PRIOR-BASED THREE-STAGE UNSUPERVISED INVERTIBLE NEURAL NETWORK FOR HYPERSPECTRAL AND MULTISPECTRAL IMAGE FUSION

Mengnan Jin[1,2], Wenjuan Zhang[1],✉, Yongchuan Cui[1,2], Jie Pan[1], Dailiang Peng[1],
[1]*Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China*
[2]*School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China*
✉Corresponding Author, *zhangwj@aircas.ac.cn*

***Abstract*—Hyperspectral images (HSI) inherently face a trade-off between spatial resolution and spectral resolution due to the limitations of imaging principles. To rapidly obtain remote sensing images with both high spatial and high spectral resolution, unsupervised deep learning methods for fusing HSI and multispectral images (MSI) have achieved remarkable progress in recent years. However, existing studies often overlook the issue of high-frequency information loss in the fused images. To address this limitation, we propose a Prior-based three-stage unsupervised Invertible neural Fuse Network (PIFNet). Specifically, the framework consists of three key modules: prior information extraction, spectral channel mapping, and detail feature fuse. In particular, the detail feature fuse module leverages an invertible neural network to prevent information loss through mutual generation of input and output features. Experimental results on simulated datasets demonstrate that the PIFNet outperforms existing unsupervised deep learning methods, highlighting its potential and effectiveness in HSI-MSI fusion tasks.**

***Index Terms*—remote sensing, hyperspectral image,image fusion,super-resolution, deep learning, invertible neural network**

I. Introduction

Hyperspectral remote sensing images possess a unique data structure that integrates spatial and spectral information, enabling a wide range of applications [1]. However, due to the inherent limitations of imaging principles, there exists an inevitable trade-off between spatial resolution and spectral resolution, resulting in HSI generally having low spatial resolution [2]. This limitation significantly hinders the practical applications of HSI. Therefore, fusing high spatial resolution MSI (HrMSI) with low spatial resolution HSI (LrHSI) presents an effective solution to address this challenge.

Current fusion methods can be broadly categorized into two types: traditional methods and deep learning-based methods. Traditional methods, including pan-sharpening, Bayesian fusion, and matrix decomposition [3]. Ren *et, al.* developed a novel joint fusion model based on spectral unmixing, tailored for different typical ground objects [4]. Rely heavily on extensive prior knowledge, which limits their generalization capability. Deep learning-based methods involve constructing deep neural networks and training the models using large amounts of data. Ultimately, the input image is fed into the network to obtain the target image. Deep learning-based methods can be further divided into supervised and unsupervised learning approaches. Considering that the target images are typically unavailable in real-world applications, supervised deep learning fusion methods face the challenge of lacking appropriate training datasets [5], which restricts their applicability. Consequently, unsupervised deep learning methods have emerged as the mainstream approach in recent years. For instance, inspired by spectral unmixing, Yao *et, al.* designed a novel coupled unmixing network with a cross-attention mechanism, referred to as CUCaNet, to enhance the spatial resolution of LrHSI using HrMSI [6]. Zheng *et, al.* proposed an unsupervised deep learning method named HyCoNet to address the fusion problem under unknown Spatial Response Function (SRF) and Point Spread Function (PSF) conditions [3]. Similarly, Liu *et, al.* designed a spectral diffusion model to capture the spectral distribution of HSI and leveraged the prior knowledge of both MSI and HSI to optimize the generative direction of the model [7]. However, existing models primarily rely on the forward propagation of Convolutional Neural Network (CNN) for image fusion or generation, which often results in the irreversible compression or loss of high-frequency information within the images [8].

To address this issue, we propose a three-stage unsupervised invertible neural network based on image prior knowledge. Specifically, in the first stage, similar to previous works, the model learns the PSF and SRF information of the fused images to facilitate the design of the loss function for subsequent fusion networks. In the second stage, spectral mapping from MSI to HSI is performed through weight sharing. In the third stage, an Invertible Neural Network (INN) is employed to extract detailed features from the fused images, enabling lossless information transmission and effectively preserving the high-frequency information of the original images. Through these three stages, our network achieves high-quality image reconstruction within an unsupervised framework.

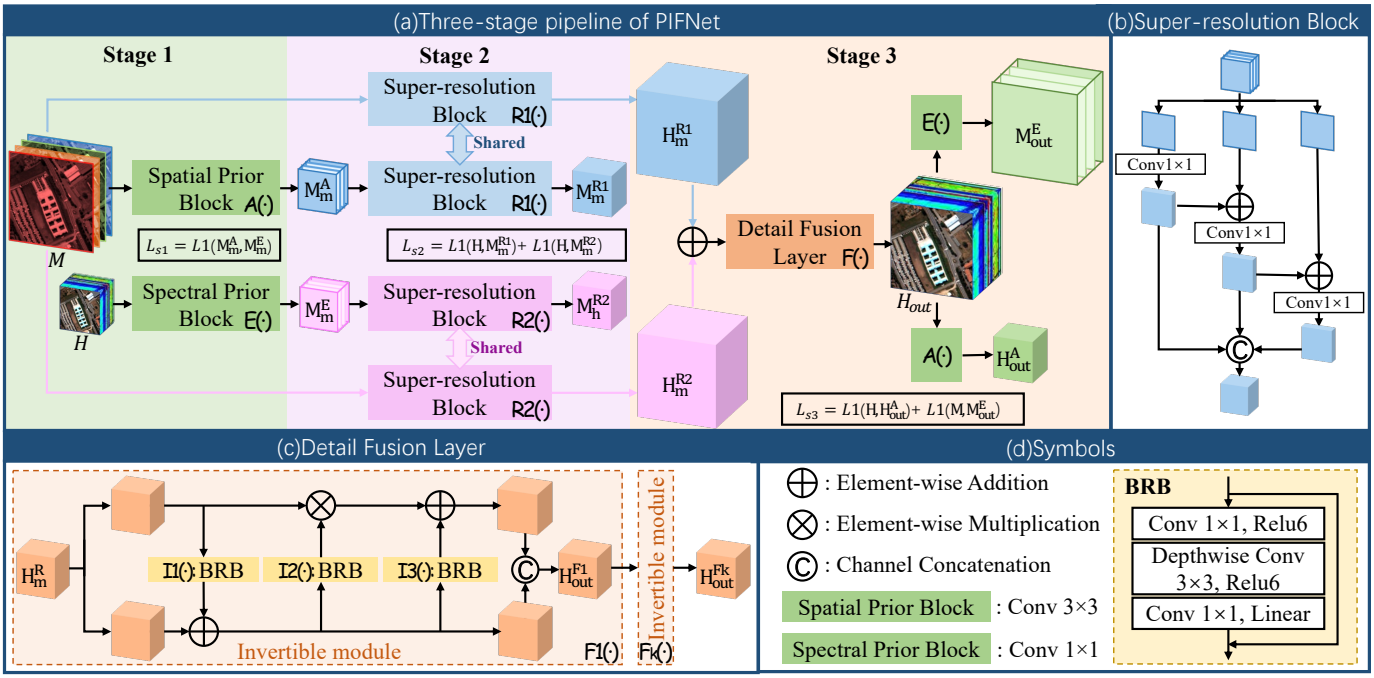

Fig. 1: The overall architecture of PIFNet, (a) the three-stage workflow diagram of the PIFNet, (b) the detailed architecture of the super-resolution module in the second stage, (c) the detailed architecture of the detail fusion layer in the third stage, and (d) the explanation of some symbols used in the workflow diagram.

II. Methodology

Our proposed PIFNet consists of three stages: the prior information extraction stage, the spectral channel mapping stage, and the detail feature fusion stage. Fig.1(a) illustrates the detailed workflow. Specifically, the LrHSI and the HrMSI are used as inputs to the model. The fusion process generates a high-spatial-resolution hyperspectral image (HrHSI) as the output.

A. Prior Information Extraction Stage

The first stage primarily consists of two components, namely the spatial prior block and the spectral prior block, which are denoted as $A(\cdot)$ and $E(\cdot)$, respectively. These two blocks are designed to learn the prior information of SRF and PSF required for spatial and spectral downsampling of the images. Specifically, the HrMSI $M \in \mathbb{R}^{C_m \times H_m \times W_m}$ is spatially downsampled to obtain a low-spatial-resolution multispectral image (LrMSI) $M_m^A \in \mathbb{R}^{C_m \times H_h \times W_h}$, while the LrHSI $M \in \mathbb{R}^{C_h \times H_h \times W_h}$ is spectrally downsampled to produce another LrMSI$M_m^E \in \mathbb{R}^{C_m \times H_h \times W_h}$:

$$M_m^A = A(M), \quad M_m^E = E(H) \tag{1}$$

Among them, $A(\cdot)$ and $E(\cdot)$ are two learnable weight functions. Referring to the idea of the degradation model, the input LrHSI and HrMSI are highly registered images of the same scene. Therefore, the two generated LrMSI should remain consistent. Based on this condition, the loss function for the first stage is designed as follows:

$$L_{S_1} = \frac{1}{N} \sum_{i=1}^{N} \left| (M_m^A - M_m^E) \right| \tag{2}$$

Where N represents the total number of data points in the three-dimensional image. The L1 loss function is employed to constrain the absolute error between the two images, ensuring their consistency.

B. Spectral Channel Mapping Stage

Inspired by the work of Li *et, al.* [9], the fusion problem is reformulated as a spectral mapping problem. The specific process of spectral super-resolution is illustrated in Fig.1(b). Two consistent spectral super-resolution networks are utilized to perform spectral super-resolution on the two LrMSI generated in the first stage, reconstructing the LrHSI $M_m^{R1} \in \mathbb{R}^{C_h \times H_h \times W_h}$ and $M_m^{R2} \in \mathbb{R}^{C_h \times H_h \times W_h}$.

$$M_m^{R1} = R1(M_m^A), \quad M_m^{R2} = R2(M_m^E) \tag{3}$$

Where $R1(\cdot)$ and $R2(\cdot)$ represent the spectral super-resolution networks. The two super-resolved images are constrained with the initially input LrHSI, and the loss function for the second stage is defined as follows.

$$L_{S_2} = \frac{1}{N} \sum_{i=1}^{N} \left| (H - M_m^{R1}) \right| + \frac{1}{N} \sum_{i=1}^{N} \left| (H - M_m^{R2}) \right| \tag{4}$$

After completing the training of the network in the stage 2 using the aforementioned process, the initial HrMSI

M is super-resolved into two simulated HrHSI $H_m^{R1} \in \mathbb{R}^{C_h \times H_m \times W_m}$ and $H_m^{R2} \in \mathbb{R}^{C_h \times H_m \times W_m}$ through the shared weights of these two networks.

$$H_m^{R1} = R1(M), \quad H_m^{R2} = R2(M) \tag{5}$$

The two simulated HRHSI will be used as inputs in the next phase, where they will undergo fusion and feature extraction to get the final result.

C. Detail Feature Fusion Stage

We utilize INN to extract high-frequency features from the images, enabling the lossless transmission of information Fig. 1(c) illustrates the specific data flow in the stage 3. Initially, the two simulated HrHSI H_m^{R1} and H_m^{R2} obtained in the second stage are fused to another HrHSI $H_m^R \in \mathbb{R}^{C_h \times H_m \times W_m}$, followed by the extraction of high-frequency detail features from the fused image.

$$H_m^R = cat(H_m^{R1}, H_m^{R2}), \quad H_{out}^{F1} = F1(H_m^R) \tag{6}$$

Among them, $cat(\cdot)$ denotes tensor concatenation, $F1(\cdot)$ represents the function of the first INN layer, and $H_{out}^{F1} \in \mathbb{R}^{C_h \times H_m \times W_m}$ is the output of the first INN layer. The input and output results of the k-th INN can be expressed as:

$$H_{out}^{F_k} = F_k(H_m^{F_{k-1}}) \tag{7}$$

Therefore, the final fused result $H_o ut \in \mathbb{R}^{C_h \times H_m \times W_m}$ obtained after passing through n INN is:

$$H_{out} = F(H_m^R) = F_n(H_m^{F_{n-1}}) \tag{8}$$

We then perform downsampling on H_{out} using the SRF and PSF downsampling information learned in the stage 1, resulting in LrHSI $H_{out}^A \in \mathbb{R}^{C_h \times H_h \times W_h}$ and the HrMSI $H_{out}^E \in \mathbb{R}^{C_m \times H_m \times W_m}$.

$$H_{out}^A = A(H_{out}), \quad M_{out}^E = E(H_{out}) \tag{9}$$

The similarity between the downsampled image and the input image is computed to constrain the fusion quality of H_{out}. Therefore, the loss function for the stage 3 is expressed as follows:

$$L_{S_3} = \frac{1}{N}\sum_{i=1}^{N}\left|\left(H_{out}^A - H\right)\right| + \frac{1}{N}\sum_{i=1}^{N}\left|\left(M_{out}^E - M\right)\right| \tag{10}$$

We employed the simple and commonly used L1 loss as the loss function and achieved good convergence without imposing excessive constraints on the network.

III. Experimental Results

A. Experimental Data and Setup

We selected the publicly available Pavia University dataset as a benchmark for comparison. To simulate the hyperspectral dataset, the original image, with a size of 610×340 pixels, was downsampled by a factor of 8 in spatial resolution. Consequently, a region of size 608×336 pixels was selected for simulation. LrHSI were generated using a Gaussian blur operation as the downsampling operator, while HrMSI were produced by downsampling the spectral information of the image using the QuickBird SRF.

Five state-of-the-art unsupervised methods were selected for comparison, including MIAE [10], UDALN [9], SURE [11], M2U-Net[12], and SDP [7]. In addition, five quantitative metrics were employed for evaluation: Spectral Angle Mapper (SAM), Peak Signal-to-Noise Ratio (PSNR), Error Relative Global Dimensional Synthesis (ERGAS), Structural Similarity Index (SSIM), and Universal Quality Index (UQI).

B. Results and Analysis

Table I presents the quantitative evaluation results. Our method, by employing INN to preserve high-frequency information within the image, provides the best overall performance, demonstrating its effectiveness in spectral fidelity, image quality, and structural preservation. Although it does not yield the best result in terms of ERGAS, the performance of our method is very close to the optimal result obtained by SURE.

TABLE I: The performance of the quantitative evaluation results. The best performance is shown in **bold** and the second best is underlined. "↓" means smaller is better for this index. "↑" means bigger is better for this index.

Method	SAM↓	PSNR↑	ERGAS↓	SSIM↑	UQI↑
MIAE	2.3855	38.68	0.6262	0.9790	0.9959
UDALN	2.9946	31.80	1.7832	0.9683	0.9792
SURE	2.5105	38.63	**0.6239**	0.9753	0.9973
M2U-Net	2.4108	42.47	0.8459	0.9752	0.9972
SDP	2.9064	37.51	0.7137	0.9743	0.9962
PIFNet	**2.3554**	**43.17**	0.7821	**0.9851**	**0.9974**

Fig.2 shows the true color image and pixel-wise SAM heatmap results on the Pavia University dataset. All methods successfully generate fusion results. The true color image of M2U-Net closely matches the ground truth, demonstrating its effective utilization of inherent information in the degraded model. Our method exhibits minimal visual differences with the other four methods. In the heatmap, different colors represent varying SAM values, with pixels corresponding to smaller SAM values appearing closer to blue. The SAM heatmap reveals that the results from our method are generally closer to blue, indicating higher global spectral fidelity. Notably, in high-frequency texture regions, such as the detailed textures

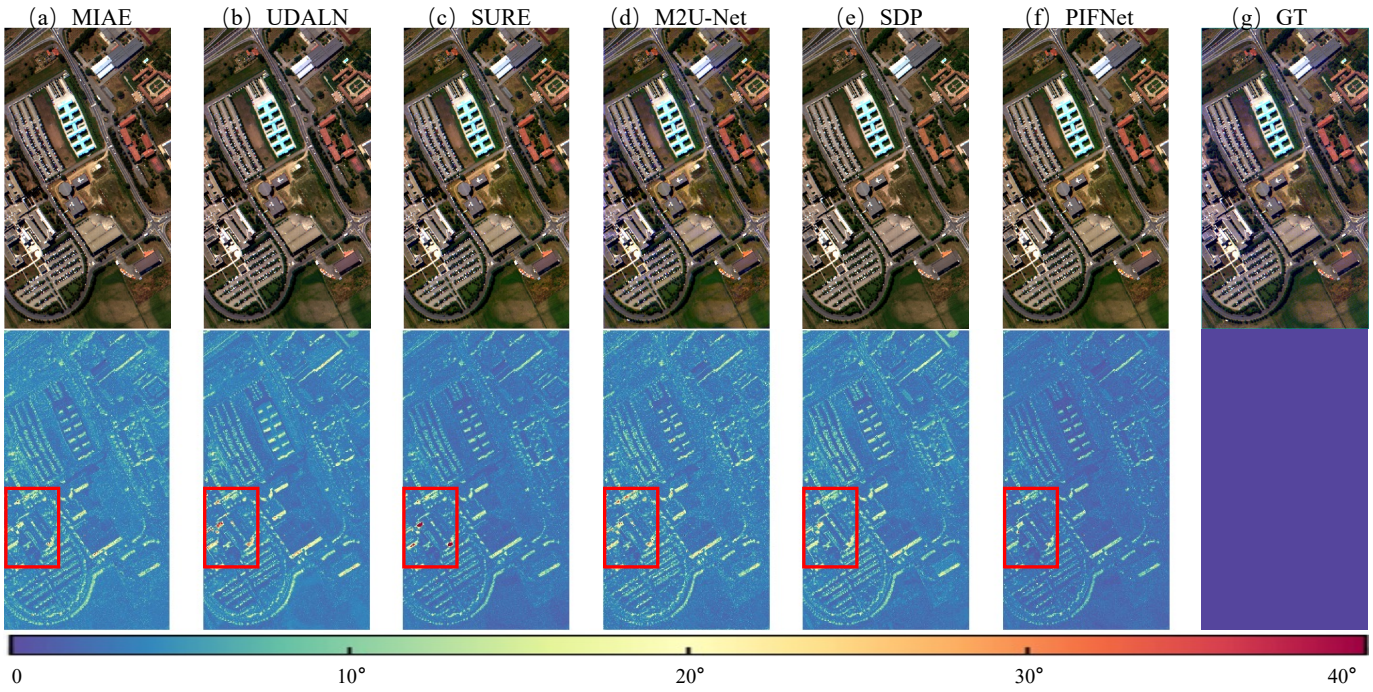

Fig. 2: The visual comparison results on the Pavia University dataset, where the first row shows the true color images generated by the six methods, and the second row displays the pixel-wise SAM heatmap comparison results. (a)MIAE, (b)UDALN, (c)SURE, (d)M2U-Net, (e)SDP, (f)PIFNet, (g)GroundTruth

marked by red boxes, our method demonstrates a more uniform and accurate color distribution, further confirming its advantage in preserving spatial high-frequency details.

IV. Conclusion

This paper proposes a prior-based three-stage unsupervised invertible neural network (PIFNet). Through the three modules of prior information mining, spectral channel mapping, and fusion feature learning, particularly the introduction of INN, Our method effectively prevents the loss of high-frequency information. Experimental results on the simulated dataset demonstrate the effectiveness of our approach in both qualitative and quantitative fusion performance.

References

[1] N. Yokoya, C. Grohnfeldt, and J. Chanussot, "Hyperspectral and multispectral data fusion: A comparative review of the recent literature," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 2, pp. 29–56, 2017.

[2] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Information Fusion*, vol. 69, pp. 40–51, 2021.

[3] K. Zheng, L. Gao, W. Liao, D. Hong, B. Zhang, X. Cui, and J. Chanussot, "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2487–2502, 2020.

[4] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, and J. Huang, "A locally optimized model for hyperspectral and multispectral images fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[5] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, "Mct-net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion," *Knowledge-Based Systems*, vol. 264, p. 110362, 2023.

[6] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*. Springer, 2020, pp. 208–224.

[7] J. Liu, Z. Wu, and L. Xiao, "A spectral diffusion prior for unsupervised hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[8] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5906–5916.

[9] J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[10] J. Liu, Z. Wu, L. Xiao, and X.-J. Wu, "Model inspired autoencoder for unsupervised hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[11] H. V. Nguyen, M. O. Ulfarsson, J. R. Sveinsson, and M. Dalla Mura, "Deep sure for unsupervised remote sensing image fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[12] J. Li, K. Zheng, L. Gao, L. Ni, M. Huang, and J. Chanussot, "Model-informed multi-stage unsupervised network for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.