FMPM-DNet: Hyperspectral Pansharpening Dynamic Network Based on Feature Modulation and Probability Mask

Xiaozheng Wang[1]*, Yong Yang[1]†, Shuying Huang[1]*, Hangyuan Lu[2], Weiguo Wan[3], Aoqi Zhao[1]
[1]Tiangong University, Tianjin, China
[2]Jinhua University of Vocational Technology, Jinhua, China
[3]Jiangxi University of Finance and Economics, Nanchang, China
xiaozhengwang95@gmail.com, greatyangy@126.com, huangshuying@tiangong.edu.cn, lhyhziee@163.com,
wanweiguo@jxufe.edu.cn, aoqizhao927@163.com

Abstract

Currently, most Hyperspectral(HS) pansharpening methods have two problems, namely the lack of consideration the spatial variations of HS images and inaccurate feature reconstruction in multi-channel complex mapping relationships, leading to spectral and spatial distortions in the fusion results. To address these issues, we propose a dynamic network based on feature modulation and probability mask (FMPM-DNet) for HS pansharpening, including two stages of spectral-spatial feature modulation and feature reconstruction. In the first stage, to increase the feature representation ability of the model, a wave function is defined based on complex transformation to convert spatial features into wave-like features. On this basis, considering the spatial variations of HS images, a dynamic feature modulation unit (DFMU) is constructed to achieve adaptive modulation and coarse fusion of features by dynamically generating spectral-spatial correction matrix. In the second stage, a feature probability mask unit (FPMU) is designed to realize global feature embedding at different depths and local feature embedding at the same depth to obtain refined fused features. Extensive experiments on three widely used datasets demonstrate that the proposed FMPM-Net achieves significant improvements in both spatial and spectral quality metrics compared to some state-of-the-art (SOTA) methods.

Code — https://github.com/EchoPhD/FMPM-DNet

Introduction

In the optical remote sensing systems, the balance among the spatial resolution, the spectral resolution, and the signal-to-noise ratio (SNR) is a key consideration factor. Single remote sensing system cannot directly acquire high-spatial-resolution hyperspectral (HRHS) images. Typically, spectroscopy imaging systems can capture hundreds of narrow spectral bands at once, yielding hyperspectral (HS) data with abundant spectral information, but often reduce the spatial resolution of HS data. In contrast, panchromatic (PAN) imaging systems can provide single-band images with high spatial resolution. To accommodate requirements for HRHS data in many practical remote sensing applications, such as environmental monitoring, target detection, and classification (Aburaed et al. 2023), one possible approach is to reconstruct HRHS images from low spatial resolution HS (LRHS) images and PAN images. This process is commonly referred to as HS pansharpening.

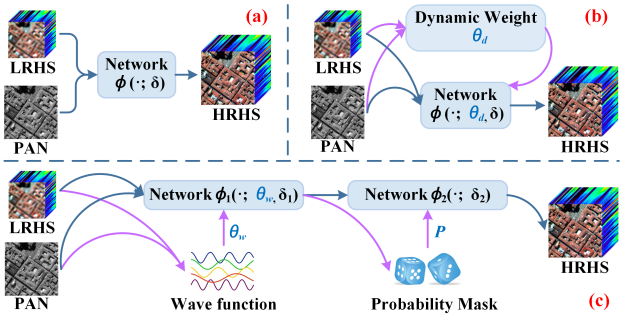


Figure 1: Different network structures. (a) static network, (b) dynamic network, and (c) FMPM-DNet.

The traditional pansharpening methods can be classified into four categories (Li et al. 2024): component substitution (CS) methods (Aiazzi et al. 2007), multiresolution analysis (MRA) methods (Aiazzi et al. 2006), Bayesian (VO) methods (Wei et al. 2015), and matrix decomposition-based methods (Yokoya, Yairi, and Iwasaki 2012). Although the traditional methods are easy to implement and physically interpretable, they often suffer from spatial and spectral distortions due to imprecise definition of prior knowledge and imprecise manual feature extraction (Zhang et al. 2020).

In HS pansharpening, the complex mapping relationship between multiple channels has always been a challenging problem that is difficult to solve. Deep learning (DL)-based methods have been proven to be more effective in handle these problems compared to traditional methods. According to the network architectures shown in Figure 1(a) and (b), DL-based methods can be roughly divided into two categories: methods based on static network and methods based on dynamic network. The methods based on static network (He et al. 2023) aims to directly map PAN and LRHS images to HRHS images by learning networks with fixed parameters. For example, Hyper-DSNet (Zhuo et al.

*These authors contributed equally.
†Corresponding author.

2022) is proposed to preserve spatial details and spectral fidelity by constructing a deep-shallow fusion structure with multi-detail extraction and spectral attention. FPFNet (Dong et al. 2023) is a dual-branch feature pyramid network structure that reconstructs HRHS images through the progressive fusion of multi-scale features. DMOEAD (Wu et al. 2024) is a multi-task, multi-objective driven HS fusion network that enhances the final fusion results by mutually optimizing the shared high-frequency information obtained by CNNs. This type of network structure that learns fixed parameters is prone to ignoring the spatial variations of HS images, resulting in a decrease in the generalization of the model.

The methods based on the dynamic network shown in Figure 1(b) dynamically adjust the learned fixed parameters by learning a set of dynamic mapping weights based on different input images. Tree-SNet (He et al. 2024) constructed a spectral-spatial tree-structured network that addresses the imbalance between spatial and spectral information by using adaptive convolutions. HyperRefiner (Zhou et al. 2023) constructed an up-sampling spectral-preserving network that achieves fine fusion of multi-scale PAN and HS image features by constructing an adaptive cross-attention weight matrix. Although dynamic networks can improve network performance, some networks typically utilize dense connection structures to achieve complex multi-channel feature mapping between HS and PAN images, which leads to feature redundancy issues, resulting in spectral and spatial distortions in the HS pansharping results (He et al. 2024).

In response to the above issues, we propose a dynamic network based on feature modulation and probability mask (FMPM-DNet) for the HS pansharpening, which adopts a dual-stage dynamic network structure shown in Figure 1(c): spectral-spatial feature modulation stage and feature reconstruction stage. In the first stage, a wave function is defined to transform spatial features into wave-like features, which represent both modal features as complex domain features(real and imaginary features). Then, a dynamic feature modulation unit (DFMU) is constructed to modulate LRHS image features using PAN image features, resulting in rough fusion features. In the second stage, to learn the precise mapping relationship between multiple channels, a feature probability mask unit (FPMU) is designed, which can achieve fine fusion feature reconstruction by selectively embedding global and local features. Our main contributions can be summarized as follows:

1. An FMPM-DNet with a spectral-spatial modulation stage and a feature reconstruction stage is proposed for obtaining fusion results possessing spatial and spectral fidelity.

2. A wave function-based DFMU is constructed to realize adaptive modulation and coarse fusion of two modal features by dynamically generating spatial and spectral modulation weights.

3. An FPMU is designed to achieve fine reconstruction of fused features by learning of global and local probability masks.

Proposed Approach

In this section, we propose a FMPM-DNet for HS pansharpening, as shown in Figure 2, which consists of two stages: a spectral-spatial feature modulation stage and a feature reconstruction stage. In the first stage, a sub-network with dual U-Nets is constructed to extract and fuse two modal features at different scales. In each scale layer, we construct a DFMU to achieve coarse fusion of current scale features. In DFMU, a wave function is first defined to transform image features into wave-like features in complex domain. Then, by dynamically generating spatial and spectral modulation weights, the modulation and integration of the two modal wave-like features are achieved. In the second stage, inspired by neural structure search (Zhang et al. 2021), multiple probabilistic reconstruction layers are constructed to gradually achieve fine reconstruction of fused features. In each reconstruction layer, an FPMU is designed to achieve global embedding and local embedding of fused features by utilizing learned global and local probability masks. Below, we provide a detailed introduction to DFMU and FPMU.

Dynamic Feature Modulation Unit (DFMU)

Considering that real and imaginary information in the complex domain can more effectively address complex image data (Tang et al. 2022), a DFMU based on wave function is developed in the complex domain, as shown in Figure 2(a). DFMU achieves rough fusion of LRHS and PAN image features at each scale by modulating the real and imaginary parts of the wave-like features corresponding to each pixel in the LRHS image. This process includes three steps: feature transformation, feature modulation, and feature fusion.

Firstly, in the feature transformation step, a wave function is designed to achieve the transformation of features from the spatial domain to the complex domain. We take the features of the input PAN and HS images as waves $|z| \odot e^{t\theta}$(t is the imaginary unit satisfying $t^2 =$ -1, and $\odot$ is the element-wise product) with amplitude $|z|$ and phase θ. The amplitude represents the intensity of spectral features and spatial details, while the phase represents the directionality of spatial structures and patterns. To embed the wave-like features from the complex domain into the proposed DFMU, allowing for feature modulation using simple addition operations, we use the Euler's formula to represent the waves with real and imaginary parts. The feature transformation of the wave function $WF(\cdot)$ is defined as:

$$WF\left(F_{(x,y)}\right) = \left|z_{(x,y)}\right| \odot \cos\theta_{(x,y)} + t\left|z_{(x,y)}\right| \odot \sin\theta_{(x,y)} \quad x = 1 \cdots, \mathrm{H}; \quad y = 1 \cdots, \mathrm{W} \tag{1}$$

where $|\cdot|$ denotes the absolute value operation. The amplitude $\left|z_{(x,y)}\right|$ is real-value feature of the input feature $F_{(x,y)}$ (F represents PAN image feature f_{pan} or HS image feature f_{hs})at spatial position (x, y). $\theta_{(x,y)}$ denotes phase. H and W denote the height and width of input images, respectively.

In the network, the amplitude is similar to real-value features in traditional model. This element-wise absolute value operation can be absorbed into the phase term, as defined below.

$$\left|z_{(x,y)}\right| e^{t\theta_{(x,y)}} = \begin{cases} z_{(x,y)} e^{t\theta_{(x,y)}}, & z_{(x,y)} > 0 \\ z_{(x,y)} e^{\mathrm{t}(\theta_{(x,y)}+\pi)}, & \text{otherwise} \end{cases} \tag{2}$$

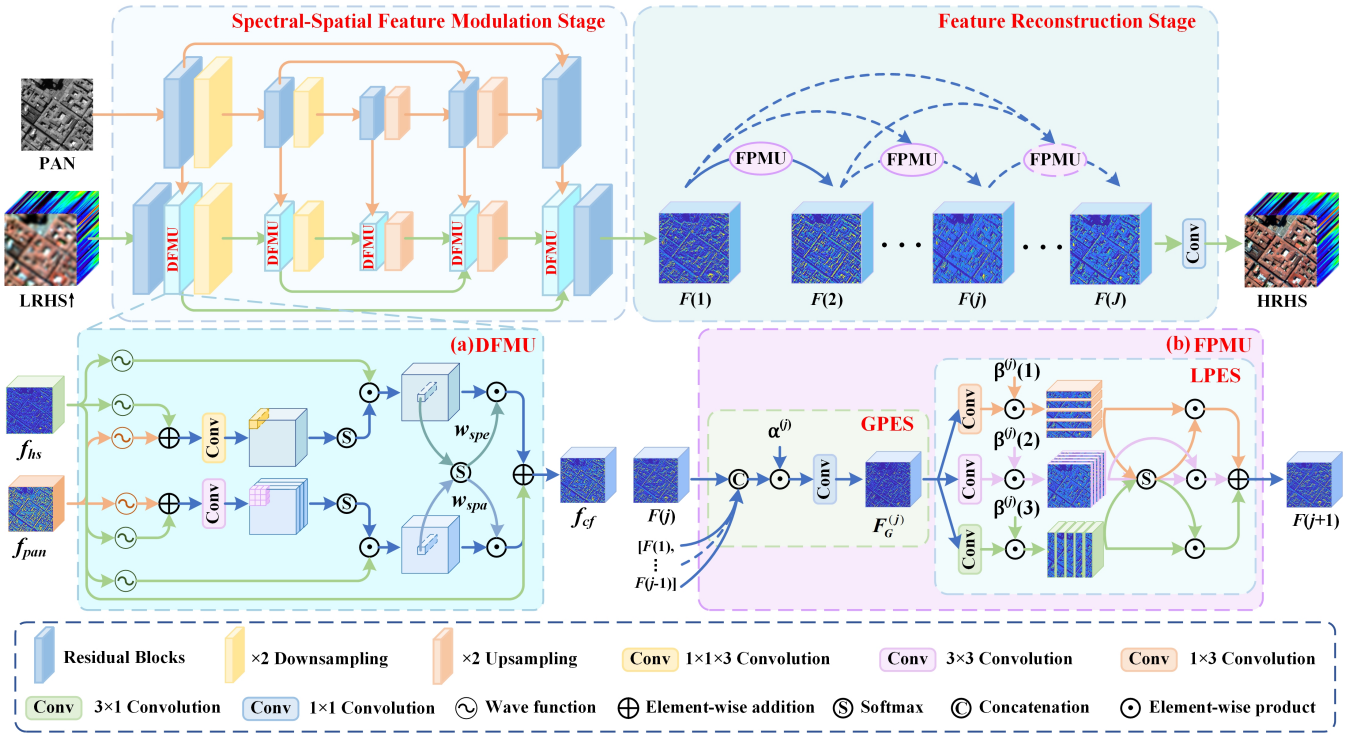


Figure 2: The overall architecture of FMPM-DNet.

Thus, the absolute value operation can be removed to simplify the calculation, and a simple convolution operation is used to obtain the amplitude of the input features. The operation is as follows:

$$z_{(x,y)} = Conv_{1\times1}(F_{(x,y)}) \tag{3}$$

where $Conv_{1\times1}(\cdot)$ indicates the convolution operation with a kernel size of 1×1. To obtain the specific attributes of different input features, we utilize an estimation module (Tang et al. 2022) to generate the phase information of the input features. The specific operation is as follows:

$$\theta_{(x,y)} = \mathrm{ReLU}\left(BN\left(Conv_{1\times1}\left(F_{(x,y)}\right)\right)\right) \tag{4}$$

where $BN(\cdot)$ represents the batch normalization, and $\mathrm{ReLU}(\cdot)$ indicates the ReLU activation function. According to equations (3) and (4), we can obtain the amplitudes and phases corresponding to the PAN and HS image features at each scale, which are substituted into the wave function equation (1) to obtain the real and imaginary parts of the wave-like features. Six wave functions are used to achieve feature transformation, and the operations are as follows.

$$Cat\left(\mathrm{Re}_p^{qi}\ \ \mathrm{Im}_p^{qi}\right) = WF^i\left(f_{pan}\right),\ i = 1,2 \tag{5}$$

$$Cat\left(\mathrm{Re}_{hs}^{ki}\ \ \mathrm{Im}_{hs}^{ki}\right) = WF^i\left(f_{hs}\right),\ i = 1,2 \tag{6}$$

$$Cat\left(\mathrm{Re}_{hs}^{vi}\ \ \mathrm{Im}_{hs}^{vi}\right) = WF^i\left(f_{hs}\right),\ i = 1,2 \tag{7}$$

where Re_p^{qi} and Im_p^{qi} represent the real and imaginary parts corresponding to the PAN features. Re_{hs}^{ki} and Re_{hs}^{vi} represent the real parts corresponding to the HS features. Im_{hs}^{ki} and Im_{hs}^{vi} represent the imaginary parts corresponding to the HS features. $Cat\left(\cdot\right)$ represents the concatenation operation.

In the feature modulation step, we first synthesize the real and imaginary parts of PAN and HS image features, and employ different convolution operations to dynamically generate spatial and spectral modulation matrices $SpaM$ and $SpeM$. Then, these two matrices are multiplied with the real and imaginary parts of the HS image features to achieve the corrected HS image features. This process can be expressed as follows.

$$SpaM = S\left(Conv_{3\times3}\left(Cat\left(\mathrm{Re}_{hs}^{k2}+\mathrm{Re}_p^{q2}, \mathrm{Im}_{hs}^{k2}+\mathrm{Im}_p^{q2}\right)\right)\right) \tag{8}$$

$$SpeM = S\left(Conv_{1\times1\times3}\left(Cat\left(\mathrm{Re}_{hs}^{k1}+\mathrm{Re}_p^{q1}, \mathrm{Im}_{hs}^{k1}+\mathrm{Im}_p^{q1}\right)\right)\right) \tag{9}$$

$$F_{spa} = SpaM \odot Cat\left(\mathrm{Re}_{hs}^{v1}, \mathrm{Im}_{hs}^{v1}\right) \tag{10}$$

$$F_{spe} = SpeM \odot Cat\left(\mathrm{Re}_{hs}^{v2}, \mathrm{Im}_{hs}^{v2}\right) \tag{11}$$

where $Conv_{3\times3}(\cdot)$ and $Conv_{1\times1\times3}(\cdot)$ indicate the convolution operation with a kernel size of 3×3 and 1×1×3, respectively. $S(\cdot)$ represents the softmax operation. F_{spa} and F_{spe} represent the modulation features in the spectral and spatial dimensions, respectively.

Next, in the feature fusion step, we design a soft cross-attention mechanism to adaptively fuse two modulation features F_{spa} and F_{spe}. Specifically, these two modulation features are fed into the softmax operation to calculate two adaptive weight matrices w_{spa} and w_{spe}.

$$w_x = \frac{\exp\left(F_x\right)}{\exp\left(F_{spe}\right)+\exp\left(F_{spa}\right)},\ x = \{spe, spa\} \tag{12}$$

where $\exp(\cdot)$ refers to the exponential function. Finally, the weight matrices w_{spa} and w_{spe} are multiplied with two modulation features to obtain the fused modulation features, which are then added to the input HS image features of DFMU to obtain the coarse fusion features f_{cf} at the current scale. The specific operation is as follows.

$$f_{cf} = f_{hs} + (w_{spe} \odot F_{spe} + w_{spa} \odot F_{spa}) \tag{13}$$

Feature Probability Mask Unit (FPMU)

At present, most networks integrate features of different depths by directly concatenating or summing all features, which may result in inaccurate feature reconstruction when learning complex mapping relationships between multiple channels. Therefore, in each feature reconstruction layer in the second stage, an FPMU is constructed by introducing learnable probability masks, as shown in Figure 2(b), which selectively utilizes features to obtain accurate reconstructed features. FPMU consists of two learnable probabilistic mask structures: global probabilistic embedding structure (GPES) and local probabilistic embedding structure (LPES). GPES achieves the embedding and integration of features from different reconstruction layers through defining adaptive global probability masks. LPES achieves feature embedding and integration in both channel and spatial dimensions by defining adaptive local probability masks. Taking the j-th FPMU as an example, we provide a detailed introduction to the construction of two structures.

In GPE, the input features and global probability masks are considered as elements in the global probability space SG, as represented below.

$$SG = \{(\alpha^{(1)}, F^{(1)}),(\alpha^{(2)}, F^{(2)}), \cdots ,(\alpha^{(j)}, F^{(j)})\} \tag{14}$$

where the binary vector $\alpha^{(j)} \in \{0,1\}^{\mathrm{j}}$ with length j indicates whether to use the features from the previous j layers. $F^{(j)}$ denotes the aggregated features from the previous j layers. $\alpha^{(j)}$ follows the Bernoulli distribution $B(X)$, also known as the 0-1 distribution, which is a discrete probability distribution defined as:

$$B(X)= \begin{cases} 1-p, & X=0 \\ p, & X=1 \end{cases} \tag{15}$$

where X represents a random variable, equivalent to $\alpha^{(j)}(m)$, $m \in [1, 2, \cdots, j]$ in the space SG, and p is the probability parameter, which satisfies the condition $0 \leq p \leq 1$. p can be obtained by training a mask network with a learnable independent Bernoulli distribution. However, Bernoulli distribution is not differentiable and cannot be directly used to calculate gradients. To solve this problem, a differentiable continuous random variable needs to be introduced to approximate the discrete random variable, allowing gradient calculation during training. Gumbel softmax (Jang, Gu, and Poole 2017) can relax a discrete Bernoulli distribution into a continuous space, which can be defined as:

$$M(p) = sigmoid(\frac{1}{\tau}(log(\frac{p}{1-p}) + log(\frac{log(r_1)}{log(r_2)}))) \tag{16}$$

where r_1 and r_2 are random noises with standard uniform distribution in the range of [0,1], τ is a temperature. $\mathrm{sigmod}(\cdot)$ represents the sigmoid activation function.

According to Formula (16), the feature selection and integration operations for each reconstruction layer can be represented by the following equations.

$$\begin{aligned} T^{(m,\,j)} &= F^{(m)} \odot M(B(\alpha^{(j)}(m))) \\ F_G^{(j)} &= Conv_{1\times1}(T^{(0,\,j)}, T^{(1,\,j)}, \cdots, T^{(j,j)}) \end{aligned} \tag{17}$$

where $T^{(m,j)}$ indicates the features selected from $F^{(m)}$, and $F_G^{(j)}$ represents the global aggregated features.

In LPES, local feature embedding is achieved by adaptively learning local probability masks in the horizontal (h), vertical (v), and channel (c) directions. Firstly, the global aggregated features are processed through three 2D convolutions $C_n^{(j)}(\cdot) = \left[C_h^{(j)}(\cdot),\, C_v^{(j)}(\cdot), C_c^{(j)}(\cdot)\right], n \in \{h, v, c\}$, which extracts features in the horizontal, vertical, and channel dimensions. Then, according to Formula (16), three learnable probability masks are defined to achieve the selection of these three features. The above operations can be represented by:

$$O^{(n,j)} = C_n^{(j)}(F_G^{(j)}) \odot M(B(\beta^{(j)}(n))) \tag{18}$$

where $O^{(n,j)}$ represents the selected features in three directions. $\beta^{(j)} \in \{0,1\}^3$ with length of 3 indicates whether to use the features from the horizontal, vertical, and channel dimensions. Finally, the soft cross-attention mechanism is adopted to embed local features into the output of FPMU, and the operations are defined as follows.

$$w_n = \frac{\exp\left(O^{(n,j)}\right)}{\exp(O^{(h,j)}) + \exp(O^{(v,j)}) + \exp(O^{(c,j)})} \tag{19}$$

$$F^{(j+1)} = \sum_n w_n \odot O^{(n,j)} \tag{20}$$

where w_h, w_v, and w_c represent the embedding weights for the features $O^{(h,\,j)}$, $O^{(v,\,j)}$, and $O^{(c,\,j)}$, respectively. $F^{(j+1)}$ represents the output of the j-th FPMU.

Loss Function

L_1 loss is used to calculate the reconstruction loss, which is commonly used in HS pansharpening tasks(Zhuo et al. 2022). The reconstruction loss is defined as:

$$L_{ref} = \frac{1}{CHW} \| H_{GT} - I_{HRHS} \|_1 \tag{21}$$

where H_{GT} denotes the ground truth (GT) HRHS images, I_{HRHS} is the network-predicted HRHS images, and (C, H, W) denotes the dimensions of H_{GT}: the number of spectral channels, image height, and width.

Experimental Results and Analysis

Dataset and Metrics

To validate the effectiveness of the proposed FMPM-DNet, extensive experiments were conducted on three publicly available and widely used HSI datasets, including Pavia center(Plaza et al. 2009), Botswana (Ungar 2002), and Chikusei(Yokoya and Iwasaki 2016). The datasets were processed

Datasets	Methods	PSNR(↑)	SAM(↓)	SSIM(↑)	SCC(↑)	RMSE$\times 10^{-2}$(↓)	ERGAS(↓)
Pavia center	CNMF (Yokoya et al. 2012)	32.0244	7.2364	0.8913	0.9462	2.7090	4.8195
	DARN (Zheng et al. 2020)	37.1592	5.2685	0.9545	0.9824	1.5299	2.8876
	Hyperkite (Bandara et al. 2022)	36.9586	5.3630	0.9527	0.9809	1.5661	2.9363
	Hyper-DSNet (Zhuo et al. 2022)	37.1725	5.1541	0.9539	0.9822	1.5245	2.8929
	FPFNet (Dong et al. 2023)	37.2121	5.3962	0.9520	0.9822	1.5297	2.8992
	HyperRefiner (Zhou et al. 2023)	37.7193	5.0461	0.9588	0.9856	1.4265	2.7748
	Tess-SNet (He et al. 2024)	38.1569	4.8628	0.9594	0.9855	1.3815	2.6432
	FMPM-DNet (Ours)	**38.8193**	**4.6083**	**0.9633**	**0.9874**	**1.2829**	**2.4837**
Botswana	CNMF (Yokoya et al. 2012)	33.6006	2.3018	0.9193	0.9549	4.5080	3.8151
	DARN (Zheng et al. 2020)	41.1629	1.9877	0.9453	0.9818	1.3589	2.5892
	Hyperkite (Bandara et al. 2022)	42.4773	1.7728	0.9576	0.9825	1.2897	1.8724
	Hyper-DSNet (Zhuo et al. 2022)	43.7004	1.8418	0.9554	0.9809	1.2998	1.4724
	FPFNet (Dong et al. 2023)	44.1665	1.8960	0.9621	0.9843	1.2396	1.4639
	HyperRefiner (Zhou et al. 2023)	44.2455	1.6751	0.9663	0.9864	1.1655	1.3997
	Tess-SNet (He et al. 2024)	45.2278	1.5787	0.9671	0.9875	1.0738	1.3039
	FMPM-DNet (Ours)	**45.6029**	**1.5171**	**0.9692**	**0.9886**	**1.0343**	**1.2521**
Chikusei	CNMF (Yokoya et al. 2012)	35.6701	3.7933	0.8970	0.9013	1.9730	6.6018
	DARN (Zheng et al. 2020)	41.0419	2.4217	0.9686	0.9743	0.9837	4.2614
	Hyperkite (Bandara et al. 2022)	41.6503	2.3207	0.9702	0.9761	0.9246	4.0896
	Hyper-DSNet (Zhuo et al. 2022)	41.6990	2.3315	0.9705	0.9765	0.9197	4.0150
	FPFNet (Dong et al. 2023)	42.3698	2.3260	0.9759	0.9826	0.8243	3.8711
	HyperRefiner (Zhou et al. 2023)	42.9958	2.1112	0.9777	0.9829	0.8046	3.5126
	Tess-SNet (He et al. 2024)	43.2155	2.0608	0.9795	0.9847	1.7487	3.4806
	FMPM-DNet (Ours)	**44.0595**	**1.9074**	**0.9826**	**0.9874**	**0.6809**	**3.1860**

Table 1: The average quantitative results on the Pavia center, Botswana, and Chikusei datasets.

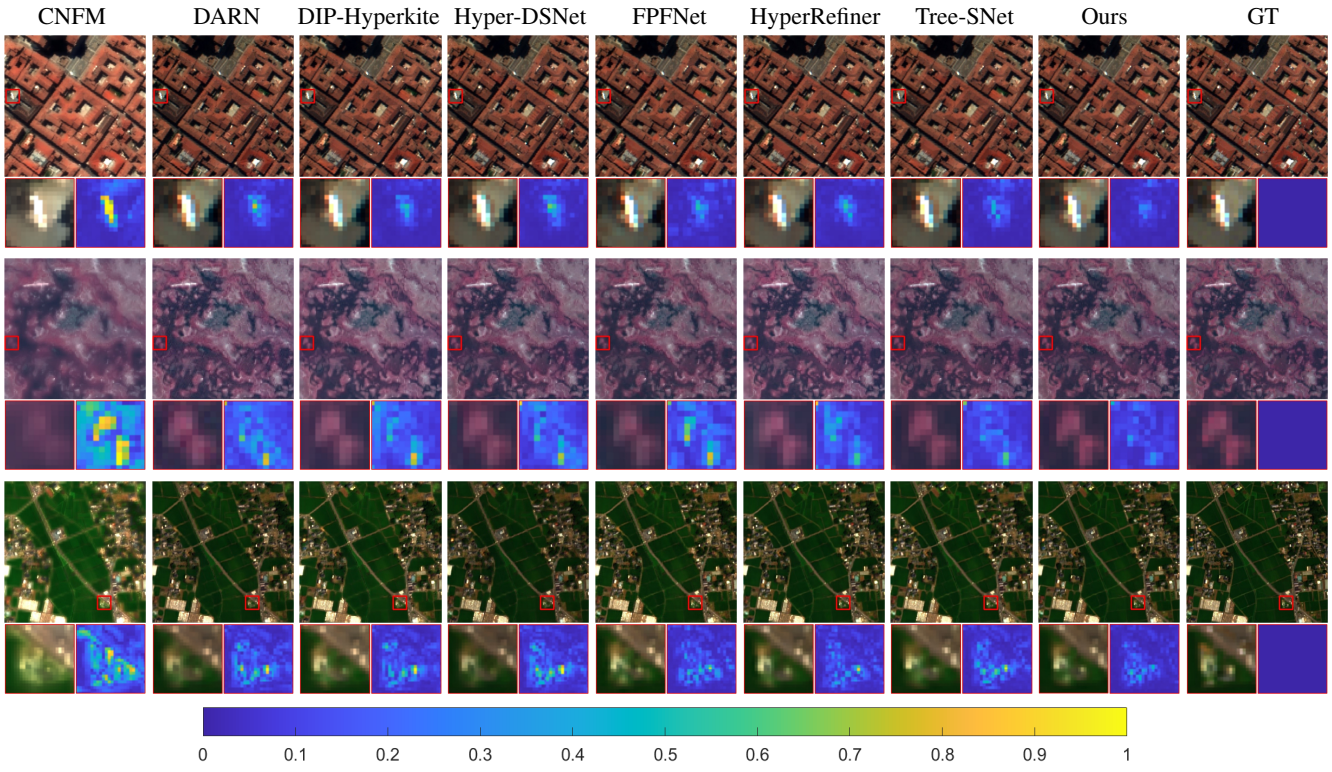
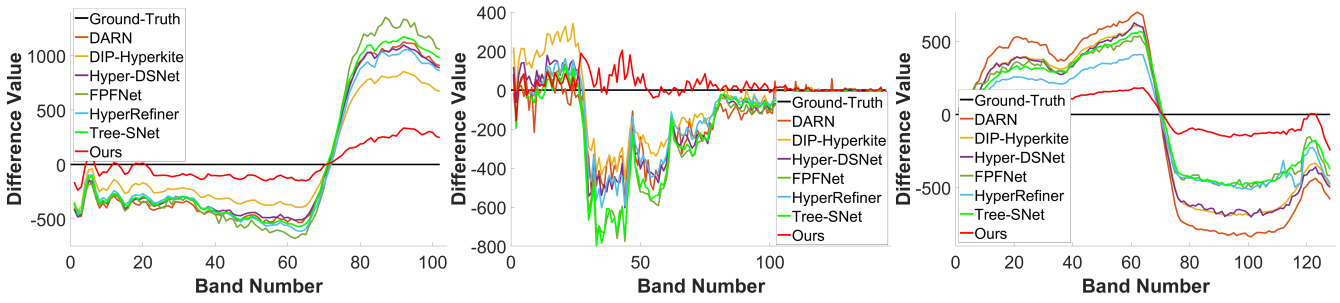


Figure 3: Fusion results on the three datasets.(The first row presents the results on the Pavia center dataset, the second row shows the results on the Botswana dataset, and the third row displays the results on the Chikusei dataset.)



(a) Pavia center (155, 49) (b) Botswana (26, 26) (c) Chikusei (171, 215)

Figure 4: Spectral difference curves at three positions in three images from three datasets.

One stage	Two stage	PSNR↑	SAM↓	SSIM↑	SCC↑	ERGAS↓
✓	×	38.4756	4.8016	0.9616	0.9866	2.5618
×	✓	38.0810	4.8951	0.9589	0.9850	2.5618
✓	✓	**38.8193**	**4.6083**	**0.9633**	**0.9874**	**2.4837**

Table 2: Ablation study on the dual-stage structure.

Methods	PSNR↑	SAM↓	SSIM↑	SCC↑	REGAS↓
w/o DFMU	38.6033	4.6914	0.9662	0.9867	2.5263
w/ DFMU	**38.8193**	**4.6083**	**0.9633**	**0.9874**	**2.4837**

Table 3: Ablation study on DFMU structure.

Methods	PSNR↑	SAM↓	SSIM↑	SCC↑	REGAS↓
w/o Wave	38.4758	4.7491	0.9616	0.9866	2.5593
w/ Wave	**38.8193**	**4.6083**	**0.9633**	**0.9874**	**2.4837**

Table 4: Ablation study on the wave function in DFMU structure.

following the Wald's protocol (Wald 2000) and set according to Bandara(Bandara et al. 2022). We compared FMPM-DNet with several SOAT methods, including one traditional methods: CNMF(Yokoya, Yairi, and Iwasaki 2012), and six DL-based methods: DARN(Zheng et al. 2020), DIP-Hyperkite(Bandara et al. 2022), Hyper-DSNet(Zhuo et al. 2022) , FPFNet(Dong et al. 2023), HyperRefiner(Zhou et al. 2023), and Tree-SNet(He et al. 2024). To objectively evaluate the performance of all comparative methods, six objective metrics were adopted, including Spectral Cross Correlation (SCC), Spectral Angle Mapping (SAM), Structural SIMilarity (SSIM), Root Mean Square Error (RMSE), Erreur Relative Globale Adimensionnelle de Synthese (ERGAS), and Peak Signal-to-Noise Ratio (PSNR)(Zheng et al. 2020; Bandara et al. 2022).

Experimental Setup

We retrained all DL-based methods using Python 3.9 and PyTorch 1.13 on Ubuntu 20.04 system with a NVIDIA GTX A6000. The initial learning rate, epoch, and batch size are set to 0.0001, 8000, and 4, respectively. Adam is selected as the optimizer, and the learning rate decays by 0.8 every 2000 epochs.

GPES	LPES	PSNR↑	SAM↓	SSIM↑	SCC↑	REGAS↓
×	×	38.4795	4.7636	0.9609	0.9862	2.5613
×	✓	38.7094	4.5948	0.9634	0.9871	2.5050
✓	×	38.6997	4.6188	0.9634	0.9871	2.5059
✓	✓	**38.8193**	**4.6083**	**0.9633**	**0.9874**	**2.4837**

Table 5: Ablation study on the probability masks in FPMU structure.

Methods	DIP-Hyperkite	Hyper-DSNet	FPF Net	Hyper Refiner	Tree-SNet	Ours
#Params(M)	0.97	0.53	22.34	19.32	9.12	5.84
FLOPs(G)	212.64	21.89	174.92	85.00	205.00	100.16

Table 6: Comparision of parameters and FLOPS of DL-based methods on the Chikusei dataset.

Objective and Subjective Comparison

Table 1 presents the objective evaluation results of the fusion results on the three datasets. The best results are highlighted in bold, while the second-best results are underlined. The table shows that our proposed method outperforms the other methods in all evaluation indicators. Specifically, the PSNR values of the proposed method are 0.6624 dB, 0.3751 dB, and 0.8440 dB higher than those of the second-best comparison methods, respectively.

Figure 3 shows the fused images obtained by the different methods on the three datasets. To observe the differences between the fusion results more clearly, the Mean Absolute Error (MAE) maps between the fusion results and the GT are calculated and displayed in the lower right corner, with magnified local areas shown in the lower left corner. It can be seen that our results are visually closest to the GT images and the corresponding MAE maps contain the least residual information. This also indicates that our results have better spatial and spectral fidelity.

To further assess the spectral preservation capability of the DL-based methods, Figure 4 presents the spectral difference curves of three randomly selected pixels in three images from these datasets (Pavia center (155, 49), Botswana (26, 26), and Chikusei (171, 215)). Our method exhibits the least spectral difference values, demonstrating its excellent spectral preservation capability.

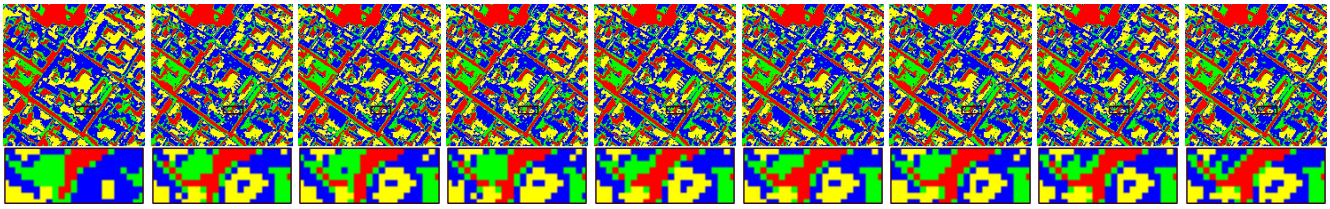


Figure 5: The classification results of fused images from comparison methods on the Pavia center dataset.

Metrics	CNMF	DRAN	DIP-Hyperkite	Hyper-DSNet	FPFNet	HyperRefiner	Tree-SNet	Ours
OA	0.7795	0.9451	0.9405	0.9435	09459	0.9465	0.9519	**0.9564**
Kappa	0.9393	0.9246	0.9184	0.9225	0.9258	0.9265	0.9340	**0.9402**

Table 7: Objective comparison of classification results on the Pavia center dataset.

Ablation Study

Several ablation experiments were conducted on the Pavia center dataset to demonstrate the effectiveness of the proposed components: the dual-stage structure, DFMU, and FPMU.

Effectiveness of Dual-stage Structure. We conducted ablation experiments on networks with first stage, second stage, and dual stages, and the results are shown in Table 2. From the table, we can observe that compared with the suboptimal methods in Table 1, the network with first stage achieves better performance, with an increase of approximately 0.32dB in PSNR value, indicating that the network with this stage has good feature correction ability. The network with dual stages exhibits better performance compared to networks with only one stage, indicating that the dual-stage structure is effective.

Effectiveness of DFMU. To validate the effectiveness of DFMU, we directly removed the DFMU structure and replaced it with a concatenation operation on PAN and HS feature maps. The results shown in Table 3 indicate that the dynamic network using DFMU achieves better results.

Effectiveness of Wave Function. To validate the effectiveness of the wave function in DFMU, we conducted ablation experiments by replacing the wave functions in DFMU with 1×1 convolutions. Table 4 displays the comparison results. From the table, we can observe that the network using the wave function have better performance.

Effectiveness of Probability Masks in GPES and LPES of FPMU. The ablation experiments were conducted to validate the effectiveness of probability masks in GPES and LPES of FPMU. In the experiments, without GPES or LPES means removing the probability masks in GPES or LPES. The results shown in Table 5 indicate that removing the probability masks in both GPES and LPES obtain the lowest metric values, while retaining the probability masks in GPES or LPES obtains better results. Our method achieves the best performance, which indicates the probability masks in GPES and LPES of FPMU are effective in improving the performance of the model.

Model Complexity. We compared the parameters and computational complexity of five DL-based models , and the results are shown in Table 6. Although the proposed method has more parameters than DIP-Hyperkite and Hyper-DSNet, our method achieves higher PSNR values. Compared with FPFNet, HyperRefinert, and Tree-SNet, our method achieves better results with fewer parameters. Overall, our method has better performance.

Classification Application

To further demonstrate the effectiveness of the proposed method, we conducted downstream ground object classification experiments on the fused results. We used the k-means algorithm in the software ENVI on satellite images to evaluate the fusion results of different methods. The number of classification categories was set to 4, and the maximum number of iterations was set to 5. The classification results are shown in Figure 5. The figure shows that the classification result of our fused images is the closest to that of the GT. The accuracy metrics in Table 7 also indicate that our method achieves better fusion results compared to other methods.

Conclusion

In this paper, we propose a novel dual-stage dynamic fusion network, called FMPM-DNet. The first stage consists of two U-Net branches, which are used to extract features of different scales from PAN and LRHS images, and to modulate LRHS image features using PAN image features. At each scale layer, a DFMU based on wave function is constructed, which generates coarse fusion features by adaptively modulating the features of LRHS images. The second stage is constructed to obtain the final pansharpening results by gradually reconstructing refined fusion features. At each reconstruction layer in this stage, an FPMU is designed to achieve feature selection and embedding by defining learnable probability masks. Experiments conducted on three widely used HS datasets demonstrate that our FMPM-DNet outperforms other SOTA methods in both quantitative and qualitative evaluations.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62072218 and No. 62261025).

References

Aburaed, N.; Alkhatib, M. Q.; Marshall, S.; Zabalza, J.; and Al Ahmad, H. 2023. A Review of Spatial Enhancement of Hyperspectral Remote Sensing Imaging Techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16: 2275–2300.

Aiazzi, B.; Alparone, L.; Baronti, S.; Garzelli, A.; and Selva, M. 2006. MTF-tailored multiscale fusion of high-resolution MS and Pan imagery. *Photogrammetric Engineering & Remote Sensing*, 72(5): 591–596.

Aiazzi, B.; Baronti, S.; Selva, M.; and 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.

Bandara, W. G. C.; Valanarasu, J. M. J.; Patel, V. M.; and Patel, V. M. 2022. Hyperspectral Pansharpening Based on Improved Deep Image Prior and Residual Reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Dong, W.; Yang, Y.; Qu, J.; Li, Y.; Yang, Y.; and Jia, X. 2023. Feature Pyramid Fusion Network for Hyperspectral Pansharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

He, L.; Xi, D.; Li, J.; Lai, H.; Plaza, A.; and Chanussot, J. 2023. Dynamic Hyperspectral Pansharpening CNNs. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–19.

He, L.; Ye, H.; Xi, D.; Li, J.; Plaza, A.; and 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.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations (ICLR)*.

Li, J.; Cai, Y.; Li, Q.; Kou, M.; and Zhang, T. 2024. A review of remote sensing image segmentation by deep learning methods. *International Journal of Digital Earth*, 17(1): 2328827.

Plaza, A.; Benediktsson, J. A.; Boardman, J. W.; Brazile, J.; Bruzzone, L.; Camps-Valls, G.; Chanussot, J.; Fauvel, M.; Gamba, P.; Gualtieri, A.; et al. 2009. Recent advances in techniques for hyperspectral image processing. *Remote sensing of environment*, 113: S110–S122.

Tang, Y.; Han, K.; Guo, J.; Xu, C.; Li, Y.; Xu, C.; and Wang, Y. 2022. An image patch is a wave: Phase-aware vision mlp. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10935–10944.

Ungar, S. G. 2002. Overview of the Earth Observing One (EO-1) Mission. In *IEEE International Geoscience and Remote Sensing Symposium*, volume 1, 568–571. IEEE.

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

Wei, Q.; Bioucas-Dias, J.; Dobigeon, N.; and Tourneret, J.-Y. 2015. Hyperspectral and Multispectral Image Fusion Based on a Sparse Representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7): 3658–3668.

Wu, X.; Feng, J.; Shang, R.; Wu, J.; Zhang, X.; Jiao, L.; and Gamba, P. 2024. Multi-task multi-objective evolutionary network for hyperspectral image classification and pansharpening. *Information Fusion*, 108: 102383.

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

Yokoya, N.; Yairi, T.; and Iwasaki, A. 2012. Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2): 528–537.

Zhang, L.; Nie, J.; Wei, W.; Zhang, Y.; Liao, S.; and Shao, L. 2020. Unsupervised adaptation learning for hyperspectral imagery super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3073–3082.

Zhang, X.; Huang, Z.; Wang, N.; Xiang, S.; and Pan, C. 2021. You Only Search Once: Single Shot Neural Architecture Search via Direct Sparse Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9): 2891–2904.

Zheng, Y.; Li, J.; Li, Y.; Guo, J.; Wu, X.; and Chanussot, J. 2020. Hyperspectral pansharpening using deep prior and dual attention residual network. *IEEE transactions on geoscience and remote sensing*, 58(11): 8059–8076.

Zhou, B.; Zhang, X.; Chen, X.; Ren, M.; and 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.

Zhuo, Y.-W.; Zhang, T.-J.; Hu, J.-F.; Dou, H.-X.; Huang, T.-Z.; and Deng, L.-J. 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.