$ to prepare for generating complementary details.

Thus, in the proposed additive hybrid attention (AHA), the LRMS multiscale embedding sequence $\mathbf{E}_{\mathrm{L}}^{i}$ ($\mathbf{E}_{\mathrm{L}}^{0}$ for W-AHAT, $\mathbf{E}_{\mathrm{L}}^{1}$ for SW-AHAT) is linearly projected to the query matrix $\mathbf{Q}$. Both $\mathbf{E}_{\mathrm{L}}^{i}$ and $\mathbf{E}_{\mathrm{P}}$ should serve as keys and values to extract spectral and spatial information related to $\mathbf{Q}$. In the vanilla ViT [35], the positional information of the image patches is embedded in a position embedding sequence and added to the patch embedding sequence. Similarly, the information of PAN patches, i.e., $\mathbf{E}_{\mathrm{P}}$, is linearly transformed to PAN keys $\mathbf{K}_{\mathrm{P}}$ and PAN values $\mathbf{V}_{\mathrm{P}}$, and they are separately added to the LRMS keys $\mathbf{K}_{\mathrm{L}}$ and LRMS values $\mathbf{V}_{\mathrm{L}}$ for the injection of PAN information, as shown in Fig. 3. It is noteworthy that $\mathbf{E}_{\mathrm{P}}$ mainly contains spatial information and $\mathbf{E}_{\mathrm{L}}^{i}$ is rich of spectral information. Thus, the linear layers that transform them into keys and values have different weights. The calculation of the proposed AHA can be summarized as follows:

$$\begin{aligned}&\mathrm{AHA}\left(\mathbf{Q},\mathbf{K}_{\mathrm{L}},\mathbf{V}_{\mathrm{L}},\mathbf{K}_{\mathrm{P}},\mathbf{V}_{\mathrm{P}}\right)\\&=\mathrm{softmax}\left(\frac{\mathbf{Q}\left(\mathbf{K}_{\mathrm{L}}+\mathbf{K}_{\mathrm{P}}\right)^{\mathrm{T}}}{\sqrt{d}}\right)\left(\mathbf{V}_{\mathrm{L}}+\mathbf{V}_{\mathrm{P}}\right)\end{aligned}\tag{5}$$

where d is the dimension of query vectors in $\mathbf{Q}$.

The goal of the AHA is to extract long-range fused features from $\mathbf{E}_{\mathrm{L}}^{i}$ and $\mathbf{E}_{\mathrm{P}}$. However, the local fine-grained detail features

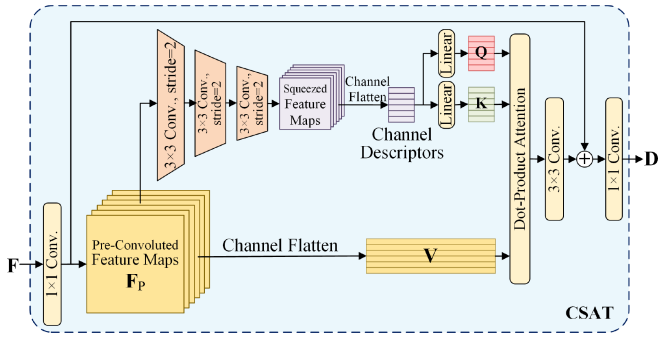


Fig. 4. Structure of the CSAT block.

are also vital to the pansharpening task. The convolution layer is naturally good at local reception. Thus, a convolutional detail enhancement short cut (DESC) is designed to add local detail features. Specifically, $\mathbf{E}_{\mathrm{P}}$ is reshaped back to feature maps and propagated to the DESC. It has a 3×3 depth-wise convolution layer between two point-wise convolution layers, which can save parameters and makes it lightweight [52]. The output of the DESC is flattened to a vector sequence and added to the output of AHA. Finally, the output of the AHAT $\mathbf{E}_{\mathrm{L}}^{i+1}$ is obtained through a standard multilayer perceptron (MLP).

In addition, the computation complexity of the attention is quadratic to the length of embedding sequences, i.e., HW. To reduce the computational burden, the shifted window partitioning approach in [39] is employed. Specifically, in the W-AHAT block, regular window partitioning and merging configurations are used. In the SW-AHAT block, the window partitioning and merging configurations are shifted half the window size of the W-AHAT. The rest of the two blocks are exactly the same.

D. CSAT-Based Detail Generation Module

The AHAT-based feature fusion module only captures the dependencies along the spatial dimension. However, the correlations along the channel dimension are not captured, which will affect the generation of details $\mathbf{D}$. Thus, we propose a channel self-attention (CSA) in our detail generation module.

Before input to the CSA, $\mathbf{E}_{\mathrm{L}}^{2}$ is reshaped to feature maps $\mathbf{F} \in \mathbb{R}^{H\times W\times sl}$ and preconvoluted via a 1×1 convolution. The CSA learns residuals to enhance the preconvoluted feature maps $\mathbf{F}_{\mathrm{P}}$. Specifically, as shown in Fig. 4, three 3×3 convolution layers with a stride of 2 are used to squeeze the information of $\mathbf{F}_{\mathrm{P}}$ into smaller feature maps. Then, each channel of the squeezed feature maps is flattened to a vector, which is a channel descriptor containing the global information of the channel. The channel descriptors are linearly projected to queries and keys. Each channel of $\mathbf{F}_{\mathrm{P}}$ is directly flattened to a value to avoid spatial information loss in squeezing and projecting. Subsequently, through dot-product attention, the self-similarity among $\mathbf{F}_{\mathrm{P}}$ channels is captured and the residuals are obtained by a 3×3 convolution layer to enhance $\mathbf{F}_{\mathrm{P}}$. Finally, the details $\mathbf{D}$ are generated from the enhanced $\mathbf{F}_{\mathrm{P}}$ via a 1×1 convolution layer.

IV. Experimental Results

A. Datasets

The experiments are conducted on two datasets collected by QB and WV3 satellites. The QB data have four MS bands, while the WV3 data have eight MS bands. Since the real HRMS images are unavailable, we follow Wald's protocol [53] to spatially degrade the LRMS and PAN images by a factor of 4 (the spatial-resolution ratio between them), and the original LRMS image can be used as the reference image for supervised learning and reduced-resolution evaluation. We also perform full-resolution evaluation with the original LRMS and PAN images, but there are no reference images for assessment.

All the reduced-resolution and full-resolution images are randomly cropped into LRMS patches with a size of 32×32 and PAN patches with a size of 128×128. We crop the images from the top left to the bottom right with a fixed stride that is greater than the patch size. As a result, 11 216 QB and 11 160 WV3 reduced-resolution patch pairs are extracted. To generate the training, validation and reduced-resolution testing sets, the QB and WV3 patch pairs are divided into 8974/1121/1121 and 8928/1116/1116 patch pairs in a ratio of 8:1:1, respectively. For full-resolution testing, we also crop 1121 QB and 1116 WV3 full-resolution patch pairs.

B. Implementation Details

The proposed method is implemented using the PyTorch framework and trained with an NVIDIA GeForce RTX 3090 GPU. Our model is trained for 500 epochs by optimizing the ℓ_1 loss between the fused image and the reference image. To implement the optimization, we employ an AdamW [54] optimizer with an initial learning rate of 0.0005, a momentum of 0.9, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay coefficient of 0.05. The minibatch size is set to 16.

Considering the tradeoff between computational complexity and performance, the number of scales is set to $s = 5$ and the embedding dimension of each scale is set to $l = 12$ by default. Thus, in the AHAT blocks, the dimension of embedding vectors, queries, keys, and values is $sl = 60$. The window size in the shifted window partitioning approach of the AHAT blocks is set to 8. Since the dimension of embedding vectors $sl = 60$ is relatively low, the number of heads of the AHA is set to 2. In the CSAT, the channel number of feature maps is also set to 60.

C. Compared Methods and Quantitative Metrics

The MDPNet is compared with nine representative methods, including two CS algorithms: GSA [43], BDSD [13]; one MRA method: MTF-GLP-FS [55]; one VO-based method: TV [20], three CNN-based methods: PNN [27], MSDCNN [30], and PanCSC-Net [56]; and two transformer-based methods: Zhou et al. [37] and DR-NET [38].

Five widely used metrics are adopted for the quantitative assessment of the methods. The metrics can be grouped into four full-reference indicators and one no-reference indicator according to whether they require a reference HRMS image in their calculations. For reduced-resolution assessment,

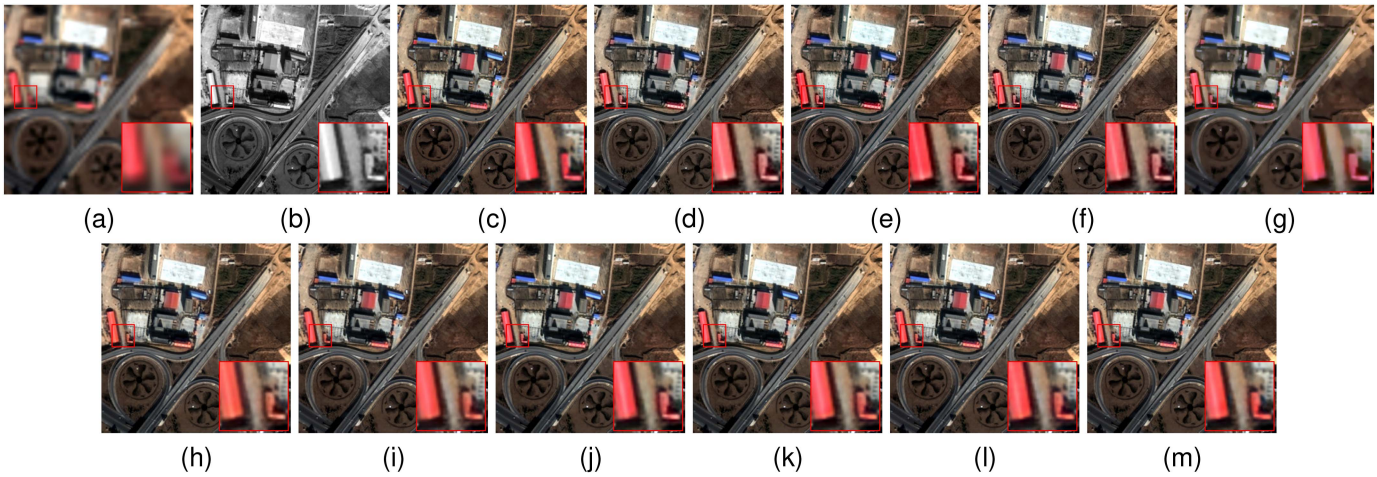


Fig. 5. Visual results on the QB reduced-resolution testing set. (a) Degraded LRMS image. (b) Degraded PAN image. (c) Reference image. (d) GSA. (e) BDSD. (f) MTF-GLP-FS. (g) TV. (h) PNN. (i) MSDCNN. (j) PanCSC-Net. (k) Zhou et al. [37] (l) DR-NET. (m) MDPNet.

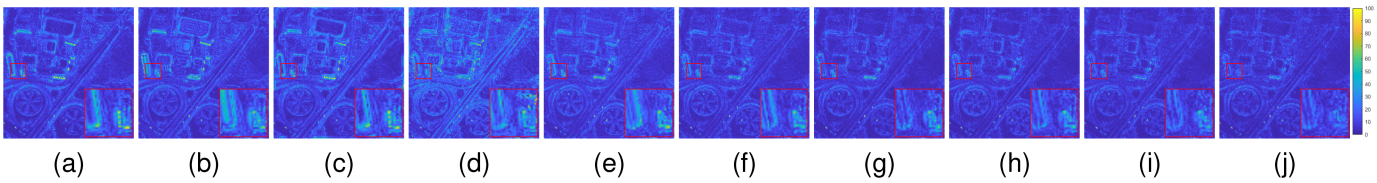


Fig. 6. Residual maps between the results and the reference in Fig. 5. (a) GSA. (b) BDSD. (c) MTF-GLP-FS. (d) TV. (e) PNN. (f) MSDCNN. (g) PanCSC-Net. (h) Zhou et al. [37] (i) DR-NET. (j) MDPNet.

we measure the four full-reference indexes including spectral angle mapper (SAM) [57], Erreur Relative Globale Adimensionnelle de Synthese (ERGAS) [58], spatial correlation coefficient (sCC) [59], and the $Q2^n$ [60], [61] index (i.e., Q4 for four-band data and Q8 for eight-band data). ERGAS and $Q2^n$ evaluate the overall quality of pansharpening results, SAM estimates spectral distortions, and sCC measures the quality of spatial details. For full-resolution assessment, we employ the no-reference index hybrid quality with no reference (HQNR) [62] with its spectral distortion component D_λ and spatial distortion component D_S to measure the pansharpening quality in the absence of the reference HRMS image.

D. Reduced Resolution Assessment

Fig. 5 shows the visual results of the compared algorithms on a QB reduced-resolution testing patch pair. The red box area is enlarged in the corner of the image to provide clearer visual comparison. To highlight the differences, residual maps between the pansharpening results and the reference image are visualized in Fig. 6. A pixel with a small mean absolute error (MAE) is shown in blue and a pixel with a big MAE is displayed in yellow. From the red buildings in the enlarged view, it can be observed that the results of GSA and MTF-GLP-FS are lighter in color than the reference image. In the results of BDSD and TV, the red buildings appear a little redder and a little pinker, respectively. The fusion image of the PNN shows an apparent yellow tint. In the result of the MSDCNN, the reflection of the sun on the red rooftop is missing and slight blurring effects appear on the building edges. For the outcomes of PanCSC-Net, Zhou et al. [37] and DR-NET, the color of the little red building in the enlarged view is lighter than that of the reference image. According to the enlarged view in the residual maps, it can also be found that the residuals of PanCSC-Net, DR-NET, and Zhou et al. [37] are larger than those of the proposed MDPNet. These findings prove that our method possesses a better visual effect and fewer errors than the compared methods.

Fig. 7 displays the visual results of the compared methods on a WV3 reduced-resolution testing patch pair. Fig. 8 shows the corresponding residual maps. In the enlarged view of GSA, BDSD, and MTF-GLP-FS fusion results, the color of the rooftop is apparently whiter than that in the reference image, which is an obvious spectral distortion. Also suffering from evident spectral distortion, the rooftop in the enlarged view of the TV result appears darker in color than that in the reference image. According to the enlarged view in the residual maps, it can be found that the residual maps of PNN and MSDCNN have larger yellow areas than the transformer-based methods. In the result of PanCSC-Net, the color near building borders is lighter than the reference image. Among the transformer-based approaches, the residual maps of Zhou et al. [37] and DR-NET have more yellow points with a large MAE than the proposed MDPNet in the enlarged view, which demonstrates that our method achieves more accurate prediction.

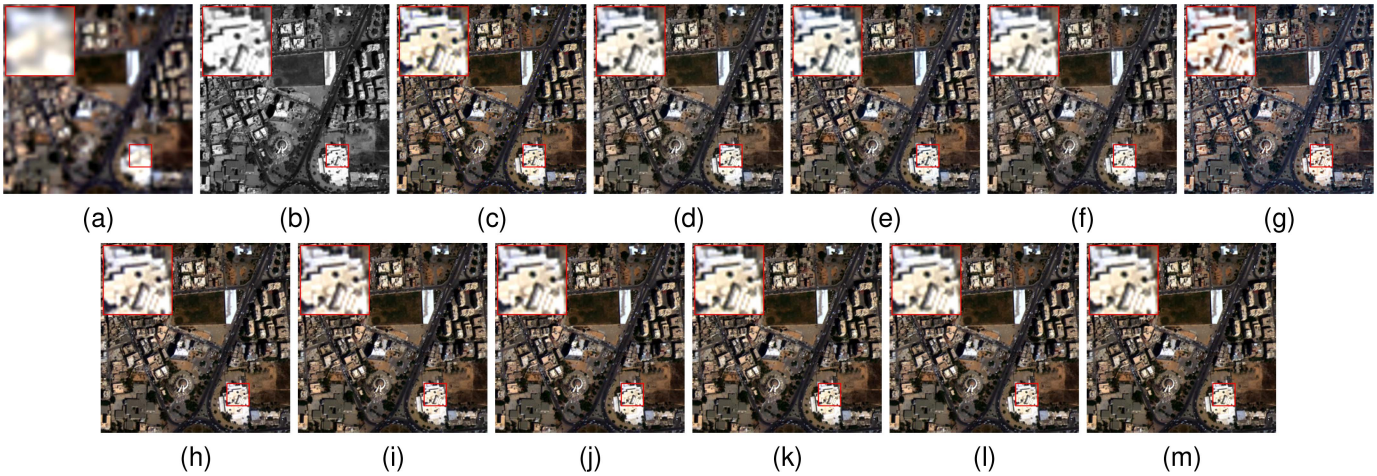


Fig. 7. Visual results on the WV3 reduced-resolution testing set. (a) Degraded LRMS image. (b) Degraded PAN image. (c) Reference image. (d) GSA. (e) BDSD. (f) MTF-GLP-FS. (g) TV. (h) PNN. (i) MSDCNN. (j) PanCSC-Net. (k) Zhou et al. [37] (l) DR-NET. (m) MDPNet.

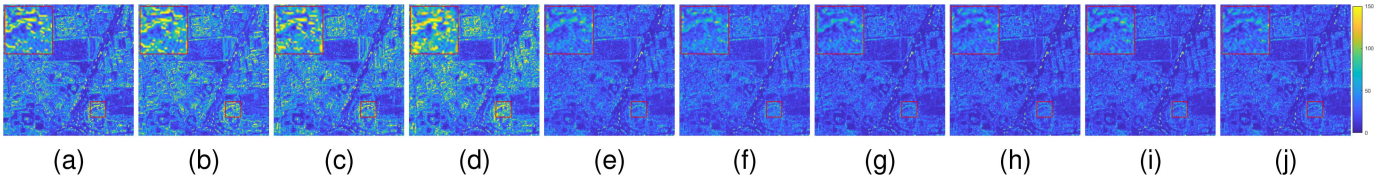


Fig. 8. Residual maps between the results and the reference in Fig. 7. (a) GSA. (b) BDSD. (c) MTF-GLP-FS. (d) TV. (e) PNN. (f) MSDCNN. (g) PanCSC-Net. (h) Zhou et al. [37] (i) DR-NET. (j) MDPNet.

TABLE I
AVERAGE VALUES AND STANDARD DEVIATIONS (STD) OF THE QUANTITATIVE COMPARISON ON 1121 QB REDUCED-RESOLUTION TESTING PATCHES

Method	SAM (±STD)	ERGAS (± STD)	sCC (± STD)	Q4 (± STD)
GSA	1.9273±0.8478	1.4668±0.6854	0.9678±0.0251	0.8788±0.0931
BDSD	1.9180±0.8243	1.4297±0.6622	0.9736±0.0193	0.8888±0.0829
MTF-GLP-FS	1.8771±0.8068	1.4589±0.6426	0.9691±0.0234	0.8812±0.0871
TV	1.9741±0.8194	1.8726±0.9622	0.9656±0.0191	0.8245±0.1316
PNN	1.4217±0.5216	1.0570±0.3631	0.9850±0.0096	0.9107±0.0866
MSDCNN	1.2629±0.4563	0.9011±0.3213	0.9886±0.0064	0.9328±0.0747
PanCSC-Net	1.2009±0.4481	0.8500±0.3237	0.9895±0.0062	0.9426±0.0670
Zhou et al. [37]	1.1601±0.4196	0.8130±0.3062	0.9907±0.0054	0.9447±0.0675
DR-NET	1.1280±0.4101	0.7855±0.2957	0.9918±0.0048	0.9475±0.0662
MDPNet	**1.1058±0.4022**	**0.7650±0.2876**	**0.9920±0.0048**	**0.9490±0.0647**
Ideal value	0	0	1	1

Table I lists the quantitative assessment results of all the compared methods across the 1121 pairs of QB reduced-resolution testing patches, including the mean value and standard deviation (STD) on each evaluation index. The best result of each index is shown in boldface, and the second-best result is underlined. From the full-reference indicators, it can be found that the traditional methods generally fall behind the DL-based methods over the QB reduced-resolution data. Among the DL-based methods, using deep multiscale features, the MSDCNN yields much better quantitative results than the simple three-layer PNN. PanCSC-Net shows better performance than PNN and MSDCNN, but is slightly inferior to Zhou et al. [37] and DR-NET. The transformer-based Zhou et al. [37] and DR-NET slightly surpass the CNN-based methods on all the metrics, while the proposed MDPNet achieves better quantitative results than the compared transformer-based approaches on the QB reduced-resolution testing data.

Table II reports the quantitative assessment results of all the compared methods across the 1116 pairs of WV3 reduced-resolution testing patches. Compared to the QB data with four MS bands, the WV3 data contain eight MS bands and are more challenging, especially in spectral preservation. For the spectral distortion indicator SAM, the DL-based methods yield obviously better values than the traditional algorithms. On the other metrics, the DL-based approaches also show superior performance. The CNN-based PNN and MSDCNN have close quantitative results in terms of all the indicators on the WV3 reduced-resolution testing data, and MSDCNN is slightly better. PanCSC-Net surpasses the classical CNN-based methods PNN and MSDCNN, and yields results close to Zhou

TABLE II
AVERAGE VALUES AND STD OF THE QUANTITATIVE COMPARISON ON 1116 WV3 REDUCED-RESOLUTION TESTING PATCHES

Method	SAM (± STD)	ERGAS (± STD)	sCC (± STD)	Q4 (± STD)
GSA	5.1082±2.3376	3.8059±1.7799	0.9455±0.0395	0.8779±0.1224
BDSD	5.6257±2.6578	4.0366±1.8778	0.9536±0.0281	0.8688±0.1314
MTF-GLP-FS	5.0723±2.2696	3.8957±1.7825	0.9466±0.0365	0.8746±0.1176
TV	5.2736±2.0988	4.1111±1.5417	0.9435±0.0309	0.8538±0.1356
PNN	3.4963±1.2863	2.6023±1.2673	0.9810±0.0132	0.9126±0.1228
MSDCNN	3.4077±1.1944	2.4896±1.1989	0.9829±0.0123	0.9218±0.1213
PanCSC-Net	3.0840±1.2134	2.2823±1.1937	0.9853±0.0122	0.9281±0.1157
Zhou et al. [37]	3.0180±1.1999	2.2440±1.1637	0.9861±0.0114	0.9286±0.1156
DR-NET	2.9762±1.2064	2.2273±1.2094	0.9871±0.0119	0.9298±0.1152
MDPNet	**2.9724±1.1921**	**2.1783±1.0993**	**0.9874±0.0103**	**0.9307±0.1168**
Ideal value	0	0	1	1

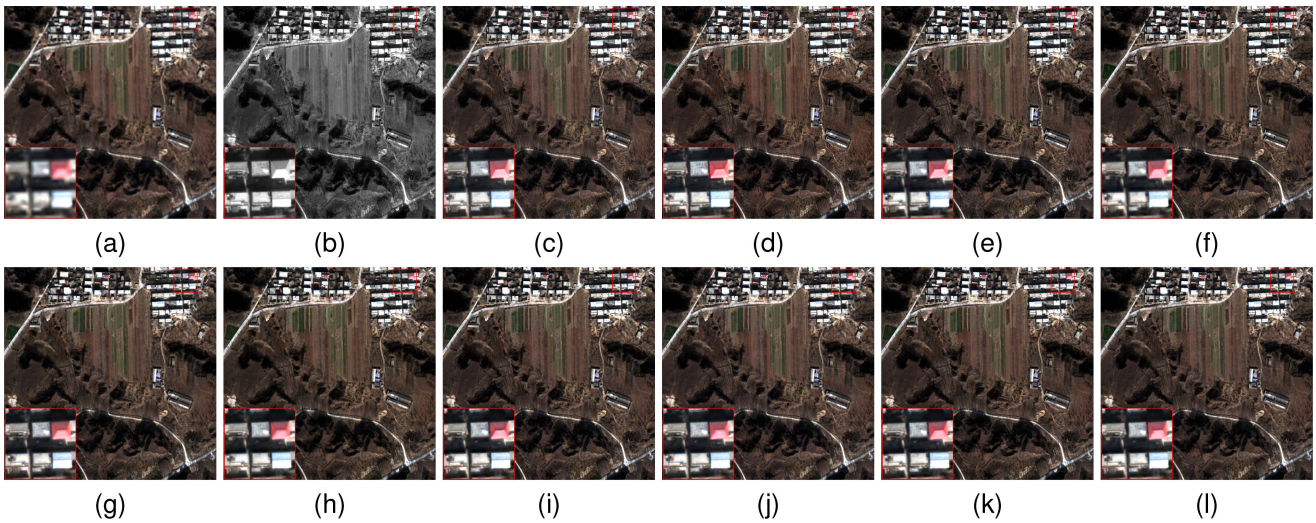


Fig. 9. Visual results on the QB full-resolution testing set. (a) LRMS image. (b) PAN image. (c) GSA. (d) BDSD. (e) MTF-GLP-FS. (f) TV. (g) PNN. (h) MSDCNN. (i) PanCSC-Net. (j) Zhou et al. [37] (k) DR-NET. (l) MDPNet.

et al. [37] and DR-NET. The transformer-based methods slightly outperform PanCSC-Net on the WV3 reduced-resolution testing data. The proposed MDPNet yields slightly better quantitative results than Zhou et al. [37] and DR-NET on all the metrics.

E. Full-Resolution Assessment

Fig. 9 shows the visual results of the compared approaches on a QB full-resolution testing patch pair. It can be observed that the red rooftop in the enlarged view of the GSA fusion result is obviously lighter in color. The edges of the red rooftop in the fusion image of the BDSD are oversaturated. The outcome of MTF-GLP-FS suffers from both lighter colors and oversaturated edges. In the enlarged view of the TV fusion result, obvious color artifacts can be found on the rooftops. The red rooftop in the result of the PNN has a yellowish hue, and that in the result of the MSDCNN suffers from blurring effects. The shadow over the red rooftop in the result of PanCSC-Net is so dark that it looks unnatural. In the outcome of Zhou et al. [37] the red rooftops present slight color distortion and artifacts. The result of DR-NET has an oversaturation problem on the edges of the rooftop. By comparison, it can be found that the fusion result of the proposed MDPNet presents the best spatial and spectral fidelity.

Fig. 10 displays the visual results of the compared approaches on a WV3 full-resolution testing patch pair. Through the enlarged view, it can be observed that the building in the fusion results of GSA, MTF-GLP-FS, and MSDCNN has a lighter color, while that in the fusion image of BDSD is too dark. The fusion results of TV, PanCSC-Net and DR-NET suffer from some color artifacts. In the enlarged view of the fusion results of the PNN and Zhou et al. [37] a small number of pixels with abnormal colors also appear on the rooftop. The fusion image of the proposed MDPNet has clear spatial details and higher spectral fidelity.

Table III lists the quantitative results across the 1121 pairs of QB full-resolution testing patches. The GSA and TV have a higher spatial distortion index D_S and a higher spectral distortion index D_λ, respectively. The BDSD shows a balance on the metrics, and MTF-GLP-FS yields a much better D_λ value than other traditional methods. On the whole, the traditional methods fall behind the DL-based methods. The MSDCNN performs much better on the spectral index D_λ than the PNN. The transformer-based Zhou et al. [37] and DR-NET slightly outperform the CNN-based methods on all the metrics and Zhou

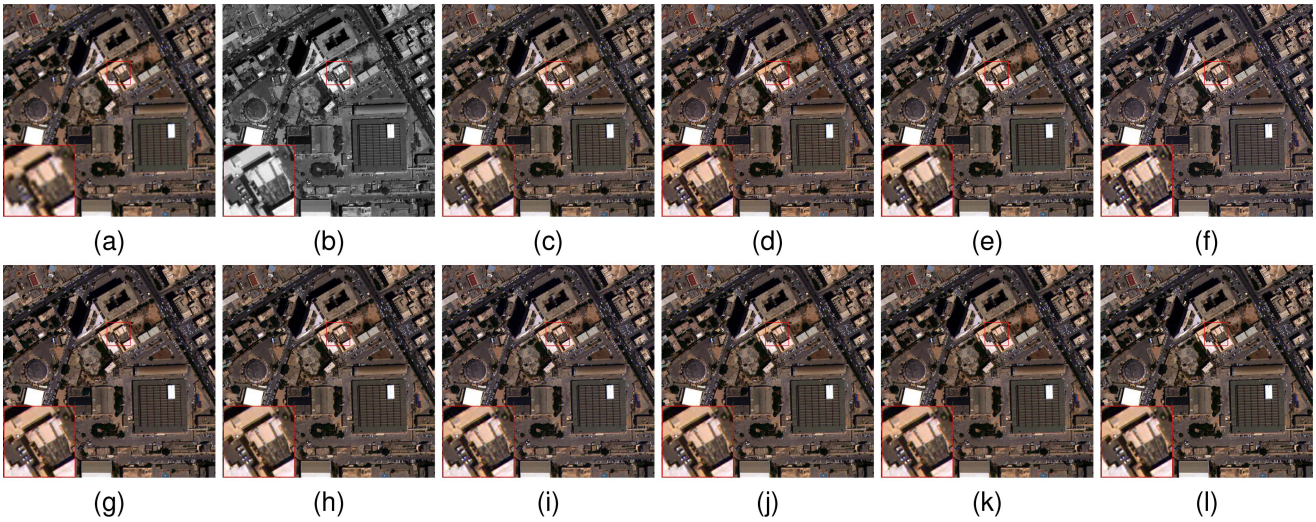


Fig. 10. Visual results on the WV3 full-resolution testing set. (a) LRMS image. (b) PAN image. (c) GSA. (d) BDSD. (e) MTF-GLP-FS. (f) TV. (g) PNN. (h) MSDCNN. (i) PanCSC-Net. (j) Zhou et al. [37] (k) DR-NET. (l) MDPNet.

TABLE III
AVERAGE VALUES AND STD OF THE QUANTITATIVE COMPARISON ON 1121 QB FULL-RESOLUTION TESTING PATCHES

Method	D_λ ($\pm$ STD)	D_S ($\pm$ STD)	HQNR ($\pm$ STD)
GSA	0.0796±0.0713	0.1216±0.0980	0.8146±0.1328
BDSD	0.0701±0.0470	0.0734±0.0647	0.8636±0.0908
MTF-GLP-FS	0.0413±0.0337	0.0887±0.0786	0.8748±0.0895
TV	0.1398±0.1483	0.0692±0.0390	0.8024±0.1511
PNN	0.0619±0.0700	0.0554±0.0572	0.8887±0.0986
MSDCNN	0.0498±0.0582	0.0514±0.0507	0.9031±0.0862
PanCSC-Net	0.0281±0.0210	0.0492±0.0434	0.9246±0.0559
Zhou et al. [37]	0.0333±0.0290	0.0486±0.0493	0.9208±0.0675
DR-NET	0.0498±0.0636	0.0495±0.0467	0.9050±0.0891
MDPNet	**0.0270±0.0197**	**0.0437±0.0413**	**0.9311±0.0540**
Ideal value	0	0	1

TABLE IV
AVERAGE VALUES AND STD OF THE QUANTITATIVE COMPARISON ON 1116 WV3 FULL-RESOLUTION TESTING PATCHES

Method	D_λ ($\pm$ STD)	D_S ($\pm$ STD)	HQNR ($\pm$ STD)
GSA	0.0829±0.0885	0.0924±0.0929	0.8392±0.1390
BDSD	0.1446±0.1084	0.0734±0.0639	0.7969±0.1370
MTF-GLP-FS	0.0484±0.0481	0.0681±0.0751	0.8889±0.0982
TV	**0.0395±0.0471**	0.0771±0.0768	0.8883±0.0980
PNN	0.0819±0.0988	0.0610±0.0897	0.8696±0.1435
MSDCNN	0.0798±0.1025	0.0584±0.0870	0.8737±0.1443
PanCSC-Net	0.0512±0.0471	0.0524±0.0748	0.9021±0.1031
Zhou et al. [37]	0.0558±0.0519	0.0509±0.0784	0.8996±0.1081
DR-NET	0.0698±0.0654	0.0571±0.0843	0.8819±0.1219
MDPNet	0.0424±0.0408	**0.0455±0.0694**	**0.9164±0.0931**
Ideal value	0	0	1

et al. [37] performs better than DR-NET. Although PanCSC-Net is inferior to the transformer-based methods on the reduced-resolution data, it shows better results on the full-resolution QB data. The proposed MDPNet achieves even better quantitative results than PanCSC-Net on all three metrics.

Table IV reports the quantitative results across the 1116 pairs of WV3 full-resolution testing patches. From the no-reference indicators, it can be found that GSA has unsatisfactory overall fusion quality. BDSD performs poorly on the D_λ index. MTF-GLP-FS and TV have much better quantitative results than other traditional methods and even surpass the CNN-based methods. The D_λ values of PNN and MSDCNN are relatively poor. The transformer-based Zhou et al. [37] yields the second-best spatial distortion index D_S and is slightly superior to DR-NET. PanCSC-Net performs satisfactorily on D_λ and D_S indexes, and has the second-best HQNR value. The proposed MDPNet yields the second-best D_λ value, while its D_S and HQNR values are better than all the compared methods.

F. Ablation Study

Since the multiscale embedding blocks, AHAT-based feature fusion module and CSAT-based detail generation module are the core of the proposed MDPNet, a series of ablation experiments are conducted to verify their effectiveness. The results of the ablation experiments will be presented and analyzed in the following.

1) Multiscale Embedding Blocks: The multiscale embedding block embeds $s = 5$ different sizes of image patches centered on each pixel into a multiscale embedding vector. To test the effect of each scale and the total number of scales s, a variety of module settings are tested under the condition that the dimension of a multiscale embedding vector is kept as $sl = 60$ for the fairness of comparison. Specific module settings are reported in Table V. MDPNet represents the full proposed method. To study the effect of only using the information at one scale, model variants with suffixes o1–o5 are tested. To study the effect of scale number s, model variants with suffixes s4–s2 reduce one scale per step. Note that MDPNet-o1 is equivalent to MDPNet-s1, and the proposed MDPNet is equivalent to MDPNet-s5.

Table VI reports the quantitative evaluation results corresponding to the experimental settings in Table V. The SAM, ERGAS, sCC, and Q4 indexes are measured on 1121 QB reduced-resolution testing patch pairs. The D_λ, D_S, and HQNR

TABLE V
EXPERIMENTAL SETTINGS ON THE ABLATION STUDY OF THE MULTISCALE EMBEDDING BLOCKS

Model	s	l	1×1	3×3	5×5	7×7	9×9
MDPNet-o1	1	60	✓				
MDPNet-o2	1	60		✓			
MDPNet-o3	1	60			✓		
MDPNet-o4	1	60				✓	
MDPNet-o5	1	60					✓
MDPNet-s2	2	30	✓	✓			
MDPNet-s3	3	20	✓	✓	✓		
MDPNet-s4	4	15	✓	✓	✓	✓	
MDPNet	5	12	✓	✓	✓	✓	✓

TABLE VI
AVERAGE ABLATION STUDY RESULTS OF THE MULTISCALE EMBEDDING BLOCKS ON THE 1121 QB REDUCED-RESOLUTION AND 1121 QB FULL-RESOLUTION TESTING PATCHES

Model	SAM	ERGAS	sCC	Q4	D_λ	D_S	HQNR
MDPNet-o1	1.1992	0.8427	0.9901	0.9426	0.0342	0.0516	0.9169
MDPNet-o2	1.1845	0.8359	0.9903	0.9438	0.0268	0.0459	0.9292
MDPNet-o3	1.1772	0.8171	0.9907	0.9445	0.0321	0.0545	0.9165
MDPNet-o4	1.1752	0.8274	0.9904	0.9438	0.0268	0.0453	0.9297
MDPNet-o5	1.1673	0.8167	0.9907	0.9447	0.0288	0.0451	0.9280
MDPNet-s2	1.1741	0.8234	0.9905	0.9445	0.0270	0.0457	0.9292
MDPNet-s3	1.1435	0.7889	0.9915	0.9468	0.0272	0.0474	0.9273
MDPNet-s4	1.1291	0.7884	0.9914	0.9472	**0.0262**	0.0463	0.9294
MDPNet	**1.1058**	**0.7650**	**0.9920**	**0.9490**	0.0270	**0.0437**	**0.9311**
Ideal value	0	0	1	1	0	0	1

indexes are measured on 1121 QB full-resolution testing patch pairs. From MDPNet-o1 to MDPNet-o5, the fusion performance on the reduced-resolution data improves slightly in general as the size of the embedded patch grows larger (i.e., scale is increased). But the performance on the full-resolution data fluctuates greatly. This indicates that adopting a larger embedded patch size makes the embedding vector contain more useful information in the reduced-resolution case, while the single-scale embedding tends to be less robust on the full-resolution data. As for the influence of the number of scales s, compared with MDPNet-o1, MDPNet-s2 adds the embedding of 3×3 patches, and the performance is significantly improved at both reduced and full resolutions. From MDPNet-s2 to MDPNet-s4, the fusion effect is constantly promoted on the reduced-resolution data as the embedding of larger patches is added, and the increments are larger than those from MDPNet-o1 to MDPNet-o5. The results of MDPNet-s2 are slightly inferior to those of ACPMT-o5. However, the fusion results of MDPNet-s3, MDPNet-s4, and MDPNet are far superior to those of MDPNet-o5, which indicates that the performance enhancement brought by multiscale information is far greater than that by only increasing a single scale. On the full-resolution data, MDPNet-s2 to MDPNet-s4 show smaller performance fluctuations. Furthermore, the general performance of MDPNet-s2 to MDPNet-s4 is slightly better than that of MDPNet-o1 to MDPNet-o5. The proposed MDPNet uses five scales to obtain the best fusion results, which proves the superiority of the proposed multiscale embedding blocks over standard single-scale patch embedding.

2) AHAT-Based Feature Fusion Module: To verify the effectiveness of the AHAT-based feature fusion module, a series of model variants are designed and tested. Fig. 11 shows the structural changes that we made in the model variants. As shown in Fig. 11(a), the model w/o AHAT removes the entire AHAT-based feature fusion module and directly adds $\mathbf{E}_{\mathrm{P}}$ to $\mathbf{E}_{\mathrm{L}}^{0}$ for fusion. Table VII lists the quantitative results of the ablation experiments for the feature fusion module. Among them, the results of w/o AHAT are far worse than those of the full MDPNet, which proves the key role that the AHAT-based feature fusion module plays in our MDPNet. The AHA and the DESC are two core operations of the AHAT. As shown in Fig. 11(b), the variant Self-Attention+DESC (SA+DESC) replaces the AHA with the standard self-attention to verify the superiority of the AHA. On the other hand, as shown in Fig. 11(c), the model Only AHA removes the DESC and retains only the AHA to verify the effectiveness of the DESC. It can be seen from Table VII that the performances of SA+DESC and Only AHA are significantly inferior to the full MDPNet, which confirms the importance of the AHA and the DESC. Adding PAN keys and values to MS keys and values apparently improves the fusion performance, and a CNN-based short cut is also helpful for transformers in feature fusion.

TABLE VII
AVERAGE ABLATION STUDY RESULTS OF THE FEATURE FUSION AND DETAIL GENERATION MODULES ON THE 1121 QB REDUCED-RESOLUTION AND 1121 QB FULL-RESOLUTION TESTING PATCHES

Model	SAM	ERGAS	sCC	Q4	D_λ	D_S	HQNR
w/o AHAT	1.3250	0.9454	0.9871	0.9345	0.0356	0.0499	0.9174
SA+DESC	1.1570	0.8061	0.9910	0.9455	0.0281	0.0474	0.9265
Only AHA	1.1644	0.8079	0.9909	0.9455	0.0277	0.0471	0.9271
w/o CSAT	1.1789	0.8277	0.9904	0.9438	0.0269	0.0459	0.9290
w/o CSA	1.1763	0.8231	0.9906	0.9441	**0.0266**	0.0453	0.9299
MDPNet	**1.1058**	**0.7650**	**0.9920**	**0.9490**	0.0270	**0.0437**	**0.9311**
Ideal value	0	0	1	1	0	0	1

3) CSAT-Based Detail Generation Module: To validate the CSAT-based detail generation module, two ablation experiments are conducted. As shown in Fig. 11(d), the variant w/o CSAT replaces the entire CSAT-based detail generation module with a simple 3×3 convolution layer to prove the necessity of the module. Furthermore, as shown in Fig. 11(e), the variant w/o CSA verifies the significance of the CSA by removing the CSA from the detail generation module. Table VII lists the quantitative results of the ablation experiments. It can be found that the results of w/o CSA are slightly better than those of w/o CSAT, which proves the positive effect of the CSAT-based detail generation module. Besides, the performance of w/o CSAT and w/o CSA on the reduced-resolution data is significantly inferior to that of the full MDPNet, but on the full-resolution data, w/o CSAT and w/o CSA yield remarkable performance close to the MDPNet, especially on the spectral distortion index D_λ. This indicates that the CSAT tends to improve the spatial quality of pansharpening outcomes at full resolution. An underlying cause may be that the spectral preservation is mainly guaranteed by the skip connection that inject the details $\mathbf{D}$ to the up-sampled LRMS image $\widetilde{\mathbf{L}}$, while the CSAT is primarily responsible for refining the detail of features to generate high-quality $\mathbf{D}$.

G. Computational Efficiency

To further evaluate the computational efficiency and model complexity of the proposed method, the network parameters

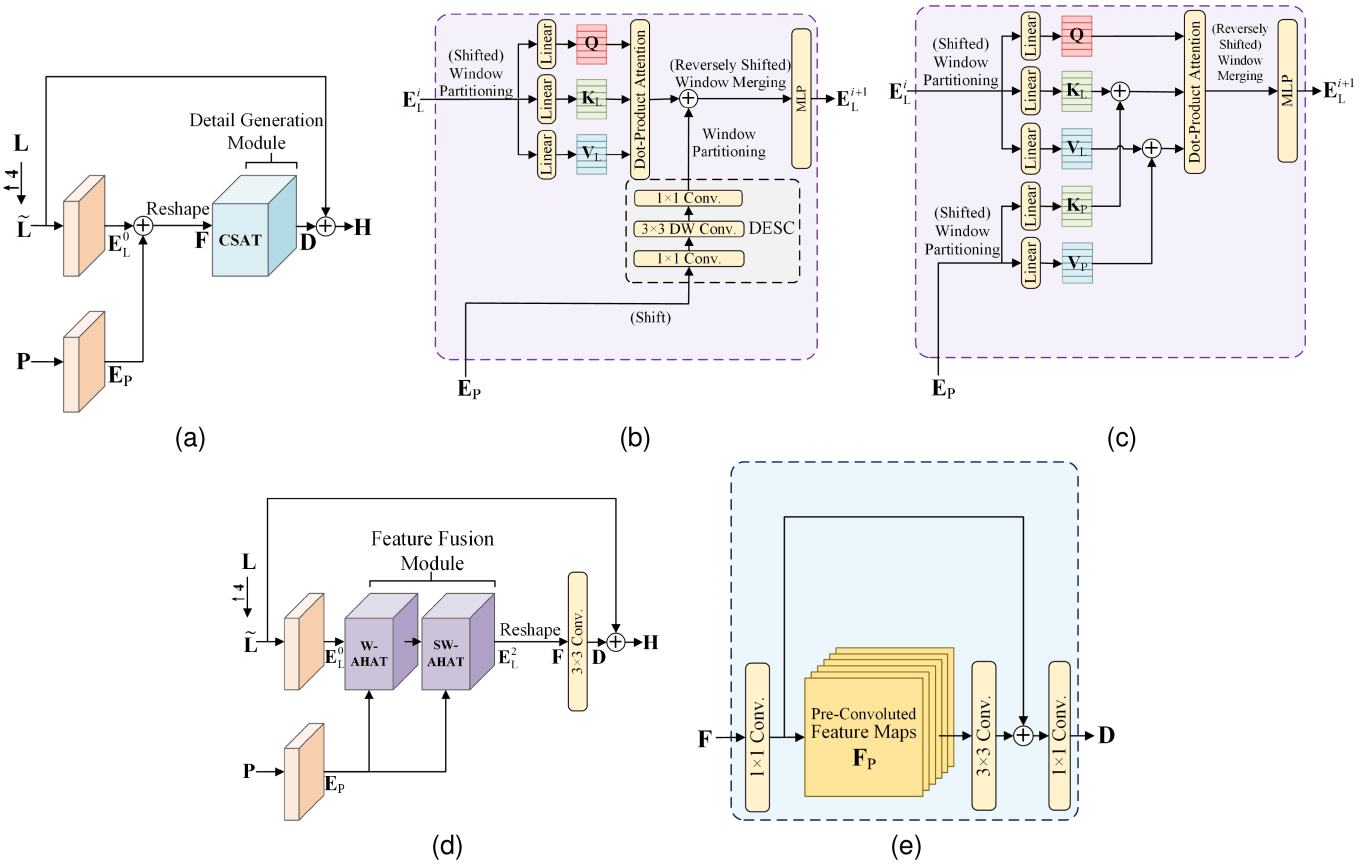


Fig. 11. Structural change diagrams of the model variants in the ablation study for the feature fusion and detail generation modules. (a) W/o AHAT. (b) SA+DESC. (c) Only AHA. (d) W/o CSAT. (e) W/o CSA.

TABLE VIII
NUMBER OF PARAMETERS (#PARAM.) AND AVERAGE RUNNING TIME OF DIFFERENT METHODS ON THE 1121 QB REDUCED-RESOLUTION TESTING PATCHES

Method	#Param.	Time (s)
GSA	-	0.0067
BDSD	-	0.0121
MTF-GLP-FS	-	0.0211
TV	-	0.4026
PNN	80,420	0.0013
MSDCNN	189,852	0.0038
PanCSC-Net	58,300	0.0117
Zhou et al. [37]	70,600	0.0235
DR-NET	2,619,017	0.0120
MDPNet	322,540	0.0093

and the average running time of all the compared methods on the 1121 QB reduced-resolution testing images are measured in Table VIII. The traditional methods are tested on a 2.6-GHz Intel Core i7-10750H CPU, while the DL-based methods are tested on an NVIDIA GeForce RTX 2060 GPU. By comparison, it can be found that the running time of TV is significantly longer than those of other methods, while the running times of PNN and MSDCNN are significantly shorter than those of other DL-based methods due to their simple network architecture. Compared to other transformer-based methods, the proposed MDPNet has a much shorter average running time, which demonstrates the greater efficiency of our MDPNet. As for the number of parameters, DR-NET has significantly more parameters than other methods. This is because the quantity of feature maps throughout the DR-NET is large, which inevitably leads to a large number of parameters in the network layers. PanCSC-Net has the fewest parameters, but its running time is relatively long. Note that the number of parameters is not directly related to the running time. There could be a lot of time-consuming operations without parameters in a DL-based method. Although DR-NET has a large number of parameters, it shows a relatively short running time. The number of parameters in our MDPNet is acceptable, much lower than that of DR-NET but higher than those of MSDCNN and Zhou et al. [37] Although the proposed MDPNet has more parameters, its running time is relatively short.

V. CONCLUSION

In this article, we propose a pansharpening network based on multiscale embedding and dual attention transformers, termed MDPNet. To avoid the inefficiency caused by directly combining the transformer with the classical multiscale network architecture, we propose the multiscale embedding block to embed multiscale information of the images into two embedding sequences. Then, the transformers only need to process these two embedding sequences to make full use of multiscale information.

Moreover, considering domain-specific knowledge, we propose the AHAT, in which the PAN spatial information is added to the MS keys and values for long-range feature extraction and information fusion. Finally, the CSAT is proposed to capture the correlations along the channel dimension and further enhance the fused feature maps. Experimental results on QB and WV3 datasets demonstrate that the proposed MDPNet outperforms the different kinds of pansharpening methods in terms of both visual effects and quantitative metrics, and its running time is shorter than the compared transformer-based methods. Moreover, ablation studies verified the effectiveness of the multiscale embedding block, AHAT, and CSAT.

In the future, we will make efforts to reduce the number of values and keys in the proposed AHAT while maintaining its effectiveness on the pansharpening task, which might result in fewer network parameters and higher efficiency. Moreover, there is bound to be some redundant information among the multiscale embedding vectors. Efforts will be made to remove redundancy within the embedding process.

REFERENCES

[1] P. Ghamisi et al., "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[2] F. Bovolo, L. Bruzzone, L. Capobianco, A. Garzelli, S. Marchesi, and F. Nencini, "Analysis of the effects of pansharpening in change detection on VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 53–57, Jan. 2010.

[3] M. T. Amare, S. T. Demissie, S. A. Beza, and S. H. Erena, "Land cover change detection and prediction in the Fafan catchment of Ethiopia," *J. Geovisualization Spatial Anal.*, vol. 7, no. 2, 2023, Art. no. 19.

[4] J. K. Gilbertson, J. Kemp, and A. Van Niekerk, "Effect of pan-sharpening multi-temporal landsat 8 imagery for crop type differentiation using different classification techniques," *Comput. Electron. Agriculture*, vol. 134, pp. 151–159, 2017.

[5] Y. Zhang, "Understanding image fusion," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 6, pp. 657–661, 2004.

[6] V. Arora, E. Yin-Kwee Ng, and A. Singh, "Machine learning and its applications," in *Smart Electrical and Mechanical Systems*, R. Sehgal, N. Gupta, A. Tomar, M. D. Sharma, and V. Kumaran, Eds. New York, NY, USA: Academic Press, 2022, ch. 1, pp. 1–37. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323907897000026

[7] Y. Bouslihim, M. H. Kharrou, A. Miftah, T. Attou, L. Bouchaou, and A. Chehbouni, "Comparing pan-sharpened Landsat-9 and Sentinel-2 for land-use classification using machine learning classifiers," *J. Geovisualization Spatial Anal.*, vol. 6, no. 2, Art. no. 35, 2022.

[8] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[9] G. Vivone et al., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.

[10] G. Vivone, M. Dalla Mura, A. Garzelli, and F. Pacifici, "A benchmarking protocol for pansharpening: Dataset, preprocessing, and quality assessment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6102–6118, Jun. 2021.

[11] W. Carper, T. Lillesand, and R. Kiefer, "The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data," *Photogrammetric Eng. remote Sens.*, vol. 56, no. 4, pp. 459–467, 1990.

[12] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6 011 875, Jan. 4, 2000.

[13] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, Jan. 2008.

[14] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1204–1211, May 1999.

[15] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, 2000, doi: 10.1080/014311600750037499.

[16] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2006.

[17] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[18] Y. Zhang, S. De Backer, and P. Scheunders, "Noise-resistant wavelet-based Bayesian fusion of multispectral and hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3834–3843, Nov. 2009.

[19] M. R. Vicinanza, R. Restaino, G. Vivone, M. Dalla Mura, and J. Chanussot, "A pansharpening method based on the sparse representation of injected details," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 180–184, Jan. 2015.

[20] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[21] X. Otazu, M. Gonzalez-Audicana, O. Fors, and J. Nunez, "Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376–2385, Oct. 2005.

[22] M. Ghahremani and H. Ghassemian, "Remote-sensing image fusion based on curvelets and ICA," *Int. J. Remote Sens.*, vol. 36, no. 16, pp. 4131–4143, 2015.

[23] H. Shen, M. Jiang, J. Li, Q. Yuan, Y. Wei, and L. Zhang, "Spatial–spectral fusion by combining deep learning and variational model," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6169–6181, Aug. 2019.

[24] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Art. no. 113856, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425723004078

[25] D. Hong et al., "SpectralGPT: Spectral foundation model," 2023, *arXiv:2311.07113*.

[26] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[27] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.

[28] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1753–1761.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[30] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[31] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10227–10242, Dec. 2021.

[32] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520302591

[33] H. Zhou, Q. Liu, and Y. Wang, "PGMAN: An unsupervised generative multiadversarial network for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6316–6327, 2021.

[34] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[35] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.

[36] J. Li, W. Fan, T. Lian, and F. Liu, "Cross-attention-based common and unique feature extraction for pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, Oct. 2023.

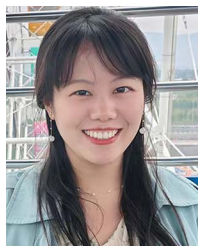
[37] M. Zhou, J. Huang, Y. Fang, X. Fu, and A. Liu, "Pan-sharpening with customized transformer and invertible neural network," in *Proc. AAAI Conf. Art. Intell.*, 2022, pp. 3553–3561. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20267
[38] X. Su, J. Li, and Z. Hua, "Transformer-based regression network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5407423.
[39] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[40] P. Zhang, Y. Mei, P. Gao, and B. Zhao, "Cross-interaction kernel attention network for pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Jun. 2023, Art. no. 5001505.
[41] W. Dou, Y. Chen, X. Li, and D. Z. Sui, "A general framework for component substitution image fusion: An implementation using the fast image fusion method," *Comput. Geosci.*, vol. 33, no. 2, pp. 219–228, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098300406001245
[42] L.-J. Deng et al., "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279–315, Sep. 2022.
[43] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS +pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.
[44] L. He et al., "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.
[45] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6995–7010, Aug. 2021.
[46] 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.
[47] C. Jin, L.-J. Deng, T.-Z. Huang, and G. Vivone, "Laplacian pyramid networks: A new approach for multispectral pansharpening," *Inf. Fusion*, vol. 78, pp. 158–170, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001809
[48] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
[49] D. Hong et al., "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 5518615.
[50] W. G. C. Bandara and V. M. Patel, "Hypertransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1767–1777.
[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.
[52] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 14600–14609.
[53] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997. [Online]. Available: https://hal.science/hal-00365304
[54] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.
[55] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.
[56] X. Cao, X. Fu, D. Hong, Z. Xu, and D. Meng, "PanCSC-Net: A model-driven deep unfolding method for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5404713.
[57] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147–149.
[58] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France: Les Presses de l'École des Mines, 2002.
[59] J. Zhou, D. L. Civco, and J. Silander, "A wavelet transform method to merge landsat TM and spot panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.
[60] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.
[61] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, Oct. 2009.
[62] B. Aiazzi, L. Alparone, S. Baronti, R. Carlá, A. Garzelli, and L. Santurri, "Full-scale assessment of pansharpening methods and data products," in *Proc. Image Signal Process. Remote Sens.*, 2014, pp. 1–12.

Wensheng Fan received the B.S. degree in network engineering from the North University of China, Taiyuan, China, in 2018, and the M.S. degree in software engineering from the Taiyuan University of Technology, Jinzhong, China, in 2023. He is currently working toward the Ph.D. degree in electrical engineering with the College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan.

His research interests include deep learning, intelligent control theory and application, and remote sensing image processing.

Fan Liu (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xian, China, in 2014.

She is currently an Associate Professor with the College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Jinzhong, China. Her main research interests include remote sensing image processing and machine learning.



Jingzhi Li received the B.S. degree in software engineering in 2020 from the Taiyuan University of Technology, Jinzhong, China, where she is currently working toward the M.S. degree in software engineering with the College of Computer Science and Technology (College of Data Science).

Her research interests include deep learning and remote sensing image processing.