ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 6 July 2025; Accepted: 30 September 2025

https://doi.org/10.17559/TV-20250706002800
Original scientific paper

HPN-ICE: Information Cross Embedding for Hyperspectral Pansharpening

Yan JIN

Abstract: Hyperspectral (HS) pansharpening aims to generate high-spatial-resolution hyperspectral (HRHS) images by fusing panchromatic (PAN) images with low-spatial-resolution hyperspectral (LRHS) images. However, many existing HS pansharpening methods fail to capture global dependencies between cross-modal features, leading to spectral and spatial distortions.To address this issue, we propose a hyperspectral pansharpening network based on information cross embedding (HPN-ICE). The model progressively fuses HS and PAN image features through two modules: the global feature fusion module (GFFM) and the multi-directional feature enhancement module (MFEM). In GFFM, a feature embedding fusion module (FEFM) is firstly designed based on the information cross embedding, which efficiently fuses spectral and spatial features by establishing cross dependencies between two modal features. Then, a frequency-domain channel attention module (FCAM) is constructed to enhance the global spectral information in the frequency domain. MFEM is constructed to enhance the local details of fused features in multi-dimensional directions. Extensive experiments conducted on three widely used datasets demonstrate that HPN-ICE achieves significant improvements in both spatial and spectral quality metrics over some state-of-the-art (SOTA) methods. The code will be released on GitHub.

Keywords: frequency-domain; hyperspectral pansharpening; mamba model; multi-directional feature

1 INTRODUCTION

Hyperspectral (HS) images usually comprise hundreds of narrow spectral bands, offering high spectral resolution, with each band representing the radiance information of the scene within specific wavelength intervals. However, due to the physical limitations of hyperspectral sensors, only low-spatial-resolution hyperspectral (LRHS) images can be acquired, which limits the application of HS images in numerous fields [1, 2]. In contrast, panchromatic (PAN) imaging systems can provide single-band images with high spatial resolution. One possible approach is to reconstruct high-spatial-resolution hyperspectral (HRHS) images from LRHS images and PAN images. This process is commonly referred to as HS pansharpening.

The traditional methods are roughly classified into four categories, such as component substitution (CS)-based methods [3-7], multi-scale resolution analysis (MRA)-based methods [8-11], Bayesian-based methods [12-15], and Matrix decomposition-based methods [16-18]. The CS-based methods obtain HR-HS images by substituting the spatial components of HS images with the spatial components of PAN images, mainly including Principal Component Analysis (PCA) [4], Intensity-Hue-Saturation (IHS) [5], and Gram-Schmidt Adaptive (GSA) [6, 7] methods. The MRA-based methods generate HR-HS images by injecting the spatial details of PAN images at different scales into LR-HS images. The multi-scale decomposition approaches used in these methods mainly include the Wavelet Transform (WT)-based methods [8, 9] and Laplacian Pyramid (LP)-based methods [10, 11]. In addition, some hybrid methods, such as Guided Filter PCA (GFPCA) [19] based on CS and MRA were also proposed to balance the quality of spatial and spectral information. Bayesian-based methods solve fusion problems by constructing regularization terms based on Bayesian priors, such as Bayesian Naive [12] and Bayesian Sparse Promoting Gaussian Prior (BSP) [13, 14]. However, the solutions of these methods are complex and require precise prior definitions. Matrix factorization-based approaches regularize the fusion problem via utilizing the priors of spectral unmixing, and the coupled non-negative matrix factorization (CNMF) [16] is a commonly used and typical method. Although traditional methods enhance spatial details, they often suffer from severe spectral distortion due to limited feature extraction, imprecise priors, and spectral gaps between LR-HS and PAN images.

In recent years, convolutional neural networks (CNNs) have been developed in pansharpening tasks due to their excellent feature learning capacities [20, 21]. Bandar et al. [22] adopted the DIP approach to upsample LR-HS images and designed an over-complete network called HyperKite to effectively capture image edge details. However, since the difference between the two image features is not considered, directly extracting spatial and spectral features from the two images may cause spectral and spatial distortions in the fusion results. To reduce the aliasing of spatial and spectral features, Qu et al. [23] proposed a two-branch detail extraction pansharpening method, which uses a pre-trained network to sharpen LR-HS images. He et al. [24] developed a spectrum prediction convolutional neural network (HyperPNN), which improves the spectral and spatial prediction capabilities by introducing a spectrum prediction structure. Dong et al. [25] proposed a feature pyramid fusion network (FPFNet) for pansharpening, which extracts multi-resolution features from PAN and HS images by constructing two branches and gradually injects them into the fused features. Wang et al. [26] proposed DISPNet, an interpretable deep unfolded network with intrinsic supervision for pansharpening. The method leverages spatial consistency and spectral projection priors to enhance spatial quality and modality correlation. However, its reliance on a variational framework and iterative algorithm introduces significant computational complexity. Zhuo et al. [27] designed a Hyper-DSNet that ensures the spatial and spectral fidelity of fused images by constructing a deep-shallow fusion structure with multi-detail extraction and spectral attention. Dong et al. [28] developed a deep CNN within a Gaussian Laplacian pyramid for pansharpening, which reconstructs HR-HS image by injecting sub-band residuals extracted from PAN images into upsampled HS images. However, due to the limitation of kernel size in convolution

operations, these methods mainly focus on learning local features of the image and lack exploration of long-range features, leading to spatial and spectral distortions in the fusion results.

Due to the long-range feature learning capability of the Transformer architecture, it has been widely applied in computer vision tasks. Considering the advantages of Transformer and CNN, researchers have attempted to explore the learning of global and local contextual information in images by combining the two structures. For example, Liu et al. [29] designed an Interactformer that is a dual-branch network consisting of Transformer and 3D-CNN network to extract global and local features. Zhou et al. [30] constructed a deep learning fusion network called HyperRefiner based on autoencoders and self-attention mechanism. Bandara et al. [31] designed a HyperTransformer to improve the reconstruction quality of spatial and spectral features by effectively capturing the feature relationship between PAN and LR-HS images. Shang et al. [32] modeled the panshapening task as an optimization model and utilized transformer and CNN structures to optimize and solve the model. Zhou et al. [33] designed a self-attention dual-stream network architecture based on Transformer, which extracts two modal features from Multispectral and PAN images and fuses spatial and spectral features by constructing a cross-attention module.

However, due to the high-dimensional data property of HS images, these methods usually have high computational complexity. Recently, researchers have developed structured state space models (SSM) [34] with the capacity to learn global features for the purpose of reducing computational complexity. Then, there are some challenges when applying the Mamba model, as an outstanding SSM, to the task of HS pansharpening. The Mamba model only supports linear input at the pixel level. Therefore, when it is used for cross-modal image fusion, it is rather difficult to effectively capture and establish long-range dependencies between cross-modal images. In addition, converting HS images to one-dimensional sequences for modeling may overlook spectral correlation info among channels in the Mamba model. In HS pansharpening tasks, frequency domain processing methods [35, 36] have drawn wide attention for their unique global modeling capabilities. Image transformation from spatial to frequency domain enables effective spectral feature separation or enhancement via HS images frequency characteristics.

To address the above challenges, we capitalize on the advantages of the Mamba model and frequency-domain computational methods, and propose a hyperspectral pansharpening network based on information cross embedding (HPN-ICE) that aims to simultaneously preserve spectral fidelity and spatial detail. The feature embedding fusion module (FEFM), which is constructed based on Mamba model, carries out the cross-modal fusion of LRHS and PAN images through pixel-wise cross combination calculations. Subsequently, the fused images are fed into the Mamba model for feature learning. This design effectively overcomes the limitations of the Mamba model in establishing cross-modal image mapping relationships in HS pansharpening. In addition, we proposed a frequency channel attention module (FCAM) to enhance spectral features in the frequency domain. The FCAM combines channel attention mechanisms with the fast fourier transform (FFT), enabling the interaction of input features at the spectral level and thereby strengthening the representation of spectral information.

Finally, considering the rich land-cover information in HS images, we designed a multi-directional feature enhancement module (MFEM) to extract and enhance the details and structural information in the fusion image through multi-directional attention mechanisms. Our main contributions can be summarized as follows:

1) An HPN-ICE composed of multiple GFFMs and MFFM is proposed to achieve fusion of PAN and LRHS images, in order to obtain HRHS images with both spectral and spatial fidelity.

2) In GFFM, a FEFM is designed based on the Mamba model to achieve the fusion of two image features by establishing long-range dependencies across modal features. A FCAM is constructed to enhance the spectral and detail features in the frequency domain.

3) A MFEM is constructed to enhance the detail information in the fused features by designing multi-directional attention mechanisms.

2 PROPOSED METHOD

In this section, as shown in Fig. 1a, we propose an HPN-ICE for HS pansharpening, the specific execution process of HPN-ICE is as follows.

Firstly, the LRHS image $I_{\mathrm{LH}} \in R^{H/4 \cdot W/4 \cdot C}$ is upsampled to the same spatial size as PAN image to obtain the upsampled LR-HS image $UI_{\mathrm{LH}} \in R^{H \cdot W \cdot C}$, and the PAN image $I_P \in R^{H \cdot W \cdot 1}$ is extended to the same number of channels as the LRHS image through a copy operation to obtain a fake hyperspectral PAN $FI_{\mathrm{HS\_P}} \in R^{H \cdot W \cdot C}$ (H and W represent the height and width of PAN image, and C represents the channel number of LRHS image). Multiple cascaded GFFMs are constructed to achieve progressive fusion of two image features, with each GFFM consisting of a FEFM and an FCAM. FEFM is designed to achieve interactive embedding of local features between two images and learning of global features, while FCAM is constructed to enhance spectral features by computing a frequency channel attention. The input of the first GFFM is UI_{LH} and $FI_{\mathrm{HS\_P}}$, and the subsequent GFFM takes the fused features obtained from the previous GFFM and $FI_{\mathrm{HS\_P}}$. The above operations are as follows.

$$UI_{\mathrm{LH}} = \mathrm{Up}\left(I_{\mathrm{LH}}\right), FI_{\mathrm{HS\_P}} = \mathrm{Exp}\left(I_P\right) \tag{1}$$

$$\begin{aligned} FI_G^l &= \mathrm{GFFM}\left(FI_G^{l-1}, FI_{\mathrm{HS\_P}}\right) \\ &= \mathrm{FCAM}\left(\mathrm{FEFM}\left(FI_G^{i-1}, FI_{\mathrm{HS\_P}}\right)\right) \\ l &= 1, 2, ..., n, FI_G^0 = UI_{\mathrm{LH}} \end{aligned} \tag{2}$$

where Up(·) and Exp(·) represent upsampling and channel expansion operations, respectively. GFFM(·), FEFM(·), and FCAM(·) denote GFFM, FEFM, and FCAM. FI_G^l denotes the output of the l-th (l = 1, 2, …, n) FEFM.

Then, the output of the last GFFM is fed into an MFEM to enhance spatial detail features by designing a multi-directional attention mechanism. The enhanced features are fed into a convolutional layer to achieve feature integration. Finally, a jump link is used between the integrated features and UI_{LH} to obtain the final fusion result $\hat{I}_{\mathrm{HS}}$. The above operations can be expressed as follows.

$$\hat{I}_{\mathrm{HS}} = \mathrm{Conv}B\left(\mathrm{MFEM}\left(FI_G^n\right)\right) + UI_{\mathrm{LH}} \tag{3}$$

where Conv$B(\cdot)$ denotes a convolution block that includes a 3 × 3 convolution operation, a LeakReLU function, and a 3 × 3 convolution operation, and MFEM$(\cdot)$ denotes the MFEM.

Next, we will provide a detailed introduction to the main components in the network.

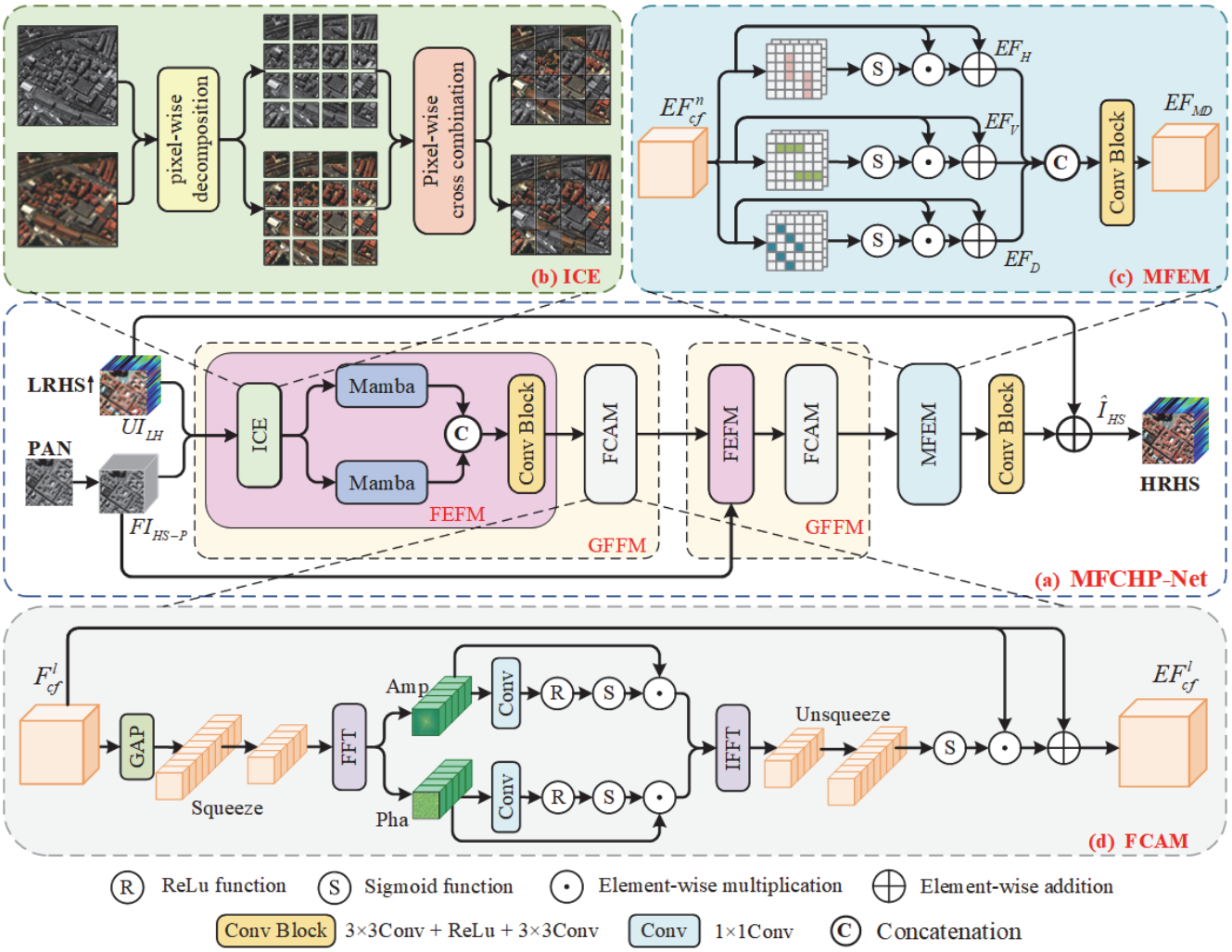


Figure 1 Overall architecture of HPN-ICE, showing ICE for cross-modal feature embedding, FCAM for spectra lenhancement in the frequency domain, and MFEM for multi-directional spatial detail extraction

2.1 Feature Embedding Fusion Module (FEFM)

At present, most HS pansharpening networks extract features from two types of images separately, which makes it difficult to establish cross dependencies between two modal features, resulting in inaccurate feature fusion. To address this issue, we constructed an FEFM consisting of information cross embedding (ICE) strategy and two Mamba blocks. ICE is designed to achieve the embedding of pixel information between two modal features FI_G^{l-1} and $FI_{\mathrm{HS\_P}}$, in order to obtain two features $CF_{\mathrm{HS\_P}}^{l}$ and $CF_{\mathrm{P\_HS}}^{l}$ with cross-modal information. $CF_{\mathrm{HS\_P}}^{l}$ represents the cross-modal features obtained by embedding the LRHS image information in the PAN image features, and $CF_{\mathrm{P\_HS}}^{l}$ represents the cross-modal features obtained by embedding PAN image information in the LRHS image features. Considering that the Mamba model [37, 38] has the advantage of lower computational complexity compared to Transformer, we adopt two Mamba blocks to extract long-distance features from two cross-modal features $CF_{\mathrm{HS\_P}}^{l}$ and $CF_{\mathrm{P\_HS}}^{l}$. The outputs from two Mamba blocks are concatenated and integrated through a convolution layer to obtain the coarse fusion features F_{cf}^{l}. The operation of FEFM can be expressed as follows.

$$\left\{CF_{\mathrm{HS\_P}}^{l}, F_{\mathrm{P\_HS}}^{l}\right\} = \mathrm{ICE}\left(FI_G^{l-1}, FI_{\mathrm{HS\_P}}\right) \tag{4}$$

$$F_{cf}^{l} = \mathrm{Conv}B\left(\mathrm{Cat}\left(M\left(CF_{\mathrm{HS\_P}}^{l}\right), M\left(CF_{\mathrm{P\_HS}}^{l}\right)\right)\right) \tag{5}$$

where ICE(·) and $M(\cdot)$ denote the ICE and Mamba block, and Cat(·) denotes the concatenation operation. The execution process of FEFM is as follows.

The ICE strategy shown in Fig. 1b is defined to execute the mutual embedding of two types of image information. Because PAN and LRHS images are obtained from the same region, their corresponding positions represent the same structural information. In response to this characteristic, we define the ICE strategy for the interaction of features between two modal images, Moreover, it helps to establish cross dependencies in later feature extraction. Specifically, sampling information from odd rows and columns in $FI_{\mathrm{HS\_P}}$ is embedded into the corresponding positions in $FI_{\mathrm{HS\_P}}$ or F_{cf}^{l} to obtain the cross-modal features $CF_{\mathrm{P\_HS}}^{l}$. Similarly, the sampling information from even rows and columns in $FI_{\mathrm{HS\_P}}$ or F_{cf}^{l} is embedded into the corresponding positions in $FI_{\mathrm{HS\_P}}$ to obtain the cross-modal features $CF_{\mathrm{HS\_P}}^{l}$. The above operations can be expressed as follows.

$$\begin{aligned} CF_{\mathrm{P\_HS}}^{l} &= FI_{G}^{l}(i,j) \\ &= \begin{cases} FI_{\mathrm{HS\_P}}^{l}(i,:) \text{ if } i = 1, ..., H \text{ and } i\mathrm{mod}2 = 1 \\ FI_{\mathrm{HS\_P}}^{l}(:,j) \text{ if } j = 1, ..., W \text{ and } j\mathrm{mod}2 = 1 \end{cases} \end{aligned} \tag{6}$$

$$\begin{aligned} CF_{\mathrm{HS\_P}}^{l} &= FI_{\mathrm{HS\_P}}^{l}(i,j) \\ &= \begin{cases} FI_{G}^{l}(i,:) \text{ if } i = 1, ..., H \text{ and } i\mathrm{mod}2 = 0 \\ FI_{G}^{l}(:,j) \text{ if } j = 1, ..., W \text{ and } j\mathrm{mod}2 = 0 \end{cases} \end{aligned} \tag{7}$$

Eq. (6) and Eq. (7) show how ICE embeds cross-modal features, ensuring each modality captures structural details from the other, which in turn enhances cross-dependency between the two modalities during subsequent feature extraction.

2.2 Frequency Channel Attention Module (FCAM)

We construct an FCAM, as shown in Fig. 1d, which employs channel attention mechanisms for amplitude and phase components in the frequency domain. Specifically, firstly, a global average pooling operation is applied to the input feature maps F_{cf}^{l}. to obtain a vector of size $C \times 1 \times 1$. To reduce redundancy channel information, a channel compression operation is performed to reduce the number of channels in the vector. In addition, FFT is used to obtain the amplitude and phase components $\mathcal{A}\left(F_{cf}^{l}\right)$ and $\mathcal{P}\left(F_{cf}^{l}\right)$ of the vector. The operations can be represented as follows.

$$\left\{\mathcal{A}\left(F_{cf}^{l}\right), \mathcal{P}\left(F_{cf}^{l}\right)\right\} = O\left(\mathrm{Seq}\left(\mathrm{GAP}\left(F_{cf}^{l}\right)\right)\right) \tag{8}$$

where GAP(·) denotes the global average pooling, Seq(·) denotes the channel compression operation, and O denotes the FFT operation.

Then, channel attention operation is used for amplitude and phase components to enhance image features in the channel dimension. These operations are as follows.

$$\mathcal{EA}\left(F_{cf}^{l}\right) = S\left(\sigma\left(\mathrm{Conv}_{1\text{-}1}\left(\mathcal{A}\left(F_{cf}^{l}\right)\right)\right)\right) \odot \mathcal{A}\left(F_{cf}^{l}\right) \tag{9}$$

$$\mathcal{EP}\left(F_{cf}^{l}\right) = S\left(\sigma\left(\mathrm{Conv}_{1\text{-}1}\left(\mathcal{P}\left(F_{cf}^{l}\right)\right)\right)\right) \odot \mathcal{P}\left(F_{cf}^{l}\right) \tag{10}$$

where $S(\cdot)$ and σ represent the Sigmoid and LeakyReLU activation functions, respectively. $\mathrm{Conv}_{1\times1}(\cdot)$ denotes the 1×1 convolution operation, and $\odot$ represents the element-wise multiplication. $\mathcal{EA}(\cdot)$ and $\mathcal{EP}(\cdot)$ denote the enhanced amplitude and phase components.

Finally, the inverse fast fourier transform (IFFT) is used for the enhanced amplitude and phase components to reconstruct the vector, in order to obtain enhanced channel features in the spatial domain. Channel expansion operation is adopted to expand the channel dimension of constructed vector, and the sigmoid function is used to generate the weighting coefficients. These coefficients are used to weight the input features in the channel dimension, and an addition operation is employed to achieve spectral enhancement of the input features. These operations are as follows.

$$EF_{cf}^{l} = \sigma\left(\mathrm{Exp}\left(O_{I}\left(\mathcal{A}\left(F_{cf}^{l}\right), \mathcal{P}\left(F_{cf}^{l}\right)\right)\right)\right) \odot F_{cf}^{l} + F_{cf}^{l} \tag{11}$$

where $O_I(\cdot)$ represents the IFFT operation, and EF_{cf}^{l} represents the enhanced features.

2.3 Multi-Directional Feature Enhancement Module (MFEM)

To enhance the detail features in the fused features, we design a MFEM to enhance the spatial detail features by designing multi-directional attention weights. As shown in the Fig. 1c, three directional convolution operations are firstly performed on the input feature maps to achieve feature extraction in the horizontal, vertical, and diagonal directions, and then the sigmoid function is used to generate three directional attention maps. These attention maps are used to enhance features in three directions. The feature enhancement operations in three directions are as follows.

$$EF_{x} = S\left(\mathrm{Conv}_{x}\left(EF_{cf}^{n}\right)\right) \cdot EF_{cf}^{n} + EF_{cf}^{n}, x = H, V, D \tag{12}$$

where $\mathrm{Conv}_{H}(\cdot)$, $\mathrm{Conv}_{V}(\cdot)$ and $\mathrm{Conv}_{D}(\cdot)$ convolution operations with a kernel size of 3 in the horizontal, vertical, and diagonal directions, respectively.

Finally, the enhanced features in three directions are concatenated and are passed through a convolutional layer to generate the final enhanced features EF_{MD}. The operations are as follows.

$$EF_{\mathrm{MD}} = \mathrm{Conv}B\left(\mathrm{Concat}\left(EF_H, EF_V, EF_D\right)\right) \tag{13}$$

where Concat(·) represents the concatenation operation.

2.4 Loss Function

To maintain both spatial and spectral fidelity in the fusion results, a joint loss function $\mathcal{L}_{\mathrm{total}}$ consisting of reconstruction loss $\mathcal{L}_1$ and spectral loss $\mathcal{L}_{\mathrm{SAM}}$ is defined to guide the network training. This function $\mathcal{L}_{\mathrm{total}}$ can be expressed as follows.

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_1 + \lambda \mathcal{L}_{\mathrm{SAM}} \tag{14}$$

$$\mathcal{L}_1 = \frac{1}{N}\sum_{n}^{N}\left\| I_{\mathrm{GT}}^{n} - I_f^{n} \right\|_1 \tag{15}$$

$$\mathcal{L}_{\mathrm{SAM}} = \frac{1}{N}\sum_{n}^{N}\frac{1}{\pi}\arccos\left(\frac{I_{\mathrm{GT}}^{n}\cdot I_f^{n}}{\left\|I_{\mathrm{GT}}^{n}\right\|_2 \cdot \left\|I_f^{n}\right\|_2}\right) \tag{16}$$

where λ is a weighting factor set to 0.001 based on empirical evidence, N is the batch size, I_{GT}^{n} and I_f^{n} represent the ground truth (GT) image and fused image, and $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ norm, arccos(·) denotes the inverse cosine function, π is the symbol for PI, and $\|\cdot\|_2$ denotes the $\mathcal{L}_2$ norm.

3 EXPERIMENTS

3.1 Datasets, Metrics, and Training Details

To demonstrate the performance of the proposed algorithm, we conducted extensive experiments on three commonly used datasets, including Pavia Center dataset [39], Botswana dataset [40], and Chikusei dataset [41]. The datasets were processed following the Wald's protocol [42] and set according to Bandara [43]. We compared HPN-ICE with several SOAT methods, including one traditional method: CNMF [16] and several deep leaning (DL)-based methods: DHP-DARN [44], DIP-Hyperkite [43], Hyper-DSNet [27], FPFNet [25], HyperRefiner [30], and Tree-SNet [45]. Four objective metrics were used on the simulated datasets: SSIM, SAM, ERGAS, and PSNR [44]. We retrained all DL-based methods with Python 3.9 and PyTorch 1.13 on Ubuntu 20.04 system with a NVIDIA RTX A6000. The proposed structure was trained for 1000 epochs using the Adam optimizer. The initial learning rate was set to 0.001, and inference was performed every five rounds of training. When the loss of the verification set no longer decreased for five consecutive rounds, the learning rate was attenuated by half.

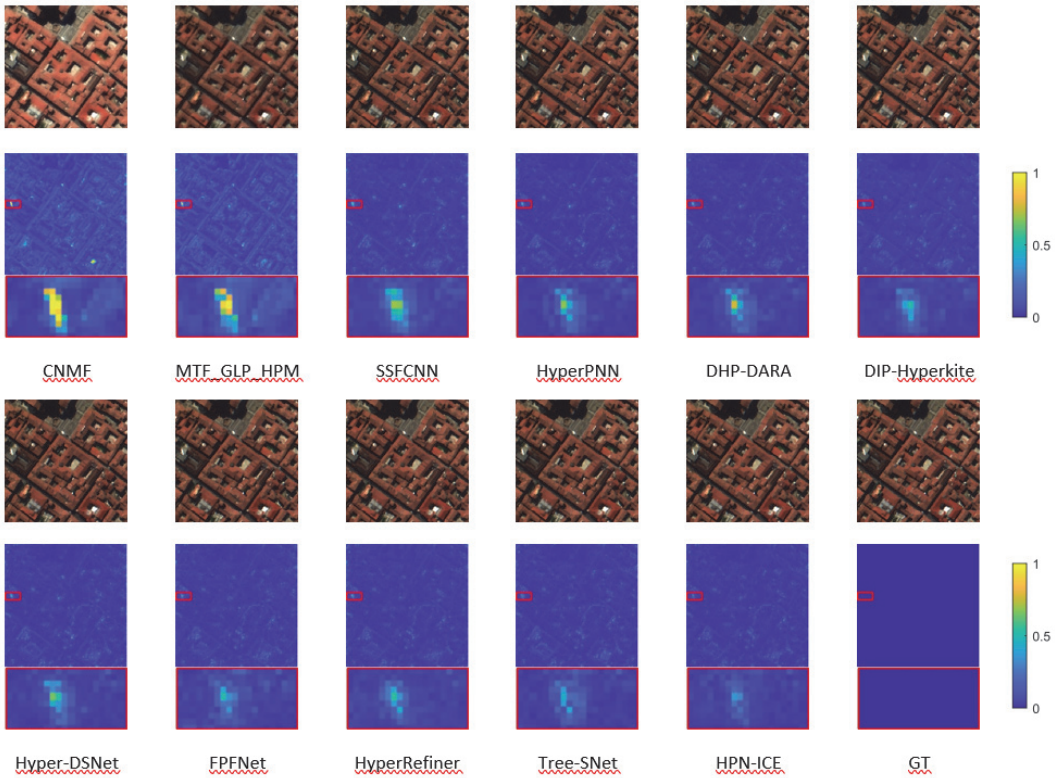


Figure 2 Fusion results on the Pavia Center dataset

Tab. 1 summarizes the objective evaluation metrics of the fusion results for different methods across three datasets. The best results are highlighted in bold, while the second-best results are underlined. It can be observed that our proposed method outperforms other methods in most evaluation metrics. As shown in Fig. 2, Fig. 3, and Fig. 4, we present the fusion results of different methods on three public datasets. To more clearly illustrate the differences between the fusion results, we display a zoomed-in local region in the bottom-left corner and the Mean Absolute Error (MAE) maps between the fusion results and the GT in the bottom-right corner. It can be observed that our

method is visually closest to the GT images, demonstrating better spatial and spectral fidelity.

To validate the spectral fidelity of the proposed method, as shown in Fig. 5, we randomly selected three points from three datasets (Pavia Center (12,47), Botswana (60,10), and Chikusei (8,191)) to illustrate their spectral difference curves across different methods. From the figure, it is evident that our method exhibits the smallest spectral differences, demonstrating its superior spectral fidelity.

Table 1 The average quantitative results on the Pavia Center, Botswana, and Chikusei datasets

Dataset	Methods	SSIM(↑)	SAM(↓)	SCC(↑)	RMSE × 10^{-2}(↓)	ERGAS(↓)	PSNR(↑)
PaviaCenter	CNMF	0.8913	7.2364	0.9462	2.7090	4.8195	32.0244
	MTF_GLP_HPM	0.8911	8.8859	0.9267	3.1100	5.9379	31.3672
	SSFCNN	0.9426	5.7787	0.9761	1.8104	3.3362	35.5897
	HyperPNN	0.9520	5.2954	0.9806	1.5656	2.9419	37.0370
	DHP-DARN	0.9545	5.2685	0.9824	1.5299	2.8876	37.1592
	DIP-Hyperkite	0.9527	5.3630	0.9809	1.5661	2.9363	36.9586
	Hyper-DSNet	0.9539	5.1541	0.9822	1.5245	2.8929	37.1725
	FPFNet	0.9520	5.3962	0.9822	1.5297	2.8992	37.2121
	HyperRefiner	0.9588	5.0461	0.9856	1.4265	2.7748	37.7193
	Tree-SNet	0.9594	4.8628	0.9855	1.3815	2.6432	38.1569
	HPN-ICE	0.9642	4.5331	0.9879	1.2617	2.4530	38.9527
Botswana	CNMF	0.9193	2.3018	0.9549	4.5080	3.8151	33.6006
	MTF_GLP_HPM	0.9352	2.2292	0.9637	3.1770	2.7787	36.9958
	SSFCNN	0.9386	1.9164	0.9818	1.3880	4.1864	42.3542
	HyperPNN	0.9522	1.9909	0.9778	1.3955	1.8251	42.5861
	DHP-DARN	0.9453	1.9877	0.9818	1.3589	2.5892	41.1629
	DIP-Hyperkite	0.9576	1.7728	0.9825	1.2897	1.8724	42.4773
	Hyper-DSNet	0.9554	1.8418	0.9809	1.2998	1.4724	43.7004
	FPFNet	0.9621	1.8960	0.9843	1.2396	1.4639	44.1665
	HyperRefiner	0.9663	1.6751	0.9864	1.1655	1.3997	44.2455
	Tree-SNet	0.9671	1.5787	0.9875	1.0738	1.3039	45.2278
	HPN-ICE	0.9689	1.4764	0.9888	1.0102	1.2641	45.6354
Chikusei	CNMF	0.8970	3.7933	0.9013	1.9730	6.6018	35.6701
	MTF_GLP_HPM	0.8458	7.7090	0.7052	4.5440	32.0045	30.6859
	SSFCNN	0.9656	2.4829	0.9712	1.0188	4.3537	40.8220
	HyperPNN	0.9644	2.5668	0.9699	1.0318	4.5882	40.6471
	DHP-DARN	0.9686	2.4217	0.9743	0.9837	4.2614	41.0419
	DIP-Hyperkite	0.9702	2.3207	0.9761	0.9246	4.0896	41.6503
	Hyper-DSNet	0.9705	2.3315	0.9765	0.9197	4.0150	41.6990
	FPFNet	0.9759	2.3260	0.9826	0.8243	3.8711	42.3698
	HyperRefiner	0.9777	2.1112	0.9829	0.8046	3.5126	42.9958
	Tree-SNet	0.9795	2.0608	0.9847	0.7487	3.4806	43.2155
	HPN-ICE	0.9797	1.9205	0.9844	0.7539	3.3000	43.5860

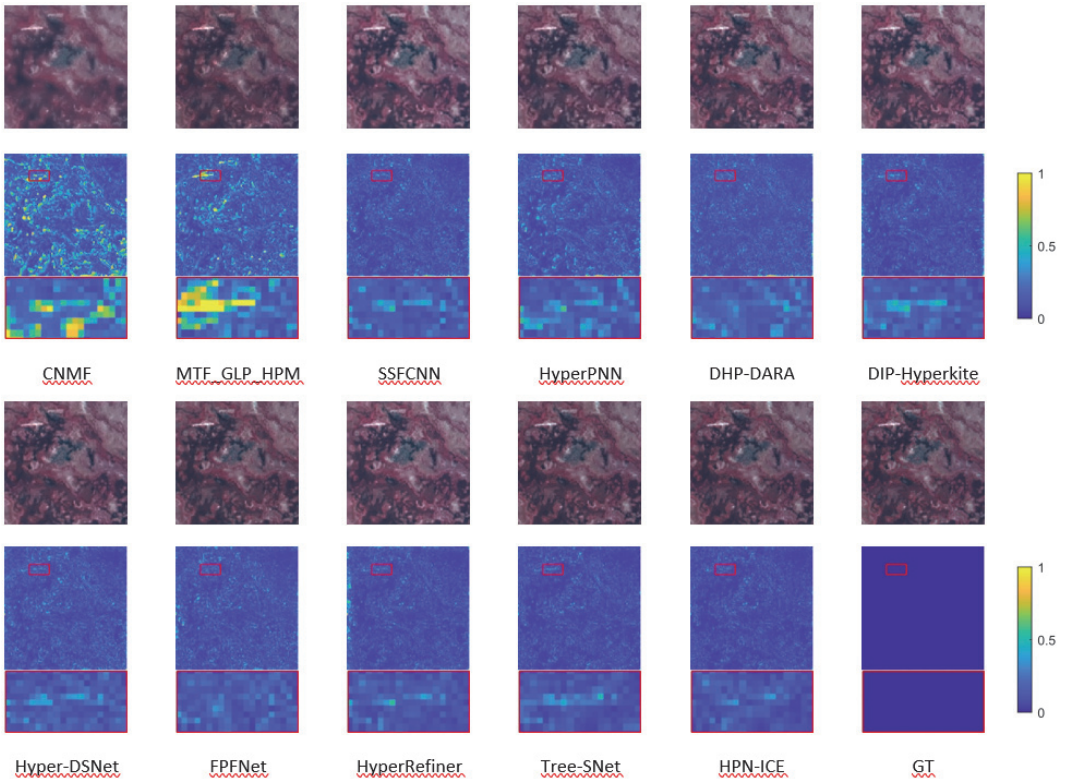


Figure 3 Fusion results on the Pavia Botswana dataset

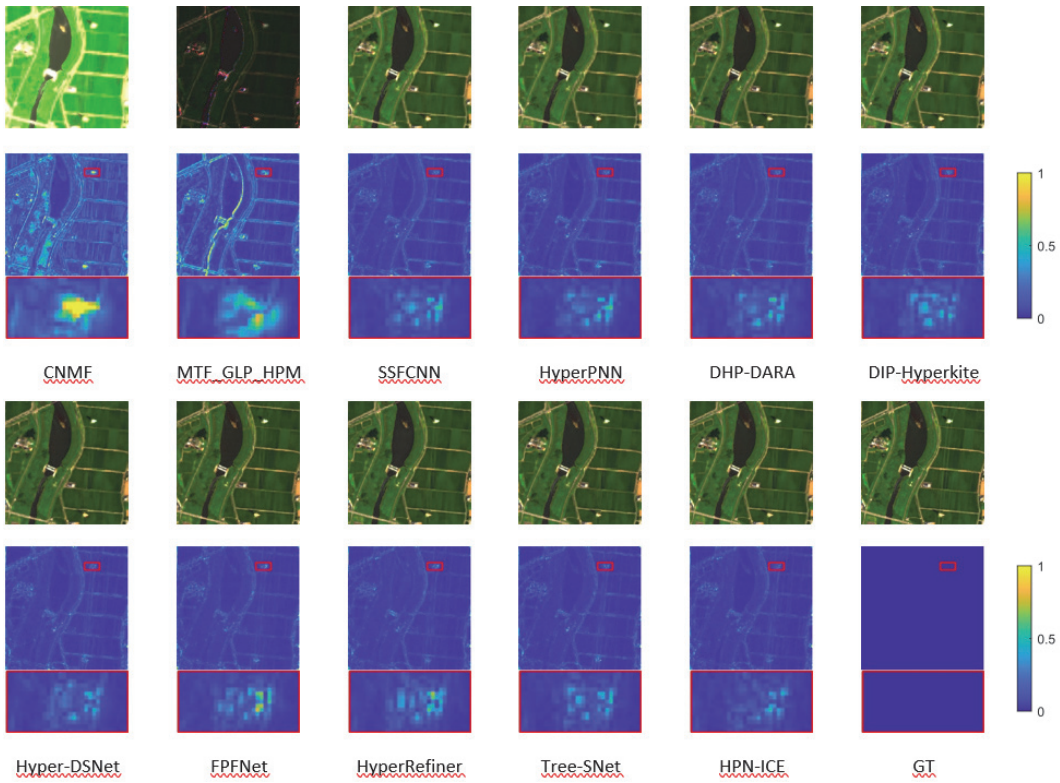
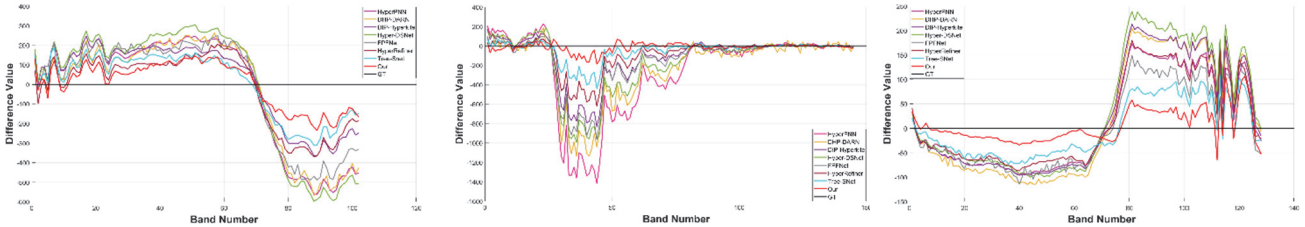


Figure 4 Fusion results on the Chikusei Dataset



Pavia Center(12,47) Botswana(60,10) Chikusei(8, 191)

Figure 5 Spectral difference curves between GT images and pansharpening results

3.2 Ablation Study

To validate the effectiveness of components in HPN-ICE, we conducted ablation studies on the Pavia Center dataset.

3.2.1 Number of GFFM Layers

We conducted ablation experiments on the number of layers of the GFFM. The results are shown in Tab. 2. The network performance reaches its optimum when the number of GFFM is 2. Therefore, the number of GFFM is set to 2 in our work.

3.2.2 Effectiveness of ICE

We conducted ablation experiments by replacing the ICE structure with the addition operation between PAN and HS images (Method A) or the concatenation operation (Method B), respectively. As shown in Tab. 3, using the proposed ICE module can significantly improve the performance of the network. This confirms that ICE contributes more effectively than simple addition or concatenation by strengthening cross-modal dependencies.

3.2.3 Effectiveness of FCAM

We replaced the FCAM with a convolutional layer with a kernel size of 3×3 and a channel attention (CA) module [46], respectively. As shown in Tab. 4, the model using the FCAM outperforms other models in all performance metrics.

3.2.4 Effectiveness of Ablation of MFEM

We replaced the convolutional operations of the MFEM with multiple convolution operations with a kernel size of 3 × 3, as shown in Tab. 5. The network containing the MFEM achieves better performance in objective metrics, especially in terms of spectral fidelity and spatial fidelity.

3.2.5 Model Complexity

We compared HPN-ICE with five DL-based methods in terms of parameter count and computational complexity. The detailed results are shown in Tab. 6. Although our proposed method has higher computational complexity compared to DHP-DARN, Hyper-DSNet, and HyperRefiner, it demonstrates superior PSNR performance, showcasing significantly better HS pansharpening results.

This highlights a clear computation trade-off, namely that although HPN-ICE incurs marginally higher computational cost, the significant gains in spectral fidelity and spatial detail make this trade-off worthwhile, especially for tasks demanding high reconstruction quality.

3.2.6 Classification Application

To further demonstrate the effectiveness of the proposed method, we performed classification experiments on the HS pansharpening results from the Chikusei dataset, obtained using DL-based methods. The fusion results of different methods were classified using the K-Means algorithm in the ENVI software. The classification process was configured with six categories and a maximum iteration count of 10. The classification results are presented in Fig. 6. From the magnified regions, it is evident that the classification results of our method align most closely with those of the GT image. Additionally, overall accuracy (OA↑) and the kappa coefficient (K↑) metrics in Tab. 7 confirm the superior fusion performance of our proposed method. Importantly, classification accuracy serves as a meaningful proxy for fusion quality, as reliable spectral-spatial preservation in the pansharpened images directly improves downstream classification performance, thereby highlighting the practical relevance of our method.

Table 2 Ablation study of gffm layers numbers

	SSIM(↑)	SAM(↓)	SCC(↑)	RMSE × 10^{-2}(↓)	ERGAS(↓)	PSNR(↑)
w/o GFFM	0.9619	4.6582	0.9864	1.3256	2.5532	38.5219
GFFM × 1	0.9633	4.5824	0.9873	1.2875	2.4938	38.7685
GFFM × 2	0.9642	4.5331	0.9879	1.2617	2.4530	38.9527
GFFM × 3	0.9637	4.5668	0.9875	1.2773	2.4779	38.8421

Table 3 Ablation study of ice

	SSIM(↑)	SAM(↓)	SCC(↑)	RMSE × 10^{-2}(↓)	ERGAS(↓)	PSNR(↑)
A	0.9640	4.5535	0.9875	1.2759	2.4761	38.8606
B	0.9632	4.6114	0.9871	12954	2.5106	38.7217
PCFM	0.9642	4.5331	0.9879	1.2617	2.4530	38.9527

Table 4 Ablation study of fcam

	SSIM(↑)	SAM(↓)	SCC(↑)	RMSE × 10^{-2}(↓)	ERGAS(↓)	PSNR(↑)
Conv	0.9635	4.5905	0.9874	1.2825	2.4893	38.8075
CA	0.9622	4.6542	0.9868	1.3120	2.5278	38.6106
w/FCAM	0.9642	4.5331	0.9879	1.2617	2.4530	38.9527

Table 5 Ablation study of mfem

	SSIM(↑)	SAM(↓)	SCC(↑)	RMSE × 10^{-2}(↓)	ERGAS(↓)	PSNR(↑)
w/o MFEM	0.9620	4.6371	0.9866	1.3173	2.5344	38.5806
HPN-ICE	0.9642	4.5331	0.9879	1.2617	2.4530	38.9527

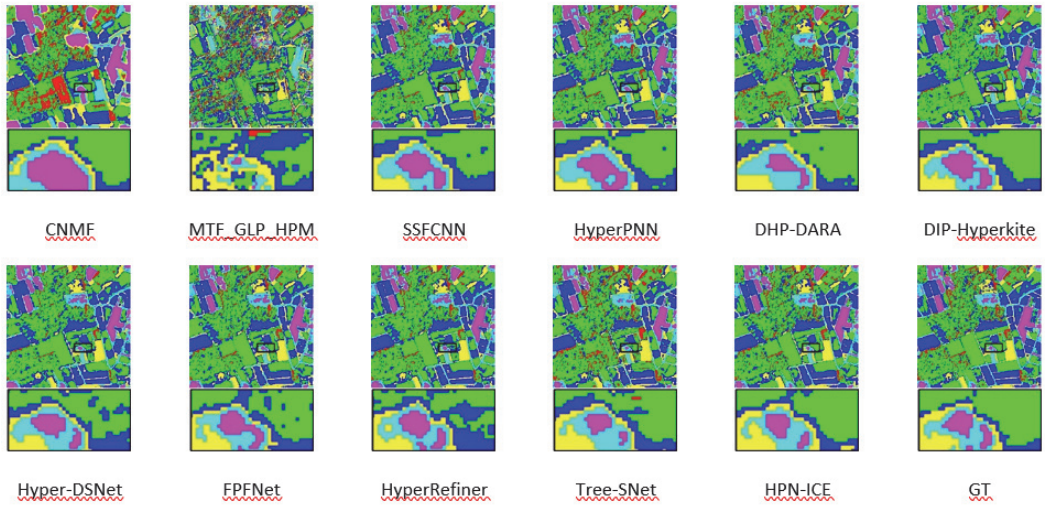


Figure 6 Results of classification experiments on Chikusei dataset

Table 6 Comparison of params and flops of different methods

Methods	Parameters (*M*)	FLOPS (*G*)	PSNR(↑)
SSFCNN	0.71	46.25	40.8220
HyperPNN	0.52	8.86	40.6471
DHP-DARN	0.45	29.06	41.0419
DIP-Hyperkite	0.97	212.64	41.6503
Hyper-DSNet	0.53	21.89	41.6990
FPFNet	22.34	174.92	42.3698
HyperRefiner	19.32	85.00	42.9958
Tree-SNet	9.12	205.00	43.2155
HPN-ICE	4.08	163.23	43.5860

Table 7 Objective evaluation of classification results

Methods	OA(↑)	K(↑)
CNMF	0.6062	0.4844
MTF_GLP_HPM	0.4654	0.2655
SSFCNN	0.7875	0.7069
HyperPNN	0.7708	0.6843
DHP-DARN	0.7756	0.6936
DIP-Hyperkite	0.8087	0.7329
Hyper-DSNet	0.7942	0.7142
FPFNet	0.8102	0.7373
HyperRefiner	0.8392	0.7773
Tree-SNet	0.7978	0.7243
HPN-ICE	0.8556	0.8007

4 CONCLUSION

This paper proposed HPN-ICE, a hyperspectral pansharpening network that integrates three key modules: the Feature Embedding Fusion Module (FEFM), the Frequency Channel Attention Module (FCAM), and the Multi-directional Feature Enhancement Module (MFEM). The FEFM, based on an information cross embedding strategy, enables effective interaction between hyperspectral and panchromatic features, ensuring that cross-modal dependencies are captured. The FCAM enhances spectral fidelity by modeling features in the frequency domain, while the MFEM reinforces structural and textural details through multi-directional attention. Extensive experiments conducted on three benchmark datasets, Pavia Center, Botswana, and Chikusei, demonstrated that HPN-ICE consistently outperforms traditional and state-of-the-art deep learning methods in both quantitative metrics (SSIM, SAM, ERGAS, PSNR) and visual quality. Ablation studies further confirmed the effectiveness of the proposed modules, showing that ICE improves cross-modal feature fusion, FCAM strengthens spectral representation, and MFEM enhances spatial detail fidelity. Classification experiments additionally validated the practical utility of the fused hyperspectral images for downstream applications. In summary, HPN-ICE achieves a strong balance between spectral preservation and spatial detail enhancement, offering a robust solution for hyperspectral pansharpening. Beyond benchmark testing, the method shows promise for real-world applications such as environmental monitoring, mineral exploration, precision agriculture, and urban analysis, where high-resolution hyperspectral data can significantly improve decision-making. Future work will focus on reducing computational complexity, extending the model to real-time scenarios, and exploring its adaptability to other multi-modal remote sensing tasks.

5 REFERENCES

[1] Ghrefat, H., Awawdeh M., Howari F., & AlRawabdeh A. (2023). Mineral exploration using multispectral and hyperspectral remote sensing data. G*eoinformatics for Geosciences*, 197-222. https://doi.org/10.1016/B978-0-323-98983-1.00013-2

[2] Li, J., Cai, Y., Li, Q., Kou, M., & Zhang, T. (2024). A review of remote sensing image segmentation by deep learning methods. *International Journal of Digital Earth*, *17*(1). https://doi.org/10.1080/17538947.2024.2328827

[3] Dong, W., Liang, J., & Xiao, S., (2020). Saliency analysis and gaussian mixture model-based detail extraction algorithm for hyperspectral pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(8), 5462-5476. https://doi.org/10.1109/TGRS.2020.2966550

[4] Shah, P., Younan, H., & King, R. (2008). An efficient pan-sharpening method via a combined adaptive pca approach and contourlets. *IEEE transactions on Geoscience and Remote Sensing*, *46*(5), 1323-1335. https://doi.org/10.1109/TGRS.2008.916211

[5] Tu, T., Su S., Shyu, H., & Huang, S. (2001). A new look at ihs-like image fusion methods. *Information fusion*, *2*(3), 177-186. https://doi.org/10.1016/S1566-2535(01)00036-7

[6] Aiazzi, B., Baronti, S., & Selva, M., (2007). Improving component substitution pansharpening through multivariate regression of ms + pan data. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(10), 3230-3239. https://doi.org/10.1109/TGRS.2007.901007

[7] Laben, A. & Brower, V. (2000). *Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening*. U.S. Patent 6,011,875. 2000-1-4.

[8] Jinju, J., Santhi, N., Ramar, K., & Bama S. (2019). Spatial frequency discrete wavelet transform image fusion technique for remote sensing applications. *Engineering Science and Technology*, *22*(3), 715-726. https://doi.org/10.1016/j.jestch.2019.01.004

[9] Otazu X., Gonzalez-Aud, M., Fors, O., & Nu'nez, J. (2005). Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods. I*EEE Transactions on Geoscience and Remote Sensing*, *43*(10), 2376-2385. https://doi.org/10.1109/TGRS.2005.856106

[10] Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., & Selva M. (2006). Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. *Photogrammetric Engineering & Remote Sensing*, *72*(5), 591-596. https://doi.org/10.14358/PERS.72.5.591

[11] Vivone, G., Restaino, R., Mura, D., Licciardi, Giorgio., & Chanussot, J. (2013). Contrast and error-based fusion schemes for multispectral image pansharpening. *IEEE Geoscience and Remote Sensing Letters*, *11*(5), 930-934. https://doi.org/10.1109/LGRS.2013.2281996

[12] Wei, Q., Dobigeon, N., & Tourneret J. (2015). Fast fusion of multi-band images based on solving a sylvester equation. *IEEE Transactions on Image Processing*, *24*(11), 4109-4121. https://doi.org/10.1109/TIP.2015.2458572

[13] Lin, B., Tao, X., Li, S., Dong, L., & Lu, J. (2016). Variational bayesian image fusion based on combined sparse representations. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1432-1436. https://doi.org/10.1109/ICASSP.2016.7471913

[14] Hardie, C., Eismann, M., & Wilson, L., (2004). Map estimation for hyperspectral image resolution enhancement using an auxiliary sensor. *IEEE Transactions on Image Processing*, *13*(9), 1174-1184. https://doi.org/10.1109/TIP.2004.829779

[15] Lin, H. & Zhang, A. (2017). Fusion of hyperspectral and panchromatic images using improved hysure method. *2nd international conference on image, vision and computing*, 489-493. https://doi.org/10.1109/ICIVC.2017.7984604

[16] Yokoya, N., Yairi, T., & Iwasaki, A., (2011). Coupled nonnegative matrix factorization (cnmf) for hyperspectral and multispectral data fusion: Application to pasture classification. *IEEE International Geoscience and Remote Sensing Symposium*, 1779-1782. https://doi.org/10.1109/IGARSS.2011.6049465

[17] Karoui, S., Djerriri, K., and Boukerch, I. (2016). Pansharpening multispectral remote sensing data by multiplicative joint nonnegative matrix factorization. *International Journal of Remote Sensing*, *37*(4), 805-818. https://doi.org/10.1080/01431161.2015.1137650

[18] Karoui, S., Deville, Y., Benhalouche, Z., & Boukerch, I., (2016). Hypersharpening by joint-criterion nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, *55(*3), 1660-1670. https://doi.org/10.1109/TGRS.2016.2628889

[19] Qu, J., Li, Y., & Dong, W. (2018). Guided filter and principal component analysis hybrid method for hyperspectral pansharpening. *Journal of Applied Remote Sensing*, *12*(1), 015003. https://doi.org/10.1117/1.JRS.12.015003

[20] Kranjčić, N. & Župan, R.,(2018). Satellite-based hyperspectral imaging and cartographic visualization of bark beetle forest damage for the city of Čabar. *Tehnički glasnik*, *12*(1), 39-43. https://doi.org/10.31803/tg-20171219085721

[21] Hao, Z. & Ma, J. (2021). Gtp-pnet: A residual learning network based on gradient transformation prior for pansharpening. *ISPRS Journal of Photogrammetry and Remote Sensing*, *172*, 223-239. https://doi.org/10.1016/j.isprsjprs.2020.12.014

[22] Li, J., Cui, R., Li, B., Song, R., Li, Y., Dai, Y., & Du, Q. (2020). Hyperspectral image super-resolution by band attention through adversarial learning. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(6), 4304-4318. https://doi.org/10.1109/TGRS.2019.2962713

[23] Qu, J., Hou, S., Dong, W., Xiao, S., Du, Q., & Li Y. (2021). A dual-branch detail extraction network for hyperspectral pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-13. https://doi.org/10.1109/TGRS.2021.3130420

[24] He, L., Zhu, J., Li, J., Plaza, A., Chanussot, J., & Li, B. (2019). Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(8), 3092-3100. https://doi.org/10.1109/JSTARS.2019.2917584

[25] Dong, W., Yang, Y., Qu, J., Li, Y., Yang, Y., & Jia, X. (2023). Feature pyramid fusion network for hyperspectral pansharpening. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(1), 1555-1567. https://doi.org/10.1109/TNNLS.2023.3325887

[26] Wang, H., Gong, M., Mei, X., Zhang, H., & Ma, J. (2024). Deep unfolded network with intrinsic supervision for pan-sharpening. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*, 5419-5426. https://doi.org/10.1609/aaai.v38i6.28350

[27] Zhuo, Y., Zhang, T., Hu, J., Dou, H., Huang, T., & Deng, L. (2022). A deep-shallow fusion network with multidetail extractor and spectral attention for hyperspectral pansharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 7539-7555. https://doi.org/10.1109/JSTARS.2022.3202866

[28] Dong, W., Zhang, T., Qu, J., Xiao, S., Liang, J., & Li, Y. (2021). Laplacian pyramid dense network for hyperspectral pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-13. https://doi.org/10.1109/TGRS.2021.3076768

[29] Liu, Y., Hu, J., Kang, X., Luo, J., & Fan, S. (2022). Interactformer: Interactive transformer and cnn for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-15. https://doi.org/10.1109/TGRS.2022.3183468

[30] Zhou, B., Zhang, X., Chen, X., Ren, M., & Feng, Z. (2023). Hyperrefiner: a refined hyperspectral pansharpening network based on the autoencoder and self-attention. *International Journal of Digital Earth*, *16*(1), 3268-3294. https://doi.org/10.1080/17538947.2023.2246944

[31] Bandara W. & Patel V. (2022). Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1767-1777. https://doi.org/10.1109/CVPR52688.2022.00181

[32] Shang, Y., Liu, J., Yang, J., & Wu, Z. (2022). A modelinspired approach with transformers for hyperspectral pansharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 7187-7202. https://doi.org/10.1109/JSTARS.2022.3199207

[33] Zhou, H., Liu, Q., & Wang, Y., (2022). Panformer: A transformer based model for pan-sharpening. *2022 IEEE International Conference on Multimedia and Expo*, 1-6. https://doi.org/10.1109/ICME52920.2022.9859770

[34] Gu, A., Goel, K., & Re, C. (2021). Efficiently modeling long sequences with structured state spaces. *arXiv preprint, arXiv:2111.00396*.

[35] Zhou, M., Huang, J., Li, C., Yu, H., Yan, K., Zheng, N., & Zhao, F. (2022). Adaptively learning low-high frequency information integration for pan-sharpening. *Proceedings of the 30th ACM International Conference on Multimedia*, 3375-3384. https://doi.org/10.1145/3503161.3547924

[36] Hwang, K., Yoon, G., Song, J., & Yoon, S. (2024). Fusing bi-directional global-local features for single image superresolution. *Engineering Applications of Artificial Intelligence*, *127*, 107336. https://doi.org/10.1016/j.engappai.2023.107336

[37] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision mamba: Efficient visual representation learning with bidirectional state space model. *Proceedings of the 41st International Conference on Machine Learning*, 62429-62442.

[38] He, X., Cao, K., Zhang, J., Yan, K., Wang, Y., Li, R., Xie, C., Hong, D., & Zhou, M. (2024). Pan-mamba: Effective pan-sharpening with state space model. *Information Fusion*, 102779. https://doi.org/10.1016/j.inffus.2024.102779

[39] Plaza, A., Benediktsson, J., Boardman, J. et al. (2009). Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, *113*, S110-S122. https://doi.org/10.1016/j.rse.2007.07.028

[40] Ungar, S., Pearlman, J., Mendenhall, J., & Reuter, D. (2003). Overview of the earth observing one (eo-1) mission. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(6), 1149-1159. https://doi.org/10.1109/TGRS.2003.815999

[41] Yokoya, N. & Iwasaki, A. (2016). *Airborne hyperspectral data over Chikusei*. Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep.SAL-2016-05-27, 5(5), 5.

[42] Wald, L., (2000). Quality of high resolution synthesised images: Is there a simple criterion? Third conference Fusion of Earth data: merging point measurements, raster maps and remotely sensed images. *SEE/URISCA*, 99-103.

[43] Bandara, W., Valanarasu, J., & Patel, V. (2021). Hyperspectral pansharpening based on improved deep image prior and residual reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-16. https://doi.org/10.1109/TGRS.2021.3139292

[44] Zheng, Y., Li, J., Li, Y., Guo, J., Wu, X., & Chanussot, J. (2020). Hyperspectral pansharpening using deep prior and dual attention residual network. *IEEE transactions on Geoscience and Remote Sensing*, *58*(11), 8059-8076. https://doi.org/10.1109/TGRS.2020.2986313

[45] He, L., Ye, H., Xi, D., Li, J., Plaza, A., & Zhang, M. (2024). Tree-structured neural network for hyperspectral

pansharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 2516-2530. https://doi.org/10.1109/JSTARS.2023.3344117

[46] Woo, S., Park, J., Lee J., & Kweon I. (2018). Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

Contact information:

Yan JIN
Tiangong University,
School of Computer Science and Technology,
Tianjin 300387, China
E-mail: jinyan@tiangong.edu.cn