



# Efficient Pansharpening by Joint-Modality Recursive Training

Qilei Li , Graduate Student Member, IEEE, Wenhao Song , Mingliang Gao , Senior Member, IEEE, Wenzhe Zhai, Jianhao Sun, and Gwanggil Jeon , Senior Member, IEEE

**Abstract**—Multispectral images captured by remote sensing systems usually have low spatial resolution. Pansharpening offers a promising solution by enhancing the resolution of these low-resolution multispectral images to a high-resolution multispectral without the need for costly hardware upgrades. Existing methods employ either CNN or Transformer as the feature extractor backbone, however, CNN-based methods are weak in capturing long-distance correlation, and Transformer-based methods are limited to extracting fine-grain detail. Moreover, these models achieve impressive results with numerous learnable parameters, which makes them impractical for integration into remote sensing systems. In this work, a parameter-efficient pansharpening model, named joint-modality association network, is built by leveraging complementary information from multiple modalities and recursive training. It aims to improve the resolution of remote-sensing images. Specifically, we efficiently leverage the complementary information from different modalities, including the transformer and CNN joint block, and employ a hierarchical association mechanism to create a more distinctive and informative representation by associating intramodality and cross-modality. Furthermore, the parameter-sharing mechanism of recursive training can effectively reduce the number of parameters in the model. Benefiting from its lightweight design and effective information fusion strategy, the proposed method can generate faithful super-resolved multispectral images that excel in both spectral and spatial resolution. Experimental results show the superiority of the proposed method over extensive benchmarks.

**Index Terms**—Multimodalities, multispectral images, pansharpening, parameter-efficient, spatial resolution.

## I. INTRODUCTION

**O**WING to the robust grounding capabilities exhibited by satellites, remote sensing images obtained through

Manuscript received 9 October 2023; revised 2 April 2024; accepted 3 July 2024. Date of publication 16 July 2024; date of current version 5 August 2024. (Qilei Li and Wenhao Song contributed equally to this work.) (Corresponding authors: Mingliang Gao; Gwanggil Jeon.)

Qilei Li is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: qilei@iee.org).

Wenhao Song, Mingliang Gao, Wenzhe Zhai, and Jianhao Sun are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: 18110403087@stumail.sdut.edu.cn; mlgao@sdut.edu.cn; 20404020495@stumail.sdut.edu.cn; 22504030001@stu.mail.sdut.edu.cn).

Gwanggil Jeon is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China, and also with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2024.3429423

sensor systems encapsulate much substantive ground information. Low-resolution multispectral (LMS) images and panchromatic (PAN) images are two modalities frequently captured by satellites [1]. The LMS image exhibits high spectral resolution yet comparatively constrained spatial resolution, whereas the PAN image shows inverse attributes in this regard [2]. Pansharpening is significant in remote sensing systems because it enhances the spatial resolution of LMS images without the need for expensive hardware upgrades. In response to the requirements posed by various practical applications, such as land surveying [3], environmental monitoring [4], and object detection [5], pansharpening techniques fuse acquired LMS images and PAN images to generate a high-resolution multispectral (HMS) image.

Over the past few decades, traditional approaches have been introduced to address the pansharpening task [6]. The traditional methods of pansharpening can be categorized into four primary categories, i.e., component substitution (CS)-based methods [7], multiresolution analysis (MRA)-based methods [8], CS-MRA hybrid-based methods [9], and model-based methods [10]. Nonetheless, considering the diverse sensors' different spectral responses and terrestrial objects' intricate nature, it is difficult to formulate the correlation between source images and HMS image representations through traditional methods [11].

In recent years, the prominent feature extraction capabilities and inherent nonlinearity exhibited by deep neural networks [12], [13], [14] have propelled the development of pansharpening. For example, Masi et al. [15] proposed a convolutional neural network (CNN) for pansharpening (PNN), which adapts the architecture of the super-resolution technique SRCNN [16] by incorporating domain-specific insights from remote sensing. In addition, Zhong et al. [17] introduced a hybrid pansharpening method, which leverages a three-layer CNN architecture to enhance the spatial resolution of LMS images. Furthermore, Wei et al. [18] introduced the DRPNN model. It employs a deeper neural network to capture the residuals between the LMS image and the ground truth. While the GAN-based methods incorporate a generator and a discriminator to facilitate the fusion procedure through a game, without reliance on ground truth [19]. For example, Liu et al. [20] introduced PSGAN, which aims to enhance similarity by emphasizing the alignment of probability distributions. Ma et al. [21] introduced an unsupervised method in terms of PAN-GAN, with an architecture designed to enable the retention of abundant spectral details from MS images and spatial characteristics from PAN images.

In particular, PAN-GAN’s generator is designed to engage in adversarial games with both the spectral discriminator and the spatial discriminator independently. Zhang et al. [22] introduced a bidirectional pyramid network that combines the traditional approach with deep learning, processing MS and PAN images through distinct branches at different levels.

Despite the notable successes of current pansharpener techniques, three unresolved challenges merit attention. First, most existing pansharpener methods may involve large model parameters and high-computational costs, which are detrimental to practical deployment and execution. Second, different remote-sensing image modalities contain different complementary information features. For example, PAN images possess strong spatial information, whereas multispectral images contain strong spectral information. Existing methods are unable to effectively utilize these complementary pieces of information to enhance the spectral fidelity and spatial resolution of fused images. Furthermore, many transformer models require a significant amount of computation to achieve high performance. Their limitations in terms of hardware resources and battery life make them unsuitable for mobile applications.

To cope with the above issues, we propose a model with the recursive neural network (RNN) design. By introducing a parameter-sharing mechanism within the RNN, we effectively reduce the parameter count of the model. Besides, CNN and transformer blocks are used to extract local and nonlocal features, respectively. For better local and nonlocal feature fusion, we design a joint modality module that combines these two. Through comprehensive evaluations of multiple satellite benchmark datasets, the proposed method demonstrates superior performance compared to state-of-the-art methods. It highlights the potential applications of deep learning in the field of environmental monitoring and assessment. By employing a RNN design with a parameter-sharing mechanism, joint-modality association network (JMAN) significantly reduces the number of model parameters. This makes the network more efficient and portable. The efficient design of the proposed method allows for better utilization of hardware resources and battery life, which is particularly beneficial for mobile applications in remote sensing.

We summarize the contribution of this article as follows:

- 1) We introduced a parameter-sharing mechanism, where the same module is repeatedly used in RNNs to construct a deep network, effectively reducing the number and complexity of model parameters, and thereby improving the model’s efficiency and portability.
- 2) We designed a unified multimodal module that combines CNNs and visual transformers. By employing the local receptive fields of CNNs and the nonlocal characteristics of transformers, we aim to extract and fuse both local and global information from the source images.
- 3) We introduced two modality knowledge association mechanisms, namely, intramodal knowledge association and cross-modal knowledge association, to enhance the feature complementarity and distinctiveness between different modalities through mutual referencing and interactive learning.

## II. RELATED WORK

### A. Deep Learning-Based Pansharpener Methods

With the development of CNNs [23], the deep-learning methods have achieved dominant performance in pansharpener. In general, deep learning-based pansharpener methods can be categorized into two principal classes, namely, CNN-based methods and generative adversarial network (GAN)-based methods. The pioneering utilization of a CNN in the domain of pansharpener is exemplified by a convolutional neural network for PNN [15]. PNN adopts streamlined network architectures while achieving commendable performance benchmarks. Inspired by PNN, many CNN-based methods have been subsequently formulated and presented. For example, Liu et al. [24] introduced TFNet, which formulates an encoder-decoder network to execute a comprehensive process encompassing feature extraction, fusion, and reconstruction within the CNN framework. Yang et al. [25] built a dual-stream network architecture for amplifying residual information, facilitating the transfer of inter-resolution information. The GAN-based methods consist of a generator component and a discriminator component. By adversarial game, the GAN-based approaches can synthesize images of elevated quality, thereby effectively cheating discriminators of the framework. For example, the pioneer of the GAN-based pansharpener approach is pansharpener by the generative adversarial network (PSGAN) [20], which incorporates both the synthesized image and the ground truth as input to the discriminator component. Subsequently, Shao et al. [26] built a residual architecture, termed residual encoder-decoder conditional generative adversarial network (RED-cGAN). The discriminator within this framework is devised to enhance spatial information within the ultimate outcomes. Moreover, Ma et al. [21] built a dual-discriminator architecture. This unsupervised method operates in the absence of ground truth, where the two discriminators promote the output to contain the spatial attributes of the PAN image and the spectral characteristics of the LMS image. Although the aforementioned methods have achieved visually favorable results, these methods are unable to effectively utilize global and local information.

### B. Transformer

The transformer model has gained extensive prominence in the domain of natural language processing (NLP) [27]. Benefiting from its potent aptitude for comprehensive contextual feature exploration, a plethora of methods rooted in transformer architecture have emerged, specifically tailored to address diverse computer vision (CV) challenges. In the work conducted by Zheng et al. [28], a transformer architecture is utilized to address the semantic segmentation task. In this context, the semantic segmentation task was reformulated as a sequence-to-sequence prediction problem. Concurrently, Wang et al. [29] introduced the pyramid vision transformer, an innovative framework tailored for a spectrum of dense prediction tasks. Recently, transformer-based pansharpener methods have emerged in the research landscape. Inspired by the vision transformer (ViT) in

image classification, Meng et al. [30] proposed a transformer-based model for pansharpening. Su et al. [31] introduced an architecture named DR-NET, based on a transformer regression framework. It consists of three stages, i.e., feature extraction to capture spectral and spatial details, feature fusion to integrate the extracted features, and image reconstruction to yield images with balanced spectral distribution and comprehensive spatial details. Zhu et al. [32] proposed a method named MHATP-Net, which combined different levels of attention and a special loss function to address the challenges of spectral distortion and insufficient spatial detail from remote sensing. The Transformer model has strong global information capture capability. Therefore, this work employs the Transformer model to capture the global information of the source image.

### C. Lightweight Network

To meet the requirements of practical engineering, lightweight model has been explored by many scholars [33], [34], [35]. These models are expected to apply fewer parameters to get a satisfactory result. The MobileNet series has pioneered the concept of leveraging depthwise separable convolutions for lightweight networks. Howard et al. [36] introduced MobileNet V1, decomposing standard convolutions into depthwise and pointwise convolutions, effectively curtailing parameter count and computational overhead. Subsequently, Sandler et al. [37] proposed MobileNet V2, which incorporates residual structures and linear bottlenecks to enhance performance. MobileNet V3 [38] introduced adaptive width and activation function design, yielding improved efficiency and effectiveness. Besides, efficientNet [39] has presented a scaling approach, achieving enhanced performance by harmonizing network depth, width, and resolution. This technique has propelled models to achieve superior efficiency without imposing additional computational burdens. Beyond architectural design, quantization and pruning constitute pivotal techniques for enhancing lightweight network efficiency. Quantization reduces parameter representation to fewer bits, mitigating memory and computation demands. On the other hand, it reduces model size and computational complexity by eliminating nonessential connections and parameters [40]. Lightweight networks reduce the need for processing power by reducing model parameters and simplifying computation. This allows them to run on devices with lower computational power, such as smartphones, embedded systems, and IoT devices.

### D. Recursive Neural Network

RNNs were introduced as an attempt to capture hierarchical structures in NLP. They aimed to address the limitations of standard RNNs and feedforward neural networks in handling structured data like syntax trees. Socher et al. [41] introduced the recursive neural tensor network (RNTN), a prominent early work in NLP utilizing recursive structures. RNTN extended the recursive autoencoder model to capture compositional semantics by representing phrases as binary-branching trees. The model employed tensor-based composition functions to compute

phrase representations. RNNs found success in sentiment analysis tasks. Socher et al. [41] demonstrated that RNTNs could effectively capture sentiment information through compositional representations, improving the sentiment classification accuracy of phrases and sentences. This opened up new possibilities for handling syntactic and semantic nuances in text analysis.

RNNs have been extended to the field of CV in recent decades. For example, Sadr et al. [42] proposed a method that integrates CNNs with RNNs for sentiment analysis. In this study, a RNN is employed, leveraging its tree-like structure to substitute the pooling layer in convolutional networks. This results in a performance surpassing that of both the conventional CNN and RNN with fewer parameters. McLaughlin et al. [43] combined CNN and RNN architectures to propose a network to identify individuals in videos. The incorporation of RNN into CNN's foundation enables the network to process video sequences. Furthermore, the inclusion of a temporal pooling layer over the RNN layer facilitates the handling of videos with arbitrary lengths. This alleviates the strong discriminative state shift of RNN toward previous frames and addresses the multiscale challenges within video sequences [44]. The structure of RNN facilitates the reduction of the number of model parameters, therefore, the proposed method uses the RNN structure to implement a parameter-sharing network.

## III. PROPOSED METHOD

### A. Overview

The objective of pansharpening is to improve the spatial resolution of the LMS images with guidance from the corresponding PAN images, which has higher spatial resolution. This process is expected to be achieved by a learnable mapping function  $f_\theta$ , where  $\theta$  is the parameter optimized by supervised training. In this work, we proposed JMAN to enhance the feature discrimination by exploring the synergy of different feature modalities with two types of knowledge associations, and enable the network to be parameter-efficient by recursive training.

The overview of the proposed JMAN is shown in Fig. 1. Given an LMS image  $M \in \mathcal{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$  and the corresponding high-resolution PAN image  $P \in \mathcal{R}^{1 \times H \times W}$ , we align the spatial resolution of  $M$  and  $P$  and upsample with LMS by a scale of 4, and denote the upsampled MS image as  $M_\uparrow$ . We follow the conventional residual learning diagram to focus on modeling missed high-frequency detail texture in the MS image. We eliminate the low-frequency from the upsampled MS image by differentiating it with the PAN image as  $x = M_\uparrow \ominus P$ , where  $\ominus$  is the broadcasted subtraction operator. Consecutively,  $x$  is fed into a head module to extract low-level features, as follows:

$$\mathbf{F}_{\text{low}} = H_{\text{head}}(x) \quad (1)$$

where the head module  $H_{\text{head}}$  consists of two convolution layers. Next, the low-level features  $\mathbf{F}_{\text{low}}$  are passed to the main body module, composed by the associated multimodal blocks, to extract and fuse the hierarchical intermediate representations learned jointly by two encoders, namely, the conventional CNN for local information and the recent ViT for global information. We abstract this process in (2) and provide a detailed illustration

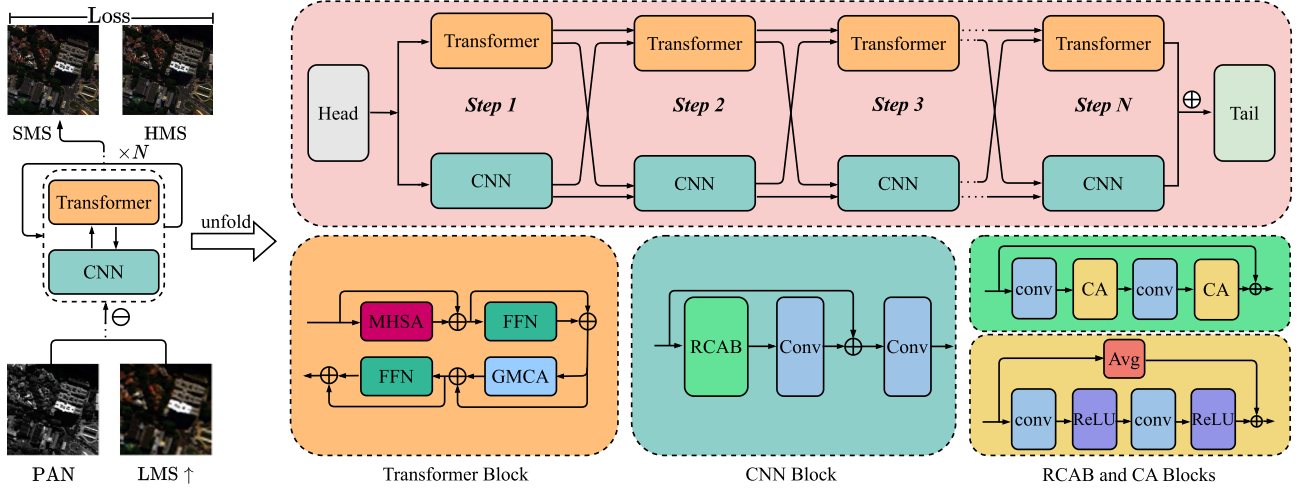


Fig. 1. Framework of the proposed method. The network is designed as a recurrent neural network for parameter efficiency. The multimodal blocks, including a transformer block and a CNN block, are interactive with each other by the proposed inter-modality knowledge association. Further complemented by the intermodality knowledge association, the learned representation is informative for reconstructing a HMS image.  $\oplus$  is the broadcasted concatenation operator.

in Section III-B.

$$\mathbf{F}_{\text{high}} = H_{\text{body}}(\mathbf{F}_{\text{low}}) + \mathbf{F}_{\text{low}} \quad (2)$$

where  $H_{\text{body}}$  is the function of the body module and  $\mathbf{F}_{\text{high}}$  is the extracted high-level feature, which encodes semantic information of the input image pair. Subsequently, a tail module is designed to reconstruct the missed high-frequency component to the MS image and produce a super-resolved MS image as follows:

$$\begin{aligned} \hat{M} &= H_{\text{tail}}(\mathbf{F}_{\text{high}}) \\ M_s &= M \uparrow + \hat{M} \end{aligned} \quad (3)$$

where  $M_s$  is the super-resolved image that has both merits of rich spectral information and high-spatial resolution.

### B. Multimodality Joint Learning

We designed the body part as a joint modality module to incorporate both CNN layers and transformer blocks module, to benefit from the dedicated local receptive field of CNN, and the nonlocal property of the transformer. The feature representations extracted by both blocks are mutually fused for exploring the local-global synergy. To build a deeper network that is capable of extracting hierarchical intermediate activation while maintaining the network parameter-efficient, inspired by the recent progress on RNNs, we designed the foldable joint-modality module by sharing the parameters in different steps. Assuming there are  $S$  steps in total, our design enables the number of learnable parameters to be one-quarter compared with that of conventional design. The proposed method leverages complementary information from multiple modalities in the pansharpening process by efficiently utilizing the complementary information from different networks, including transformer and CNN blocks. The proposed model extracts global and local information through the interaction of CNN and Transformer modules and performs multistep transmission to extract features of different depths. This allows JMAN to effectively fuse local

and global features, enhancing the spectral fidelity and spatial resolution of the fused images.

*Global perception by Transformer:* Considering that the usage of the recursive pattern of input images makes the feature extraction more efficient, we employ transformer blocks for nonlocal information exploring to model the contextual correlations. As illustrated in Fig. 1, the transformer block is composed of cascaded layers, including a multihead self-attention (MHSA) layer, two feed-forward networks (FFN), and a grouped multihead cross-attention (GMCA) layer. The MHSA layer is designed to model nonlocal correlations from the tokenized input  $\mathbf{F}_{\text{Tok}}^t$ . The GMCA layer consists of two paralleled multihead cross-attention (MHCA) blocks, which is designed to explore the intramodality knowledge by mutual-referenced attentive learning. The FFNs following both attention layers are used for feature refinement and dimension alignment. Finally, the output of the transformer block is decoded as the CNN feature to perform intermodality knowledge interaction. Given the input feature  $\mathbf{F}_{\text{T}}^s$  at step  $s$ , the process of Transformer block can be mathematically formulated as follows:

$$\begin{aligned} \mathbf{F}_{\text{tok}}^s &= H_{\text{tok}}(\mathbf{F}_{\text{T}}^s) \\ \mathbf{F}_{\text{out}}^s &= H_{\text{tsfm}}(\mathbf{F}_{\text{Tok}}^s), \quad \mathbf{F}_{\text{T}}^{s+1} = H_{\text{dec}}(\mathbf{F}_{\text{out}}^s) \end{aligned} \quad (4)$$

where  $H_{\text{tok}}(\cdot)$  and  $H_{\text{dec}}(\cdot)$  are the tokenizer and the reversed decoder,  $\mathbf{F}_{\text{T}}^{s+1}$  is the decoded representation yielded by the transformer block, which has the same dimension as the CNN feature.  $H_{\text{tsfm}}^s(\cdot)$  is the operator of the transformer block, including the attention layers and the FFNs.

*Local concentration by CNN:* In contrast to the nonlocal awarded transformers, CNN layers concentrate on specific regions using 2-D kernels. To explore the complementary knowledge offered by local attentive learning, we design the CNN-based cascaded residual channel attention blocks, which is shown in Fig. 1. Similar to the Transformer branch, the CNN branch is unfolded into  $S$  steps, and the feature activation is

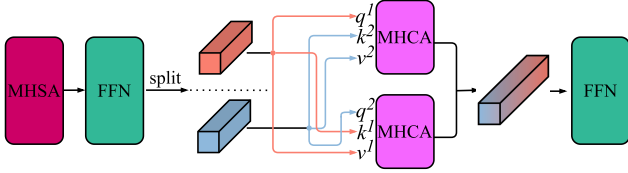


Fig. 2. Illustration of intramodality knowledge association. It is achieved by mutually referring to the other counterpart as the query (Q), key (K), and value (V) in MHCA.

progressively refined as follows:

$$\mathbf{F}_C^{s+1} = H_{\text{CNN}}(\mathbf{F}_C^s) \quad (5)$$

where  $H_{\text{CNN}}(\cdot)$  denotes the function of the CNN block,  $\mathbf{F}_C^s$  and  $\mathbf{F}_C^{s+1}$  are the feature maps extracted by the CNN block at  $s$  step and  $(s+1)$  step.

The combination of transformer and CNN architectures is used to leverage the strengths of both. Transformers are good at capturing long-range dependencies in data, whereas CNNs excel at extracting local features. This dual approach can enhance the model's understanding of complex patterns and improve overall performance.

### C. Intramodality Knowledge Association

We introduce the intramodality knowledge association mechanism to explore the feature interaction within the same modality. For the transformer block at each step, as illustrated in Fig. 2, the nonlocal feature extracted by the MHSA block is sliced evenly into two separate parts and, respectively, fed into a GMCS module, which consists of two parallel MHCS blocks. The intramodality interaction is achieved by mutually referring the input to another. Assuming a feature map  $\mathbf{F}_{\text{TF}}^s$  from the first FFN, we omit the step-index  $s$  for notion simplicity, it is sliced into two parts across the channel dimension and mapped into query (Q), key (K), and value (V) spaces by the corresponding projector as follows:

$$\begin{aligned} \{\mathbf{F}_{\text{TF}}^1, \mathbf{F}_{\text{TF}}^2\} &= H_{\text{split}}(\mathbf{F}_{\text{TF}}) \\ \{q^1, k^1, v^1\} &= \{H_{q1}(\mathbf{F}_{\text{TF}}^1), H_{k1}(\mathbf{F}_{\text{TF}}^1), H_{v1}(\mathbf{F}_{\text{TF}}^1)\} \\ \{q^2, k^2, v^2\} &= \{H_{q2}(\mathbf{F}_{\text{TF}}^2), H_{k2}(\mathbf{F}_{\text{TF}}^2), H_{v2}(\mathbf{F}_{\text{TF}}^2)\} \end{aligned} \quad (6)$$

where  $H_q$ ,  $H_k$ , and  $H_v$  are the projectors. Recall the cross-attention operation is formulated as follows:

$$\text{Cross-Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \quad (7)$$

We perform the intradomain knowledge association by mutually referring each to the other as follows:

$$\begin{aligned} \mathbf{F}_{\text{TF}}^{1\leftarrow 2} &= \text{Cross-Attention}(q^1, k^2, v^2) + \mathbf{F}_{\text{TF}}^1 \\ \mathbf{F}_{\text{TF}}^{2\leftarrow 1} &= \text{Cross-Attention}(q^2, k^1, v^1) + \mathbf{F}_{\text{TF}}^2 \end{aligned} \quad (8)$$

where  $\mathbf{F}_{\text{TF}}^{1\leftarrow 2}$  and  $\mathbf{F}_{\text{TF}}^{2\leftarrow 1}$  are the associated intermediate representations, where are then concatenated and followed by an FFN to produce the transformer output  $\mathbf{F}_{\text{out}}^s$  as follows:

$$\mathbf{F}_{\text{out}}^s = H_{\text{FFN}}(H_{\text{cat}}(\mathbf{F}_{\text{TF}}^{1\leftarrow 2}, \mathbf{F}_{\text{TF}}^{2\leftarrow 1})) \quad (9)$$

where  $H_{\text{FFN}}(\cdot)$  and  $H_{\text{cat}}(\cdot)$  refer to the operation of FFN and the concatenation.

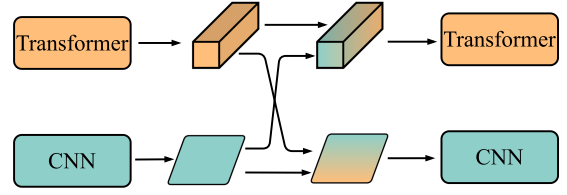


Fig. 3. Illustration of intermodality knowledge association. The features extracted by the network in different modalities are cross-referenced by each other for feature augmentation.

### D. Intermodality Knowledge Association

In contrast to intermodal knowledge association, we explore the global information provided by the transformer block, and the local information provided by the CNN block through additively aggregating them to augment the discrimination ability of the representation of learned features. The diagram of the intermodality knowledge association module is shown in Fig. 3.

Specifically, given the output  $\{\mathbf{F}_{\text{T}}^s, \mathbf{F}_{\text{C}}^s\}$ , respectively, from the transformer block and the CNN block before step  $s$ , we explore the synergy through an adaptive fusion block, which takes the concatenated representations as the input, and perform local-global association to produce the output. The process is denoted as follows:

$$\mathbf{F}^s = H_{\text{cat}}(\mathbf{F}_{\text{T}}^s, \mathbf{F}_{\text{C}}^s), \quad \mathbf{F}_{\text{fuse}}^s = H_{\text{fuse}}(\mathbf{F}^s) + \mathbf{F}^s \quad (10)$$

where  $\mathbf{F}_{\text{fuse}}^s$  is the fused representation, and  $H_{\text{fuse}}(\cdot)$  is the fusion block composed of a 3-layer CNN. In order to utilize the complementary information in  $\mathbf{F}_{\text{fuse}}^s$ , we inject the cross-modality knowledge into the corresponding block to enhance unimodality representation as follows:

$$\begin{aligned} \{\mathbf{F}_{\text{C}\leftarrow\text{T}}^s, \mathbf{F}_{\text{T}\leftarrow\text{C}}^s\} &= H_{\text{split}}(\mathbf{F}_{\text{fuse}}^s) \\ \mathbf{F}_{\text{C}}^s &= H_{\text{C\_ref}}(\mathbf{F}_{\text{C}\leftarrow\text{T}}^s) + \mathbf{F}_{\text{C}}^s \\ \mathbf{F}_{\text{T}}^s &= H_{\text{T\_ref}}(\mathbf{F}_{\text{T}\leftarrow\text{C}}^s) + \mathbf{F}_{\text{T}}^s \end{aligned} \quad (11)$$

where  $H_{\text{C\_ref}}$  and  $H_{\text{T\_ref}}(\cdot)$  are 1-layer CNN for feature refinement. Then, the output  $\mathbf{F}_{\text{T}}^s$  and  $\mathbf{F}_{\text{C}}^s$  will be used as the input for the next step.

### E. Model Training

We take  $\mathbf{F}_{\text{fuse}}^s$ , which is the output of the last step, as the high-level feature  $\mathbf{F}_{\text{high}}$  to produce the super-resolved MS image by (3). Since all the designed modules are differentiable, the proposed JMAN model can be trained end-to-end by conventional gradient update. We use the mean squared error (mse) loss to supervise the training, and set the high-resolution HMS as the ground truth. The loss function is formulated as follows:

$$\mathcal{L} = \sum_{i=1}^N |M_s - M_{\text{GT}}|_2 \quad (12)$$

where  $N$  is the number of training samples, and  $M_{\text{GT}}$  is the HMS image.  $||_2$  denotes  $L_2$  norm (Euclidean norm) of the vector difference.

#### IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we first introduce the benchmark details and evaluation metrics. Qualitative and quantitative experiments are then conducted to elucidate the superiority of the proposed method. Finally, we conduct ablation experiments to validate the necessity of certain structural designs.

##### A. Benchmark Details

To validate the performance of different methods in pansharpening, this experiment is conducted using three benchmark datasets, i.e., WorldView 3 (WV3), Gaofen 2 (GF2), and QuickBird (QB). The WV3 dataset is a high-resolution and multispectral remote sensing data with eight bands. The QB dataset is a reliable high-resolution remote sensing data with four bands. The GF2 dataset also has four bands but with a larger spatial sampling interval compared to QB. The benchmark information for this experiment is based on the performance of PAN sharpening on the datasets. Therefore, 20 simulated data and 20 real data are selected from each dataset, containing representative objects, for comparison. The best method is selected based on various quantification metrics.

##### B. Evaluation Metrics

To measure the pansharpening images, the current design employed different quantitative metrics to assess simulation data and real data. For the simulated data, this study utilizes a reduced-resolution approach to validate its synthesis performance. In this scenario, the quality of fused images can be evaluated using four commonly used pansharpening metrics, named spectral angle mapper (SAM), relative dimensionless global error in synthesis (ERGAS), universal image quality index ( $Q2^n$ ), and spatial correlation coefficient (SCC).  $Q2^n$  as a metric used for evaluating the quality of fused  $2^n$ -band images. It is employed to measure the quality of the fused images. Higher values indicate better visual consistency with the ground truth. SAM is used to assess the spectral similarity of the fused images to the ground truth. Lower values are better. ERGAS evaluates the spectral distortion and spatial distortion of fused images. Lower values indicate less error. SCC reflects the spatial consistency of the fused images. Higher values denote better correlation with the reference.

##### C. Comparison With State-of-the-Art Methods

In this section, we conducted qualitative and quantitative experiments on three publicly available datasets, i.e., GaoFen2 (GF2), QuickBird (QB), and WorldView III (WV3) datasets, and compared the proposed method with eight state-of-the-art methods, e.g., BT-H [45], BDS-D-PC [46], MTF-GLP-HPM-R [47], C-GSA [48], MSDCNN [49], BDPN [22], DiCNN [50], MTF-GLP-FS [51], SFITNet [52], and LAG-Conv [53]. The hyperparameter  $s$  for the time step was set to 2 for this test.

1) *Qualitative Evaluation*: To qualitatively assess the performance of various methods in remote sensing image fusion tasks, we compared the fusion results of the JMAN method

TABLE I  
COMPARE WITH SOTA METHODS ON WV3 DATASET

Method	Q8 $\uparrow$	SAM $\downarrow$	ERGAS $\downarrow$	SCC $\uparrow$
LMS $\uparrow$	0.6199	5.7534	7.1220	0.7434
BT-H [45]	0.8310	4.8734	4.5496	0.9253
BDS-D-PC [46]	0.8280	5.4024	4.6766	0.9075
C-GSA [48]	0.8228	5.6111	4.8442	0.8945
SR-D [54]	0.8178	4.9190	4.6397	0.9166
MTF-GLP-HPM-R [47]	0.8332	5.2971	5.1301	0.8909
MTF-GLP-FS [51]	0.8316	5.2791	4.6776	0.9007
TV [55]	0.7952	5.6540	4.8292	0.8977
DRPNN [18]	0.8815	3.7251	2.7496	0.9740
MSDCNN [49]	0.8531	4.7124	3.6765	0.9541
BDPN [22]	0.8560	4.5996	3.6016	0.9506
DiCNN [50]	0.9002	3.5441	2.6607	0.9763
PNN [15]	0.8959	3.5998	2.6332	0.9764
FusionNet [56]	0.9040	<b>3.3277</b>	<b>2.4490</b>	0.9807
SFITNet [52]	0.8329	4.5605	3.2343	0.9630
LAGConv [53]	0.8629	4.4474	3.1548	0.9659
JMAN (Ours)	<b>0.9335</b>	4.4633	3.6415	<b>0.9825</b>

The best results are shown in red.

with eight other methods on three publicly available datasets and visualized them in Figs. 4–6. Although all experimental methods yielded satisfactory fusion results on the GaoFen-2 dataset, our method retained better spatial and spectral effects compared to other methods, as depicted in the red box in Fig. 4. For comparison, Fig. 6 displays the pansharpening results on the WV3 dataset. It can be observed that the results generated by BT-H, BDS-D, MSDCNN, and MTF-GLP-FS exhibit significant blurriness and fail to effectively enhance the spatial resolution of MS images. While C-GSA, BDPN, MTF-GLP-HPM-R, and DiCNN produced clear and sharp images but introduced severe spectral distortions. In contrast, the proposed method effectively balanced spectral and spatial information. Qualitative results on the QB dataset are shown in Fig. 5. The majority of the QB dataset comprises densely populated urban areas, which consequently render sharpened images susceptible to spectral distortions. BT-H, BDS-D, MTF-GLP-HPM-R, C-GSA, and MTF-GLP-FS all exhibit pronounced spectral distortions over extensive regions, whereas our method demonstrated the ability to consistently match the spectral characteristics of the ground truth.

2) *Quantitative Evaluation*: For quantitative evaluation, four metrics, i.e., the  $Q2^n$ , SAM, ERGAS, and the SCC were employed to evaluate the fusion performance of the proposed method and competitors. Table I reveals the quantitative results of the proposed method and eight comparison methods on the WV3 dataset. It proves that our method performs best in Q8 and SCC. This demonstrates that our method can effectively enhance the spectral similarity and spatial consistency of fused images. Overall, our method demonstrates excellent performance on the WV3 dataset and can generate high-quality fused images. Table II shows that the proposed method ranks first in Q8, SAM, ERGAS, and SCC on the GF2 dataset. The results exhibiting the highest Q4 and lowest ERGAS signify a higher visual consistency with the ground truth. This demonstrates that our method can effectively enhance the spectral similarity, spatial consistency, and spatial resolution of fused images while also

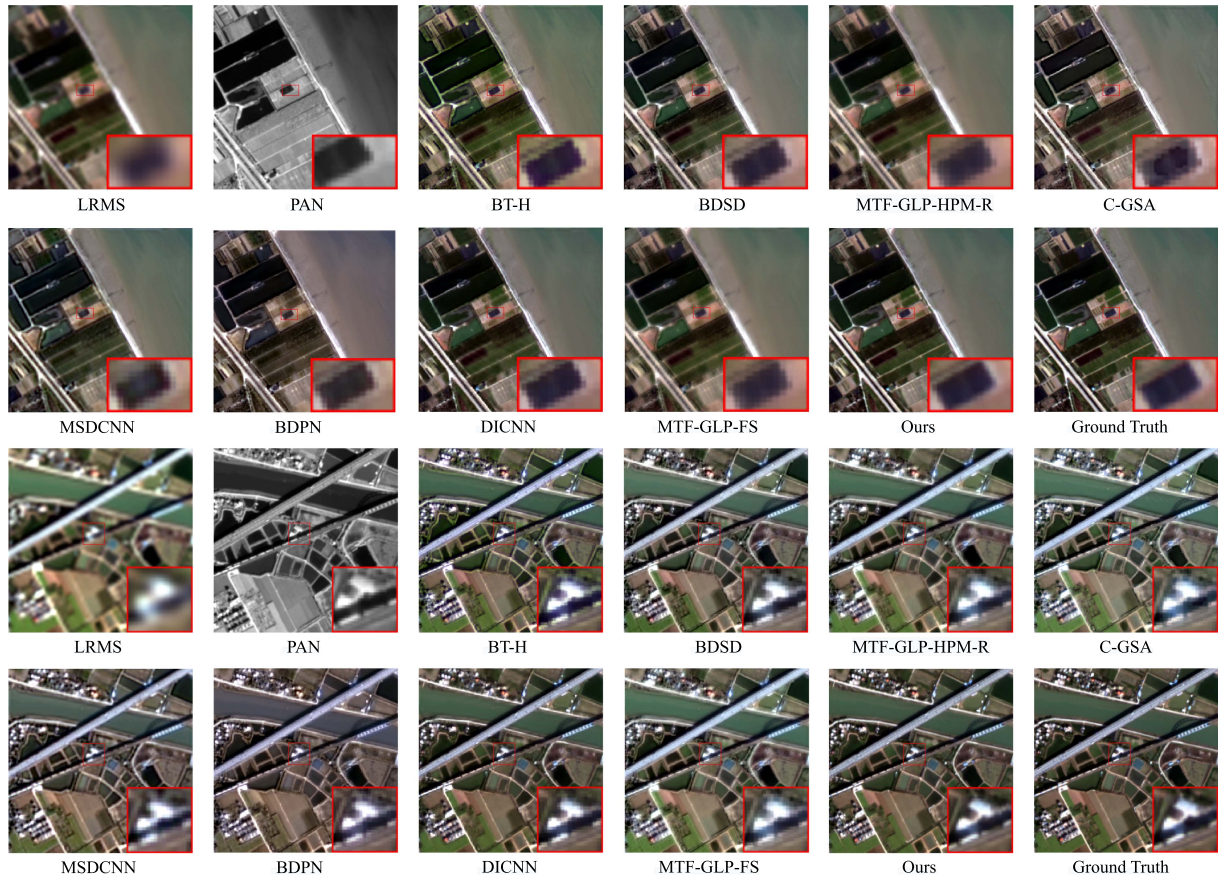


Fig. 4. Pansharpening results obtained by different models on GF2 dataset.

TABLE II  
COMPARE WITH SOTA METHODS ON GF2 DATASET

Method	Q4 $\uparrow$	SAM $\downarrow$	ERGAS $\downarrow$	SCC $\uparrow$
LMS $\uparrow$	0.8089	1.8233	2.3662	0.8643
BT-H [45]	0.9156	1.6531	1.5300	0.9549
BDSD-PC [46]	0.8890	1.6886	1.6693	0.9499
C-GSA [48]	0.9028	1.6818	1.6228	0.9494
SR-D [54]	0.9320	1.4210	1.3614	0.9595
MTF-GLP-HPM-R [47]	0.8969	1.6559	1.5941	0.9454
MTF-GLP-FS [51]	0.8938	1.6616	1.5964	0.9439
TV [55]	0.7185	3.2917	4.0400	0.8723
DRPNN [18]	0.9703	0.9387	0.8514	0.9848
MSDCNN [49]	0.9054	1.7057	1.5341	0.9497
BDPN [22]	0.9024	1.8553	1.6171	0.9535
DiCNN [50]	0.9582	1.0568	1.0817	0.9761
PNN [15]	0.9563	1.0845	1.1090	0.9737
FusionNet [56]	0.9631	0.9850	0.9938	0.9794
SFITNet [52]	0.9460	1.2136	1.2335	0.9675
LAGConv [53]	0.9327	1.3387	1.3501	0.9622
JMAN (Ours)	<b>0.9864</b>	<b>0.7113</b>	<b>0.6299</b>	<b>0.9894</b>

The best results are shown in red.

effectively reducing the spectral distortion and spatial distortion of fused images.

Table III illustrates that the proposed method performs best in the SAM, ERGAS, and SCC on the QB dataset. This demonstrates that the proposed method can effectively reduce the spectral distortion and spatial distortion of fused images while simultaneously enhancing the spatial consistency of fused images.

TABLE III  
COMPARE WITH SOTA METHODS ON QB DATASET

Method	Q4 $\uparrow$	SAM $\downarrow$	ERGAS $\downarrow$	SCC $\uparrow$
LMS $\uparrow$	0.5832	8.4593	11.9074	0.7164
BT-H [45]	0.8308	7.2112	7.4567	0.9148
BDSD-PC [46]	0.8288	8.1025	7.5668	0.9050
C-GSA [48]	0.8344	7.2577	7.4262	0.9117
SR-D [54]	0.8276	7.4305	7.5415	0.9047
MTF-GLP-HPM-R [47]	0.8404	7.7808	9.7803	0.8602
MTF-GLP-FS [51]	0.8335	7.8010	7.4135	0.9019
TV [55]	0.8197	7.5266	7.7492	0.8986
DRPNN [18]	<b>0.9276</b>	4.7546	3.9123	0.9791
MSDCNN [49]	0.8426	7.6073	7.0210	0.9193
BDPN [22]	0.8635	7.1181	6.3893	0.9340
DiCNN [50]	0.9016	5.3732	5.1823	0.9614
PNN [15]	0.9167	5.1703	4.4387	0.9717
FusionNet [56]	0.9217	4.8699	4.2126	0.9749
SFITNet [52]	0.9008	5.5886	5.1281	0.9608
LAGConv [53]	0.8903	5.9136	5.6209	0.9523
JMAN (Ours)	0.9118	<b>2.9913</b>	<b>2.2410</b>	<b>0.9853</b>

The best results are shown in red.

Although our method did not reach the highest value in the Q4 metric, it still exhibits a significant advantage compared to other methods. This indicates that the proposed method can effectively enhance the spectral similarity of fused images. In conclusion, the proposed method demonstrates excellent performance on the QB dataset.

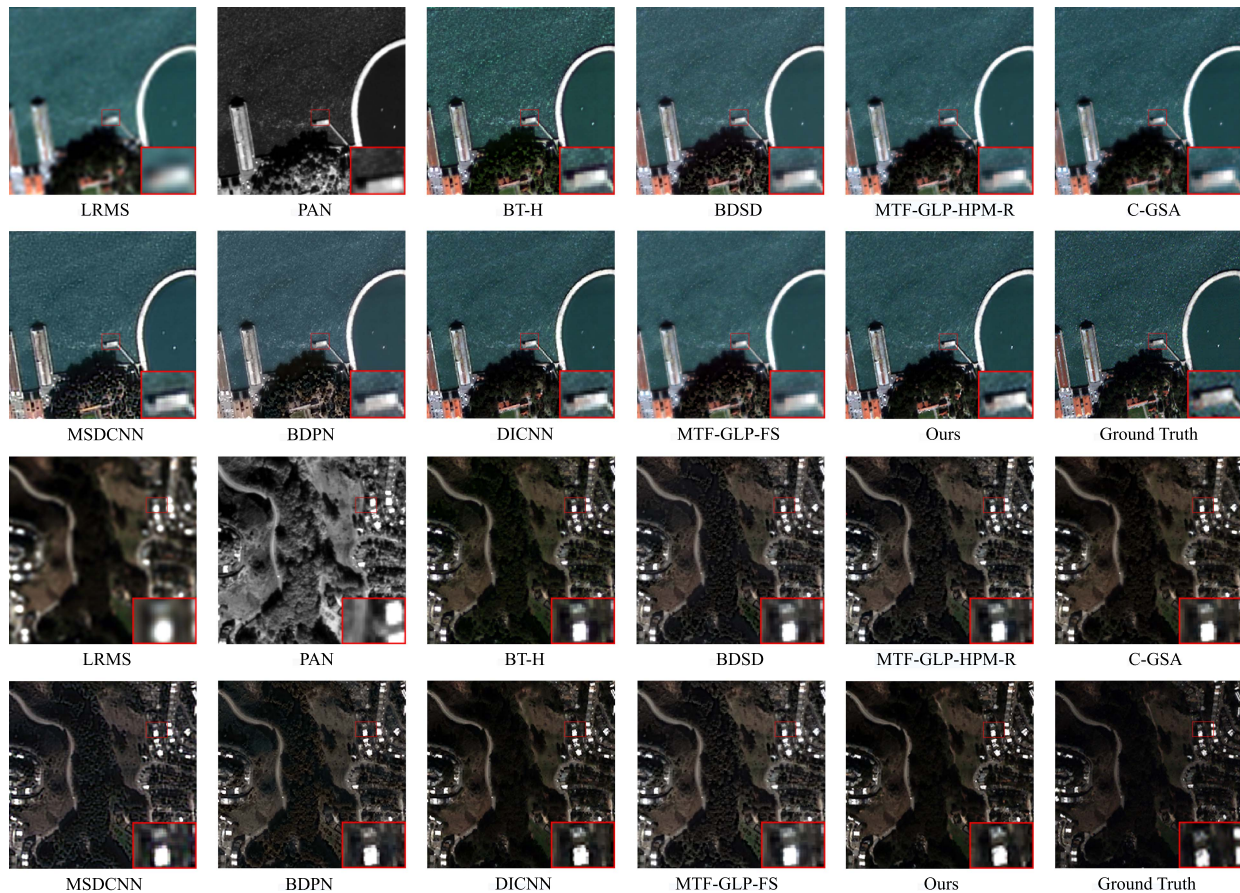


Fig. 5. Pansharpening results obtained by different models on QB dataset.

TABLE IV  
QUANTITATIVE ABLATION RESULTS OF INTRAMODALITY KNOWLEDGE ASSOCIATION AND INTERMODALITY KNOWLEDGE ASSOCIATION ON THE QB DATASET

Intra	Inter	Q4 $\uparrow$	SAM $\downarrow$	ERGAS $\downarrow$	SCC $\uparrow$
$\times$	$\times$	0.9281	4.6012	4.0680	0.9792
$\times$	$\checkmark$	0.9302	4.6044	4.0049	0.9798
$\checkmark$	$\times$	0.9299	4.5511	3.9864	0.9801
$\checkmark$	$\checkmark$	<b>0.9335</b>	<b>4.4633</b>	<b>3.6415</b>	<b>0.9825</b>

The best results are shown in red.

## V. ABLATION STUDY

To investigate the contributions and necessity of different components in our proposed method, we conducted a series of ablation experiments. We examined the impact of intramodality knowledge association, intermodality knowledge association, and multistep recursive training on model performance, respectively.

### A. Multimodality Knowledge Association Ablation Experiment

We designed four different configurations to evaluate the roles of intermodality knowledge association and intramodality knowledge association. Table IV presents the quantitative results on the QB dataset. From the table, it can be seen that when both intermodality knowledge association and intramodality

TABLE V  
QUANTITATIVE ABLATION RESULTS OF MULTISTEP ON THE QB DATASET

Step	Q4 $\uparrow$	SAM $\downarrow$	ERGAS $\downarrow$	SCC $\uparrow$
1	0.9282	4.6462	4.0691	0.9789
2	0.9335	4.4633	3.6415	0.9825
3	0.9328	4.4953	3.6753	0.9822
4	0.9340	4.4574	3.6545	0.9824

knowledge association are used simultaneously, our approach achieves the best results in four metrics, i.e., Q4, SAM, ERGAS, and SCC. This indicates that both of these knowledge association mechanisms are effective and beneficial. Furthermore, we found that intermodality knowledge association has a significant impact on improving the Q4 and SCC metrics, whereas intramodality knowledge association has a significant impact on reducing the SAM and ERGAS metrics, which suggests that the internal modality knowledge association can enhance the complementarity and distinctiveness of features between different modalities, while the cross-modality knowledge association can enhance feature fusion and coordination among different modalities.

### B. Multistep Ablation Experiment

We ablated different numbers of steps to evaluate the impact of multistep recursive training on model performance. Table V shows the quantitative results on the QB dataset. It can be seen



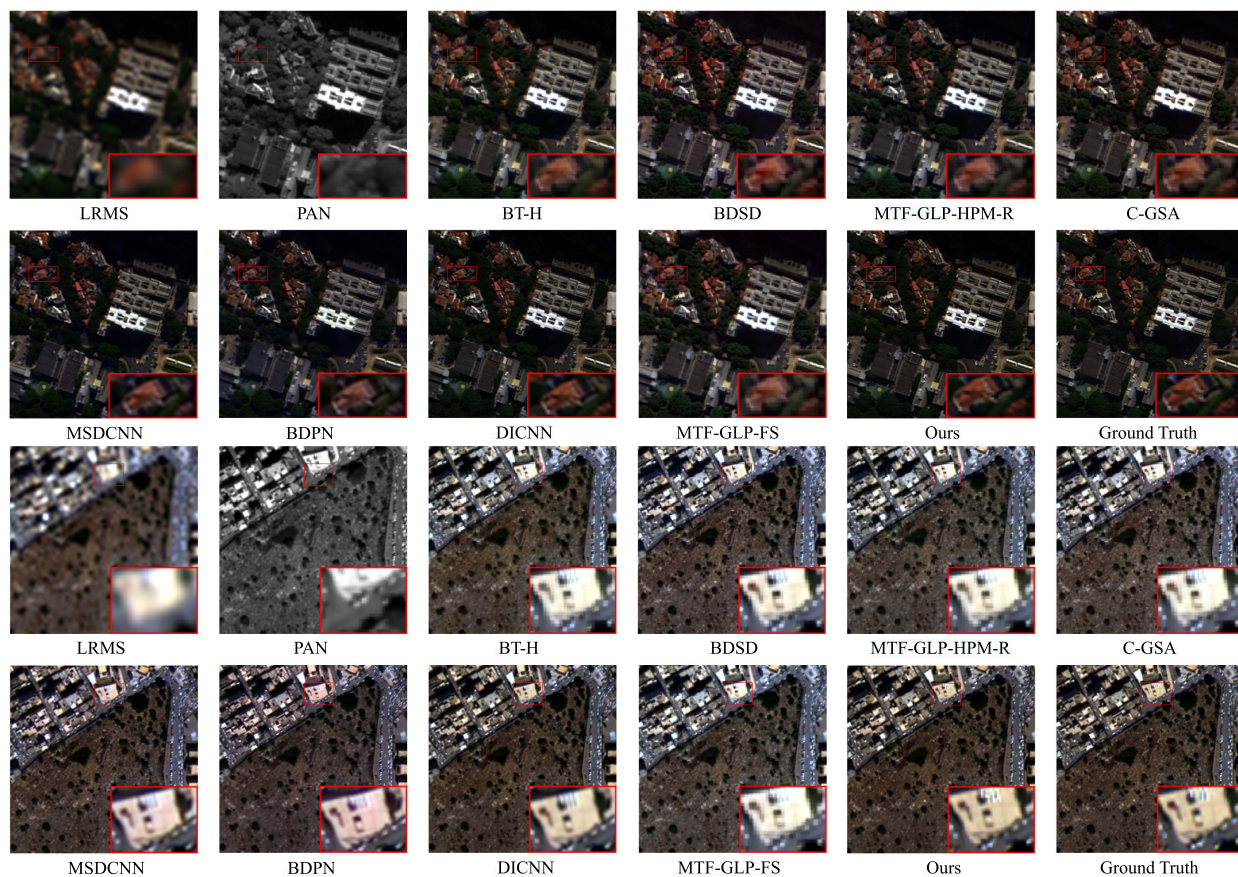


Fig. 6. Pansharpening results obtained by different models on WV3 dataset.

that when the number of steps is set to 2, our method achieves relatively ideal results in four metrics, indicating that a two-step recursive training is sufficient to extract and fuse effective feature information. When the number of steps increases to 3 and 4, there is a significant increase in the computational resources required for training. Therefore, considering both model performance and computational resource requirements, we choose a step number of 2 as the optimal configuration. Therefore, our method requires only a half of the parameters compared with the traditional CNN counterpart.

## VI. CONCLUSION

In this work, we introduced a parameter-efficient pansharpening model inspired by recent advances in recurrent neural networks. We effectively leverage complementary information from different modalities to create a distinctive and informative representation. Specifically, we employed the transformer block to explore the nonlocal global knowledge, and the CNN block for local information. By associating the two complementary knowledge, our method outperformed existing benchmarks, and produced faithful pansharpening multispectral images with enhanced spectral and spatial resolution. Hence, this work addressed a critical challenge in remote sensing systems and offered a practical solution for LMS data.

However, the effectiveness of the proposed method largely depends on high-quality, labeled training data. In the field of remote sensing, acquiring such data is both expensive and time-consuming. This limits the training and generalizability of the model, especially in situations where data is scarce. Moreover, the model may be sensitive to noisy inputs, necessitating complex preprocessing steps to reduce errors and increase accuracy. Therefore, future research will primarily explore how to utilize semisupervised learning, transfer learning, or GANs to reduce the need for large amounts of labeled data or to develop more effective data augmentation techniques to improve the robustness and adaptability of the model.

## REFERENCES

- [1] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, 2021.
- [2] F. D. Javan, F. Samadzadegan, S. Mehravar, A. Toosi, R. Khatami, and A. Stein, "A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 171, pp. 101–117, 2021.
- [3] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.
- [4] E. L. Bullock, C. E. Woodcock, and P. Olofsson, "Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis," *Remote Sens. Environ.*, vol. 238, 2020, Art. no. 110968.

- [5] Y. Gong et al., "Context-aware convolutional neural network for object detection in VHR remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 34–44, Jan. 2020.
- [6] Q. Li, X. Yang, W. Wu, K. Liu, and G. Jeon, "Pansharpening multispectral remote-sensing images with guided filter for monitoring impact of human behavior on environment," *Concurrency Computation: Pract. Experience*, vol. 33, no. 4, 2021, Art. no. e5074.
- [7] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang, "A new look at IHS-like image fusion methods," *Inf. Fusion*, vol. 2, no. 3, pp. 177–186, 2001.
- [8] K. Amolins, Y. Zhang, and P. Dare, "Wavelet based image fusion techniques—an introduction, review and comparison," *ISPRS J. Photogrammetry Remote Sens.*, vol. 62, no. 4, pp. 249–263, 2007.
- [9] M. G.-Audićana, J. L. Saleta, R. G. Catalán, and R. García, "Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1291–1299, Jun. 2004.
- [10] X. Tian, Y. Chen, C. Yang, and J. Ma, "Variational pansharpening by exploiting cartoon-texture similarities," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [11] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, and X. Guo, "SDPNet: A deep network for pan-sharpening with enhanced information representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4120–4134, May 2021.
- [12] K. Sumathi and S. K. KS, "A systematic review of fundus image analysis for diagnosing diabetic retinopathy," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 16s, pp. 167–181, 2024.
- [13] J. Prasad, A. Jain, D. Velho, and K. K. Sendhil Kumar, "Covid vision: An integrated face mask detector and social distancing tracker," *Int. J. Cogn. Comput. Eng.*, vol. 3, pp. 106–113, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666307422000110>
- [14] A. Ramachandran and K. K. Sendhil Kumar, "Tiny criss-cross network for segmenting paddy panicles using aerial images," *Comput. Elect. Eng.*, vol. 108, 2023, Art. no. 108728.
- [15] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.
- [16] 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.
- [17] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, "Remote sensing image fusion with convolutional neural network," *Sens. Imag.*, vol. 17, pp. 1–16, 2016.
- [18] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multi-spectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.
- [19] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–7.
- [20] 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. 2020.
- [21] 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.
- [22] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.
- [23] M. Anbarasi, K. S. Kumar, R. Balamurugan, and Thejasswini, "Disease prediction using hybrid optimization methods based on tuning parameters," in *Proc. 2020 10th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, 2020, pp. 643–648.
- [24] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, 2020.
- [25] Y. Yang, W. Tu, S. Huang, H. Lu, W. Wan, and L. Gan, "Dual-stream convolutional neural network with residual information enhancement for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5402416.
- [26] Z. Shao, Z. Lu, M. Ran, L. Fang, J. Zhou, and Y. Zhang, "Residual encoder-decoder conditional generative adversarial network for pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1573–1577, Sep. 2020.
- [27] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.
- [28] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6877–6886.
- [29] 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. 548–558.
- [30] X. Meng, N. Wang, F. Shao, and S. Li, "Vision transformer for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409011.
- [31] X. Su, J. Li, and Z. Hua, "Transformer-based regression network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5407423.
- [32] W. Zhu, J. Li, Z. An, and Z. Hua, "Mutiscale hybrid attention transformer for remote sensing image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5400416.
- [33] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, Mar. 2021.
- [34] Y. Zhou, S. Chen, Y. Wang, and W. Huan, "Review of research on lightweight convolutional neural networks," in *Proc. 2020 IEEE 5th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, IEEE, 2020, pp. 1713–1720.
- [35] M. Ye et al., "A lightweight model of VGG-16 for remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6916–6922, 2021.
- [36] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
- [37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
- [38] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
- [39] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. PMLR*, 2019, pp. 6105–6114.
- [40] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," 2015, *arXiv:1510.00149*.
- [41] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. 2013 Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.
- [42] H. Sadr, M. M. Pedram, and M. Teshnehlab, "A robust sentiment analysis method based on sequential combination of convolutional and recursive neural networks," *Neural Process. Lett.*, vol. 50, pp. 2745–2761, 2019.
- [43] N. McLaughlin, J. M. D. Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1325–1334.
- [44] S. Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "A fully trainable network with RNN-based pooling," *Neurocomputing*, vol. 338, pp. 72–82, 2019.
- [45] S. Lollu, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, Dec. 2017.
- [46] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019.
- [47] G. Vivone, R. Restaino, and J. Chanussot, "A regression-based high-pass modulation pansharpening approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 984–996, Feb. 2018.
- [48] R. Restaino, M. D. Mura, G. Vivone, and J. Chanussot, "Context-adaptive pansharpening based on image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 753–766, Feb. 2016.
- [49] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.
- [50] 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.
- [51] 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.
- [52] M. Zhou et al., "Spatial-frequency domain information integration for pan-sharpening," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 274–291.
- [53] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1113–1121.
- [54] M. R. Vicinanza, R. Restaino, G. Vivone, M. D. 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.

- [55] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpener algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.
- [56] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpener," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6995–7010, Aug. 2021.



**Qilei Li** (Graduate Student Member, IEEE) received the M.S. degree in signal and information processing from Sichuan University, Chengdu, China, in 2020. He is currently working toward the Ph.D. degree in computer vision, under the supervision of Prof. Shaogang (Sean) Gong, with Queen Mary University of London, London, U.K.

From 2022 to 2024, he was a Machine Learning Scientist with Veritone Inc. His research outcome has been recognized as ESI Highly Cited Paper (Top 1%). He is an evaluator for the ELLIS PhD Program. His

research interest focuses on developing a scalable person search framework for retrieving individuals at different locations and times, as captured by various cameras.

Dr. Li is a Reviewer for numerous journals and conferences.



**Wenhao Song** is working toward the M.S. degree in detection technology and automatic equipment with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China.

His research interests include information fusion, image super-resolution and deep learning.



**Mingliang Gao** (Senior Member, IEEE) received the Ph.D. degree in communication and information systems from Sichuan University, Chengdu, China, in 2013.

He is currently an Associate Professor with the Shandong University of Technology, Zibo, China. He was a visiting Lecturer at the University of British Columbia during 2018 to 2019. He has been the Principal investigator for a variety of research funding, including the National Natural Science Foundation, the China Postdoctoral Foundation, National Key Research Development Project, etc. His research interests include computer vision, machine learning, and intelligent optimal control. He has published over 150 journal/conference papers in IEEE, Springer, Elsevier, and Wiley. He is an IEEE Senior Member.



**Wenzhe Zhai** is working toward the M.S. degree in detection technology and automatic equipment with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China.

His research interests include smart city systems, information fusion, crowd analysis, and deep learning.

Mr. Zhai is also a Reviewer for numerous journals, including *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, *Neurocomputing*, *Engineering Applications of Artificial Intelligence*, and *Multimedia Systems*.



**Jianhao Sun** is working toward the M.S. degree in control engineering with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China.

His research interests include image super-resolution and deep learning.



**Gwanggil Jeon** (Senior Member, IEEE) received the B.S., M.S., and Ph.D. (summa cum laude) degrees in electronics and computer engineering from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2003, 2005, and 2008, respectively.

From 2009 to 2011, he was a Postdoctoral Fellow with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada. From September 2011 to February 2012, he was an Assistant Professor with the Graduate School of Science and Technology, Niigata University, Niigata, Japan. From December 2014 to February 2015 and June 2015 to July 2015, he was a Visiting Scholar with Centre de Mathématiques et Leurs Applications, École Normale Supérieure Paris-Saclay, France. From 2019 to 2020, he was a Prestigious Visiting Professor with Dipartimento di Informatica, Università degli Studi di Milano Statale, Italy. He was a Visiting Professor with Sichuan University, China, Universitat Pompeu Fabra, Barcelona, Spain, Xinjiang University, China, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, and University of Burgundy, Dijon, France. He is currently a Full Professor with Incheon National University, Incheon, Korea.

Dr. Jeon was a recipient of the IEEE Chester Sall Award in 2007, the ETRI Journal Paper Award in 2008, and Industry-Academic Merit Award by the Ministry of SMEs and Startups of Korea Minister in 2020. He is an Associate Editor of *Sustainable Cities and Society*, *IEEEACCESS*, *Real-Time Image Processing*, *Journal of System Architecture*, and *MDPI Remote Sensing*.