This ACCV 2024 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2024
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Spectral Modality-Aware Interactive Fusion Network for HSI Super-Resolution

Meng Xu, Jiayou Mao, Ziqian Mo, Xiyou Fu, and Sen Jia*

College of Computer Science and Software Engineering, Shenzhen University, China
senjia@szu.edu.cn

Abstract. Due to the limitations of current spectral imaging equipment in acquiring high-resolution hyperspectral images (HR-HSIs), a common approach is to fuse low-resolution hyperspectral images (LR-HSIs) with high-resolution multispectral images (HR-MSIs). However, most existing methods have not fully taken into account the correlation and discrepancy in modality between hyperspectral images (HSIs) and multispectral images (MSIs). To address this limitation, we propose an innovative spectral modality-aware interactive fusion network (SMIF-NET) for comprehensive extraction of spectral information and seamless feature fusion. First, we introduce the spectral modality-aware transformer (SMAT) with a dual-attention mechanism to compute spectral self-similarity and cross-spectral correlation. Second, we apply the interactive spatial-spectral feature fusion (IS2F2) to fuse the acquired high-level spectral and spatial features. This fusion technique combines spatial-wise and channel-wise squeeze and excitation to achieve seamless integration of spatial-spectral information. Finally, the extensive experiments on three datasets demonstrate the superior performance of SMIF-NET in both visual and quantitative assessments compared to eight state-of-the-art (SOTA) fusion-based methods.

Keywords: hyperspectral image · multispectral image · image fusion · super-resolution

1 Introduction

Hyperspectral images (HSIs) provide an extensive range of spectral information, spanning tens to hundreds of narrow bands that capture distinctive characteristics of diverse materials. However, the pursuit of high spectral resolution in hyperspectral imaging often leads to compromised spatial resolution, resulting in low-resolution hyperspectral images (LR-HSIs). This trade-off is unavoidable in current times. In contrast, a multispectral system can generate images with heightened spatial resolution and coarse spectral resolution, known as high-resolution multispectral images (HR-MSIs). To obtain high-resolution hyperspectral images (HR-HSIs), fusion-based HSI super-resolution methods have been proposed by merging LR-HSIs and HR-MSIs. The advantage of featuring both high spatial and spectral resolution makes it more suitable for various

applications, including classification [1–3], object detection [4], spectral unmixing [5, 6], and HSI inversion [7].

In recent years, deep learning has demonstrated remarkable success in the field of HSI super-resolution by employing deep neural networks to learn intricate mappings between input and output data. For instance, to prevent the loss of crucial low-level structural details, generative adversarial network (GAN)-based edge-enhancement networks were proposed to improve the spatial resolution and visual quality of the generated images by enhancing the edges between objects [8, 9]. Moreover, many fusion-based methods leverage convolutional neural networks (CNNs) to extract image features and tackle the fusion challenge between HSIs and multispectral images (MSIs) [10–14]. The exceptional performance of CNN-based fusion methods can be attributed to their robust inferential capabilities. However, the relatively smaller receptive fields of CNNs limit their ability to effectively capture global features, impacting the overall performance of CNN-based methods. Therefore, benefiting from its excellent capability to capture long-range dependencies, transformer [15] has also been extensively applied in hyperspectral super-resolution [16, 17].

However, the above-mentioned methods overlooked the substantial modality differences between HSIs and MSIs. For example, different types of sensors result in variations in noise and lighting conditions within images. Additionally, HSIs offer detailed spectral information but exhibit low spatial resolution, thereby limiting the capture of surface features. Whereas MSIs can provide higher spatial resolution information for intricate ground feature capture but face constraints in spectral resolution. On the other hand, the aforementioned models have not fully exploited the modal correlation features, leading to a certain degree of spectral and spatial distortion. Since images of the same target scene captured as both HSI and MSI must be precisely registered to achieve fusion-based HSI super-resolution tasks, the specific spatial and spectral correlations between the two multimodal data should be taken fully into consideration. It is imperative to enhance the utilization of modal correlation characteristics to mitigate the observed distortions in both spectral and spatial dimensions.

To address the aforementioned challenges, we introduce an innovative spectral modality-aware interactive fusion network (SMIF-NET) for HSI super-resolution. First, we utilize a CNN to extract spatial texture information from HR-MSIs, capitalizing on the CNN's proficiency in capturing detailed local information. Specifically, we combine the multiscale convolution and dilated convolution with varying rates to expand the receptive field of the convolutional kernel, addressing features at different scales. Second, we employ a dual-attention spectral-dimensional transformer to extract spectral features from LR-HSIs. This design not only enhances the representation capability of spectral features but also reduces computational burden. Finally, we design an interactive spatial-spectral feature fusion (IS2F2) module to seamlessly integrate distinct features. In particular, the cross-directional feature propagation is strategically implemented during the process of feature fusion to enhance the feature interaction, effectively addressing challenges stemming from modality differences.

The main contributions of this paper are detailed as follows:

- We present a spectral modality-aware interactive fusion network (SMIFNET) that contains the spectral modality-aware transformer (SMAT) layers and interactive spatial-spectral feature fusion (IS2F2) module to generate HR-HSIs with high spectral and spatial fidelity.
- The dual-attention mechanism containing spectral modality-aware multi-head attention (SMMA) is designed to comprehensively capture spectral features. A multiscale-dilated spatial feature extraction (MSFE) module is proposed to extract spatial features.
- The spectral and spatial feature maps are integrated into the IS2F2 module, employing channel-wise squeeze and excitation (CSE) and spatial-wise squeeze and excitation (SSE) to enhance the interaction.

2 Related work

2.1 CNN-based HSI Super-resolution Methods

With the development of deep learning, it exhibits a strong ability to capture the intricate features of HSIs and MSIs. For example, Dian et al. [10] proposed a deep HSI sharpening method for the fusion of the LR-HSI with the HR-MSI, which directly learns the image priors through a deep CNN-based residual learning method. Subsequently, Dian et al. [13] presented an HSI-MSI fusion method, leveraging subspace representation and incorporating a CNN denoiser. Notably, this fusion method is applicable to any HSI and MSI without requiring retraining. To fully leverage the spatial-spectral dependency, the multiscale spatial-spectral joint feature learning approaches have been developed to enhance spatial resolution [18, 19]. Considering factors such as blurring, down-sampling, and spectral response function, Wang et al. [20] introduced a two-stream fusion network, integrating it with a physical model and deep prior information.

Given that HSIs are represented as 3-dimensional tensors, Palsson et al. [12] proposed an approach that employs a 3D CNN and incorporates dimensionality reduction for HSIs before fusion. This strategy not only significantly reduces computation time but also enhances the method's robustness to noise. To obtain multiscale contextual features at a fine-grained level, Fu et al. [21] proposed a grouped multiscale dilated network structure, strategically designed to enlarge the receptive fields efficiently. Additionally, Dong et al. [22] suggested utilizing the U-net architecture as a substitute for the ResNet to obtain a denoising prior for HSI and MSI fusion. In summary, CNNs present an effective way of tackling HSI super-resolution challenges, capitalizing on their proficiency in extracting local features and performing non-linear mappings. This implies that CNNs hold significant potential for enhancing the resolution of the HSI through fusion techniques.

2.2 Transformer-based HSI Super-resolution Methods

Transformer, a deep learning model featuring the self-attention mechanism initially introduced by Vaswani et al. [15], possesses a formidable ability to capture

long-range dependencies. Given the abundant spatial and spectral information typically present in HSIs and MSIs, the utilization of transformer facilitates the establishment of global correlations. This capability allows the model to capture correlations from various regions and bands of the image, surpassing the performance of CNNs in effectively fusing the bimodal information. Cai et al. [17] first employed the transformer for the task of HSI reconstruction. Specifically, a mask-guided spectral transformer (MST) model was proposed by incorporating spectral-guided multiple self-attention, treating individual spectral features as markers, and calculating self-attention along the spectral dimension. Ma et al. [16] replaced U-net with transformer based on a model-guided deep hyperspectral image super-resolution framework (MoG-DCN) [22].

To address the limitations of vision transformer in transferring spatial features at different scales, Jia et al. [23] proposed a multiscale spatial-spectral transformer network (MSST-Net), which integrates the spatial-spectral self-attention mechanism into two multiscale branches. Li et al. [24] introduced a pyramid shuffle-and-reshuffle transformer (PSRT) to reduce the quadratic complexity of transformer by facilitating efficient information interaction among global patches. Additionally, the integration of a 3D-CNN along with the transformer layer enhances the model's capability to capture the spatial-spectral correlation in HSIs. Introducing the swin transformer [25], a hierarchical transformer with representations computed using shifted windows has demonstrated state-of-the-art performance across a wide range of reconstruction tasks. Building on this progress, Li et al. [26] designed a cross-spatial scale nonlocal attention network across spectral scales and shift windows based on transformer to effectively fuse HSIs and MSIs. In conclusion, while CNNs excel in local feature extraction and nonlinear mapping, transformer-based networks offer unique advantages, such as global context understanding. This allows the network to learn long-range dependencies in both spatial and spectral dimensions of the hyperspectral data, which is crucial for HSI super-resolution.

3 Methodology

In this section, we will introduce the proposed method SMIF-NET for fusing HSIs and MSIs, providing both an overview of the framework and a detailed description of the structure.

3.1 Network Framework

The primary objective of the proposed method is to tackle the two challenges mentioned earlier, e.g., the inaccurate modeling of cross-modality correlations and the neglect of inter-modality discrepancies. To fulfill this objective, we design an innovative spectral modality-aware interactive fusion network, termed as SMIF-NET, which is depicted in Figure. 1. The SMIF-Net primarily consists of two stages. In the spectral feature extraction stage, we employ the spectral modality-aware transformer to compute cross-spectral channel attention. This

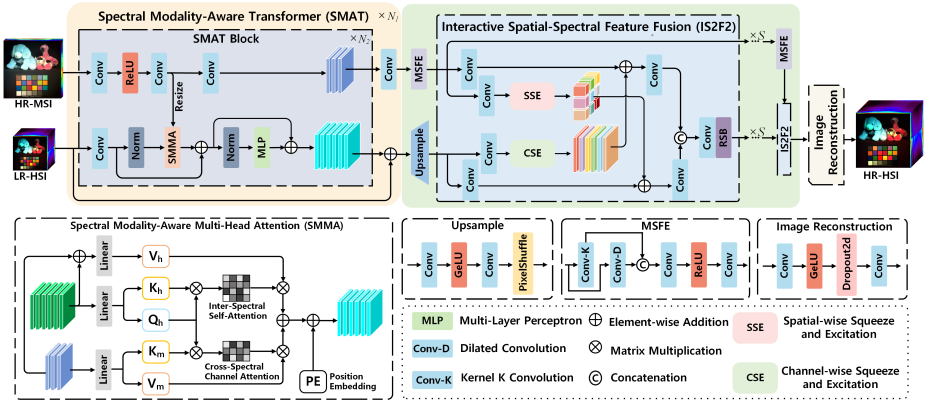

Fig. 1: The overall framework of the SMIF-NET.

method integrates spectral information from HSI and captures spectral correlations between the HSI and MSI, thereby enhancing the reconstruction of advanced spectral features. In the spatial feature insertion stage, we propose an interactive feature fusion approach, employing squeeze and excitation modules on both the spatial and spectral dimensions to fine-tune the features from distinct modalities, achieving a more harmonious integration of cross-modal information.

Initially, the paired LR-HSI and HR-MSI are fed into N_1 SMAT layers to extract spectral features. Each layer is comprised of N_2 SMAT blocks, where a dual-branch structure is applied to extract bimodal features independently and accurately model correlations within the spectral dimension. While introducing the intermediate connection through the SMMA, the feature map extracted from the HR-MSI is resized to match the spatial dimension of the extracted features from the LR-HSI. Following this, the extracted multiscale spatial features by MSFE are fed into IS2F2 to obtain the advanced spectral features. We employ cross-directional squeeze and excitation to achieve smooth bimodal information fusion. Finally, an image reconstruction module is utilized to refine the fused features, ultimately generating the estimated HR-HSI.

3.2 Spectral Modality-Aware Transformer

HSIs typically record surface reflectance or radiance data over a continuous range of wavelengths in tens or hundreds of contiguous spectral bands. Hence, the adjacent bands exhibit a certain degree of similarity and high correlation, which should be maintained during the HR-HSI reconstruction process. Moreover, in the fusion-based HSI super-resolution task, the input LR-HSIs and HR-MSIs are precisely co-registered ensuring spatial and spectral correspondence and facilitating the network to better learn the relationship between low- and high-resolution images. They are expected to demonstrate specific spectral correlations within their respective bands. Therefore, the SMAT is proposed to better preserve band-wise similarity throughout the reconstruction process.

As depicted in Figure. 1, the primary constituents of the SMAT block include a spectral modality-aware multi-head attention (SMMA) module, a feed-forward network, and two layers of normalization. Residual connections are used to prevent information loss within the SMAT block.

To obtain a comprehensive spectral feature representation, the SMMA is conducted within the spectral domain. First, we denote the two input feature maps of SMMA, which are extracted from the HSI and MSI, as $\mathbf{H}_{in} \in \mathbb{R}^{h\times w\times C}$ and $\mathbf{M}_{in} \in \mathbb{R}^{h\times w\times c}$, where h, w represent the height and width, C and c denote the number of bands of $\mathbf{H}_{in}$ and $\mathbf{M}_{in}$, respectively. $\mathbf{H}_{in}$ and $\mathbf{M}_{in}$ are then reshaped into $\mathbf{H} \in \mathbb{R}^{hw\times C}$ and $\mathbf{M} \in \mathbb{R}^{hw\times c}$. $\mathbf{H}$ is linearly projected to obtain the query $\mathbf{Q}_h \in \mathbb{R}^{hw\times C}$ and key $\mathbf{K}_h \in \mathbb{R}^{hw\times C}$. $\mathbf{M}$ is linearly projected to derive the key $\mathbf{K}_m \in \mathbb{R}^{hw\times c}$ and value $\mathbf{V}_m \in \mathbb{R}^{hw\times c}$. Notably, the value $\mathbf{V}_h \in \mathbb{R}^{hw\times C}$ is obtained by linearly projecting the summation of $\mathbf{H}$ and $\mathbf{M}$.

$$\mathbf{Q}_h = \mathbf{H}\mathbf{W}_h^{\mathrm{Q}}, \quad \mathbf{K}_h = \mathbf{H}\mathbf{W}_h^{\mathrm{K}}, \quad \mathbf{V}_h = (\mathbf{H}{+}\mathbf{M})\ \mathbf{W}_h^{\mathrm{V}}, \tag{1}$$

$$\mathbf{K}_m = \mathbf{M}\mathbf{W}_m^{\mathrm{K}}, \quad \mathbf{V}_m = \mathbf{M}\mathbf{W}_m^{\mathrm{V}}, \tag{2}$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{C\times C}$ and $\mathbf{W}_m^K, \mathbf{W}_m^V \in \mathbb{R}^{c\times c}$ are projection matrices with learnable weights and biases.

To capture diverse patterns within the data, we adopt the multi-head mechanism. $\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h, \mathbf{K}_m$ and $\mathbf{V}_m$ are partitioned into N heads along the spatial dimension, with the dimension of each head as $\frac{hw}{N}$. Thus, the dual-attention based on both the inter-spectral self-attention (ISSA) and the cross-spectral channel attention (CSCA) for the ith head (denoted as head_i) is computed as follows:

$$\mathrm{ISSA}_i = \mathbf{V}_h^i \left(\mathrm{softmax}\left(\frac{{\mathbf{K}_h^i}^T \mathbf{Q}_h^i}{\sqrt{D_{spe1}}}\right)\right), \tag{3}$$

$$\mathrm{CSCA}_i = \mathbf{V}_m^i \left(\mathrm{softmax}\left(\frac{{\mathbf{K}_m^i}^T \mathbf{Q}_h^i}{\sqrt{D_{spe2}}}\right)\right), \tag{4}$$

$$\mathrm{head}_i = \mathrm{ISSA}_i + \mathrm{CSCA}_i, \tag{5}$$

$$\mathrm{SMMA}\ (\mathbf{H}_{in},\ \mathbf{M}_{in}) = \mathrm{Concat}_{i=1}^{N}\ (\mathrm{head}_i)\ \mathbf{W}, \tag{6}$$

where ISSA_i and CSCA_i are the ISSA and CSCA computation of head_i. $\mathbf{Q}_h^i$, $\mathbf{K}_h^i$, $\mathbf{V}_h^i$, $\mathbf{K}_m^i$ and $\mathbf{V}_m^i$ denote the projection matrices for head_i. Additionally, two scaling factors D_{spe1} and D_{spe2} are employed to ensure numerical stability. $\mathbf{W}$ denotes a projection matrix with learnable parameters. The dual-attention mechanism SMMA $(\cdot)$ is formed by concatenating N heads.

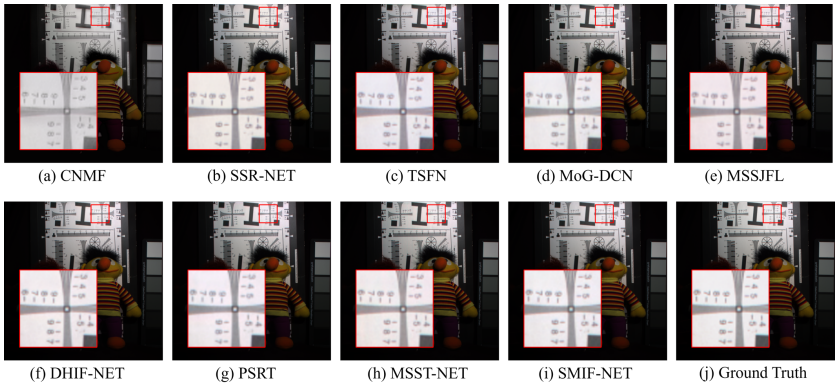

Fig. 2: Qualitative comparison of visual quality with a downsampling ratio of 8 on the CAVE dataset.

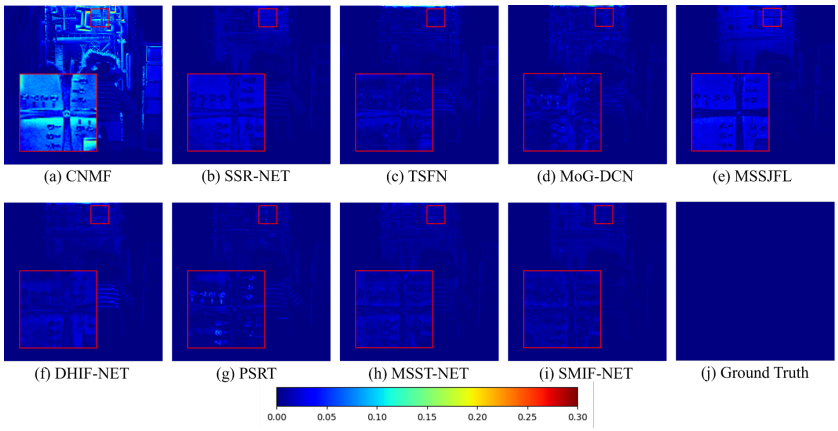

Fig. 3: Qualitative comparison of error maps with a downsampling ratio of 8 on the CAVE dataset.

3.3 Interactive Spatial-Spectral Feature Fusion

After acquiring high-level spectral features, our approach leverages the MSFE to extract advanced spatial features and then feeds the obtained features into the IS2F2 module for seamless feature integration. MSFE primarily includes multiple kernel convolutions, dilated convolutions, along with two layers of 3×3 convolutions, and a ReLU activation function. IS2F2 primarily comprises two sub-modules: the channel-wise squeeze and excitation (CSE) and the spatial-wise squeeze and excitation (SSE).

IS2F2 aims to address the challenge posed by modality differences in fusion tasks. Before feeding the two types of feature maps separately into CSE and SSE, they undergo individual convolution operations. Subsequently, spectral features

are directed into the CSE, while spatial features are routed into the SSE. The CSE module serves to model inter-channel dependencies within spectral features, utilizing these relationships to recalibrate the feature response intensity across the channels of spatial features. Conversely, the SSE module is designed to construct representative texture information based on spatial features, employing this information to the spatial dimension of spectral features. The formulations for these modules are as follows:

$$\text{Excitation}_{spe} = \text{Sigmoid}\ ((\text{ReLU}\ (\text{GAP}\ (X)\ \mathbf{W}^1_{spe}))\ \mathbf{W}^2_{spe}), \tag{7}$$

$$\text{Excitation}_{spa} = \text{Sigmoid}\ (\mathbf{W}^1_{spa}\ (\text{CAP}\ (Y))), \tag{8}$$

$$\text{CSE}\ (X) = \mathbf{W}_{cse}X\ \odot\ \text{Excitation}_{spe}, \tag{9}$$

$$\text{SSE}\ (Y) = \mathbf{W}_{sse}Y\ \odot\ \ \text{Excitation}_{spa}, \tag{10}$$

where X and Y represent high-level spectral and spatial features, respectively. The function GAP $(\cdot)$ denotes global average pooling, and $\mathbf{W}^1_{spe}$, $\mathbf{W}^2_{spe}$, and $\mathbf{W}^1_{spa}$ correspond to different projection matrices. CAP $(\cdot)$ stands for channel-wise adaptive pooling and the symbol $\odot$ signifies element-wise multiplication.

In the end, we employ cross-modality addition (CMA) to achieve the intended outcome:

$$\text{CMA}_1 = \mathbf{W}_{m1}(\text{CSE}\ (X) + \mathbf{W}_y Y), \tag{11}$$

$$\text{CMA}_2 = \mathbf{W}_{m2}(\text{SSE}\ (Y) + \mathbf{W}_x X), \tag{12}$$

$$\text{IS2F2}(X,\ Y) = \text{RSB}\ (\text{Conv}_{3\times3}\ (\text{Concat}\ (\text{CMA}_1,\ \text{CMA}_2))), \tag{13}$$

where CMA_1 and CMA_2 represent two forms of cross-modality addition. The projection matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ pertain to high-level spectral and spatial features, respectively. $\mathbf{W}_{m1}$ and $\mathbf{W}_{m2}$ serve as projection matrices for refined features. $\text{Conv}_{3\times3}$ denotes a convolutional operation with a 3×3 kernel. By concatenating CMA_1 and CMA_1 to the residual block (RSB), IS2F2 can effectively integrate both spatial and spectral information into the super-resolution process, resulting in improved performance and quality of the output HR-HSI.

3.4 Loss Function

To enhance the robustness of the model, we utilize mean absolute error as the loss function $\mathcal{L}$ to optimize the parameters of the network:

$$\mathcal{L} = \frac{1}{m}\sum_{j=1}^{m} \|\mathcal{F}(\mathbf{H}_j, \mathbf{M}_j) - \overline{\mathbf{H}}_j\|_1, \tag{14}$$

where $\mathbf{H}_j$ and $\mathbf{M}_j$ denote the jth pair of HR-HSI and LR-MSI input data, respectively. $\overline{\mathbf{H}_j}$ corresponds to the corresponding reference image, $\mathcal{F}$ represents the proposed SMIF-NET. The notation $\|\cdot\|_1$ represents the ℓ_1 norm. m represents the total number of samples within the training dataset.

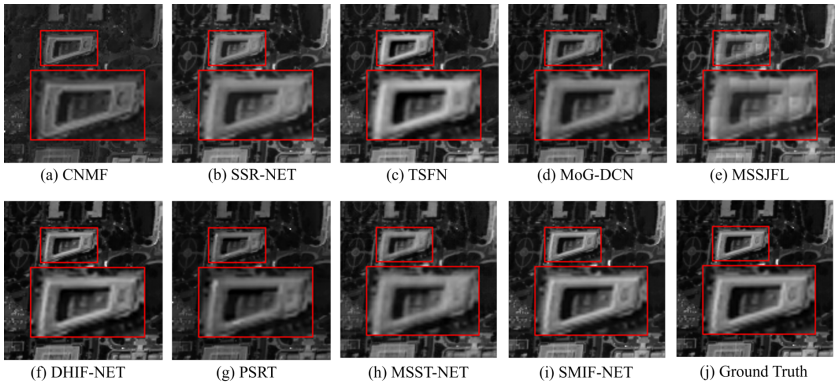

Fig. 4: Qualitative comparison of visual quality with a downsampling ratio of 8 on the WDCM dataset.

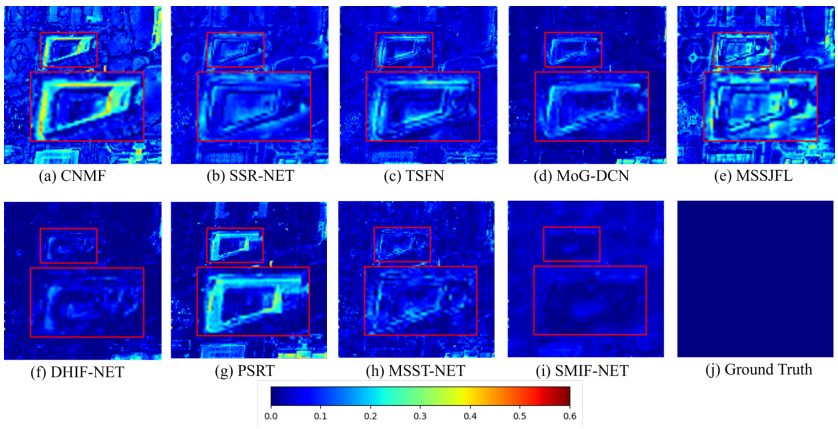

Fig. 5: Qualitative comparison of error maps with a downsampling ratio of 8 on the WDCM dataset.

4 Experiments

4.1 Experimental Settings

Datasets. We comprehensively assessed the performance of our model across three distinct datasets, including the Columbia imaging and vision laboratory (CAVE) dataset [27], Harvard dataset [28], and Washington DC Mall (WDCM) dataset [29].

The CAVE dataset includes 32 indoor HSIs, each featuring a high resolution of 512 × 512 pixels and 31 spectral bands. We selectively utilized the first 22 HSIs for model training, allocating the next 5 HSIs for validation and reserving the remaining for testing. The Harvard dataset contains 50 HSIs, with each image having dimensions of 1040 × 1392 pixels and 31 spectral bands. We employed

the first 34 HSIs for training, set aside 8 HSIs for validation, and allocated the remaining 8 HSIs for testing. The HSI image in the WDCM dataset comprises 191 spectral bands and 1280 $\times$ 307 pixels with a 2.5-meter spatial resolution. We utilized two 128 $\times$ 128 sub-images from the lower-left corner for validation and testing, with the remainder dedicated to effective model training.

Evaluation Metrics. In this paper, we utilized seven extensively acknowledged metrics to thoroughly evaluate the quality of the fused HR-HSI. These metrics encompass PSNR [30], SSIM [31], SAM [32], RASE [33], RMSE, as well as ERGAS [34].

Benchmarks. We systematically compared our model with a range of classical and deep learning-based SOTA methods, including CNMF [35], SSR-NET [18], TSFN [20], MoG-DCN [22], MSSJFL [19], DHIF-NET [14], PSRT [24], MSST-NET [23]. Configuration parameters for the compared methods were determined based on either the original authors' implementations or recommendations provided in the respective reference articles.

Implementation Details. Before implementing the proposed method on the CAVE, Harvard, and WDCM datasets, we aimed to simulate LR-HSIs and HR-MSIs from the HR-HSIs following Wald's protocol [36]. First, Gaussian filtering was applied to the HR-HSIs in all datasets, creating blurred images. Second, we simulated different spatial resolutions by downsampling the blurred HSIs with reduction ratios of 4 and 8, which led to the generation of LR-HSIs. For the CAVE and Harvard datasets, HR-MSIs with three bands were generated using the spectral response matrix of the Nikon D700 camera. In the WDCM dataset, the HR-MSIs with ten bands were generated using the spectral response matrix of the Sentinel-2 A instrument.

In the training stage, we utilized the adaptive moment estimation optimizer with a learning rate set at 4.0e-4 and a batch size of 64 for a total of 2000 epochs. For each dataset, we partitioned the HSIs and MSIs into smaller patches to minimize network memory consumption. The compared method CNMF was evaluated using MATLAB (R2013a) on a Windows Server 2012 platform equipped with two Intel Xeon E5-2650 processors and 128 GB of RAM. The other deep learning-based methods were assessed using PyTorch 2.0.0 in a Python 3.9 environment, leveraging an NVIDIA A40 GPU.

4.2 Experimental Results

Quantitative Evaluations. Tables. 1, 2 and 3 show the evaluation metrics and floating point operations (FLOPs) of SMIF-NET and other comparison methods. The best result for each metric is highlighted in bold. The results demonstrate that, when compared to deep learning methods, the traditional approach CNMF

Table 1: Experimental results involving various methods on the CAVE dataset under downsampling ratios of 4 and 8.

Ratio	Method	CAVE						
		PSNR ↑	SSIM ↑	SAM ↓	RASE↓	RMSE ↓	ERGAS ↓	FLOPs (10^{12})
4	CNMF [Yokoya et al., 2011]	36.62	0.9850	6.313	11.05	0.0139	5.8879	—
	SSR-NET [Zhang et al., 2020]	44.82	0.9976	1.607	5.020	0.0053	1.2550	0.006
	TSFN [Wang et al., 2021]	44.37	0.9967	1.699	5.309	0.0058	1.3274	1.964
	MoG-DCN [Dong et al., 2021]	45.27	0.9974	1.582	5.118	0.0054	1.2588	0.516
	MSSJFL [Min et al., 2021]	45.99	0.9975	1.415	4.520	0.0046	1.1300	0.304
	DHIF-Net [Huang et al., 2022]	48.92	0.9988	1.039	3.287	0.0034	0.8219	3.507
	PSRT [Deng et al., 2023]	48.49	0.9987	1.070	3.423	0.0035	0.8558	0.067
	MSST-NET [Jia et al., 2023]	48.66	0.9985	1.049	3.311	0.0034	0.8279	2.972
	SMIF-NET (Ours)	**50.63**	**0.9989**	**0.836**	**2.635**	**0.0028**	**0.6612**	1.200
8	CNMF [Yokoya et al., 2011]	35.97	0.9842	6.786	11.69	0.0149	6.3013	—
	SSR-NET [Zhang et al., 2020]	43.62	0.9969	1.866	5.894	0.0062	1.4736	0.006
	TSFN [Wang et al., 2021]	43.67	0.9961	1.815	5.871	0.0061	1.4678	1.964
	MoG-DCN [Dong et al., 2021]	44.28	0.9968	1.763	5.035	0.0058	1.4012	0.516
	MSSJFL [Min et al., 2021]	44.45	0.9964	1.719	5.445	0.0055	1.3612	0.304
	DHIF-Net [Huang et al., 2022]	47.65	0.9983	1.229	3.927	0.0041	0.9819	3.507
	PSRT [Deng et al., 2023]	46.79	0.9981	1.338	4.272	0.0043	1.0682	0.067
	MSST-NET [Jia et al., 2023]	47.08	0.9980	1.277	4.070	0.0042	1.0176	2.972
	SMIF-NET (Ours)	**49.01**	**0.9985**	**1.032**	**3.273**	**0.0034**	**0.8180**	1.200

Table 2: Experimental results involving various methods on the Harvard dataset under downsampling ratios of 4 and 8.

Ratio	Method	Harvard						
		PSNR ↑	SSIM ↑	SAM ↓	RASE↓	RMSE ↓	ERGAS ↓	FLOPs (10^{12})
4	CNMF [Yokoya et al., 2011]	45.48	0.9967	2.094	5.471	0.0044	1.5739	—
	SSR-NET [Zhang et al., 2020]	47.74	0.9979	1.653	3.916	0.0032	0.9790	0.027
	TSFN [Wang et al., 2021]	47.64	0.9979	1.666	3.951	0.0032	0.9878	2.064
	MoG-DCN [Dong et al., 2021]	48.30	0.9980	1.527	3.619	0.0030	0.9048	17.58
	MSSJFL [Min et al., 2021]	47.12	0.9976	1.769	4.202	0.0034	1.0505	1.218
	DHIF-Net [Huang et al., 2022]	47.85	0.9979	1.623	3.843	0.0034	0.9609	14.02
	PSRT [Deng et al., 2023]	47.85	0.9979	1.623	3.842	0.0031	0.9606	0.268
	MSST-NET [Jia et al., 2023]	48.32	0.9981	1.527	3.616	0.0030	0.9042	11.89
	SMIF-NET (Ours)	**48.55**	**0.9981**	**1.480**	**3.504**	**0.0029**	**0.8757**	4.801
8	CNMF [Yokoya et al., 2011]	44.55	0.9958	2.588	6.258	0.0051	1.8546	—
	SSR-NET [Zhang et al., 2020]	47.49	0.9978	1.672	4.093	0.0034	0.9901	0.027
	TSFN [Wang et al., 2021]	46.48	0.9974	1.874	4.650	0.0038	1.1249	2.064
	MoG-DCN [Dong et al., 2021]	47.79	0.9978	1.598	3.912	0.0032	0.9472	17.58
	MSSJFL [Min et al., 2021]	46.69	0.9974	1.856	4.502	0.0037	1.1057	1.218
	DHIF-Net [Huang et al., 2022]	47.71	0.9978	1.628	3.934	0.0033	0.9646	14.02
	PSRT [Deng et al., 2023]	47.35	0.9977	1.700	4.123	0.0034	1.0077	0.268
	MSST-NET [Jia et al., 2023]	47.30	0.9979	1.612	3.883	0.0032	0.9572	11.89
	SMIF-NET (Ours)	**47.92**	**0.9979**	**1.572**	**3.738**	**0.0032**	**0.9296**	4.801

falls short in achieving the desired results across three datasets. The deep learning methods SSR-NET and TSFN outperform CNMF, exhibiting an approximate increase of 8 dB, 2 dB, and 1 dB in PSNR on CAVE, Harvard, and WDCM datasets, respectively. PSRT and MSST-NET, benefiting from the Transformer's exceptional ability to capture long-range dependencies, surpass MoG-DCN and MSSJFL methods on the CAVE dataset. Notably, DHIF-Net demonstrated the second-best performance across the CAVE and WDCM datasets, following our proposed SMIF-NET. On the Harvard dataset, MoG-DCN secured the second position in performance. Overall, our proposed SMIF-NET demonstrates the best performance across all metrics, excelling in SAM and RASE, which measure spectral similarity and relative average spectral error, respectively. This in-

Table 3: Experimental results involving various methods on the WDCM dataset under downsampling ratios of 4 and 8.

Ratio	Method	WDCM						
		PSNR ↑	SSIM ↑	SAM ↓	RASE↓	RMSE ↓	ERGAS ↓	FLOPs (10^{12})
4	CNMF [Yokoya et al., 2011]	36.83	0.9902	4.843	9.784	0.0089	3.3375	—
	SSR-NET [Zhang et al., 2020]	37.47	0.9884	2.595	3.991	0.0132	1.4984	0.016
	TSFN [Wang et al., 2021]	37.83	0.9912	2.481	4.742	0.0127	1.4371	0.035
	MoG-DCN [Dong et al., 2021]	39.47	0.9922	2.054	3.844	0.0105	1.1902	0.628
	MSSJFL [Min et al., 2021]	40.14	0.9938	1.901	4.275	0.0097	1.1015	0.036
	DHIF-Net [Huang et al., 2022]	43.61	0.9973	1.274	2.755	0.0065	0.7390	3.992
	PSRT [Deng et al., 2023]	39.56	0.9934	2.038	4.386	0.0104	1.1774	0.005
	MSST-NET [Jia et al., 2023]	41.00	0.9943	1.725	3.803	0.0088	0.9971	1.115
	SMIF-NET (Ours)	**46.79**	**0.9985**	**0.928**	**2.143**	**0.0045**	**0.5358**	0.079
8	CNMF [Yokoya et al., 2011]	34.19	0.9834	6.651	10.67	0.0121	4.5247	—
	SSR-NET [Zhang et al., 2020]	36.73	0.9878	2.827	4.466	0.0144	1.6317	0.016
	TSFN [Wang et al., 2021]	36.91	0.9890	2.766	5.293	0.0141	1.5977	0.035
	MoG-DCN [Dong et al., 2021]	36.99	0.9880	2.731	4.082	0.0140	1.5821	0.628
	MSSJFL [Min et al., 2021]	37.72	0.9911	2.519	5.557	0.0128	1.4547	0.036
	DHIF-Net [Huang et al., 2022]	42.58	0.9970	1.438	3.056	0.0073	0.8313	3.992
	PSRT [Deng et al., 2023]	37.87	0.9908	2.474	4.947	0.0126	1.4310	0.005
	MSST-NET [Jia et al., 2023]	39.69	0.9930	2.007	4.292	0.0102	1.1601	1.115
	SMIF-NET (Ours)	**45.23**	**0.9980**	**1.104**	**2.506**	**0.0054**	**0.6407**	0.079

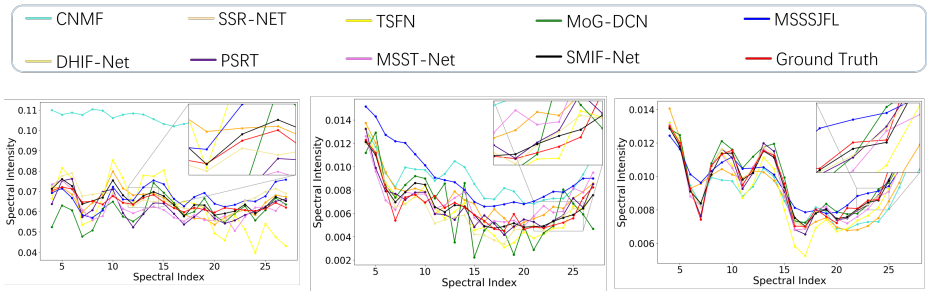

Fig. 6: Spectral curves based on three random selected locations

dicates the superior capability of our method in preserving spectral information compared to alternative approaches. In addition, our model has fewer FLOPs than MoG-DCN, DHIF-NET and MSST-NET, yet it outperforms all the comparison methods in terms of the 6 metrics, indicating that our model achieves the best balance between computational efficiency and performance.

Qualitative Evaluations. The qualitative results with a downsampling ratio of 8 on the CAVE dataset are given to illustrate the perceptual quality of each method. With the objective of representing the captured scenes in a more intuitive manner, we present pseudo-color visualizations and error maps of the output results, as shown in Figure. 2 and Figure. 3. Our approach stands out with the error map exhibiting the lowest brightness in the boxed region, indicating superior performance. We also provide grayscale images of the experimental results with a downsampling ratio of 8 on the WDCM dataset, along with the corresponding error maps in Figure. 4 and Figure. 5. The red-boxed regions in the images highlight buildings. It can be seen from the images that only the

results obtained by MSST-NET, DHIF-NET, and our proposed SMIF-NET are satisfactory, while other methods introduce severe distortions. Among the three methods, our approach stands out with the lowest error value in the highlighted region, indicating superior performance. Furthermore, we illustrate three randomly selected spectral curves in Figure. 6. spectral curves provide a complete representation of each pixel's spectral data within a hyperspectral image, capturing the detailed spectral characteristics. The accuracy of these curves is essential for achieving high-quality hyperspectral image reconstruction. For a more intuitive comparison, we have added a zoomed-in view of a particular region in the upper right corner of the figure. This magnified section makes it clear that our method most accurately mirrors the ground truth, highlighting its effectiveness in preserving spectral integrity.

Table 4: Ablation study for the number of SMAT blocks N_2 and fusion stages S

Metrics	N_2			S		
	4	5	6	3	4	5
PSNR ↑	48.99	**49.01**	48.93	48.93	**49.01**	48.95
SAM ↓	1.033	**1.032**	1.039	1.039	**1.032**	1.036
ERGAS ↓	0.8184	**0.8180**	0.8236	0.8236	**0.8180**	0.8200

Table 5: Ablation study for the CSCA, MSFE and IS2F2

CSCA	MSFE	IS2F2	PSNR ↑	SSIM ↑	SAM ↓	RMSE ↓	ERGAS ↓
✗	✗	✗	46.82	0.9978	1.344	0.0044	1.068
✗	✔	✔	48.52	0.9983	1.095	0.0036	0.8659
✔	✗	✔	48.80	0.9985	1.058	0.0035	0.8387
✔	✔	✗	48.20	0.9982	1.136	0.0038	0.8988
✔	✔	✔	**49.01**	**0.9985**	**1.032**	**0.0034**	**0.8180**

4.3 Ablation Studies

To assess the significance of the key modules integrated into our proposed network, a series of ablation studies was conducted on the CAVE dataset. First, we investigate how the performance of the proposed SMIF-NET is impacted by varying the number of SMAT blocks N_2 and fusion stages S. Each stage consists of an IS2F2 and an MSFE. Second, we systematically removed the CSCA, MSFE, and IS2F2 modules and utilized the complete version of the model for comparison.

Table. 4 presents the mean quantitative results obtained from variations of the SMIF-NET with different N_2 and S on the CAVE datasets. First, we keep $S = 4$ while adjusting N_2 to observe variations in performance. From columns 2, 3, and 4 of Table. 4, it is evident that the peak performance is achieved when N_2 is set to 5. Then, $N_2 = 5$ is set as constant, we observe the performance of the model when S is set as 3, 4, 5. It should be noted that the configuration with $S = 5$ involved the incorporation of an additional sub-module (consisting of two convolutions with kernel sizes of 7 and dilation factors of 4, respectively)

into MSFE. The results indicate that the performance is optimal when S is set to 4.

Table. 5 shows the ablation study of three key modules in SMIF-NET. Obviously, the model with all modules exhibits the best performance in all metrics. In contrast, the version without all proposed modules exhibits a significant decline in performance. The evaluation results of CSCA demonstrate significant improvements across all metrics with the incorporation of CSCA, highlighting the model's retention of band correlation and better maintenance of spectral consistency. To assess the effectiveness of the MSFE, we compare the performance of the model with and without it. It is evident that the enhanced capability in spatial information extraction provided by MSFE contributes to improved model performance. To assess the impact of the IS2F2, we conducted an ablation study by simplifying the fusion process. Instead of employing the CSE and SSE modules, we chose a straightforward concatenation of the two types of feature maps. As indicated by the metrics in Table. 5, the exclusion of the IS2F2 led to a decline in performance.

5 Conclusions

In this study, we introduce a novel spectral modality-aware interactive fusion network aimed at achieving HSI super-resolution. By leveraging the multi-layered processing of SMAT blocks and MSFE modules, advanced spectral and spatial information is acquired. The IS2F2 module is employed to facilitate the collaborative procession, which effectively mitigates modal disparities among diverse features. Therefore, SMIF-NET exhibits superior fusion performance compared with eight SOTA models, as demonstrated through the experiments conducted on the simulated natural and satellite datasets. Despite SMIF-NET having presented a remarkable performance in HSI super-resolution, its training process was implemented on the simulated degraded hyperspectral and multispectral data. Since the proposed method has not been tested on the real LR-HSI and HR-MSI data, it imposes certain limitations on its overall performance. Hence, conducting further experiments with real data sets and accordingly improving the method are the potential directions for future research.

6 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 42271336, 62271327, and 41971300; in part by Natural Science Foundation of Guangdong Province under Grant 2024A1515011079 and 2022A1515011290; in part by the Shenzhen Science and Technology Program under Grant RCJC20221008092731042 and JCYJ20220818100206015; and in part by Guangdong Province Key Laboratory of Popular High Performance Computers 2017B030314073 and Guangdong Province Engineering Center of China-made High Performance Data Computing System.

References

1. Fulin Luo, Liangpei Zhang, Bo Du, and Lefei Zhang. Dimensionality reduction with enhanced hybrid-graph discriminant learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8):5336–5353, 2020.
2. Lichao Mou, Pedram Ghamisi, and Xiao Xiang Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655, 2017.
3. Qi Wang, Xiang He, and Xuelong Li. Locality and structure regularized low rank representation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):911–923, 2018.
4. Lei Zhang, Yanning Zhang, Hangqi Yan, Yifan Gao, and Wei Wei. Salient object detection in hyperspectral imagery using multi-scale spectral-spatial gradient. *Neurocomputing*, 291:215–225, 2018.
5. Jing Yao, Deyu Meng, Qian Zhao, Wenfei Cao, and Zongben Xu. Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing. *IEEE Transactions on Image Processing*, 28(6):2991–3006, 2019.
6. Ke Zheng, Lianru Gao, Wenzhi Liao, Danfeng Hong, Bing Zhang, Ximin Cui, and Jocelyn Chanussot. Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 59(3):2487–2502, 2020.
7. Wei Li, Xicun Zhu, Xinyang Yu, Meixuan Li, Xiaoying Tang, Jie Zhang, Yuliang Xue, Canting Zhang, and Yuanmao Jiang. Inversion of nitrogen concentration in apple canopy based on uav hyperspectral images. *Sensors*, 22(9):3503, 2022.
8. Kui Jiang, Zhongyuan Wang, Peng Yi, Guangcheng Wang, Tao Lu, and Junjun Jiang. Edge-enhanced gan for remote sensing image superresolution. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):5799–5812, 2019.
9. Yuxuan Zheng, Jiaojiao Li, Yunsong Li, Jie Guo, Xianyun Wu, Yanzi Shi, and Jocelyn Chanussot. Edge-conditioned feature transform network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.
10. Renwei Dian, Shutao Li, Anjing Guo, and Leyuan Fang. Deep hyperspectral image sharpening. *IEEE transactions on neural networks and learning systems*, 29(11):5345–5355, 2018.
11. Chen Wang, Yun Liu, Xiao Bai, Wenzhong Tang, Peng Lei, and Jun Zhou. Deep residual convolutional neural network for hyperspectral image super-resolution. In *Image and Graphics: 9th International Conference, ICIG 2017, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part III 9*, pages 370–380. Springer, 2017.
12. Frosti Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 14(5):639–643, 2017.
13. Renwei Dian, Shutao Li, and Xudong Kang. Regularizing hyperspectral and multispectral image fusion by cnn denoiser. *IEEE transactions on neural networks and learning systems*, 32(3):1124–1135, 2020.
14. Tao Huang, Weisheng Dong, Jinjian Wu, Leida Li, Xin Li, and Guangming Shi. Deep hyperspectral image fusion network with iterative spatio-spectral regularization. *IEEE Transactions on Computational Imaging*, 8:201–214, 2022.
15. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

16. Qing Ma, Junjun Jiang, Xianming Liu, and Jiayi Ma. Learning a 3d-cnn and transformer prior for hyperspectral image super-resolution. *Information Fusion*, 100:101907, 2023.
17. Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17502–17511, 2022.
18. Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5953–5965, 2020.
19. Zhichao Min, Yifan Wang, and Sen Jia. Multiscale spatial-spectral joint feature learning for multispectral and hyperspectral image fusion. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 1265–1270. IEEE, 2021.
20. Xiuheng Wang, Jie Chen, Qi Wei, and Cédric Richard. Hyperspectral image super-resolution via deep prior regularization with parameter estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1708–1723, 2021.
21. Xueyang Fu, Wu Wang, Yue Huang, Xinghao Ding, and John Paisley. Deep multiscale detail networks for multiband spectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2090–2104, 2021.
22. Weisheng Dong, Chen Zhou, Fangfang Wu, Jinjian Wu, Guangming Shi, and Xin Li. Model-guided deep hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 30:5754–5768, 2021.
23. Sen Jia, Zhichao Min, and Xiyou Fu. Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion. *Information Fusion*, 96:117–129, 2023.
24. Shang-Qi Deng, Liang-Jian Deng, Xiao Wu, Ran Ran, Danfeng Hong, and Gemine Vivone. Psrt: Pyramid shuffle-and-reshuffle transformer for multispectral and hyperspectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.
25. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
26. Shuangliang Li, Yugang Tian, Cheng Wang, Hongxian Wu, and Shaolan Zheng. Hyperspectral image super-resolution network based on cross-scale non-local attention. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
27. Fumihito Yasuma, Tomoo Mitsunaga, Daisuke Iso, and Shree K Nayar. Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum. *IEEE transactions on image processing*, 19(9):2241–2253, 2010.
28. Ayan Chakrabarti and Todd Zickler. Statistics of real-world hyperspectral images. In *CVPR 2011*, pages 193–200. IEEE, 2011.
29. Naoto Yokoya, Claas Grohnfeldt, and Jocelyn Chanussot. Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, 5(2):29–56, 2017.
30. Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

31. Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
32. Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, 1992.
33. Myungjin Choi. A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter. *IEEE Transactions on Geoscience and Remote sensing*, 44(6):1672–1682, 2006.
34. Lucien Wald. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES, 2002.
35. Naoto Yokoya, Takehisa Yairi, and Akira Iwasaki. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):528–537, 2011.
36. Lucien Wald, Thierry Ranchin, and Marc Mangolini. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing*, 63(6):691–699, 1997.