Applied Intelligence (2024) 54:8041–8058
https://doi.org/10.1007/s10489-024-05580-1

Dual-branch and triple-attention network for pan-sharpening

Wenhao Song[1] · Mingliang Gao[1] · Abdellah Chehri[2] · Wenzhe Zhai[1] · Qilei Li[3] · Gwanggil Jeon[1,4]

Accepted: 31 May 2024 / Published online: 19 June 2024

Abstract
Pan-sharpening is a technique used to generate high-resolution multi-spectral (HRMS) images by merging high-resolution panchromatic (PAN) images with low-resolution multi-spectral (LRMS) images. Many existing methods face challenges in effectively balancing the trade-off between spectral and spatial information, leading to spectral and spatial structural distortion. In order to effectively tackle these issues, we propose a dual-branch and triple attention (DBTA) network. The proposed DBTA network consists of two essential modules: the Channel-spatial Attention (CSA) module and the Spectral Attention (SPA) module. The CSA module effectively captures the spatial structural information of the images by jointly using spatial and channel attention units. Meanwhile, the SPA module improves the expressive capacity of spectral information by dynamically adjusting channel weights. These two modules work in synergy to achieve comprehensive extraction and fusion of spectral and spatial information, thus resulting in more accurate and clearer reconstructed images. Extensive experiments have been conducted on various satellite datasets to evaluate the performance of the proposed DBTA method outperforms the state-of-the-art competitors in both qualitative and quantitative evaluations.

Keywords Deep learning · Image fusion · Pan-sharpening · Remote sensing · Attention mechanism

1 Introduction

The panchromatic (PAN) image is a remote sensing image characterized by a single color band. This type of image is known for its high spatial resolution but low spectral resolution [55]. Meanwhile, multi-spectral (MS) remote sensing images contain multiple bands, typically red, green, blue, near-infrared, and other bands. Each spectral band captures specific information about ground targets, reflecting their distinct features and properties. In contrast to PAN images, MS images exhibit superior spectral resolution but lower spatial resolution [61]. The pan-sharpening technique combines the spatial details from PAN images with the spectral characteristics of LRMS images to generate a high-resolution (HR) MS image. Pan-sharpening technique has been proven to be highly beneficial for various applications, *e.g.,* object detection [15], land surveying [59], and environmental monitoring [3].

Pan-sharpening algorithms can be broadly classified into two categories: traditional methods and deep learning methods. The traditional approaches can be further categorized into three groups, namely component substitution (CS)-based methods [12], multi-resolution analysis (MRA)-based methods [34], and variational optimization (VO)-based methods [33]. Considering the intricate nature of ground objects and the wide array of spectral characteristics captured by various sensors, traditional methods face difficulties in establishing a meaningful connection between the input image and the target HRMS image [20].

In recent years, deep learning has been widely used in pansharpening, benefiting from the powerful feature representation capabilities of neural networks [22]. Deep learning-based Pan-sharpening methods can be categorized broadly into two groups, *i.e.,* convolutional neural network (CNN)-based methods and generative adversarial network (GAN)-based methods. Existing CNN-based methods typ-

✉ Mingliang Gao
mlgao@sdut.edu.cn

1 School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China

2 Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, Canada

3 School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom

4 Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea

ically generate HRMS images by acquiring the mapping function among the MS image, PAN image, and HRMS image [60]. The GAN-based methods typically consist of generator and discriminator networks, wherein the two networks oppose each other. Using iterative learning, the generator network progressively acquires the ability to generate HRMS images of increased realism, eliminating the dependence on ground truth [37].

The PAN image encompasses substantial spatial texture information, while the MS image offers ample spectral information. Generating high-quality HRMS images necessitates fully utilizing data acquired from both sources. Nevertheless, the majority of current pan-sharpening methods employ a direct feature extraction and fusion approach from the cascaded MS and PAN images [54]. Disregarding the redundancy information between the MS and PAN images, as well as the distinct characteristics of each band in the MS image, these methods would significantly impede the ability to enhance the performance of these methods further. Existing pan-sharpening methods struggle to balance the preservation of spectral and spatial information, and thus, it results in spectral distortion and spatial structural distortion in the fused images.

To address these problems, we propose the dual-branch and triple-attention (DBTA) network for pan-sharpening. The architecture of the proposed DBTA is shown in Fig. 1. Specifically, a dual-branch scheme is built to extract spatial and spectral information of the input images, respectively. The spatial branch incorporates the channel-spatial attention module, enabling improved preservation of information such as spatial structure in the image. To maintain the spectral information of the MS images in the spectral feature branch, we propose the incorporation of the spectral attention module. The comparative experiments demonstrate the superiority of the proposed DBTA over the current state-of-the-art pan-sharpening method. The contributions of this work can be summarized as follows:

1. A dual-branch and triple-attention (DBTA) network is built for pan-sharpening to address the problem of spectral distortion and spatial structure degradation.
2. A channel-spatial attention (CSA) module is designed to preserve the spatial structural information in the images. Meanwhile, a spectral attention (SPA) module is built to preserve spectral information and avoid spectral distortion.
3. Comprehensive experimental results on three datasets, *i.e.,* WorldViewIII, QuickBird, and GaoFen2 datasets demonstrate the effectiveness of the DBTA. Meanwhile, the ablation experiments validate the effectiveness of the CSA and SPA modules.

The rest of this paper is structured as follows. Section 2 reviews the related work on pan-sharpening methods and attention mechanisms. Section 3 describes the proposed method in detail. Section 4 presents the experimental setup and results. Section 5 concludes this paper and points out future work.

2 Related work

2.1 Traditional pan-sharpening methods

In the last decade, a variety of pan-sharpening methods have been established, and striking results have been yielded. There are three main categories of traditional pan-sharpening methods, *i.e.,* component substitution (CS)-based methods [12], multi-resolution analysis (MRA)-based meth-

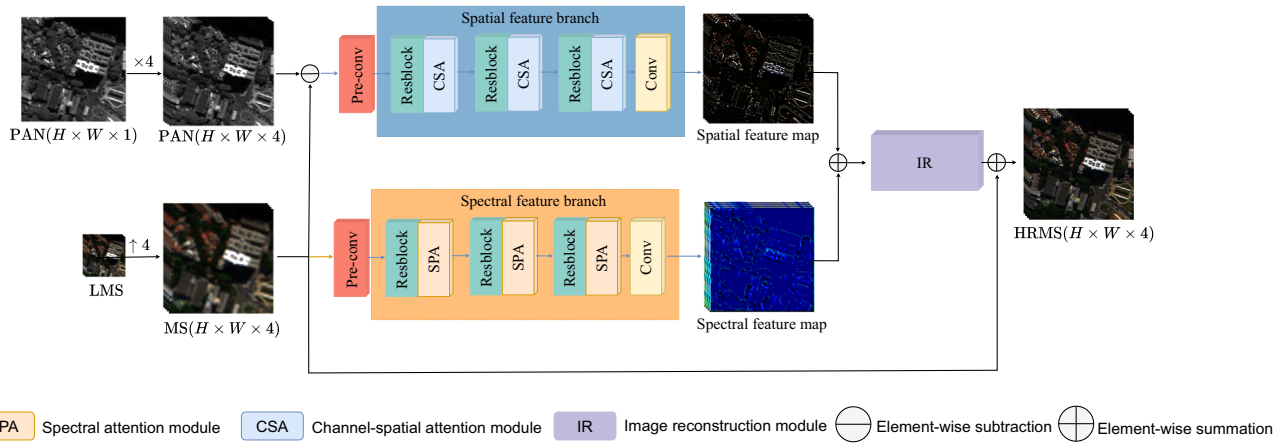

Fig. 1 Architecture of the proposed DBTA for pan-sharpening

ods [34], and variational optimization (VO)-based methods [33]. The CS-based methods decompose the MS and PAN images into multiple components. Subsequently, a specific component from the MS image is partially or fully substituted with a histogram-matched PAN image to generate an HRMS image. The representative CS-based methods include the Gram-Schmidt (GS) method [17], the principal component analysis (PCA) method [40, 43], the intensity-hue-saturation (IHS) method [26, 46], the Brovey transformation methods [13]. Nevertheless, since the MS component replaced by the PAN image also contains certain spectral information, the fused image obtained based on the CS-based method usually suffers from spectral distortion [49].

The MRA-based methods exhibit the capability to preserve favorable spectral features. The fundamental concept underlying MRA-based methods is to leverage multi-resolution decomposition technology to incorporate high-frequency details obtained from the PAN images into the up-sampled MS images. The decomposition methods typically used in the MRA-based methods include the Laplacian pyramid transform [23], the Wavelet transform [51], the Curvelet transform [36], and the Support tensor transform [50]. In most instances, the MRA-based methodology produces satisfactory spatial-spectral unified fidelity [65]. Nonetheless, the MRA-based approaches are contingent on the efficacy of multi-resolution techniques, which can result in localized spatial artifacts in the fused images [66].

The VO-based methods regard pan-sharpening as an ill-posed problem [4]. It comprises two indispensable components, *i.e.,* the formulation of the energy functional and the derivation of the optimization solution [33]. These methods always combine the energy function with appropriate regularization terms such as sparse regularization [44], low-rank recovery model [42], and variational model [11]. Although VO-based methods attempt to strike a balance between spatial enhancement and spectral preservation, their effectiveness is constrained by the utilization of shallow nonlinear expressions in their models.

2.2 Deep learning-based pan-sharpening methods

Deep learning-based methods have been rapidly adopted in remote sensing applications due to their excellent nonlinear map learning and feature extraction capabilities [22, 55]. These methods leverage the robust feature extraction capabilities of convolutional networks, which often leads to less spectral distortion and favorable fusion performance. The pan-sharpening methods based on deep learning can be broadly classified into CNN-based methods and GAN-based methods [14].

Huang et al. [19] pioneered the use of deep learning techniques for pan-sharpening. Inspired by the super-resolution convolutional neural networks (SRCNN) [10], the pan-sharpening network (PNN) [32] was proposed. However, the performance and convergence speed of PNN is limited because it only employs three convolutional layers and lacks skip connections to expedite convergence. In addition, Liu et al. [30] leveraged an encoder-decoder network to execute feature extraction, fusion, and reconstruction procedures for HRMS. Xiong et al. [52] introduced an unsupervised attention Pan-sharpening net (UAP-Net) without the need for annotated training data. Lee et al. [25] discovered that the alignment of the same object in MS and PAN images is not consistently accurate. To tackle this issue, they introduced SIPSA-Net by incorporating a feature alignment module. Su et al. [41] proposed a Transformer-Based Regression Network (DR-NET) for pan-sharpening. This approach utilizes the Transformer to construct an end-to-end network architecture for generating high-quality fused images. Ciotola et al. [7] proposed a deep learning-based full-resolution training framework for pan-sharpening. This framework is highly versatile and can be applied to deep-learning pan-sharpening models. The training process is conducted in the high-resolution domain and relies solely on the original data.

The GAN-based methods commonly employ unsupervised learning strategies to explore the underlying features of a network through the iterative interaction between the generator and discriminator [29, 31]. Liu et al. [29] introduced PSGAN, the pioneering GAN-based approach, to tackle the pan-sharpening problem. The PSGAN consists of a generator that combines the LRMS and PAN images and a discriminator to minimize the discrepancy between the fused image and the ground truth. In addition, Ma et al. [31] presented an unsupervised approach for pan-sharpening that can avoid the resolution loss caused by degradation simulation in high-resolution image fusion. Qu et al. [38] presented an unsupervised learning method for pan-sharpening by incorporating a self-attention module. To tackle the challenge of limited the resolution of the training dataset, Xu et al. [53] introduced an unsupervised pansharpening generative adversarial network model termed Unsupervised Pansharpening Based on Spectral and Spatial Loss Constrained Generative Adversarial Network (UPanGAN). This model undergoes direct training using the original panchromatic and multispectral images. In contrast to models trained on downsampled data, the UPan-GAN is well-suited for enhancing the spatial and spectral richness of original full-resolution images. The GAN-based methods exhibit a remarkable pan-sharpening effect, particularly for real data, typically of superior quality.

2.3 Attention mechanism

The attention mechanism is based on human visual attention. Attention mechanisms have wide applications and research in fields, for example, natural language processing, computer vision, and speech recognition [6, 16, 58]. The fundamen-

tal idea behind attention mechanisms is that given an input sequence and an output sequence, a model learns a weight distribution to determine which elements from the input sequence should be attended to when generating each output element [5]. Therefore, the model can dynamically adjust the focus of attention based on different tasks and contexts, and extract more useful information.

Zhang et al. [62] proposed a method based on the Residual Channel Attention Network (RCAN) by introducing the Residual in Residual (RIR) structure and Channel Attention (CA) mechanism within the residual channels. Tang et al. [45] proposed the Attention Consistent Network (ACNet) by introducing a dual-branch structure, different attention techniques, and an attention consistency module. Li et al. [27] introduced the Multi-Scale Channel Attention Residual Network (MSCARN) by incorporating techniques including multi-scale feature extraction, channel attention mechanism, and residual learning. Liang et al. [28] proposed the parallel multi-scale attention constraint network (PMACNet) for the fusion of remote sensing images by utilizing the mechanisms of transformers and employing a parallel convolutional neural network structure to extract regions of interest and residual information from low-resolution multi-spectral images and high-resolution panchromatic images.

3 Methodology

3.1 Overview

The architecture of the proposed DBTA network for pan-sharpening is shown in Fig. 1. It consists of three primary components, *i.e.*, spatial feature branch, spectral feature branch, and image reconstruction module.

Specifically, given an LRMS image $M \in R^{C\times\frac{H}{4}\times\frac{W}{4}}$ and the PAN image $P \in R^{1\times H\times W}$, the LRMS image is up-sampled to align with the resolution of the PAN image. Then, the PAN image is duplicated across the channels to match the number of channels in the MS image. The registered MS and PAN images are inputted into the proposed model.

For the spatial feature branch, the MS and PAN images are subtracted to extract the spatial and potential spectral features [8]. The spatial feature branch employs a 3×3 convolution layer to transform the image after the subtraction of elements into feature space with modality-specific features, denoted as P_0. After that, the spatial feature map P_0 undergoes feature extraction and detail learning through the multiple residual blocks and channel-spatial attention modules (CSA). The spatial branch focuses on the spatial position and content of the features of interest. This is because both the spatial location and content of features are equally important. For example, in an image containing cars, the spatial position and content of the cars are equally important because they can help the network distinguish between different cars.

For the spectral feature branch, the up-sampled MS image is employed as the input and then undergoes the transformation into the feature space using a 3×3 convolutional layer, denoted as M_0. Subsequently, the ultimate spectral feature map is obtained from M_0 by utilizing a feature extractor consisting of numerous residual blocks and spectral attention modules (SPA).

We employ the residual channel attention block [62] as the image reconstruction module (IR) to mitigate the impact of noise and artifacts. The architecture of the image reconstruction module is shown in Fig. 2. We transmit the feature F_f, which combines spatial and spectral features, to the image reconstruction module to achieve effective feature fusion and generate the reconstructed image I_f. Finally, the generation of HR-MS images is accomplished by integrating the up-sampled MS images into the transformed representation through a skip connection.

By simultaneously utilizing the channel-spatial attention module and the spectral attention module in two distinct branches, the DBTA network can learn how to emphasize or suppress the most informative features in both spatial and spectral dimensions selectively, thereby enhancing the performance of feature extraction in spatial and spectral domains. The residual blocks [21] are employed in both branches to extract image features. Figure 3 shows that the residual block consists of two 3×3 convolutional layers and

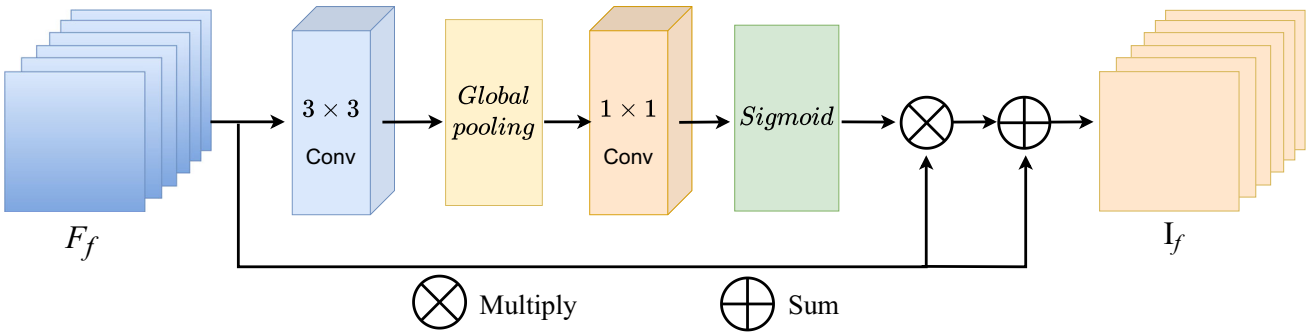

Fig. 2 Architecture of the image reconstruction module

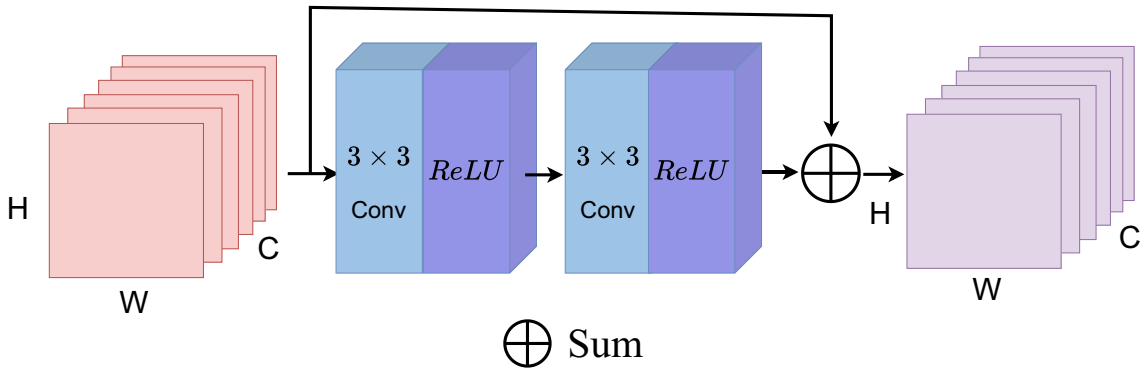

Fig. 3 Architecture of the residual block

a skip connection. The skip connection adds the input feature map to the output of the convolutional layers to form the residual output. After each convolutional layer, we apply the ReLU activation function for non-linear transformation.

3.2 Channel-spatial attention module

The architecture of the channel-spatial attention module is shown in Fig. 4. The channel-spatial attention module composes a channel attention (CA) unit and a spatial attention (SA) unit. The channel attention unit adjusts the weights of different channel features to emphasize the important channels containing object properties, colors, textures, and so on. The spatial attention unit is to modify the feature map weights in various spatial locations to emphasize crucial feature regions, such as the shape and structure of objects.

3.2.1 Channel attention unit

To effectively extract texture and structural features, we adopt the CA unit that learns the importance weights of each channel automatically. The channel attention unit is a mechanism that can adaptively recalibrate the feature maps of each channel in the CNN. It can selectively emphasize or suppress the feature maps of each channel and enable the network to focus on the most informative features. The architecture of the channel attention unit is presented in Fig. 4. The computation of the channel attention (CA) unit is defined as follows,

$$\mathrm{F_c} = \sigma\left(\mathrm{C_1}(\mathrm{MaxPool}(F) \oplus \mathrm{AvgPool}(F))\right), \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function. $\mathrm{C_1}$ is a 3×3 kernel convolution operator. The functions MaxPool$(\cdot)$ and AvgPool$(\cdot)$ denote the max pooling and average pooling operations along the channel dimensions, respectively. The symbol $\oplus$ represents the elemental addition.

3.2.2 Spatial attention unit

The utilization of a channel attention unit enables the model to concentrate on image channels to enhance the clarity and sharpness of the resultant image. However, relying solely on channel attention might overlook the unique characteris-

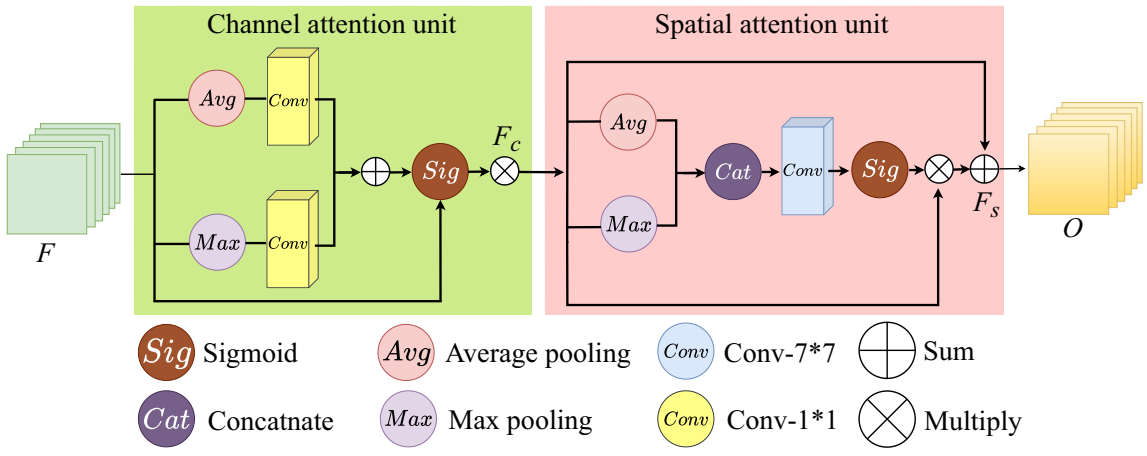

Fig. 4 Architecture of the channel-spatial attention module

tics present in various regions of the image, thereby limiting the comprehensive improvement of the image quality. To address this issue, the spatial attention (SA) unit is integrated following the channel attention unit. Figure 4 illustrates the framework of the SA unit. It is formulated as follows.

$$F_s = \sigma\left\{C_7(\text{Cat}\left[\text{MaxPool}\left(F_c\right); \text{AvgPool}\left(F_c\right)\right])\right\} \otimes F_c, \quad (2)$$

where σ represents the sigmoid function, and C_1 denotes a convolution operation with the filter size of 7×7. The symbol $\otimes$ denotes the element-wise multiplication. MaxPool$(\cdot)$ and AvgPool$(\cdot)$ are the max pooling and average pooling functions in channel dimensions, respectively.

3.3 Spectral attention module

The proposed spectral attention module enhances the capability of the method to extract spectral features. This is accomplished by recalibrating the feature maps by using global context guidance, thereby emphasizing the most salient features. In contrast to the channel attention unit, the spectral attention module is specifically designed for the spectral bands found in multi-spectral images. It can dynamically recalibrate the spectral bands within multi-spectral images, thereby selectively enhancing or suppressing spectral bands. Therefore, the spectral attention module helps the network to focus on the most informative spectral bands. Figure 5 illustrates the architecture of the spectral feature attention.

Assuming an input tensor $X \in \mathbb{R}^{H \times W \times C}$, the global context vector is calculated through global pooling $G \in \mathbb{R}^{1 \times 1 \times C}$,

$$G(:,:,k) = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i, j, k), \quad (3)$$

where k is the index of the channel dimension. To obtain the weight matrix of the spectral information, we implement two fully connected layers to conduct a nonlinear transformation of $G \in \mathbb{R}^{1 \times 1 \times C}$. Then, the sigmoid gating function is utilized to compute the vector of channel scaling factors $S \in \mathbb{R}^{1 \times 1 \times C}$,

$$S = \sigma\left(W_2 \delta\left(W_1 G\right)\right), \quad (4)$$

where δ represents the ReLU function [35]. $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. To balance the computational cost and model capacity, a reduction ratio r is introduced. To accommodate the proposed model, the value of r is set to 4. The output $Y(i, j, k)$ of the spectral attention module is defined as follows,

$$Y(i, j, k) = S(:,:,k) \cdot X(i, j, k). \quad (5)$$

3.4 Loss function

The mean squared error (MSE) is adopted to quantify the deviation between the fusion result and the ground truth [39]. It is formulated as,

$$\mathcal{L} = \sum_{i=1}^{K} \left\| \mathbf{H}_i - \mathbf{H}_{gt,i} \right\|_2, \quad (6)$$

where K represents the number of training data, $\mathbf{H}_i$ represents the output high-resolution MS image, $\| \cdot \|_2$ is l_2 norm and $\mathbf{H}_{gt,i}$ represents the corresponding ground truth, respectively.

4 Experimental results and discussion

4.1 Dataset

To demonstrate the effectiveness of the proposed method, the comparison experiments are conducted on three publicly available datasets,*i.e.*, GaoFen2 (GF2), QuickBird (QB), and WorldView III (WV3) datasets that were introduced by Deng et al. [9]. For these three datasets, the MS images

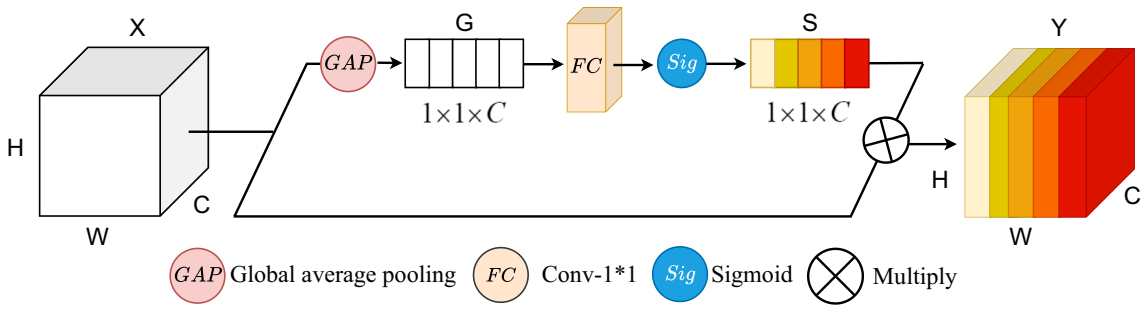

Fig. 5 Architecture of the spectral attention module

have spatial resolutions of 4 m, 2.44 m, and 1.2 m, respectively, whereas the PAN images have corresponding spatial resolutions of about 1 m, 0.61 m, and 0.3 m. The original spatial resolution ratio between the MS and PAN images in these datasets is 4. Two types of test datasets are employed, *i.e.,* reduced resolution test dataset following the Wald protocol [47], and full resolution test datasets at the original scale. A total of 40 image pairs are used for the test.

To augment the training dataset, the original image is initially divided into several image blocks. Then, down-sampling is applied to generate low-resolution LRPAN and LRMS images. The MS image before down-sampling serves as the ground truth. Finally, we obtain 22010 image patches of the GF2 dataset, 19044 image patches of the QB dataset, and 10794 image patches of the WV3 dataset. We divided the datasets of three satellites into a training set comprising 90% of the data and a validation set comprising 10% of the data for each satellite, respectively. The size of LRMS images utilized for training is 16×16, whereas the size of the LRPAN and original MS images is 64×64. Table 1 illustrates the specifics of the dataset employed in the experiment.

4.2 Training details

The training and test phases are performed utilizing the PyTorch framework on an NVIDIA RTX3080Ti GPU. For the training phase, the Adam optimizer is utilized for optimization. It consists of 1500 epochs of iterations with a batch size of 32. The learning rate is initialized to 3×10^{-4} and set as a fixed value during training. Since the real MS and PAN images are mismatched, we constructed the datasets for training and testing using the Wald protocol [47].

4.3 Qualitative evaluation

Qualitative evaluation refers to the level of satisfaction that humans have towards an image. From a human perceptual standpoint, it encompasses luminance, contrast, chromaticity, and authenticity indicators. Eight state-of-the-art competitors, *i.e.*, PNN [32], DICNN [18], MSDCNN [56], BDPN [63], DRPNN [64], SFITNet [67], LAGConv [24] and FusionNet [8] are employed for comparison.

4.3.1 Qualitative evaluation on the WorldViewIII dataset

Two types of tests, namely reduced-resolution test and full-resolution test are carried out on the WorldViewIII. The reduced-resolution results example contains numerous forests and dense houses, which can easily lead to issues such as spectral distortion and spatial structure distortion. The qualitative results on the WorldViewIII dataset are presented in Fig. 6. The comparison results prove that MSDCNN [56], BDPN [63], PNN [32], SFITNet [67], LAGConv [24] and DRPNN [64] exhibite significant distortion across extensive forested regions and buildings. Meanwhile, the DICNN [18] and FusionNet [8] methods fail to achieve a satisfactory balance between spectral and spatial information and thus result in significant spatial distortions in the output. On the contrary, the proposed DBTA model demonstrates satisfactory performance in terms of spectral distribution and spatial structure compared to alternative approaches.

The full-resolution results example contains densely houses and vehicles, where densely residential areas can lead to spectral distortion issues while moving vehicles can easily cause spatial artifacts. Figure 7 presents the results obtained from the full-resolution dataset. The images produced by MSDCNN [56], BDPN [63], and DRPNN [64], exhibit distortions in both spectral and spatial attributes. Due to the insufficient integration of spectral information by DICNN [18] and PNN [32], the generated images exhibit significant spectral distortions. By comparison, the results of the proposed DBTA are clear, and there are no blur effects. Although SFITNet [67], LAGConv [24], and FusionNet [8] show no visible spectral distortion, they have noticeable spatial artifacts compared to PAN images. The comparative evaluation illustrates the superior performance of the proposed method compared with other competitors.

Table 1 Details of the datasets used for evaluation

Size	Phase	WorldViewIII	QuickBird	GaoFen2
Patches	Training	10794	19044	22010
	Reduced_testing	20	20	20
	Full_testing	20	20	20
MS_size	Training	16×16	16×16	16×16
	Reduced_testing	64×64	64×64	64×64
	Full_testing	128×128	128×128	128×128
PAN_size	Training	64×64	64×64	64×64
	Reduced_testing	256×256	256×256	256×256
	Full_testing	512×512	512×512	512×512

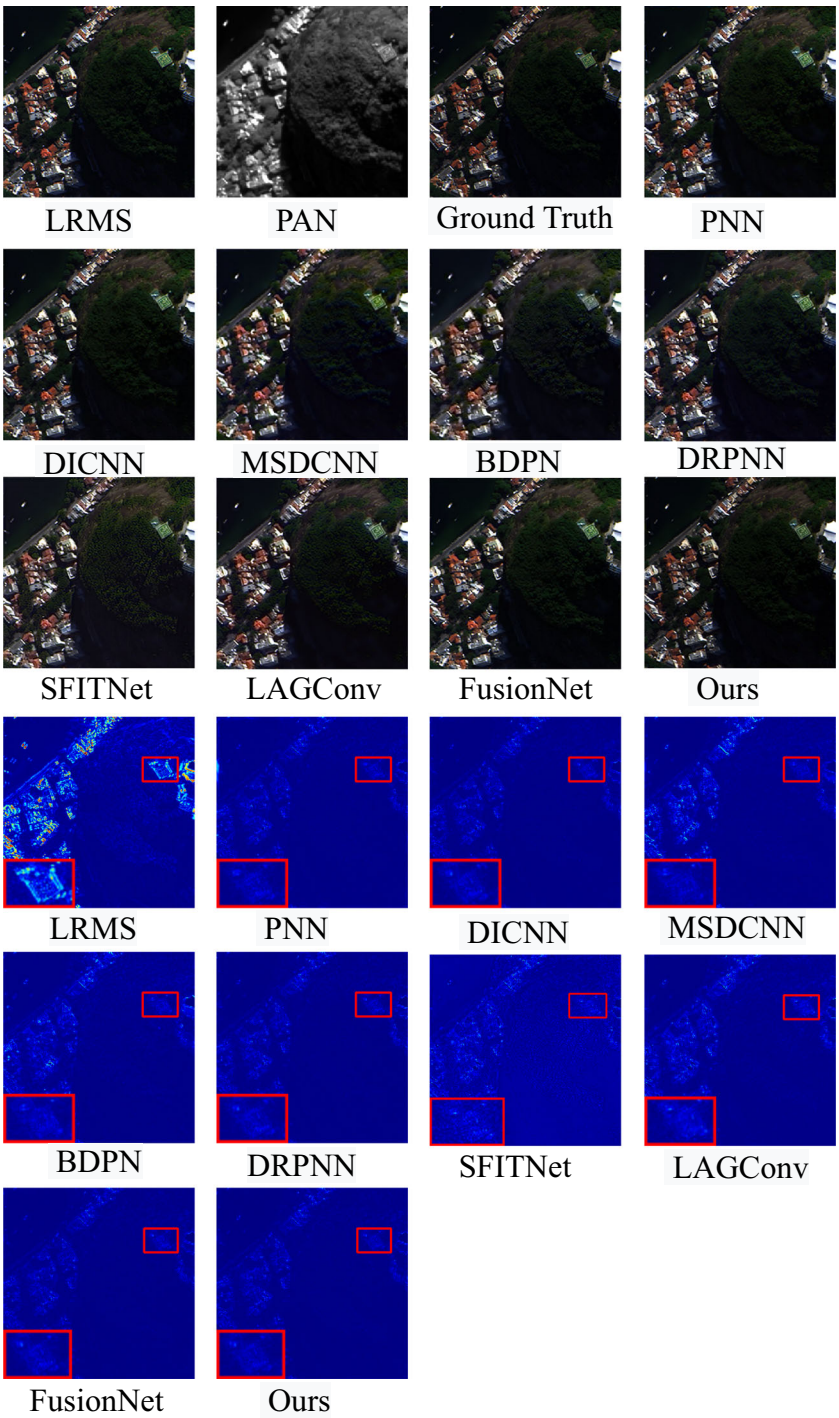

Fig. 6 Qualitative evaluation of different methods for pan-sharpening under reduced resolution test conditions on the WorldViewIII dataset. The images presented in the last three rows depict the residuals between different results and the ground truth

4.3.2 Qualitative evaluation on the QuickBird dataset

Figure 8 illustrates the results of the low-resolution testing on the QuickBird dataset. It is evident that the spatial structure in the proposed DBTA results closely resembles the ground truth. For example, in the reference image, the closely grouped boats positioned in the lower right corner maintain their spatial structure. The proposed method, FusionNet [8] and DRPNN [64] retain this feature, whereas other approaches display artifacts in the image. Furthermore, the proposed method can effectively retain spectral distribution information, such as the sea surface and road spectral distribution. However, DRPNN [64] suffers from spectral distortion in both the sea surface and road areas. Due to the influence of different channel bands, there are differences in some spatial details between PAN images and ground truth.

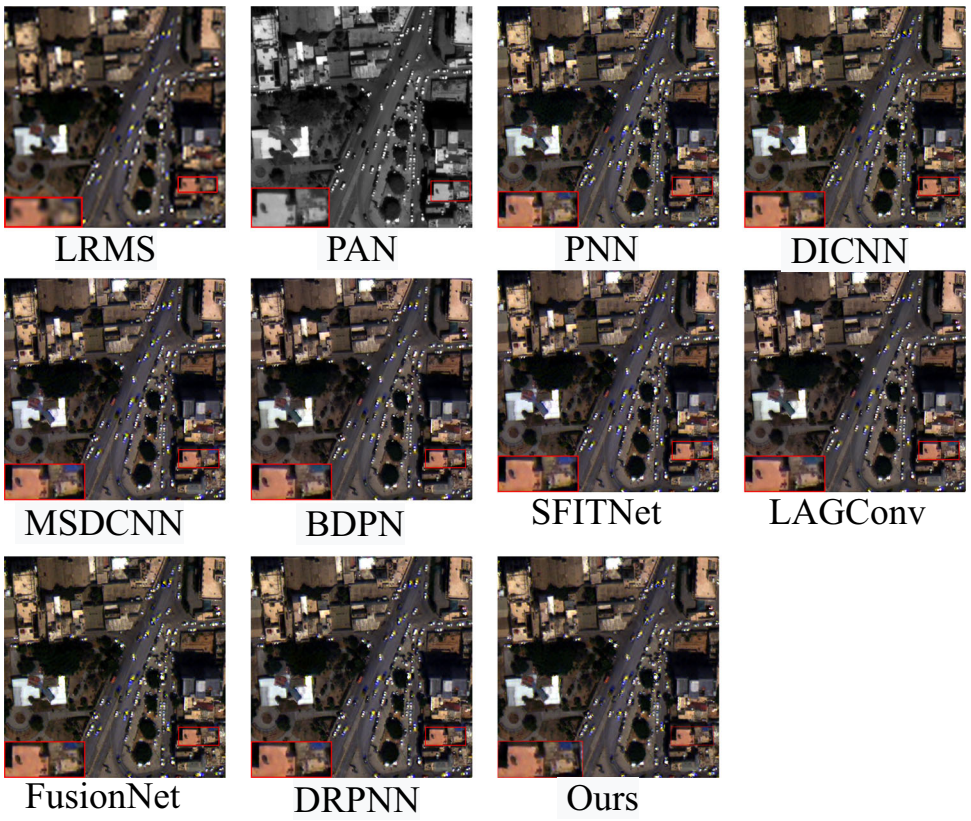

Fig. 7 Qualitative evaluations of different pan-sharpening approaches for full resolution test on the WorldViewIII dataset

Therefore, despite FusionNet [8] presenting details consistent with PAN images, there is still a significant gap compared to the reference image. The results of residuals in Fig. 8 illustrate the higher consistency between the proposed method and the ground truth than other methods.

The qualitative results of different pan-sharpening approaches for full resolution test on the QuickBird dataset are depicted in Fig. 9. It shows that the spectral distributions of DRPNN [64], PNN [32], SFITNet [67], and the proposed DBTA model closely resemble those of the LRMS images. Meanwhile, the DBTA model outperforms the DRPNN [64], PNN [32], and SFITNet [67]. For example, the proposed DBTA model yields clearer representations of dense ships and roads in the results.

4.3.3 Qualitative evaluation on the GaoFen2 dataset

The qualitative results on GaoFen2 dataset with reduced resolution are presented in Fig. 10. This example is an urban image with densely populated houses. It can be observed that the MSDCNN [56], SFITNet [67], LAGConv [24] and BDPN [63] are all suffered from significant spectral distortion. For instance, the originally red roof appears relatively dim in all these methods, tending to blend with the colours of the surrounding roads. Although the DRPNN [64], DICNN [18], FusionNet [8], and PNN [32] maintain spectral quality in the fused images, it falls short in preserving clear spatial information, particularly the structural details of buildings in densely populated areas. The proposed DBTA method effectively preserves spatial structural details, even in densely populated residential regions.

The qualitative results on GaoFen2 dataset with full resolution are presented in Figure 11. It shows that the PNN [32], DRPNN [64], and the proposed DBTA model can preserve spectral information. For example, the spectral distribution of the roof, vehicles, and land closely resembles that of the LRMS image. However, in terms of the preserving of spatial structural information, the proposed DBTA network demonstrates superior performance compared to PNN [32] and DRPNN [64].

4.4 Quantitative evaluation

To evaluate the performance of the reduced resolution images, four evaluation metrics, namely the universal image quality index ($Q2n$) [48], the spectral angle mapper (SAM)

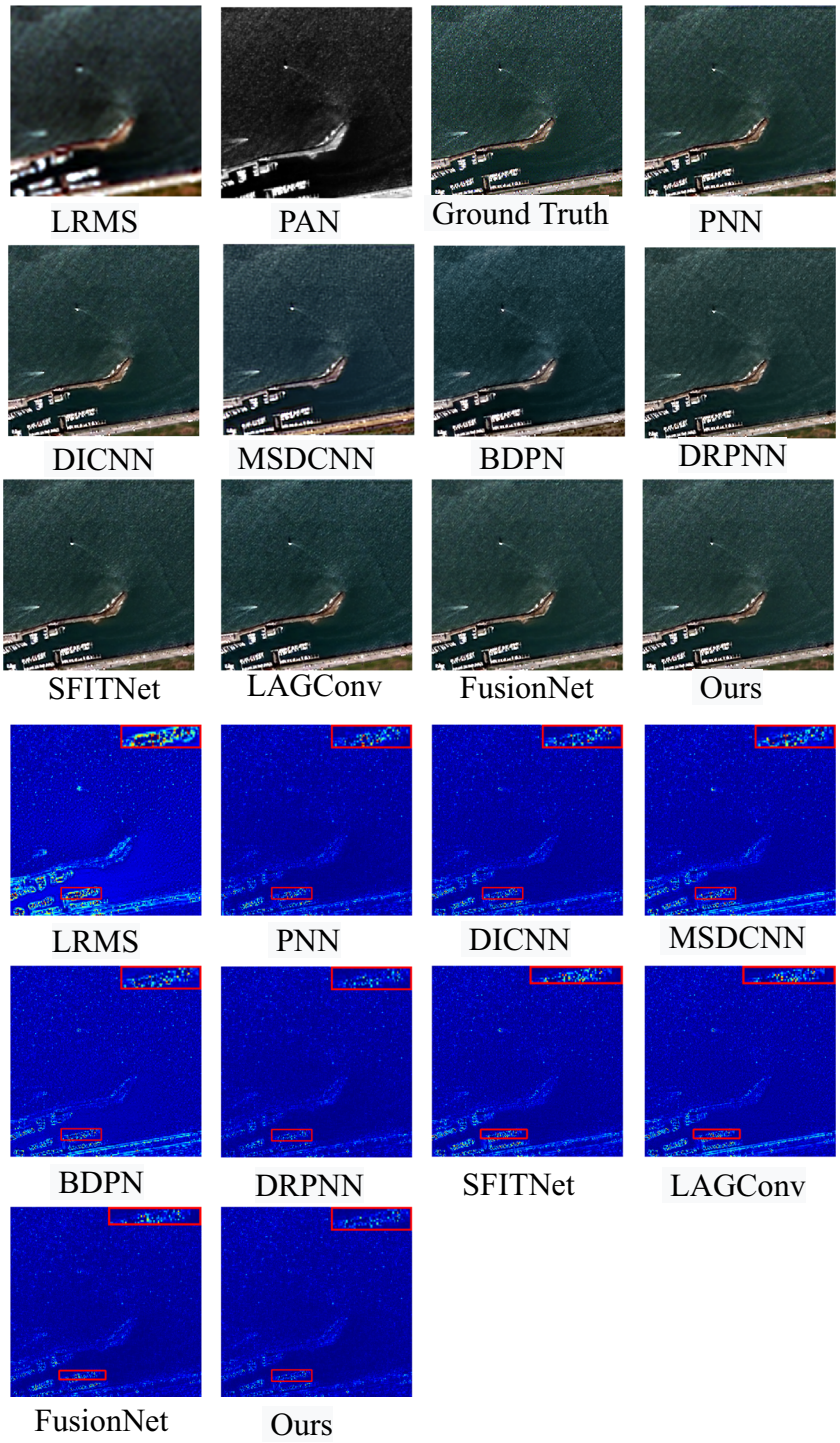

Fig. 8 Qualitative evaluation of different methods for pan-sharpening under reduced resolution test conditions on the QuickBird dataset. The images presented in the last three rows depict the residuals between different results and the ground truth

[57], the relative dimensionless global error in synthesis (ERGAS) [2], and the spatial correlation coefficient (SCC) are utilized.

Specifically, the $Q2n$ (termed Q4 when the data band is 4 and termed Q8 when the data band is 8) is a generic metric used for assessing image quality. It evaluates image quality by comparing the differences between the ground truth and the fused image. The higher value of Q2n indicates a greater similarity between the image and the ground truth, resulting in lower distortion. The SAM metric is used to measure the spectral similarity between the fused image and the multispectral image. The lower SAM value indicates a smaller angular difference between the spectral characteristics of the pixel and the ground truth spectrum. The ERGAS

Fig. 9 Qualitative evaluation of different pan-sharpening approaches for full resolution test on the QuickBird dataset

is an indicator used to compare the spatial information differences between fused images and the original images. The lower ERGAS value indicates better preservation of spectral information in the fusion result. The SCC is used to assess the spatial correlation between the fused and original images. The higher SCC value indicates better preservation performance of the fusion result in terms of spectral and spatial information. Additionally, metrics with no reference, including D_λ [1], D_s [1], and QNR [1], are employed for evaluating full-resolution images. Lower D_λ and D_s values, along with higher QNR, indicate that the algorithm performs well in terms of spectral preservation, spatial preservation, and overall quality, enabling the generation of high-quality fused images of MS and PAN images.

Quantitative comparative results on the WorldViewIII, QuickBird, and GaoFen2 datasets are summerized in Table 2. For the WorldViewIII dataset with reduced-resolution images, Table 2 shows that the DBTA performs best in Q8, SAM, ERGAS, and SCC. Meanwhile, the DBTA performs best in D_λ, D_s, and QNR when tested on full-resolution images. This indicates that DBTA demonstrates the versatility and excellent performance in image processing at different resolutions on the WorldViewIII dataset.

For the QuickBird dataset, Table 2 illustrates that the DBTA ranks first in Q4, SAM, ERGAS, and SCC when tested on reduced-resolution images. Therefore, the DBTA demonstrates rewarding performance in both spectral preservation and spatial structure preservation. For full-resolution images, the proposed method ranks fourth, second, and third in D_λ, D_s, and QNR, respectively. It is worth mentioning that the calculation of D_λ involves comparing the resulting image with the LRMS image. The LRMS image and the HRMS image have different resolutions. Therefore, considering the scale difference, it is reasonable to expect that the spectral distribution of the LRMS image may not be precisely identical to that of the generated high-resolution image. Likewise, the down-sampling process can introduce differences in the spatial distribution compared to the actual spatial distribution.

For the GaoFen2 dataset, Table 2 demonstrates that the DBTA performs best in Q4, SAM, ERGAS, and SCC indicators. Meanwhile, it demonstrates suboptimal results in the unsupervised evaluation indicators. The reasons can be explained from the following aspects. Firstly, the dataset used for training and evaluation may possess inherent characteristics, such as variations in spectral properties or sensor-

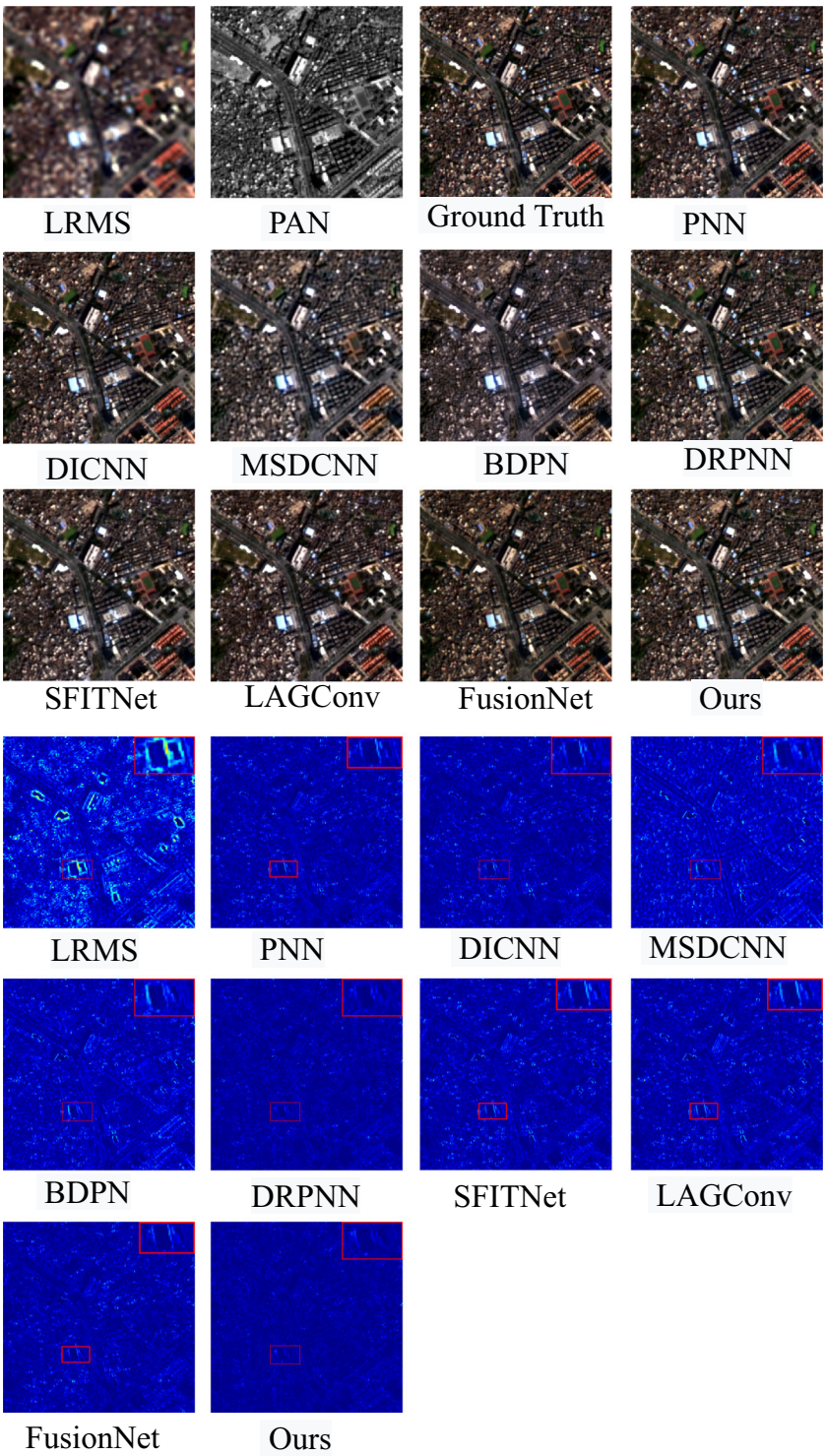

Fig. 10 Qualitative evaluation of different methods for pan-sharpening under reduced resolution test conditions on the GaoFen2 dataset. The images presented in the last three rows depict the residuals between different results and the ground truth

specific artifacts, which challenge the effectiveness of the supervised model. Furthermore, the complexity of the model architecture can also impact its ability to capture and leverage the necessary information for accurate pan-sharpening. Moreover, it is worth noting that the suitability of the evaluation metrics themselves should be considered, as they may not fully capture the perceptual quality or visual fidelity of the enhanced images. Qualitative evaluation of different pan-sharpening approaches for full resolution tested on the GaoFen2 dataset is shown in Fig. 11. It proves that the proposed model can deliver the highest quality subjective results.

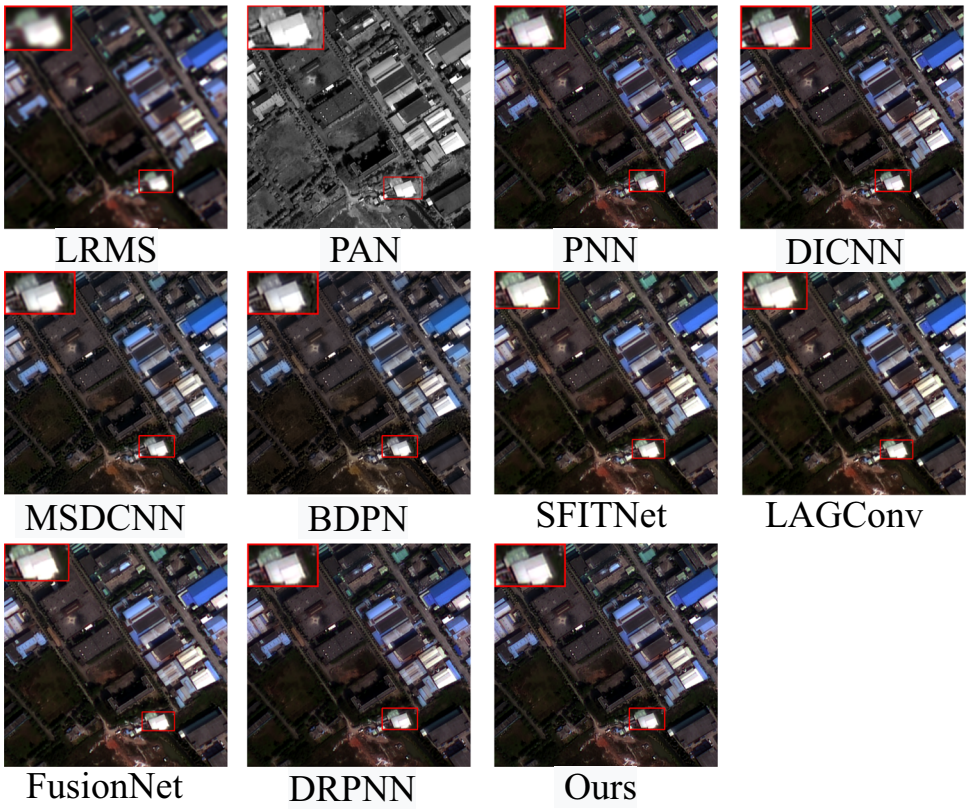

Fig. 11 Qualitative evaluation of different pan-sharpening approaches for full resolution test on the GaoFen2 dataset

4.5 Efficiency analysis

To evaluate the efficiency of the methods, the floating point operations (FLOPs), multiply-accumulate operations (MACs), and parameters of the model are evaluated. FLOPs refer to the number of floating-point operations (additions and multiplications) performed. Floating-point operations are mathematical operations used for handling real numbers, typically involving addition, subtraction, multiplication, and division of floating-point numbers. FLOPs are used to measure the computational workload required by neural network models during inference or training processes. MACs quantify the number of mathematical operations performing multiplication and accumulation. In deep learning, MACs are used to measure the computational complexity of the models. The parameters refer to the total quantity of all weights and biases in the model. The parameters are used to measure the complexity of the model.

All methods are evaluated on an NVIDIA GeForce RTX 3090Ti GPU. The efficiency results are shown in Table 3. It shows that the PNN, DICNN, and LAGConv are highly efficient due to their simple network architectures. Compared to other methods, the proposed model achieves acceptable parameter counts, complexity, and memory costs while ensuring performance.

4.6 Ablation study

A series of ablation experiments are conducted to validate the effectiveness of the CSA and SPA modules. Table 4 presents four configurations of the proposed method on three datasets.

The supervised evaluation metrics are used exclusively in the ablation experiments for two primary purposes. First and foremost, these metrics are widely used in various image-processing tasks because of their interpretability and alignment with human visual perception. Furthermore, the main objective of the proposed method is to enhance the level of detail and clarity in the generated images, aiming to closely approximate the ground truth. Hence, supervised evaluation metrics offer a quantitative assessment of the quality of the generated outcomes.

It can be seen from Table 4 that the proposed method performs best on the Q4, SAM, ERGAS, and SCC indicators on QuickBird and GaoFen2 datasets. It demonstrates the robustness and generality of the DBTA model. Nonetheless, the inclusion of either the CSA or SPA module independently resulted in a deterioration in method performance. This is due to the potential for an imbalance of spectral and spatial information in the fused image when the CSA module or SPA module is added separately. Meanwhile, the proposed DBTA model ranks first in the SAM indicator, second in

Table 2 Quantitative comparison results on the WorldViewIII, QuickBird, and GaoFen2 datasets

Datasets	Methods	Reduced-resolution Q2n ↑	SAM ↓	ERGAS ↓	SCC ↑	Full-resolution D_λ ↓	D_s ↓	QNR ↑
WorldViewIII	PNN[32]	0.8959±0.0897	3.5998±0.7335	2.6332±0.6695	0.9764±0.0072	0.0229±0.0091c	0.0442±0.0149	0.9339±0.0221
	DICNN[18]	0.9002±0.0863	3.5441±0.7357	2.6607±0.6729	0.9763±0.0070	0.0378±0.0119	0.0448±0.0173	0.9193±0.0259
	MSDCNN[56]	0.8531±0.0946	4.7124±1.1293	3.6765±0.9773	0.9541±0.0098	0.0392±0.0159	0.0653±0.0322	0.8986±0.0444
	BDPN[63]	0.8560±0.0988	4.5996±0.9629	3.6016±0.8668	0.9506±0.0173	0.0429±0.0183	0.0771±0.0358	0.8839±0.0497
	DRPNN[64]	0.8815±0.0947	3.7251± 0.7891	2.7496±0.6778	0.9740±0.0068	0.0280±0.0107	0.0518±0.0226	0.9218±0.0312
	SFITNet[67]	0.8329±0.1262	4.5605±0.9851	3.2343±0.7880	0.9630±0.0105	0.0295±0.0130	0.0674±0.0321	0.9055±0.0421
	LAGConv[24]	0.8629±0.1036	4.4474±0.8811	3.1548±0.8127	0.9659±0.0096	0.0387±0.0165	0.0637±0.0323	0.9005±0.0451
	FusionNet[8]	0.9040±0.0867	3.3277± 0.6721	2.4490±0.6126	0.9740±0.0068	0.0227±0.0085	0.0397±0.0155	0.9386±0.0219
	Ours	0.9057±0.0866	3.1178±0.6249	2.3034±0.5446	0.9832±0.0055	0.0211±0.0084	0.0301±0.0121	0.9495±0.0180
QuickBird	PNN[32]	0.9167±0.0979	5.1703±0.9208	4.4387±0.3430	0.9717±0.0124	0.0604±0.0116	0.0629±0.0255	0.8807±0.0324
	DICNN[18]	0.9016±0.0977	5.3732±0.9942	5.1823±0.4518	0.9614±0.0143	0.0931±0.0133	0.1114±0.0246	0.8062±0.0329
	MSDCNN[56]	0.8426±0.0957	7.6073±1.6438	7.021±0.6390	0.9193±0.0165	0.0772±0.0329	0.1253±0.0204	0.8078±0.0457
	BDPN[63]	0.8635±0.0938	7.1181±1.4041	6.3893±0.5095	0.934±0.0178	0.1759±0.0453	0.1955±0.0352	0.6645±0.0607
	DRPNN[64]	0.9276±0.0952	4.7546±0.8440	3.9123±0.3227	0.9791±0.0086	0.0556±0.0188	0.0463±0.0181	0.9009±0.0332
	SFITNet[67]	0.9008±0.1003	5.5886±1.0135	5.1281±0.5935	0.9608±0.0185	0.0873±0.0219	0.0933±0.0311	0.8282±0.0461
	LAGConv[24]	0.8903±0.0955	5.9136±1.0849	5.6209±0.5218	0.9523±0.0163	0.0898±0.0213	0.1310±0.0335	0.7916±0.0467
	FusionNet[8]	0.9217±0.0972	4.8699±0.8696	4.2126±0.3056	0.9749±0.0105	0.0710±0.0211	0.0688±0.0268	0.8656±0.0423
	Ours	0.9319±0.0881	4.6316±0.8264	3.8285±0.3033	0.9803±0.0079	0.0759±0.0230	0.0578±0.0227	0.8706±0.0273
GaoFen2	PNN[32]	0.9563±0.0099	1.0845±0.2362	1.1090±0.2512	0.9737±0.0066	0.0153±0.0071	0.0689±0.0115	0.9169±0.0116
	DICNN[18]	0.9582±0.0104	1.0536±0.2217	1.0821±0.2422	0.9761±0.0058	0.0361±0.0162	0.0983±0.0127	0.8692±0.0186
	MSDCNN[56]	0.9054±0.0218	1.7057±0.2974	1.5341±0.2760	0.9497±0.0088	0.0447±0.0327	0.0992±0.0202	0.8606±0.0368
	BDPN[63]	0.9242±0.0254	1.8553±0.2560	1.6171±0.2593	0.9535±0.0085	0.0773±0.0462	0.1392±0.0239	0.7943±0.0476
	DRPNN[64]	0.9703±0.0092	0.9387±0.1824	0.8514±0.1629	0.9848±0.0030	0.0236±0.0113	0.0711±0.0100	0.9070±0.0135
	SFITNet[67]	0.9460±0.0141	1.2136±0.2423	1.2335±0.2680	0.9675±0.0078	0.0326±0.0199	0.0873±0.0114	0.8828±0.0166
	LAGConv[24]	0.9327±0.0196	1.3387±0.2513	1.3501±0.2735	0.9622±0.0082	0.0439±0.0197	0.0952±0.0128	0.8650±0.0191
	FusionNet[8]	0.9631±0.0097	0.9850±0.2054	0.9938±0.2114	0.9794±0.0048	0.0374±0.0187	0.1013±0.0137	0.8651±0.0218
	Ours	0.9749±0.0087	0.8536±0.1640	0.7798±0.1408	0.9871±0.0022	0.0340±0.0171	0.0988±0.0161	0.8706±0.0227

The presented values represent the mean value ± standard deviation. (The best, second-best, and third-best results are highlighted in red, blue, and green, respectively)

Table 3 The efficiency results of eight comparison methods and the proposed method

Methods	FLOPs	MACs	Params
PNN [32]	111.10	111.10	0.08042
DICNN [18]	96.88	96.88	0.04218
MSDCNN [56]	436.76	436.76	0.18985
BDPN [63]	2,131.74	8,509.22	2.95841
DRPNN [64]	963.23	963.23	0.41837
SFITNet [67]	236.62	231.93	0.10473
LAGConv [24]	290.26	71.95	0.05370
FusionNet [8]	175.18	684.79	0.15031
Ours	412.98	3,135.38	0.25483

The unit of the displayed values is million (M)

the ERGAS and SCC indicators, and third place in the Q8 indicator on the WorldViewIII dataset. This proves that the proposed method can effectively preserve the spectral information of fused images and enhance their spatial structure.

It is worth mentioning that the Q8 metric quantifies the similarity between the fused images and the ground truth images by assessing both spectral and spatial characteristics. The proposed method may not be very similar to ground truth images in terms of pixel values. This is because ground truth images are obtained by downsampling the original MS images, which could introduce some errors and artifacts in terms of spectral and spatial information. Therefore, the pan-sharpened images generated by the proposed method are more realistic and clearer than ground truth images, but their similarity to ground truth images is lower. When the CSA and SPA modules are equipped, the performance of the method is improved. This proves that the CSA module and SPA module are beneficial in enhancing the performance of our method. The CSA module can effectively extract and preserve the spatial structural information of the image, while the SPA module can effectively extract and preserve the spectral information of the image. The combination of these two modules allows the proposed method to simultaneously focus on the most informative features in both spatial and spectral dimensions, thereby improving feature extraction and fusion.

5 Conclusion and future work

In this paper, we propose a Dual-Branch and Triple-Attention Network (DBTA) to mitigate the problems of spectral and spatial structural distortion encountered in pan-sharpening tasks. The proposed model consists of two key components, namely the channel-spatial attention (CSA) module and the spectral attention (SPA) module. Specifically, the primary objective of the CSA module is to capture spatial structural information within the image. This module enhances spatial structure details within the target area while suppressing noise and artifacts. Additionally, the SPA module is introduced to extract spectral information from MS images. Numerous qualitative and quantitative experiments have validated the effectiveness and superiority of the proposed DBTA.

In future studies, it is worth considering the integration of unsupervised learning techniques to enhance the performance of pan-sharpening. Incorporating unsupervised self-learning or adaptive mechanisms can better leverage the structure and characteristics of the images to guide the pan-sharpening process, ultimately improving the generalization capability of the model.

Table 4 Comparative results of the baseline with different compound modes on the WorldViewIII, QuickBird, and GaoFen2 datasets

Datasets	CSA	SPA	Q2n↑	SAM↓	ERGAS↓	SCC↑
WorldViewIII	✗	✗	0.9058± 0.0871	3.2282± 0.6488	2.4318± 0.6252	0.9809± 0.0063
	✓	✗	0.9064±0.0866	3.1326±0.6223	2.2967±0.5274	0.9833±0.0054
	✗	✓	0.9056± 0.0869	3.2727±0.6541	2.4531±0.6415	0.9809±0.0063
	✓	✓	0.9057±0.0866	3.1178±0.6249	2.3034±0.5446	0.9832±0.0055
QuickBird	✗	✗	0.9261±0.9005	4.8968±0.8859	4.0586±0.4213	0.9771±0.0089
	✓	✗	0.8857±0.0964	5.9455±1.0945	5.8172±0.5235	0.9484±0.0148
	✗	✓	0.8852±0.0945	5.8388±1.0863	5.8653±0.5170	0.9482±0.0137
	✓	✓	0.9319±0.0881	4.6316±0.8264	3.8285±0.3033	0.9803±0.0079
GaoFen2	✗	✗	0.9594±0.0092	1.0543±0.2160	1.0888±0.2470	0.9743±0.0069
	✓	✗	0.9329±0.0221	1.4155±0.2130	1.3255±0.2539	0.9636±0.0079
	✗	✓	0.9350±0.0167	1.3059±0.2325	1.3286±0.2640	0.9629±0.0081
	✓	✓	0.9749±0.0087	0.8536±0.1640	0.7798±0.1408	0.9871±0.0022

The presented values represent the mean value ± standard deviation. The best result is highlighted in red

Acknowledgements This work is supported in part by the National Natural Science Foundation of China (No.61601266).

Author Contributions Wenhao Song: Conceptualization and Original draft. Mingliang Gao: Supervision, Review and Editing. Abdellah Chehri: English polishing. Wenzhe Zhai: Network development. Qilei Li: Data curation and Formal analysis. Gwanggil Jeon: Methodology and English polishing.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Competing of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Alparone L, Aiazzi B, Baronti S, Garzelli A, Nencini F, Selva M (2008) Multispectral and panchromatic data fusion assessment without reference. Photogramm Eng Remote Sensing 74(2):193–200
2. Alparone L, Wald L, Chanussot J, Thomas C, Gamba P, Bruce LM (2007) Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data-fusion contest. IEEE Trans Geosci Remote Sens 45(10):3012–3021. https://doi.org/10.1109/TGRS.2007.904923
3. Bullock EL, Woodcock CE, Olofsson P (2020) Monitoring tropical forest degradation using spectral unmixing and landsat time series analysis. Remote Sens Environ 238:110968
4. Cao X, Chen Y, Cao W (2022) Proximal pannet: A model-based deep network for pansharpening. In: Proceedings of the AAAI conference on artificial intelligence vol 36, pp 176–184
5. Chen Y, Peng G, Zhu Z, Li S (2020) A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. Appl Soft Comput 86:105919
6. Choromanski KM, Likhosherstov V, Dohan D, Song X, Gane A, Sarlos T, Hawkins P, Davis JQ, Mohiuddin A, Kaiser L, Belanger DB, Colwell LJ, Weller A (2021) Rethinking attention with performers. In: International conference on learning representations
7. Ciotola M, Vitale S, Mazza A, Poggi G, Scarpa G (2022) Pansharpening by convolutional neural networks in the full resolution framework. IEEE Trans Geosci Remote Sens 60:1–17
8. Deng LJ, Vivone G, Jin C, Chanussot J (2020) Detail injection-based deep convolutional neural networks for pansharpening. IEEE Trans Geosci Remote Sens 59(8):6995–7010
9. Deng LJ, Vivone G, Paoletti ME, Scarpa G, He J, Zhang Y, Chanussot J, Plaza A (2022) Machine learning in pansharpening: A benchmark, from shallow to deep networks. IEEE Geosci Remote Sens Mag 10(3):279–315
10. Dong C, Loy CC, He K, Tang X (2015) Image super-resolution using deep convolutional networks. IEEE Geosci Remote Sens Mag 38(2):295–307
11. Fu X, Lin Z, Huang Y, Ding X (2019) A variational pan-sharpening with local gradient constraints. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10265–10274
12. Gao H, Li S, Li J, Dian R (2023) Multispectral image pansharpening guided by component substitution model. IEEE Trans Geosci Remote Sens
13. Gillespie AR, Kahle AB, Walker RE (1987) Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques. Remote Sens Environ 22(3):343–365
14. Gong M, Ma J, Xu H, Tian X, Zhang XP (2022) D2tnet: A convlstm network with dual-direction transfer for pan-sharpening. IEEE Trans Geosci Remote Sens 60:1–14
15. Gong Y, Xiao Z, Tan X, Sui H, Xu C, Duan H, Li D (2019) Context-aware convolutional neural network for object detection in vhr remote sensing imagery. IEEE Trans Geosci Remote Sens 58(1):34–44
16. Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, Zhang SH, Martin RR, Cheng MM, Hu SM (2022) Attention mechanisms in computer vision: A survey. Comput Vis media 8(3):331–368
17. Hashim F, Dibs H, Jaber HS (2022) Adopting gram-schmidt and brovey methods for estimating land use and land cover using remote sensing and satellite images. Nat Environ Pollut Technol 21(2):867–881
18. He L, Rao Y, Li J, Chanussot J, Plaza A, Zhu J, Li B (2019) Pansharpening via detail injection based convolutional neural networks. IEEE J Sel Top Appl Earth Obs Remote Sens 12(4):1188–1204. https://doi.org/10.1109/JSTARS.2019.2898574
19. Huang W, Xiao L, Wei Z, Liu H, Tang S (2015) A new pan-sharpening method with deep neural networks. IEEE Geosci Remote Sens Lett 12(5):1037–1041
20. Javan FD, Samadzadegan F, Mehravar S, Toosi A, Khatami R, Stein A (2021) A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery. ISPRS J Photogramm Remote Sens 171:101–117
21. Jian L, Yang X, Liu Z, Jeon G, Gao M, Chisholm D (2020) Sedrfuse: A symmetric encoder-decoder with residual block network for infrared and visible image fusion. IEEE Trans Instrum Meas 70:1–15
22. Jianwen H, Zeping W, Pei H (2023) A review of pansharpening methods based on deep learning. Remote Sensing for Natural Resources 35(1)
23. Jin C, Deng LJ, Huang TZ, Vivone G (2022) Laplacian pyramid networks: A new approach for multispectral pansharpening. Inf Fusion 78:158–170
24. Jin ZR, Zhang TJ, Jiang TX, Vivone G, Deng LJ (2022) Lagconv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening. Proceedings of the AAAI conference on artificial intelligence vol 36, pp 1113–1121
25. Lee J, Seo S, Kim M (2021) Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10166–10174
26. Leung Y, Liu J, Zhang J (2013) An improved adaptive intensity-hue-saturation method for the fusion of remote sensing images. IEEE Geosci Remote Sens Lett 11(5):985–989
27. Li X, Xu F, Lyu X, Tong Y, Chen Z, Li S, Liu D (2020) A remote-sensing image pan-sharpening method based on multi-scale channel attention residual network. IEEE Access 8:27163–27177
28. Liang Y, Zhang P, Mei Y, Wang T (2022) Pmacnet: Parallel multi-scale attention constraint network for pan-sharpening. IEEE Geosci Remote Sens Lett 19:1–5
29. Liu Q, Zhou H, Xu Q, Liu X, Wang Y (2020) Psgan: A generative adversarial network for remote sensing image pan-sharpening. IEEE Trans Geosci Remote Sens 59(12):10227–10242
30. Liu X, Liu Q, Wang Y (2020) Remote sensing image fusion based on two-stream fusion network. Inf Fusion 55:1–15
31. Ma J, Yu W, Chen C, Liang P, Guo X, Jiang J (2020) Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion. Inf Fusion 62:110–120
32. Masi G, Cozzolino D, Verdoliva L, Scarpa G (2016) Pansharpening by convolutional neural networks. Remote Sensing 8(7):594

33. Meng X, Shen H, Li H, Zhang L, Fu R (2019) Review of the pan-sharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges. Inf Fusion 46:102–113
34. Menon AS, Aravinth J, Veni S (2023) Pan-sharpening of multi-spectral remote sensing data using multi-resolution analysis. In: Machine intelligence techniques for data analysis and signal processing: proceedings of the 4th international conference MISP 2022, vol 1, pp 697–705. Springer
35. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
36. Nencini F, Garzelli A, Baronti S, Alparone L (2007) Remote sensing image fusion using the curvelet transform. Inf Fusion 8(2):143–156
37. Pan Z, Yu W, Yi X, Khan A, Yuan F, Zheng Y (2019) Recent progress on generative adversarial networks (gans): A survey. IEEE access 7:36322–36333
38. Qu Y, Baghbaderani RK, Qi H, Kwan C (2020) Unsupervised pan-sharpening based on self-attention mechanism. IEEE Trans Geosci Remote Sens 59(4):3192–3208
39. Shao Z, Cai J (2018) Remote sensing image fusion with deep convolutional neural network. IEEE J Sel Top Appl Earth Obs Remote Sens 11(5):1656–1669
40. Sharma KV, Kumar V, Singh K, Mehta DJ (2023) Landsat 8 lst pan sharpening using novel principal component based downscaling model. Remote Sens Appl: Soc Environ 30:100963
41. Su X, Li J, Hua Z (2022) Transformer-based regression network for pansharpening remote sensing images. IEEE Trans Geosci Remote Sens 60:1–23
42. Su Y, Zhu H, Wong KC, Chang Y, Li X (2022) Hyperspectral image denoising via weighted multidirectional low-rank tensor recovery. IEEE Trans Cybern 53(5):2753–2766
43. Suryanarayana G, Saidulu B, Priya MRH, Likhitha K, Pragathi K, Srikanth K (2022) Fusion of hyperspectral and multispectral images based on principal component analysis and guided bilateral filtering. International Journal of System Assurance Engineering and Management, pp 1–10
44. Tang A, Quan P, Niu L, Shi Y (2022) A survey for sparse regularization based compression methods. Ann Data Sci 9(4):695–722
45. Tang X, Ma Q, Zhang X, Liu F, Ma J, Jiao L (2021) Attention consistent network for remote sensing scene classification. IEEE J Sel Top Appl Earth Obs Remote Sens 14:2030–2045
46. Tu TM, Huang PS, Hung CL, Chang CP (2004) A fast intensity-hue-saturation fusion technique with spectral adjustment for ikonos imagery. IEEE Geosci Remote Sens Lett 1(4):309–312
47. Wald L, Ranchin T, Mangolini M (1997) Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogramm Eng Remote Sensing 63:691–699
48. Wang Z, Bovik AC (2002) A universal image quality index. IEEE Signal Process Lett 9(3):81–84
49. Wang Z, Ma Y, Zhang Y (2023) Review of pixel-level remote sensing image fusion based on deep learning. Inf Fusion 90:36–58
50. Xing Y, Wang M, Yang S, Zhang K (2018) Pansharpening with multiscale geometric support tensor machine. IEEE Trans Geosci Remote Sens 56(5):2503–2517
51. Xing Y, Zhang Y, Zhang Y (2022) Wavefusion: Wavelet assistant fusion model for pan-sharpening. In: IGARSS 2022-2022 IEEE international geoscience and remote sensing symposium, pp 1083–1086. IEEE
52. Xiong Z, Liu N, Wang N, Sun Z, Li W (2023) Unsupervised pansharpening method using residual network with spatial texture attention. IEEE Trans Geosci Remote Sens
53. Xu Q, Li Y, Nie J, Liu Q, Guo M (2023) Upangan: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network. Inf Fusion 91:31–46
54. Yan K, Zhou M, Liu L, Xie C, Hong D (2022) When pansharpening meets graph convolution network and knowledge distillation. IEEE Trans Geosci Remote Sens 60:1–15
55. Yilmaz CS, Yilmaz V, Gungor O (2022) A theoretical and practical survey of image fusion methods for multispectral pansharpening. Inf Fusion 79:1–43
56. Yuan Q, Wei Y, Meng X, Shen H, Zhang L (2018) A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE J Sel Top Appl Earth Obs Remote Sens 11(3):978–989. https://doi.org/10.1109/JSTARS.2018.2794888
57. Yuhas RH, Goetz AFH, Boardman JW (1992) Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In: JPL, Summaries of the third annual JPL airborne geoscience workshop. vol 1: AVIRIS Workshop
58. Zhai W, Gao M, Souri A, Li Q, Guo X, Shang J, Zou G (2023) An attentive hierarchy convnet for crowd counting in smart city. Cluster Comput 26(2):1099–1111
59. Zhang H, Li Y, Jiang Y, Wang P, Shen Q, Shen C (2019) Hyperspectral classification based on lightweight 3-d-cnn with transfer learning. IEEE Trans Geosci Remote Sens 57(8):5813–5828
60. Zhang H, Ma J (2021) Gtp-pnet: A residual learning network based on gradient transformation prior for pansharpening. ISPRS J Photogramm Remote Sens 172:223–239
61. Zhang H, Xu H, Tian X, Jiang J, Ma J (2021) Image fusion meets deep learning: A survey and perspective. Inf Fusion 76:323–336
62. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the european conference on computer vision (ECCV), pp 286–301
63. Zhang Y, Liu C, Sun M, Ou Y (2019) Pan-sharpening using an efficient bidirectional pyramid network. IEEE Trans Geosci Remote Sens 57(8):5549–5563. https://doi.org/10.1109/TGRS.2019.2900419
64. Zheng Y, Li J, Li Y, Cao K, Wang K (2019) Deep residual learning for boosting the accuracy of hyperspectral pansharpening. IEEE Geosci Remote Sens Lett 17(8):1435–1439
65. Zhong S, Zhang Y, Chen Y, Wu D (2017) Combining component substitution and multiresolution analysis: A novel generalized bdsd pansharpening algorithm. IEEE J Sel Top Appl Earth Obs Remote Sens 10(6):2867–2875
66. Zhou M, Huang J, Fu X, Zhao F, Hong D (2022) Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. IEEE Trans Geosci Remote Sens 60:1–14
67. Zhou M, Huang J, Yan K, Yu H, Fu X, Liu A, Wei X, Zhao F (2022) Spatial-frequency domain information integration for pan-sharpening. In: European conference on computer vision, pp 274–291. Springer (2022)

Wenhao Song is pursuing the M.S. degree at 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 received his Ph.D. in Communication and Information Systems from Sichuan University. He is now an IEEE Senior Member and an associate professor at the Shandong University of Technology. He was a visiting lecturer at the University of British Columbia during 2018-2019. 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 180 journal/conference papers in IEEE, Springer, Elsevier, and Wiley.

Abdellah Chehri is an Associate Professor at the Royal Military College of Canada (RMC), Kingston, Ontario. He is a co-author of 250+ peer-reviewed publications in established journals and conference proceedings sponsored by established publishers such as IEEE, ACM, Elsevier, and Springer. He has been listed among the top 2% cited scientists reported by Stanford University since 2020. His research interests include computer vision and wireless heterogeneous sensor networks.

Wenzhe Zhai is pursuing the M.S. degree at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. His research interests include smart city system, information fusion, crowd analysis and deep learning.

Qilei Li is a fourth year Ph.D. student in Computer Science, Queen Mary University of London, supervised by Prof. Shaogang (Sean) Gong. Previously, he received the M.S. degree from Sichuan University in 2020. His research interests include computer vision and deep learning, particularly focusing on person ReID, video/image enhancement. He is a student member of IEEE, and he serves as a reviewer for Information Fusion, IEEE TIM, IEEE Access, Concurrency and Computation: Practice and Experience, and Multimedia System.

Gwanggil Jeon received the B.S., M.S., and Ph.D. (summa cum laude) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2003, 2005, and 2008, respectively. From 2009.09 to 2011.08, he was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow. He is an IEEE Senior Member a Full Professor at 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.