UMNet: Uncertainty-guided Memory Network for Hyperspectral Pansharpening

Xiaozheng Wang[1*], Yong Yang[1†], Shuying Huang[1*], Nayu Liu[1], Ziyang Liu[1]

[1]Tiangong University, Tianjin, China
xiaozhengwang95@gmail.com, yangyong@tiangong.edu.cn, huangshuying@tiangong.edu.cn, 695704204@qq.com, qduliuzy@163.com

Abstract

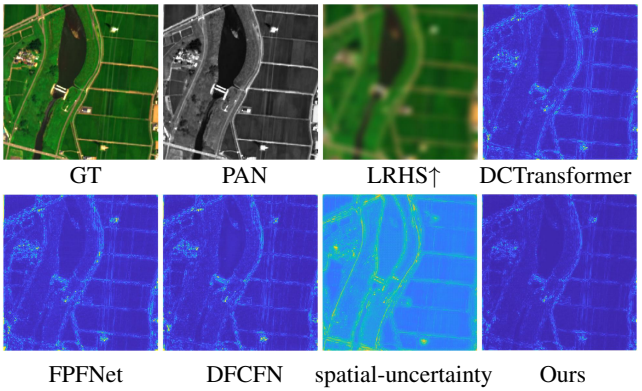
At present, most hyperspectral (HS) sharpening methods have not fully utilized the feature correlation between adjacent bands in HS images, nor have they explored the problem of feature uncertainty generated by the model during the fusion process. This may lead to inaccurate fusion features generated by the model, resulting in spatial and spectral distortions in the fusion results. To address these issues, we propose an uncertainty-guided memory network (UMNet) for HS pansharpening. A spatial-spectral recurrent fusion unit (SRFU) is designed based on the concept of temporal data modeling, which utilizes the correlation between adjacent bands to fuse spectral and spatial features from PAN and LRHS images. In SRFU, a state memory interaction unit (SMIU) is constructed based on non-negative matrix factorization (NMF) to learn the global spatial-spectral dependency of PAN and HS images in the recurrent state space. Moreover, based on uncertainty theory, we define two spatial-spectral uncertainty-guided loss functions for the HS pansharpening task to train the model step by step, ensuring that the network can reconstruct more accurate spectral and spatial features. Extensive experiments on three widely used datasets demonstrate that, compared with some state-of-the-art (SOTA) methods, the proposed UMNet has achieved significant improvements in both spatial and spectral quality metrics.

Code — https://github.com/EchoPhD/UMNet

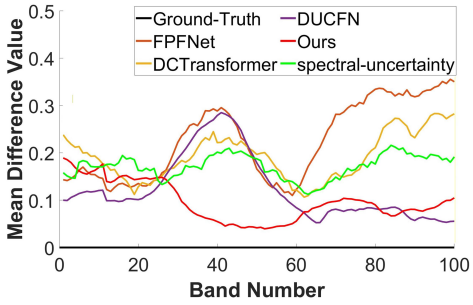
Introduction

Due to the limitations of the physical imaging system in optical remote sensing, the balance between spatial resolution and spectral resolution is a key consideration factor. single imaging system cannot directly acquire high-spatial-resolution hyperspectral (HRHS) images. Typically, hyperspectral imaging system can obtain low-spatial-resolution hyperspectral (LRHS) images with hundreds of bands that contain rich spectral information, while panchromatic (PAN) imaging system 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 LRHS and PAN images. This process is commonly referred to as HS pansharpening.

In recent years, deep learning (DL)-based methods (Wang et al. 2024; Ran et al. 2023) have been widely applied in the HS pansharpening task due to their excellent feature extraction. Dong et al. (Dong et al. 2025) proposed a dual-branch feature pyramid fusion network that reconstructs HRHS images through the progressive fusion of multi-scale features. He et al. (He et al. 2024) designed a tree-structured network that addresses the imbalance between spatial and spectral information through the use of



(a). Differences in spatial-dimension



(b). Differences in band-dimension

Figure 1: Illustration of the difference between fusion results and GT image.

*These authors contributed equally.
†Corresponding author.

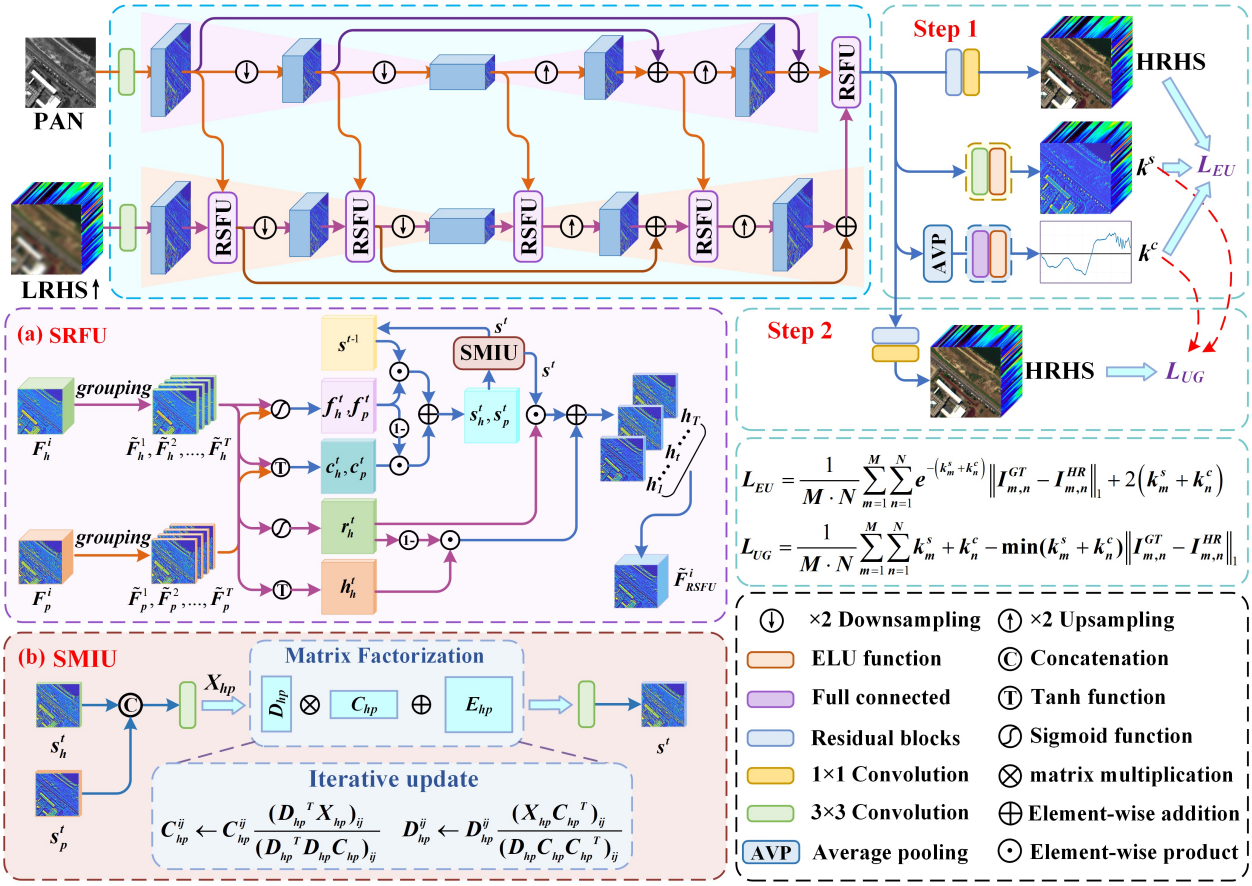


Figure 2: The overall network architecture of the proposed UMNet.

adaptive convolutions. However, most DL-based HS pansharpening methods are not specifically designed for the characteristics of HS images. For example, HS images exhibit both local and non-local similarities in the spectral dimensions. Thus, HS images can be regarded as a sequence of gray images linked by the global spectral correlation (Wang et al. 2023). The fused images produced by most methods may fail to preserve continuous spectral-spatial information of surface objects, mainly due to neglect of the sequential structure in HS images. Transformer-based methods attempt to utilize the global attention mechanism to better model the HS pansharpening. Ma et al. (Ma et al. 2024) constructed a directional paired cross-attention mechanism to achieve effective fusion between modalities. Wang et al. (Wang et al. 2025) constructed a conditional entropy Transformer within a multi-scale feature decomposition framework to correct the spectral information in coarse fusion results. Transformer-based methods yield promising fusion results but face high computational complexity. Geng et al. (Geng et al. 2021) viewed attention mechanisms as a means to capture latent pixel-wise correlations in images. Accordingly, they introduced a matrix factorization-based attention mechanism that demonstrated superior performance in several image processing tasks(Xie et al. 2021; Liu et al. 2023).

In addition, HS images contain hundreds of channels with a broader spectral range than PAN images. Modality differences induce data uncertainty during fusion (Ran et al. 2023), leading to inaccurate fusion features. To verify this uncertainty, we analyzed the spatial and spectral differences between the fusion results and ground truth (GT) images, as shown in Figure 1. Figure 1 (a) shows the mean absolute error (MAE) maps between the fusion results of several state-of-the-art (SOTA) methods and the GT image. Figure 1 (b) shows the spectral difference curve of a test image in the Chikusei dataset. We also present the spatial uncertainty map and spectral uncertainty curve predicted by our model. From Figure 1 (a), our MAE map contains less residual information compared to other methods, while the spatial uncertainty map shares similar structural information with other MAEs. From Figure 1 (b), our spectral curve is closest to the GT image, and the spectral uncertainty curve resembles the others. This indicates that uncertainty exists in the data generation process, and correcting this error can yield more accurate fused features.

Based on the above analysis, we propose an uncertainty-guided memory network (UMNet) to obtain fusion results with spatial and spectral fidelity. As shown in Figure 2, two parallel U-Nets serve as the backbone to extract and fuse multi-scale features from two image types. The U-Net in the upper branch is constructed to extract PAN image features

at different scales. The U-Net in the lower branch aims to achieve the fusion of two image features at different scales. At each scale layer, a spatial-spectral recurrent fusion unit (SRFU) is designed based on the sequence modeling mechanism, which fully utilizes the correlation between adjacent spectral channels to achieve feature selection and fusion. In SRFU, a state memory interaction unit (SMIU) is constructed based on non-negative matrix factorization, which obtains spatial-spectral fusion features by learning the global spatial-spectral dependencies between PAN and HS images in a low-dimensional state space. Compared to Transformer, SMIU can achieve better computational performance. In addition, based on Bayesian estimation uncertainty theory, a step-by-step training strategy guided by spatial-spectral uncertainty is designed to guide the network to learn more accurate spectral and spatial features. The main contributions are as follows:

- An UMNet based on sequence data modeling and uncertainty theory is proposed to achieve fine fusion of spectral and spatial features from PAN and LRHS images.
- Considering the feature correlation between adjacent bands in HS images, an SRFU is designed to achieve the fusion of two modal features in adjacent bands and the interaction of features between different bands.
- A step-by-step training strategy is designed by defining two spatial-spectral uncertainty-guided loss functions, which can guide the reconstruction of fused features in different regions by prioritizing high uncertainty regions.

Related Work

Uncertainty Estimation

This work focuses on aleatoric uncertainty in Bayesian modeling, which captures the inherent noise in observed data. In previous Bayesian modeling, this uncertainty is represented by an additional term θ_i, and the observation model can be formulated as $x_i = \bar{f}(y_i) + \varepsilon\theta_i$, where y_i and x_i represent the low-resolution image and the corresponding GT image, respectively. $\bar{f}(\cdot)$ represents any super-resolution (SR) network, ϵ represents a Laplace distribution with zero mean and unit variance. The Laplace distribution is used to characterize the likelihood function as follows:

$$p(x_i;\theta_i|y_i) = \frac{1}{2\theta_i}\exp\left(-\frac{||x_i - \bar{f}(y_i)||_1}{\theta_i}\right) \quad (1)$$

where $\bar{f}(y_i)$ and θ_i denote the SR image (mean) and uncertainty (variance) , respectively, learned by SR networks. The network is trained to estimate the logarithmic variance $s_i = \ln\theta_i$. Maximum likelihood estimation reformulates the problem as minimizing the following loss :

$$\bar{L}_{EU} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\left[\exp(-s_i)\|x_i - \bar{f}(y_i)\|_1 + 2s_i\right] \quad (2)$$

However, $\bar{L}_{EU}$ fails to significantly improve SR performance, as the variance term θ_i in the denominator reduces the penalty for high-variance pixels, weakening their contribution to the loss and causing the network to ignore them. As shown in Figure 1, however, such high-uncertainty pixels and bands often contain critical visual information such as textures and edges—a factor largely overlooked in SR tasks.

Proposed Method

In this section, we propose a UMNet for HS pansharpening, as shown in Figure 2, with parallel dual U-Net branches. The upper branch extracts features from PAN images at different scales. while the lower branch uses an SRFU to fuse spatial and spectral features from both input images. To guide the network to reconstruct accurate spatial and spectral features, a step-by-step supervision strategy consisting of two independent supervision steps is designed to train our model. Two uncertainty-based loss functions are defined to constrain the reconstruction of spatial and spectral features.

Spatial-spectral Recurrent Fusion Unit

As shown in Figure 2(a), SRFU fuses spatial and spectral features at each scale. Given the strong spectral correlation across adjacent channels, PAN and HS features are grouped along the channel dimension. These group features, reflecting spectral continuity, are treated as temporal sequences and fused recursively. Additionally, SMIU is introduced to integrate features, suppress noise, and enhance structural details. The operations of SRFU are as follows.

Taking the i-th scale of the encoder as an example, the HS and PAN feature maps $\mathrm{F}_h^i, F_p^i \in \mathbb{R}^{h_i\times w_i\times c_i}$ (where h_i, w_i, and c_i denote the height, width, and number of channels of the i-th scale feature maps, respectively.) are first grouped along the spectral channel dimension.

$$[\tilde{F}_h^1, \tilde{F}_h^2, ..., \tilde{F}_h^t, ..., \tilde{F}_h^T] = grouping(F_h^i) \quad (3)$$

$$[\tilde{F}_p^1, \tilde{F}_p^2, ..., \tilde{F}_p^t, ..., \tilde{F}_p^T] = grouping(F_p^i) \quad (4)$$

where $grouping(\cdot)$ represents the grouping operation. $\tilde{F}_h^t \in \mathbb{R}^{h_i\times w_i\times \frac{c_i}{T}}$ and $\tilde{F}_p^t \in \mathbb{R}^{h_i\times w_i\times \frac{c_i}{T}}$ represents the t-th group of feature maps of the HS and PAN images, respectively, and T represents the number of groups.

During the process of sequential fusion, each group of features $\tilde{F}_h^t$ and $\tilde{F}_p^t$ are processed through a 3×3 convolution and a Sigmoid activation function to obtain weight maps f_p^t and f_h^t, which are used as forget gates. Meanwhile, $\tilde{F}_h^t$ and $\tilde{F}_p^t$ are respectively fed into a 3×3 convolution and a Tanh function to obtain the features c_h^t and c_p^t to be integrated. The above operations can be represented by:

$$f_h^t = \bar{S}(Conv_3(\tilde{F}_h^t)),\ f_p^t = \bar{S}(Conv_3(\tilde{F}_p^t)) \quad (5)$$

$$c_h^t = \bar{T}(Conv_3(\tilde{F}_h^t)),\ c_p^t = \bar{T}(Conv_3(\tilde{F}_p^t)) \quad (6)$$

where $\bar{S}(\cdot)$ represents the Sigmoid activation function, $\bar{T}(\cdot)$ represents the Tanh activation function, $Conv_3(\cdot)$ indicates the convolution operation with a kernel size of 3×3.

Next, the forget gates are used to integrate the fusion features $s^{t\text{-}1}$ of the previous state with the group features c_h^t and c_p^t of the current state to generate the spectral spatial correlation features s_p^t and s_h^t of the t-th group. Here, s^t denotes the

fusion features obtained by fusing s_p^t and s_h^t through SMIU.

$$s_p^t = f_p^t \odot s^{t\text{-}1} + (1 - f_p^t) \odot c_p^t \tag{7}$$

$$s_h^t = f_h^t \odot s^{t\text{-}1} + (1 - f_h^t) \odot c_h^t \tag{8}$$

$$s^{t+1} = SMIU(s_p^t, s_h^t) \tag{9}$$

where $\odot$ represents the pixel-wise multiplication operation. $SMIU(\cdot)$ represents the SMIU, which fuses the PAN and HS state features in the low-dimensional space.

Subsequently, to ensure that the spectral features are not distorted, $\tilde{F}_h^t$ is used to generate a reset gate r_h^t and candidate featuresh_h^t. The reset gate r_h^t is multiplied with the state features to enhance the fusion features, and $(1\text{-}r_h^t)$ is multiplied with candidate features h_h^t to supplement spectral features. By synthesizing the enhanced fusion features and the supplementary spectral features, the spectral-spatial fusion features h_t of each group can be obtained. Therefore, the output $\tilde{F}_{SRFU}^i$ of SRFU is composed of T groups of fused features. This process can be expressed as follows:

$$r_h^t = \bar{S}(Conv_3(\tilde{F}_h^t)),\ h_h^t = \bar{T}(Conv_3(\tilde{F}_h^t)) \tag{10}$$

$$h_t = r_h^t \odot s^{t\text{-}1} + (1 - r_h^t) \odot h_h^t \tag{11}$$

$$\tilde{F}_{RSFU}^i = \mathrm{Cat}\,(h_1, h_2, ..., h_t, ..., h_T) \tag{12}$$

where $Cat(\ \cdot\ )$ denotes the concatenation operation. As shown in Eq.(12), h_t depends only on the preceding groups $(1 \sim t)$, ignoring the subsequent ones $(t \sim T)$. To capture spectral correlation across all groups, we introduce a bidirectional SRFU that models both past and future groups.

State Memory Interaction Unit

Geng et al. (Geng et al. 2021) argued that attention mechanisms aim to extract a set of concepts from unconscious pixel structures for conscious reasoning, revealing latent pixel correlations. These concepts can be viewed as a base dictionary for encoding feature matrices. In HS sharpening, LRHS, PAN, and HRHS images share structural information but differ in representation. Since LRHS and PAN contain spectral and spatial cues from HRHS, the base dictionary derived from them should match that of HRHS. Therefore, we reconstruct fused features by learning a base dictionary from the combined features of PAN and LRHS images.

Based on the above analysis, SMIU is constructed as shown in Figure 2 (b), which uses non-negative matrix factorization (NMF) method to model the global correlation of image features in low dimensional space, in order to achieve the fusion of two types of image features. Compared to the Transformer, SMIU has lower computational complexity and inference time. The specific operations are as follows. Firstly, the features s_p^t and s_h^t are concatenated and fed into a 3×3 convolution to obtain mixed state features X_{hp}. Then, NMF is performed on the mixed features to learn the base dictionary $D_{hp} \in \mathbb{R}^{\frac{c_i}{T} \times r}$ and coefficient matrix $C_{hp} \in \mathbb{R}^{r \times h_i w_i}$, in order to achieve the reconstruction of fused features. Finally, a 3×3 convolution is used for feature integration to obtain group fusion features. The above process can be represented by the following equations.

$$X_{hp} = Conv_3(Cat(s_p^t, s_h^t)) \tag{13}$$

Algorithm 1: Non-negative Matrix Factorization

Input X_{hp}. Initialize non-negative D_{hp}, C_{hp}.
for $k = 1$ **to** K **do**
$C_{hp}^{ij} \leftarrow C_{hp}^{ij} \frac{(D_{hp}^T X_{hp})_{ij}}{(D_{hp}^T D_{hp} C_{hp})_{ij}}, D_{hp}^{ij} \leftarrow D_{hp}^{ij} \frac{(X_{hp} C_{hp}^T)_{ij}}{(D_{hp} C_{hp} C_{hp}^T)_{ij}}$
end for
return Output $\overline{X}_{hp} = D_{hp} C_{hp}$

$$\begin{aligned} X_{hp} &= \overline{X}_{hp} + E_{hp} = D_{hp} C_{hp} + E_{hp} \\ \min_{D,C} &\|X_{hp} - D_{hp} C_{hp}\|_F \text{ s.t. } \mathrm{D}_{hp}^{ij} \geq 0,\ \mathrm{C}_{hp}^{jk} \geq 0 \end{aligned} \tag{14}$$

$$s^t = Conv_3(\bar{X}_{hp}) \tag{15}$$

where "One-step gradient" from (Geng et al. 2021) is adopted for iterative network updates, as detailed in Algorithm 1, which updates D_{hp} and C_{hp}.

Step-by-step Supervision Strategy

To better constrain the accuracy of the reconstructed spatial and spectral features, a step-by-step supervision strategy is designed, as shown in Figure 2. In the first step, a spatial-channel uncertainty loss L_{EU} is defined by estimating spatial and channel uncertainty of fused result constrain spatial and spectral features. In the second step, an uncertainty-guided reconstruction loss L_{UG} is defined to constrain the fusion result at the pixel level.

Spatial-channel uncertainty loss L_{EU}. Considering the characteristics of HS images, the uncertainty is decomposed into spatial and channel uncertainty, which reflect the reconstruction difficulty of spatial structures and spectral information. Assuming the GT image $I_{m,n}^{GT}$ (where m and n represent the spatial location and the channel location, respectively.) follows a Laplace distribution, the joint uncertainty is determined by the spatial uncertainty θ_m^s and the channel uncertainty θ_n^c. The observation model can be expressed as:

$$I_{m,n}^{GT} = f(\ \mathrm{LRHS} \uparrow, PAN)_{\mathrm{m,n}} + \varepsilon_{m,n}\theta_{m,n},\ \theta_{m,n} = \theta_m^s \cdot \theta_n^c \tag{16}$$

where $\varepsilon_{m,n}$ denotes the standard Laplacian noise. $f(\ \cdot\ )$ denotes the fusion network. Due to the strong coupling between the spatial and channel uncertainties, the joint uncertainty is defined as $\theta_{m,n} = \theta_m^s \cdot \theta_n^c$, with its distribution as:

$$p(I_{m,n}^{GT}, \theta_m^s, \theta_n^c | f(\cdot)_{m,n}) = p(I_{m,n}^{GT} | f(\cdot)_{m,n}, \theta_m^s, \theta_n^c) \cdot p(\theta_m^s) \cdot p(\theta_n^c) \tag{17}$$

The Laplace distribution is adopted to characterize the likelihood function:

$$p(I_{m,n}^{GT}, \theta_m^s, \theta_n^c | f(\cdot)_{m,n}) = \frac{1}{2\theta_m^s \theta_n^c} \exp\left(- \frac{\|I_{m,n}^{GT} - f(\ \cdot\ )_{m.n}\|_1}{\theta_m^s \theta_n^c} \right) \tag{18}$$

By introducing Jeffrey's prior for estimating sparse uncertainty (Figueiredo 2001), the log-likelihood of Eq.(18) is as follows:

$$\ln p(I_{m,n}^{GT}, \theta_m^s, \theta_n^c | f(\cdot)_{m,n}) = - \frac{\left\| I_{m,n}^{GT} - f(\cdot)_{\mathrm{m,n}} \right\|_1}{\theta_m^s \theta_n^c} - 2\ln\theta_m^s - 2\ln\theta_n^c \tag{19}$$

To ensure the training stability, let k_m^s=lnθ_m^s, k_n^c=lnθ_n^c. The total uncertainty loss function L_{EU} can be defined as:

$$L_{EU}=\frac{1}{M\cdot N}\sum_{m=1}^{M}\sum_{n=1}^{N}\mathrm{e}^{-(k_m^s+k_n^c)}\cdot\left\|I_{m,n}^{GT}\text{-}f(\cdot)_{\mathrm{m,n}}\right\|_1+2k_m^s+2k_n^c \tag{20}$$

The L_{EU} loss function is used to guide the first stage training of the network to obtain pre-trained network $f(\cdot)$ and data uncertainty estimates $\theta_{m,n}$.

Uncertainty-guided reconstruction loss L_{UG}. At this stage, the uncertainty estimates $\theta_{m,n}$ from pre-trained network remain fixed. Different from previous mean squared error or L_1 loss functions treating each pixel equally, the proposed uncertainty-guided reconstruction loss for the HS pansharpening task aims to better highlight the priority of pixels with high uncertainty. Since $\theta_m^s\theta_n^c=e^{k_m^s+k_n^c}$, we define the total uncertainty as $k_{m,n}=k_m^s+k_n^c$, which is directly associated with reconstruction loss. That is to say, instead of using $\mathrm{e}^{-(k_m^s+k_n^c)}$ to attenuate the importance of pixels with large uncertainty, we need to use a monotonically increasing function to prioritize them. Therefore, the linear scaling function would be a natural option for defining the loss function, which can be defined as follows:

$$L_{UG}=\frac{1}{M\cdot N}\sum_{m=1}^{M}\sum_{n=1}^{N}\tilde{k}\cdot\left\|I_{m,n}^{GT}-f(\cdot)_{\mathrm{m,n}}\right\|_1 \tag{21}$$

where $\tilde{k}_{m,n}=k_{m,n}-\min_{m,n}(k_{m,n})$ is a non-negative linear scaling function, giving higher weight to spatial regions and spectral channels with higher uncertainty.

In the network, k^s is estimated by using a convolutional layer and ELU function to the fusion result from the base network. k^c is obtained by using global average pooling, fully connected layers, and an ELU function.

Experimental Results

Datasets, Metrics, and Training Details

Extensive experiments were conducted on three simulated datasets (Pavia center (Plaza et al. 2009), Botswana (Ungar 2002), and Chikusei (Yokoya et al. 2016)) and one real dataset FR1 (Zhuo et al. 2022). The datasets were processed following the Wald's protocol (Wald 2000) and set according to Bandara (Bandara et al. 2021). We compared UMNet with some SOTA methods: DIP-Hyperkite (Bandara et al. 2021), GuideNet (Ran et al. 2023), HyperRefiner (Zhou et al. 2023), Tree-SNet (He et al. 2024), DCTransformer (Ma et al. 2024), FPFNet (Dong et al. 2025), and DFCFN (Wang et al. 2025). Five objective metrics were used on the simulated datasets (Loncan et al. 2015; Wang et al. 2004): SAM, SSIM, RMSE, ERGAS, and PSNR, three objective metrics were used in the real dataset (Zhuo et al. 2022): QNR, D_λ, and D_s. We retrained all DL-based methods with Python 3.9 and PyTorch 1.13 on Ubuntu 20.04 system with a NVIDIA RTX A6000. In training, the initial learning rate, epoch and batch size were 0.0001, 8000, and 1, respectively.

Fusion Results

Experimental Results on Simulated Datasets. Table 1 reports objective results on three datasets, with the best and second-best results in bold and underlined, respectively. Our method outperforms others across all metrics. Figure 3, the first row, shows fusion results on the Pavia Center dataset, including MAE maps (lower right) and enlarged local areas (lower left). Our results are visually closest to the GT with minimal residuals. Figure 4 illustrates spectral difference curves for three randomly selected pixels from three dataset images, showing that our method achieves the least spectral deviation, indicating superior spectral preservation.

Experimental Results on Real Dataset. We conducted experiments on the real HS dataset FR1. As shown in the second row of Figure 3, our results exhibit sharper edges and richer textures, closely resembling the PAN image, while other methods suffer from blurring. Table 2 shows our method outperforms all nonreference metrics, demonstrating good performance on the FR1 dataset.

Ablation Studies

Effectiveness of the SRFU Structure. We conducted ablation experiments on SRFU. Model A replaces SRFU with the state-space model LFMamba (Xia et al. 2024); Model B removes the sequence modeling; Model C replaces SMIU with simple addition. As shown in Table 3, incorporating both sequence modeling and SMIU yields better results.

Effectiveness of L_{EU} and L_{UG}. To validate the proposed loss, we conducted ablation experiments with different combinations of L_1, L_{EU}, and L_{UG}: using L_1 alone, L_{EU} alone, and L_{EU} + L_{UG}. We also sequentially removed spatial uncertainty θ_m^s and spectral uncertainty θ_n^c to assess their impact. As shown in Table 4, the proposed loss outperforms L_1, and combining L_{EU} with L_{UG} yields better results than L_{EU} alone. Performance drops when either uncertainty is removed, validating the loss design and training strategy. Figure 5 further confirms our method's spectral fidelity.

The number of channels within the group. We conducted ablation experiments on the channel count per group in the SRFU (Table 5), with the best performance achieved at 8 channels, which we adopt in our work.

The number of latent dimension in SMIU. We conducted ablation experiments on the latent dimension in the SMIU. As shown in Table 6, the best performance is achieved when $r=c/4$ (c is the number of channels per group).

Model Complexity. We compared the parameters and complexity of six DL-based models (Table 7). Our method uses more parameters than GuideNet and Tree-SNet but achieves higher PSNR, and delivers better results than DCTransformer and DFCFN with fewer parameters.

Classification Application To further validate our method, we performed ground object classification on the fused results using k-means in ENVI. As shown in Figure 6, our results are closest to the GT, and Table 8 confirms superior fusion quality through accuracy metrics.

Conclusion

In this paper, we propose a two-stage fusion network guided by spectral and spatial uncertainty. It consists of parallel dual U-Nets for multi-scale feature extraction and fusion from PAN and LRHS images. At each scale, an SRFU

Datasets	Methods	PSNR (↑)	SAM (↓)	SSIM (↑)	RMSE$\times 10^{-2}$ (↓)	ERGAS (↓)
Pavia center	DIP-Hyperkite (TGRS 2022)	36.9586	5.3630	0.9527	1.5661	2.9363
	GuideNet (TC 2023)	37.3485	5.2465	0.9529	1.5211	2.8542
	HyperRefiner (IJDE 2023)	37.7193	5.0461	0.9588	1.4265	2.7748
	Tree-SNet (JSTARS 2024)	38.1569	4.8628	0.9594	1.3815	2.6432
	DCTransformer (IF 2024)	38.4304	4.8032	0.9626	1.3239	2.5836
	FPFNet (TNNLS 2025)	37.2121	5.3962	0.9520	1.5297	2.8992
	DFCFN (TGRS 2025)	38.5797	4.7225	0.9621	1.3112	2.5486
	UMNet (Ours)	**39.0823**	**4.5904**	**0.9647**	**1.2478**	**2.4301**
Botswana	DIP-Hyperkite (TGRS 2022)	42.4773	1.7728	0.9576	1.2897	1.8724
	GuideNet (TC 2023)	45.1494	1.5985	0.9673	1.0665	1.3201
	HyperRefiner (IJDE 2023)	44.2455	1.6751	0.9663	1.1655	1.3997
	Tree-SNet (JSTARS 2024)	45.2278	1.5787	0.9671	1.0738	1.3039
	DCTransformer (IF 2024)	44.9387	1.5483	0.9671	1.0546	1.3320
	FPFNet (TNNLS 2025)	44.1665	1.8960	0.9621	1.2396	1.4639
	DFCFN (TGRS 2025)	44.9591	1.5767	0.9687	1.0830	1.3166
	UMNet (Ours)	**45.5051**	**1.5091**	**0.9690**	**1.0290**	**1.2909**
Chikusei	DIP-Hyperkite (TGRS 2022)	41.6503	2.3207	0.9702	0.9246	4.0896
	GuideNet (TC 2023)	42.7188	2.2169	0.9774	0.7961	3.6830
	HyperRefiner (IJDE 2023)	42.9958	2.1112	0.9777	0.8046	3.5126
	Tree-SNet (JSTARS 2024)	43.2155	2.0608	0.9795	0.7487	3.4806
	DCTransformer (IF 2024)	43.2167	2.0527	0.9802	0.7450	3.4965
	FPFNet (TNNLS 2025)	42.3698	2.3260	0.9759	0.8243	3.8711
	DFCFN (TGRS 2025)	43.4759	1.9978	0.9792	0.7680	3.3169
	UMNet (Ours)	**43.9936**	**1.9097**	**0.9824**	**0.6847**	**3.2346**

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

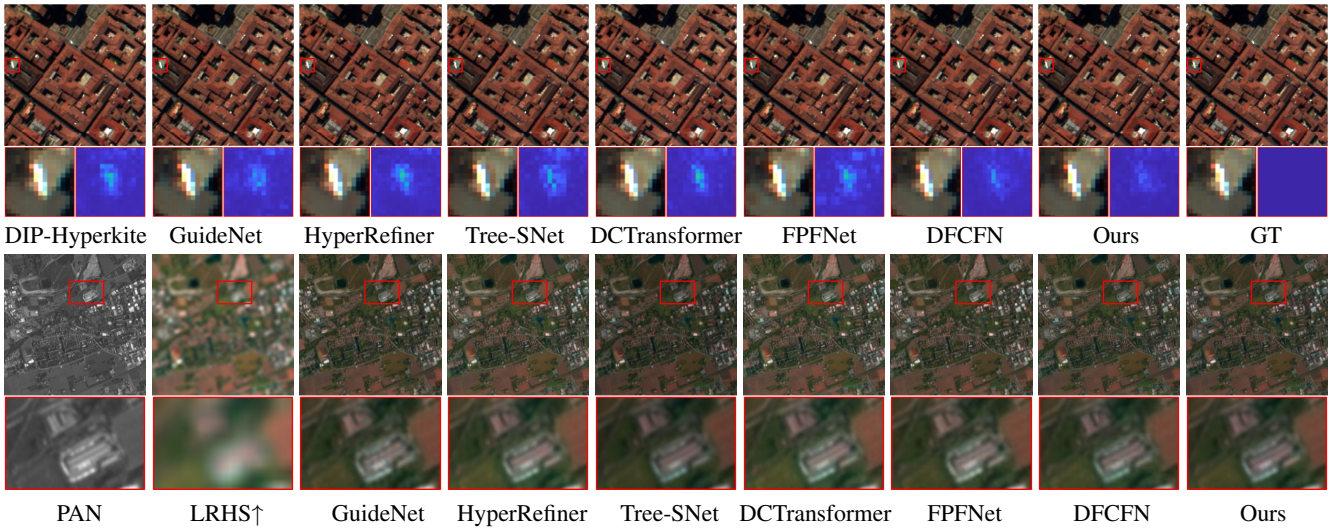
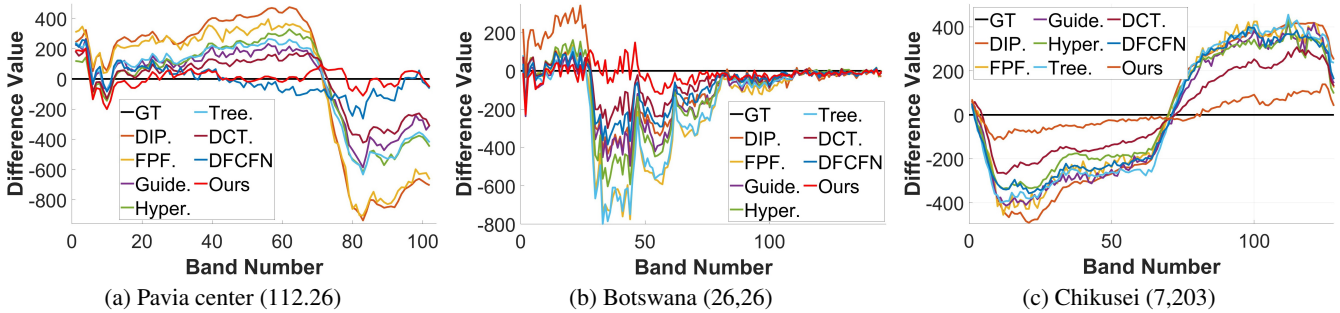
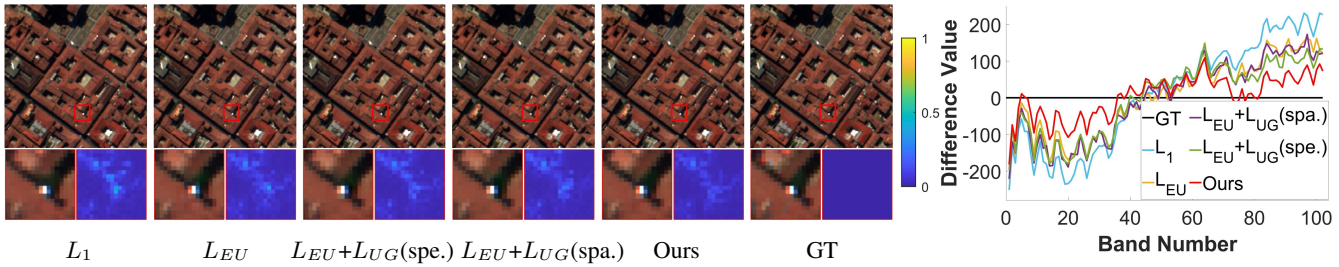


Figure 3: Visual comparison of fusion results on the Pavia Center simulated dataset and the FRI real dataset.



(a) Pavia center (112.26) (b) Botswana (26,26) (c) Chikusei (7,203)

Figure 4: Spectral difference curves. The closer the curve is to the GT, the higher spectral similarity to the GT.

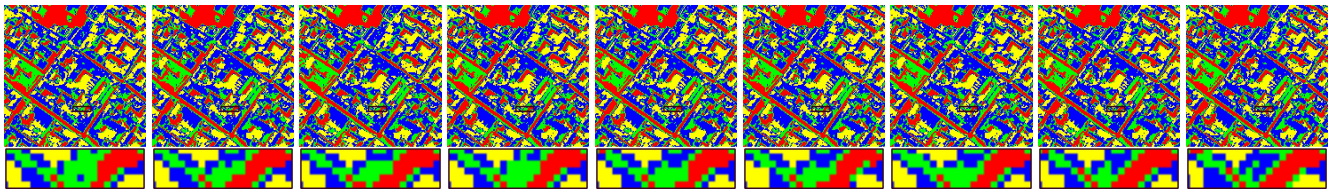


L_1 L_{EU} L_{EU}+L_{UG}(spe.) L_{EU}+L_{UG}(spa.) Ours GT

(a). The visual comparisons of results for L_{EU} and L_{UG}.

(b). The spectral curves of results for L_{EU} and L_{UG}.

Figure 5: Ablation study results of $\boldsymbol{L_{EU}}$ and $\boldsymbol{L_{UG}}$.



DIP-Hyperkite GuideNet HyperRefiner Tree-SNet DCTransformer FPFNet DFCFN Ours GT

Figure 6: Classification maps on the Pavia center dataset.

Methods	D_λ (↓)	D_s (↓)	QNR (↑)
DIP-Hyperkite	0.0272	0.0260	0.9474
GuideNet	0.0334	0.0282	0.9393
HyperRefiner	0.0257	0.0242	0.9505
Tree-SNet	0.0335	0.0243	0.9335
DCTransformer	0.0255	0.0237	0.9514
FPFNet	0.0394	0.0327	0.9293
DFCFN	0.0227	0.0202	0.9576
UMNet (Ours)	**0.0219**	**0.0178**	**0.9606**

Table 2: The average quantitative results on the FR1 dataset.

Methods	PSNR ↑	SAM ↓	SSIM ↑	RMSE ↓	ERGAS ↓
Model A	37.8800	5.0821	0.9581	1.4132	2.7945
Model B	38.7148	4.6730	0.9620	1.2978	2.5100
Model C	38.7210	4.7383	0.9628	1.2987	2.5071
Ours	**39.0832**	**4.5904**	**0.9647**	**1.2478**	**2.4300**

Table 3: Ablation study results of SRFU.

Methods	PSNR ↑	SAM ↓	SSIM ↑	RMSE ↓	ERGAS ↓
L_1	38.7714	4.6581	0.9629	1.2886	2.5006
L_{EU}	38.8343	4.7067	0.9637	1.2785	2.4788
L_{EU}+L_{UG} (θ_m^s)	38.9200	4.6457	0.9637	1.2686	2.4644
L_{EU}+L_{UG} (θ_n^c)	38.9117	4.6124	0.9638	1.2696	2.4662
L_{EU}+L_{UG}	**39.0832**	**4.5904**	**0.9647**	**1.2478**	**2.4300**

Table 4: Ablation study results of $\boldsymbol{L_{EU}}$ and $\boldsymbol{L_{UG}}$.

numbers	PSNR ↑	SAM ↓	SSIM ↑	RMSE ↓	ERGAS ↓
4	37.4640	5.3307	0.9517	1.4903	2.8404
8	**39.0832**	**4.5904**	**0.9647**	**1.2478**	**2.4300**
16	38.0425	4.9059	0.9569	1.3930	2.6776

Table 5: Ablation study on the group channel number.

r	PSNR ↑	SAM ↓	SSIM ↑	RMSE ↓	ERGAS ↓
$c/2$	38.4235	4.7969	0.9597	1.3364	2.5887
$c/4$	**39.0832**	**4.5904**	**0.9647**	**1.2478**	**2.4300**
$c/8$	38.4771	4.7782	0.9598	1.3280	2.5728

Table 6: Ablation study on the latent dimension in SMIU.

Methods	Guide Net	Hyper Refiner	Tree SNet	DCT ransformer	DFCFN	FPF Net	Ours
#Params (M)	5.1	19.3	9.1	18.5	11.0	22.3	9.2
FLOPs (G)	150.2	85.0	205.0	3220.2	83.4	174.9	137.1

Table 7: Comparison of Computational Complexity.

Methods	Guide Net	Hyper Refiner	Tree SNet	DCT ransformer	DFCFN	FPF Net	Ours
OA (↑)	0.950	0.944	0.952	0.948	0.946	0.953	**0.957**
Kappa (↑)	0.932	0.923	0.934	0.929	0.926	0.935	**0.941**

Table 8: Comparison of Computational Complexity.

treats band features as temporal sequences for recurrent fusion. To exploit shared structure, NMF is used to build an SMIU that fuses features in a low-dimensional space. Two uncertainty-based loss functions further guide the network toward high-fidelity fusion. Experiments show that UMNet achieves SOTA performance on four datasets.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62072218, in part by the Natural Science Foundation of Tianjin under Grant 24JCZDJC00130 and Grant 25JCZDJC00540, in part by the Natural Science Foundation of Hebei under Grant F2025110006 and Grant F2025110010, and in part by the Cangzhou Institute of Tiangong University under Grant TGCYY-Z-0303.

References

Aburaed, N.; Alkhatib, M.; 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.

Bandara, W. G. C.; Valanarasu, J. M. J.; Patel, V. M.; and Patel, V. M. 2021. 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. 2025. Feature Pyramid Fusion Network for Hyperspectral Pansharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 1555–1567.

Figueiredo, M. 2001. Adaptive Sparseness Using Jeffreys Prior. In Dietterich, T.; Becker, S.; and Ghahramani, Z., eds., *Advances in Neural Information Processing Systems*, volume 14, 697–704. MIT Press.

Geng, Z.; Guo, M.-H.; Chen, H.; Li, X.; Wei, K.; and Lin, Z. 2021. Is Attention Better Than Matrix Decomposition? In *International Conference on Learning Representations (ICLR)*.

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.

Liu, L.; Xu, X.; Xu, X.; and Xu, X. 2023. Self-attention Mechanism at the Token Level: Gradient Analysis and Algorithm Optimization. *Knowledge-Based Systems*, 277: 110784.

Loncan, L.; de Almeida, L. B.; Bioucas-Dias, J. M.; Briottet, X.; Chanussot, J.; Dobigeon, N.; Fabre, S.; Liao, W.; Licciardi, G. A.; Simões, M.; Tourneret, J.-Y.; Veganzones, M. A.; Vivone, G.; Wei, Q.; and Yokoya, N. 2015. Hyperspectral Pansharpening: A Review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3): 27–46.

Ma, Q.; Jiang, J.; Liu, X.; and Ma, J. 2024. Reciprocal transformer for hyperspectral and multispectral image fusion. *Information Fusion*, 104: 102148.

Peng, S.; Guo, C.; Wu, X.; and Deng, L.-J. 2023. U2net: A general framework with spatial-spectral-integrated double u-net for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 3219–3227.

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.

Ran, R.; Deng, L.-J.; Jiang, T.-X.; Hu, J.-F.; Chanussot, J.; and Vivone, G. 2023. GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution. *IEEE Transactions on Cybernetics*, 53(7): 4148–4161.

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.

Wang, J.; Lu, T.; Huang, X.; Zhang, R.; and Feng, X. 2024. Pan-sharpening via conditional invertible neural network. *Information Fusion*, 101: 101980.

Wang, Q.; Tang, Y.; Ge, Y.; Xie, H.; Tong, X.; and Atkinson, P. M. 2023. A comprehensive review of spatial-temporal-spectral information reconstruction techniques. *Science of Remote Sensing*, 8: 100102.

Wang, X.; Yang, Y.; Huang, S.; Wan, W.; Liu, Z.; Zhang, L.; and Zhao, A. 2025. DFCFN: Dual-Stage Feature Correction Fusion Network for Hyperspectral Pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–14.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Xia, W.; Lu, Y.; Wang, S.; Wang, Z.; Xia, P.; and Zhou, T. 2024. LFMamba: Light Field Image Super-Resolution with State Space Model. arXiv:2406.12463.

Xie, J.; Miao, Q.; Liu, R.; Xin, W.; Tang, L.; Zhong, S.; and Gao, X. 2021. Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition. *Neurocomputing*, 440: 230–239.

Yokoya, N.; Iwasaki, A.; Iwasaki, A.; 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.

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.