The Photogrammetric Record

ORIGINAL ARTICLE

A Multiple Depth-Level Spatial–Spectral Aggregation Network for Enhancing Multispectral and Hyperspectral Image Fusion

Bo Zhou[1] | Ziyuan Feng[1] | Miao Ren[1] | Xiaobo Zhi[1] | Xianfeng Zhang[1,2]

[1]Institute of Remote Sensing and Geographic Information System, Peking University, Beijing, China | [2]Engineering Research Centre of Earth Observation and Navigation (CEON), Ministry of Education, Beijing, China

Correspondence: Xianfeng Zhang (xfzhang@pku.edu.cn)

Received: 14 February 2025 | **Revised:** 29 July 2025 | **Accepted:** 14 August 2025

Funding: This study was financially supported by the National Natural Science Foundation of China (grant number 42171327) and the International Research Centre of Big Data for Sustainable Development Goals, China (grant number CBAS2022GSP06).

Keywords: edge spectra | feature aggregation | hyperspectral | image fusion | receptive field | spatial–spectral features

ABSTRACT

Residual convolutional neural networks (ResCNNs) have been widely utilized in hyperspectral image fusion due to their efficiency in training. However, as network depth increases, the associated reduction in receptive field gain limits the performance of ResCNNs in hyperspectral image fusion. To fully leverage global image information while maintaining the efficiency of ResCNNs, we propose a multiple depth-level spatial-spectral aggregation network (MDSSAN) based on residual learning. MDSSAN extracts spatial features from high-resolution multispectral images and spectral features from low-resolution hyperspectral images at multiple depth levels using a two-branch network structure. Subsequently, the extracted spatial and spectral features are aggregated at various depth levels through efficient self-attention blocks, thereby expanding the receptive field and enhancing feature fusion. Additionally, an edge loss function is introduced to improve fusion quality at high-frequency object edges. Compared to the existing convolutional neural fusion networks, the proposed model achieved improvements of 7.6% and 5.5%, 9.5% and 23.3%, 2.7% and 13.2%, and 1.2% and 5.5% in Peak Signal-to-Noise Ratio and Spectral Angle Mapper across Pavia Centre, Chikusei, MDAS, and Daxing datasets, respectively. Experimental results demonstrate that MDSSAN effectively and efficiently enhances spatial resolution while preserving the spectral information of low spatial resolution hyperspectral imagery. The code is available at https://github.com/zerobrave/MDSSAN_fusion.

1 | Introduction

Hyperspectral images (HSI) typically comprise hundreds or even thousands of spectral bands, providing abundant spectral information (Zhao and Du 2022). However, the inherent trade-off between spatial and spectral resolutions, stemming from sensor limitations and the high costs of acquiring large-scale hyperspectral data, poses significant challenges for remote sensing applications. High-resolution multispectral images (MSI), distinguished by their rich spatial details but limited spectral bands, have strong spatial detail perception capability. However, they do not possess the extensive spectral information available in hyperspectral images. With advancements in earth observation technology, the complementary fusion of images with different spatial and spectral resolutions has attracted considerable attention. The fusion of low-resolution hyperspectral images (LR-HSI) with high-resolution multispectral images (HR-MSI) enables the generation of high spatial resolution hyperspectral images (HR-HSI), thereby providing integrated data support for various remote

sensing tasks, such as spectral unmixing (Ehlers 1991), geotargeting (Churchill et al. 2004), and land classification (Hong et al. 2019), among others.

Current hyperspectral and multispectral image fusion methods can be broadly categorized into conventional and deep learning-based approaches. Numerous studies have focused on image pan-sharpening, which serves as a representative case for HR-MSI and LR-HSI fusion. Traditional remote sensing image fusion techniques are generally classified into three types: component substitute (CS), multi-scale resolution analysis (MRA), and model-based methods. Specifically, CS methods (Chavez and Kwarteng 1989; Carper et al. 1990; Shettigara 1992; Shah et al. 2008; Rahmani et al. 2010) replace certain components of the LR-HSI with the HR-MSI. However, this approach often leads to spectral distortion due to mismatched spectral ranges between HR-MSI and LR-HSI (Nezhad et al. 2016). MRA methods (Burt and Adelson 1983; Mallat 1989; Demirel and Anbarjafari 2011) employ spatial filters to perform multiscale decomposition of HR-MSI, extracting spatial details that are subsequently injected into LR-HSI. Nevertheless, MRA methods typically incur high computational complexity owing to the intricate design of spatial filters. Model-based methods operate under the assumption that both HR-MSI and LR-HSI can be derived from a hypothetical HR-HSI through spatial and spectral resampling. Based on this assumption, fusion models are constructed to quantitatively simulate the interrelationships among the three image types. These methods are further subdivided into matrix factorization (Nascimento and Dias 2005; Aharon et al. 2006; Dong et al. 2016) and tensor factorization (Dian et al. 2017; Li et al. 2018; Chang et al. 2020; Prévost et al. 2020; Xu et al. 2020; Borsoi et al. 2021; Liu et al. 2021) approaches. A significant drawback of model-based methods is their reliance on a priori information regarding the fusion process, which often necessitates the introduction of suitable regularization terms. This dependence limits their practical applicability and may lead to reduced fusion quality when prior assumptions do not hold in specific scenarios.

Compared to the traditional approaches, convolutional neural networks (CNNs) have achieved remarkable breakthroughs in various computer vision tasks, and numerous studies have successfully applied them to hyperspectral image fusion. Within the supervised framework, existing residual learning-based methods can be categorized into single-branch and two-branch network architectures. The single-branch architecture concatenates the upsampled images of LR-HSI and HR-MSI as network input, learning the residuals between this input and the HR-HSI. Masi et al. (2016) first proposed a three-layer convolutional network that accepts upsampled LR-HSI for image fusion. Zhang et al. (2021) introduced an interpretable spatial and spectral reconstruction network (SSRNet), which sequentially extracts spatial and spectral features from the concatenated images and employs residual connections for feature fusion. However, this network is shallow and lacks sufficient capacity to learn complex spectral and spatial information. Wei et al. (2017) proposed a multi-scale and deep convolutional neural network (MSDCNN), which performs both shallow and deep feature extraction and fusion on the concatenated images, utilizing large convolutional kernels to expand the receptive field. Although single-branch networks leverage the spatial and spectral correlations of input images, they often overlook the unique characteristics of HR-MSI and LR-HSI, resulting in suboptimal performance, particularly when there are significant modality differences between the two. In contrast, two-branch architectures address these issues by separately extracting features from HR-MSI and LR-HSI. For instance, Liu et al. (2020) proposed a two-stream remote sensing image fusion network and enhanced it by increasing depth through residual connections to achieve a larger receptive field. Xiao et al. (2022) designed a spatial detail extraction network and a spatial–spectral fusion network based on U-Net, utilizing a multiscale attention module and residual connections to integrate spatial features into HSI. Convolutional fusion networks based on residual learning typically exhibit few parameters, ease of training, and high efficiency, but their fusion performance is constrained by the local receptive field of the convolutional operations. Meanwhile, transformer-based fusion methods (Wang et al. 2022; Meng et al. 2022; Liu et al. 2023; Hu, Huang, Deng, Dou, et al. 2022; Jia et al. 2023) have been widely adopted for remote sensing image fusion due to their superior capability to capture long-range dependencies. To improve computational efficiency, Swin transformer-based fusion techniques (Deng et al. 2023; Feng et al. 2024) employ window attention within Swin blocks and extend receptive fields through shift operations, effectively approximating a global receptive field using stacked Swin transformer blocks. Nevertheless, the self-attention mechanism is inherently associated with high computational complexity, which significantly complicates the training process, especially for hyperspectral datasets that typically feature limited sample sizes and high dimensionality. These limitations can hinder the effectiveness of transformer-based approaches in such contexts.

The aforementioned neural network fusion methods primarily rely on residual connections to facilitate parameter learning and ensure rapid convergence. The residual structure enables information to propagate along shortcut paths, allowing for the network to increase depth while maintaining stable convergence. Some models, such as ResTFNet (Liu et al. 2020), seek to enhance fusion by increasing CNN depth through residual connections. However, it has been demonstrated that the receptive field gain of residual CNNs does not significantly improve with increased feature depth (Ding et al. 2022). This finding indicates that the features utilized for image fusion in these approaches remain localized, particularly in the relatively shallow residual CNNs currently in use. Typically, the performance of pixel-wise summation or concatenation during the fusion process is constrained by the size of the receptive field. To address this limitation, we enlarge the receptive field at various depth levels of the network by globally aggregating similar information prior to feature fusion. This strategy improves fusion quality while preserving the efficiency of the residual convolutional fusion network.

The primary contributions of this paper are as follows. First, we propose a multi-depth-level spatial–spectral feature aggregation network to address the limited gains in receptive field expansion achieved by simply increasing network depth in existing convolutional residual fusion networks. Our model significantly enhances the receptive field of aggregated features through a global spatial–spectral feature fusion module,

which combines similar texture features while preserving spectral information by introducing attentional interactions between HR-MSI and LR-HSI. Second, we design a residual feature extractor within a two-branch structure to separately extract spectral and spatial features. The extracted features are downsampled into smaller feature maps, thereby improving aggregation efficiency without information loss. Third, we develop a loss function to improve edge spectral perception, addressing the challenge that image reconstruction is typically more difficult at high-frequency boundaries than in spatially smooth regions. The function utilizes an edge weight mask to compute the spectral loss, allowing the model to progressively concentrate its attention on high-frequency edges throughout the training process. Additionally, the proposed model leverages LR-HSI to obtain key-value features for the self-attention mechanism, aggregates similar features of different pixel locations in HR-MSI, and transfers them into LR-HSI features within a large receptive field. Experiments conducted on the datasets with varying resolutions and certain alignment errors have demonstrated that the proposed method achieves superior fusion performance and robustness.

2 | Methods

The proposed MDSSAN mainly consists of five components: residual feature extraction module, spatial–spectral feature fusion module, feature refinement module, multiple depth-level feature fusion layer, and image reconstruction layer (Figure 1).

For the input LR-HSI and HR-MSI, several cascaded residual feature extraction modules (RFEMs) are utilized to extract spectral and spatial features at different depth levels. At each depth level, a spatial–spectral fusion module (SSFM) merges the spatial and spectral features. To preserve the original information of the input images during the fusion process, a refine module (RM) proposed in our previous work (Zhou et al. 2023) is applied after feature fusion, enabling the weighted reuse of spatial and

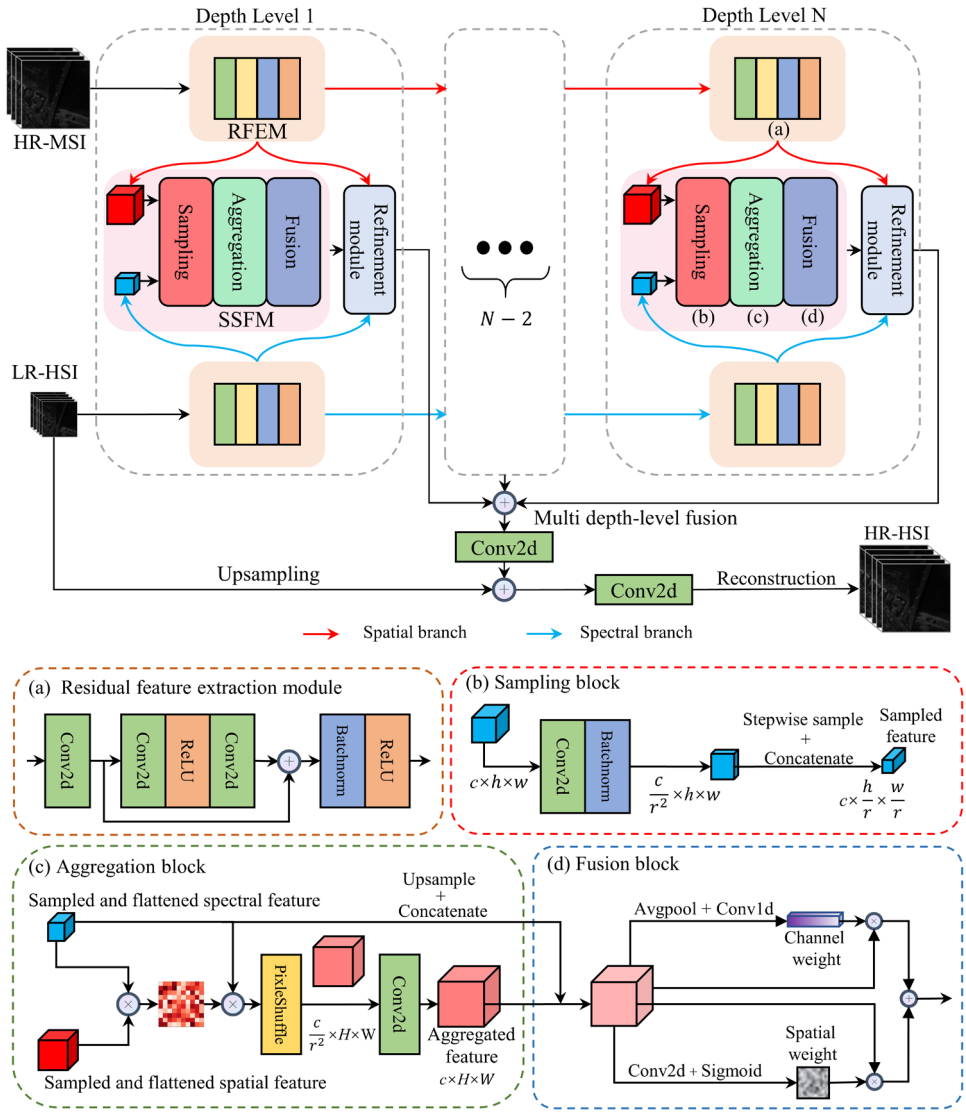

FIGURE 1 | The proposed MDSSAN network. (a) The residual feature extraction module to extract residual features in the dual spatial–spectral branches. The spatial–spectral fusion module including (b) the feature sampling, (c) aggregation, and (d) fusion blocks to perform feature fusion. The fused features are refined through a refinement module at each depth level, and the refined features from multiple depth levels are gathered to reconstruct the HR-HSI.

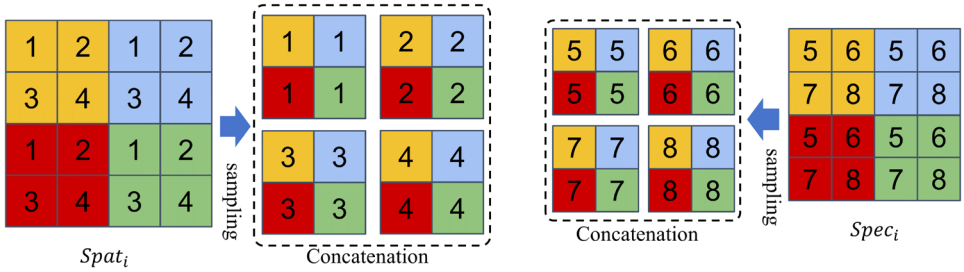

FIGURE 2 | Schematic illustration of stepwise sampling of (a) spatial and (b) spectral features at a ratio of 2 to generate four sub-feature maps.

spectral features to refine the fused features. Finally, a feature fusion layer consolidates the refined features from all depth levels and adds them to the upsampled LR-HSI, producing the final fused image.

2.1 | Residual Feature Extraction Module

This study utilizes residual feature extraction modules for feature extraction (Figure 1a). The RFEM consists sequentially of a 2-D convolutional layer, a residual block, a batch normalization layer, and an activation layer. Each residual block comprises a 2-D convolutional layer, an activation layer, and a second 2-D convolutional layer, with a skip connection between the input and output to facilitate the residual learning and accelerate model convergence.

The network's feature extraction component includes two branches: one for extracting spectral features from LR-HSI and the other for extracting spatial features from HR-MSI. Each branch contains three cascaded RFEMs. The convolutional layers of both spectral and spatial feature extraction branches use a kernel size of 3, a stride of 1, and a padding of 1 pixel. The spatial and spectral features obtained at each depth level by the RFEMs are subsequently input into the spatial–spectral fusion module.

2.2 | Spatial–Spectral Fusion Module

The residual structure ensures rapid convergence of convolutional neural networks when learning the differences between the upsampled images and HR-HSI. However, currently existing studies indicate that the effective receptive field of residual convolutional networks remains relatively unchanged as network depth increases, suggesting that simple addition or concatenation of spatial and spectral features primarily utilizes local range information. Our approach aims to fully exploit global spatial and spectral information by aggregating similar texture features from HR-MSI without introducing high computational overhead, whereas preserving spectral information. To this end, we propose a spatial–spectral fusion module that facilitates efficient self-attention between spectral and spatial features.

This module includes sampling, aggregation, and fusion blocks. Specifically, for the spatial feature $spat_i \in \boldsymbol{R}^{C\times H\times W}$ and spectral feature $spec_i \in \boldsymbol{R}^{C\times h\times w}$ obtained by the RFEMs at a given depth level in the dual-branch extraction, both are processed through distinct sampling blocks (Figure 1b). Here, C is the number of feature channels; H and W denote the height and width of the input HR-MSI; h and w denote the height and width of the input LR-HSI; and i indicates the output layer of the RFEMs. The spatial dimensions of the feature maps are reduced through stepwise sampling (Figure 2) to boost computational efficiency. The sampled feature maps are concatenated along the channel dimension to obtain $sampled\_spat_i \in \boldsymbol{R}^{C\times(H/r)\times(W/r)}$ and $sampled\_spec_i \in \boldsymbol{R}^{C\times(h/r)\times(w/r)}$, where r is the sampling scale factor. This process strategically omits some pixel-wise attention computations to improve efficiency while preserving channel-wise information within the feature maps.

The 2-D $sampled\_spat_i \in \boldsymbol{R}^{C\times S}$ and $sampled\_spec_i \in \boldsymbol{R}^{C\times s}$ (where $S = H\times W/r^2, s = h\times w/r^2$) are obtained by further flattening in the spatial dimensions. The spectral feature map incorporates both spectral information and certain spatial details, which are combined with the spatial feature map to compute the attention matrix. The feature aggregation process (Figure 1c) is calculated as follows:

$$attention_i = \mathrm{Softmax}\left(\mathrm{Matmul}\left(sampled\_spat_i^T, sampled\_spec_i\right)/\sqrt{d}\right) \tag{1}$$

$$spat\_aggre = \mathrm{Matmul}\left(sampled\_spec_i, attention_i^T\right) \tag{2}$$

where T denotes the transpose operation, $\mathrm{Matmul}(\bullet)$ represents matrix multiplication, and $\mathrm{Softmax}(\bullet)$ refers to normalization using the softmax function. Here, d is the channel number in the feature map, and $attention_i$ is the similarity matrix that quantifies the similarity between each pixel in $sampled\_spat_i$ and $sampled\_spec_i$. The aggregation feature $spat\_aggre \in \boldsymbol{R}^{C\times(H/r)\times(W/r)}$ is constructed such that each pixel contains global information from similar pixels, where $i \in \{1, 2, 3\}$.

The $sampled\_spec_i$ is upsampled to match the spatial size of $\_aggre$, and they are concatenated to obtain F_{cat_i} for spatial–spectral information fusion. The spatial–spectral fusion process (Figure 1d) employs spectral and spatial attention mechanisms to perform weighted fusion of spatial and spectral features on F_{cat_i}, which can be represented as:

$$\mathrm{CA}(\boldsymbol{x}) = \mathrm{Sigmoid}\left(w^7\mathrm{Avgpool}(\boldsymbol{x})\right) \tag{3}$$

$$\mathrm{SA}(\boldsymbol{x}) = W_2^3 \mathrm{Sigmoid}\left(W_1^3 \boldsymbol{x}\right) \tag{4}$$

$$F_{fused_i} = W^3\left(\mathrm{CA}\left(F_{cat_i}\right) \otimes F_{cat_i} + \mathrm{SA}\left(F_{cat_i}\right) \otimes F_{cat_i}\right) \tag{5}$$

where Avgpool(•) denotes the global average pooling operation and w^7 represents a 1-D convolutional layer with a kernel size of 7. Sigmoid(•) is the logistic activation function. Additionally, $\boldsymbol{x}$ represents the input feature, CA(•) is the channel attention function, SA(•) is the spatial attention function, and ⊗ indicates the dot multiplication operation. The W_1^3, W_2^3 and W^3 are 2-D convolutional layers with a kernel size of 3×3. F_{cat_i} and F_{fused_i} represent the concatenated feature and fused feature, respectively, for $i \in \{1, 2, 3\}$.

2.3 | Refinement Module

To preserve the original information and mitigate the distortion resulting from inappropriate attentional weighting, the RM (Figure 1) is incorporated into the network. The RM refines the fused features by reusing and weighting those extracted by the RFEMs. Specifically, the RM calculates channel and spatial weights for the fused features F_{fused_i}, applies these weights to the spectral and spatial features from the RFEMs, and concatenates the weighted spatial features, weighted spectral features, and fused features. A 1×1 convolution layer then adjusts the weights between each feature, completing the refinement. The refinement process of the RM can be expressed as follows:

$$F_{refined_i} = W^1\left(CA\left(Spec_i\right) \otimes F_{fused_i} \oplus SA\left(Spat_i\right) \otimes F_{fused_i} \oplus F_{fused_i}\right) \tag{6}$$

where ⊗ and ⊕ represent the dot multiply and concatenation operations, respectively. CA(•) and SA(•) denote the channel attention function (Equation 3) and the spatial attention function (Equation 4), respectively. W^1 refers to a 2-D convolutional layer of size 1×1.

2.4 | Multiple Depth-Level Fusion and HSI Reconstruction

In the dual-branch feature extraction process, multiple cascaded RFEMs yield refined features at various depth levels after fusion and refinement. To fully leverage the information from different network depth levels, these features are aggregated via a 3×3 2-D convolutional layer to produce the final fused feature. The fused feature and upsampled LR-HSI are summed and processed through a 2-D convolutional layer to reconstruct the fused image.

2.5 | Loss Function

We use the HR-HSI as the reference image to calculate loss and evaluate fusion quality. The loss function consists of spatial loss L_{spat} and edge spectral loss L_{edge_spec}, detailed as follows:

$$\lambda_{edge\_spec} = \frac{e}{E} \tag{7}$$

$$L = L_{spat} + \lambda_{edge\_spec} L_{edge\_spec} \tag{8}$$

The weight of L_{spat} is set to 1. For L_{edge_spec}, a weight λ_{edge_spec} related to the training epoch is introduced, enabling the model to gradually shift its focus from spatially smooth regions to high-frequency edges as training progresses. Specifically, e and E represent the current training epoch and the total number of training epochs, respectively.

2.5.1 | Spatial Loss

The fused image and reference image are processed through a 2-D convolutional layer with identical parameters to obtain the corresponding feature maps. The commonly used spatial loss L_{spat} is then computed by applying the L_1 loss function to both feature maps.

$$f_{\mathrm{spat}}(\boldsymbol{x}) = W^3 \boldsymbol{x} \tag{9}$$

$$\mathrm{L}_1\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = \frac{1}{c \times s} \sum_{i=1}^{n} \left|\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j\right| \tag{10}$$

$$L_{spat} = L_1\left(f_{\mathrm{spat}}(\boldsymbol{Y}), f_{\mathrm{spat}}\left(\hat{\boldsymbol{Y}}\right)\right) \tag{11}$$

where W_1^3 represents a 2-D convolutional layer with a kernel size of 3×3, s and c denote the number of pixels per channel and the number of channels in the reference HR-HSI, respectively. The number of input and output channels corresponds to the number of bands in the fused image and 64, respectively. $f_{\mathrm{spat}}(\boldsymbol{x})$ denotes the convolution operation applied to the input $\boldsymbol{x}$, whereas $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ denote the reference HR-HSI and fused HR-HSI, respectively.

2.5.2 | Edge Spectral Loss

In hyperspectral image fusion tasks, spectral reconstruction tends to be more accurate in spatially homogeneous regions, whereas it becomes challenging in areas characterized by high-frequency edges. To address this, inspired by focal loss (Lin et al. 2017), we propose a novel loss function aimed at enhancing spectral perception at edges. Specifically, we first extract edges from the input HR-MSI using the Sobel operator, generating an edge mask. This mask is normalized via a sigmoid function to create a weight mask, denoted as $\boldsymbol{M}$, which modulates the spectral loss calculation. The proposed edge spectral loss function can be expressed as follows:

$$L_{edge\_spec} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(\boldsymbol{M}(i,j) \times \left|\boldsymbol{X}(i,j) - \hat{\boldsymbol{X}}(i,j)\right|\right) \tag{12}$$

where $\boldsymbol{M}(i,j)$ is the weight at pixel position (i,j), $\boldsymbol{X}(i,j)$ and $\hat{\boldsymbol{X}}(i,j)$ denote the pixel spectra of the reference and fused images, respectively.

3 | Experiment and Results

3.1 | Experimental Datasets

We evaluated the quality of the proposed method with four datasets, Pavia Centre (Plaza et al. 2009), Chikusei (Yokoya and Iwasaki 2016), MDAS (Hu et al. 2023), and Daxing, as detailed

TABLE 1 | Description of the details of the Pavia Centre, Chikusei, and Daxing datasets.

Description	Pavia Centre	Chikusei	Daxing
Wavelength range (nm)	430–860	363–1018	400–1000
Bands	102	128	224
Spatial size (pixels)	1096×1096	2517×2335	1024×1004
Spatial resolution (m)	1.3	2.5	0.11
Downsampling ratio	4	4	4
Patch size (reference)	160	256	256
Downsampling	Gaussian kernel 8×8		
Train/validation/test split	14/6/0	61/20/0	32/12/8

TABLE 2 | Description of the details of the MDAS dataset.

	MDAS		
Description	EnMAP (LRHSI)	Sentinel-2	EnMAP (HRHSI)
Spatial resolution (m)	30	10	10
Number of bands	242	4	242
Spatial resolution (pixels)	457×296	1371×888	1371×888
Train/validation/test split	213/15/0		
Patch size	24×24×242	72×72×4	

in Tables 1 and 2. For the public Pavia Centre and Chikusei datasets, LR-HSIs were generated from HR-HSIs following Wald's protocol (Wald et al. 1997) through Gaussian blurring and downsampling, whereas HR-MSIs were obtained by spectrally resampling HR-HSIs using the spectral response function of the Landsat-8 OLI sensor. The HR-HSIs of the Daxing dataset were collected by our laboratory on June 14, 2023, in Daxing District, Beijing, China, using a DJI UAV equipped with a GaiaSky-mini-VN hyperspectral and multispectral imagers. The multispectral imager can acquire RGB images corresponding to the HR-HSIs, which serve as HR-MSIs in our study, with an average alignment error of approximately 3.2 pixels for all sample pairs. LR-HSIs were generated via Gaussian blurring and downsampling from the HR-HSIs. The MDAS dataset was acquired in Augsburg, Germany, and consists of 10 and 30 m GSD EnMAP data, as well as 10 m GSD Sentinel-2 data, making it one of the few datasets available for the full-resolution experiment. Although the training and validation sets were randomly partitioned for the Pavia Centre, Chikusei, and MDAS datasets, an additional independent test set was incorporated for the Daxing dataset to assess generalization capability.

3.2 | Assessment Metrics

In this study, four widely used evaluation metrics were employed including cross correlation (CC), spectral angle mapper (SAM), Erreur relative globale adimensionnelle de synthèse (ERGAS) (Wald 2000), and Peak Signal-to-Noise Ratio (PSNR). For clarity, $\boldsymbol{Y} \in \boldsymbol{R}^{c\times s}$ and $\hat{\boldsymbol{Y}} \in \boldsymbol{R}^{c\times s}$ represent the reference HR-HSI and the fused image, respectively. Here, c and s represent the number of spectral bands and the number of pixels per spectral band in the reference HR-HSI, respectively.

3.2.1 | CC

CC quantifies the geometric distortion in image fusion and is defined as follows:

$$\mathrm{CC}\left(\boldsymbol{Y},\hat{\boldsymbol{Y}}\right)=\sum_{i=1}^{c}\mathrm{cc}\left(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i\right)/c \tag{13}$$

where $\mathrm{cc}\left(\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i\right)$ denotes the correlation coefficient between the ith band of reference HR-HSI and the ith band of fused HR-HSI, defined as follows:

$$\mathrm{cc}(\boldsymbol{x},\hat{\boldsymbol{x}})=\sum_{j=1}^{n}\left(\boldsymbol{x}_j-\mu_{\boldsymbol{x}}\right)\left(\hat{\boldsymbol{x}}_j-\mu_{\hat{\boldsymbol{x}}}\right)/\sqrt{\sum_{j=1}^{n}\left(\boldsymbol{x}_j-\mu_{\boldsymbol{x}}\right)^2\sum_{j=1}^{n}\left(\hat{\boldsymbol{x}}_j-\mu_x\right)^2} \tag{14}$$

where $\mu_{\boldsymbol{x}}$ and $\mu_{\hat{\boldsymbol{x}}}$ denote the mean values of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, respectively. Ideally, the CC of the fused HR-HSI should be close to 1, indicating a high correlation with the reference HR-HSI.

3.2.2 | SAM

SAM calculates the angle between the fused HR-HSI spectrum and the reference HR-HSI spectrum to assess spectral similarity. The SAM is computed by applying the L_2 loss function as follows:

TABLE 3 | Model hyperparameters and training settings.

Hyperparameter and training configuration	Setup
Number of RFEM layers	3
Number of residual blocks in each RFEM	1
Kernel size and stride of Conv2d layers	3, 1
Input/output channels of Conv2d layers in RFEMs	32/32
Input/output channels of SSFMs and RMs	32/32
Sampling ratio in SSFMs	4
Initial learning rate	0.001
Weight decay	0.0001
Gamma	0.1
Training and validation batch sizes	2, 1

TABLE 4 | The metrics for the fused images of the comparative methods on the Pavia Centre dataset.

Method	PSNR (dB ↑)	CC (↑)	SAM (↓)	ERGAS (↓)
GLPHS	27.3126	0.8963	7.1375	7.3411
CNMF	27.1139	0.8808	7.6403	7.4717
HySure	26.5859	0.8280	9.3053	7.9832
ICCV15	26.6538	0.8709	9.2157	7.8649
SSRNet	27.9422	0.8852	11.7482	8.3277
MSDCNN	29.0529	0.9264	6.9595	6.0270
ResTFNet	31.9973	0.9599	5.6434	4.3509
HSRNet	30.5566	0.9468	6.0621	5.0992
MSSSHANet	31.9917	0.9618	6.5610	4.3457
MDSSAN	**34.4189**	**0.9747**	**5.3325**	**3.4094**
Improvement	7.6%	1.3%	5.5%	21.5%

Note: The best values are highlighted in bold.

$$L_2 = \|\boldsymbol{y}\|_2 = \sqrt{\sum_{i=1}^{n} {y_i}^2} \tag{15}$$

$$\mathrm{SAM}\left(\boldsymbol{y}_j, \hat{\boldsymbol{y}}_j\right) = \cos^{-1}\left({\boldsymbol{y}_j}^T \cdot \hat{\boldsymbol{y}}_j / \left(\left\|\boldsymbol{y}_j\right\|_2 \left\|\hat{\boldsymbol{y}}_j\right\|_2\right)\right) \tag{16}$$

where $\boldsymbol{y}_j$ and $\hat{\boldsymbol{y}}_j$ denote the spectral vectors of the fused HR-HSI and the reference HR-HSI at the jth pixel. SAM values close to 0 indicate better spectral fidelity, and the average SAM value of all pixels was used as the overall SAM result in the experiments.

3.2.3 | ERGAS

ERGAS considers the influence of ground sampling distance and characterizes a global fusion quality of the fused HR-HSI. It is defined as follows:

$$\mathrm{ERGAS}\left(\boldsymbol{Y}, \hat{\boldsymbol{Y}}\right) = 100\beta\sqrt{\frac{1}{c}\sum_{i=1}^{c} \left\|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\right\|_2^2 / \left({\boldsymbol{l}_s}^T \cdot \boldsymbol{y}_i / s\right)^2} \tag{17}$$

where $\boldsymbol{y}_i$ and $\hat{\boldsymbol{y}}_i$ represent the ith band of the fused HR-HSI and the ith band of the reference HR-HSI, respectively, and β represents the ratio of the spatial resolution of the HR-MSI to that of the LR-HSI, and $\boldsymbol{l}_s = [1, \ \ldots, 1] \in \boldsymbol{R}^{s\times 1}$.

3.2.4 | PSNR

PSNR is utilized to assess the quality of spatial reconstruction for each band and is defined as the ratio of the maximum power to the residual power of the signal. The PSNR can be defined as follows:

$$\mathrm{PSNR}\left(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i\right) = 10\log_{10}\left(\max\left(\boldsymbol{y}_i\right) / \left(\left\|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\right\|_2^2 / s\right)\right) \tag{18}$$

where $\max\left(\boldsymbol{y}_i\right)$ denotes the maximum value of the ith image band in the reference HR-HSI. A larger PSNR value indicates higher spatial quality of the fused HR-HSI. The average PSRN value of all bands was used as the overall PSNR result.

3.3 | Model Settings

The Adam optimizer was adopted in the model training in the study. The learning rate was adjusted every 200 training epochs by multiplying a factor, *gamma*. Detailed model hyperparameters and training settings are provided in Table 3. In both SSFMs and RMs, notably, the number of input channels in convolutional layers only changes after concatenation operations. The other convolutions maintain an input and output channel number of 32. The spatial loss weight was fixed at 1, and the edge spectral loss weight was related to the training epoch, which is calculated using Equation (7). All experiments were conducted using the PyTorch framework on an NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM.

3.4 | Experimental Results

We compared the proposed model against four conventional methods (GLPHS (Selva et al. 2015), HySure (Simoes et al. 2015), ICCV15 (Akhtar et al. 2014), and CNMF (Yokoya et al. 2012)) and five deep learning networks (SSRNet, MSDCNN, ResTFNet, HSRNet (Hu, Huang, Deng, Jiang, et al. 2022), and MSSSHANet (Zhang et al. 2025)). These comparative experiments were conducted on the Pavia Centre, Chikusei, MDAS, and Daxing datasets. To further demonstrate the effectiveness of our approach in enlarging the receptive field, results from the transformer-based method NGSTNet (Feng et al. 2024) were included for the Daxing validation and test set.

3.4.1 | Experiment on the Pavia Centre Dataset

The experimental results conducted on the Pavia Centre dataset for various quantitative metrics are displayed in Table 4. Bolded

values indicate the best performance among all methods. The proposed MDSSAN achieved optimal results across all metrics, with improvements of 1.3%, 7.6%, 5.5%, and 21.5% in CC, PSNR, SAM, and ERGAS, respectively. RGB composite images for each method are shown in Figure 3. Notably, all four traditional methods exhibited significant spatial distortions in their fusion results, with CNMF producing erroneous estimations at building edges. Additionally, all traditional methods introduced some degree of spectral distortion, particularly in vegetation reconstruction, where shadows were overemphasized, resulting in a loss of spectral and spatial information. In contrast, SSRNet, MSDCNN, ResTFNet, HSRNet, and MSSSHANet demonstrated improvements in spatial and spectral restoration compared to the traditional methods, but still failed to produce high-quality fused images. MDSSAN generated fusion results closely resembling the reference image, underscoring its effectiveness in retaining both spectral and spatial information. The mean absolute error (MAE) results for each method are depicted in Figure 4. Among all compared methods, MDSSAN excels in MAE, displaying fewer pixels with higher spectral errors, which further illustrates the superior spectral retention capability of the proposed method.

3.4.2 | Experiment on the Chikusei Dataset

The values presented in Table 5 represent the mean and standard deviation calculated from five repeated experiments for comparative methods with the Chikusei dataset. The proposed MDSSAN achieved optimal results across all metrics, with improvements of 1.5%, 9.5%, 23.3%, and 30.6% in CC, PSNR, SAM, and ERGAS, respectively. The RGB composite images fused by the comparative methods from the first experiment are displayed in Figure 5, revealing that all four traditional methods suffered significant loss of spatial details in their fused images. Among the deep learning methods, SSRNet exhibited substantial spatial information loss due to its simple structure and presented noticeable spectral aberrations. In contrast, MSDCNN,

FIGURE 3 | RGB composite images for each method on the 17th sample of the Pavia Centre dataset. In the second row are the zoom-in images of the red box in the first-row images.

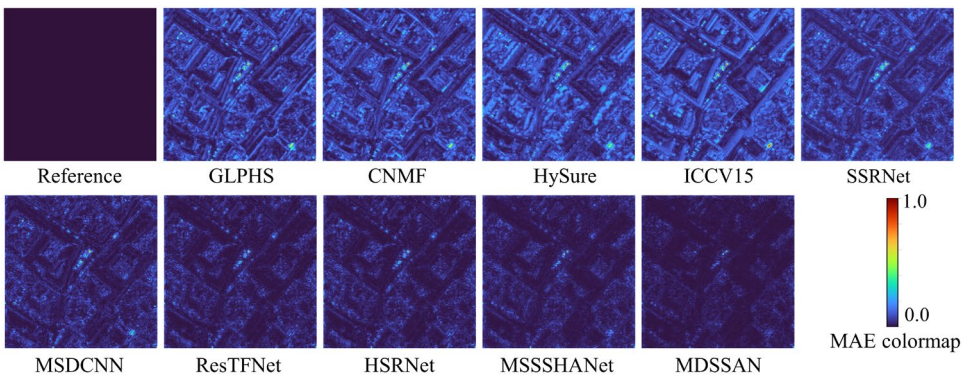

FIGURE 4 | MAE for each method on the 17th sample of the Pavia Centre dataset.

TABLE 5 | The metrics of the fused images of the comparative methods on the Chikusei dataset.

Method	PSNR (dB ↑)	CC (↑)	SAM (↓)	ERGAS (↓)
GLPHS	29.2795 ± 0.00	0.9184 ± 0.00	3.2868 ± 0.00	6.0916 ± 0.00
CNMF	28.5086 ± 0.00	0.8850 ± 0.02	3.7088 ± 0.02	6.4215 ± 0.01
HySure	26.6745 ± 0.00	0.7312 ± 0.00	4.7684 ± 0.00	7.3722 ± 0.00
ICCV15	28.3095 ± 0.05	0.9008 ± 0.00	3.7374 ± 0.02	6.8408 ± 0.06
SSRNet	28.4032 ± 0.00	0.8245 ± 0.00	8.8913 ± 0.00	11.7715 ± 0.00
MSDCNN	30.8042 ± 1.18	0.9388 ± 0.01	3.4286 ± 0.55	5.4019 ± 0.70
ResTFNet	32.0589 ± 0.10	0.9507 ± 0.00	2.5320 ± 0.15	4.3614 ± 0.13
HSRNet	32.4797 ± 0.04	0.9566 ± 0.00	2.3652 ± 0.01	4.2771 ± 0.01
MSSSHANet	34.0893 ± 0.12	0.9666 ± 0.00	2.7530 ± 0.01	3.7243 ± 0.07
MDSSAN	**37.3522** ± 0.09	**0.9813** ± 0.00	**1.8136** ± 0.01	**2.5839** ± 0.02
Improvement	9.5%	1.5%	23.3%	30.6%

Note: The best values are highlighted in bold.

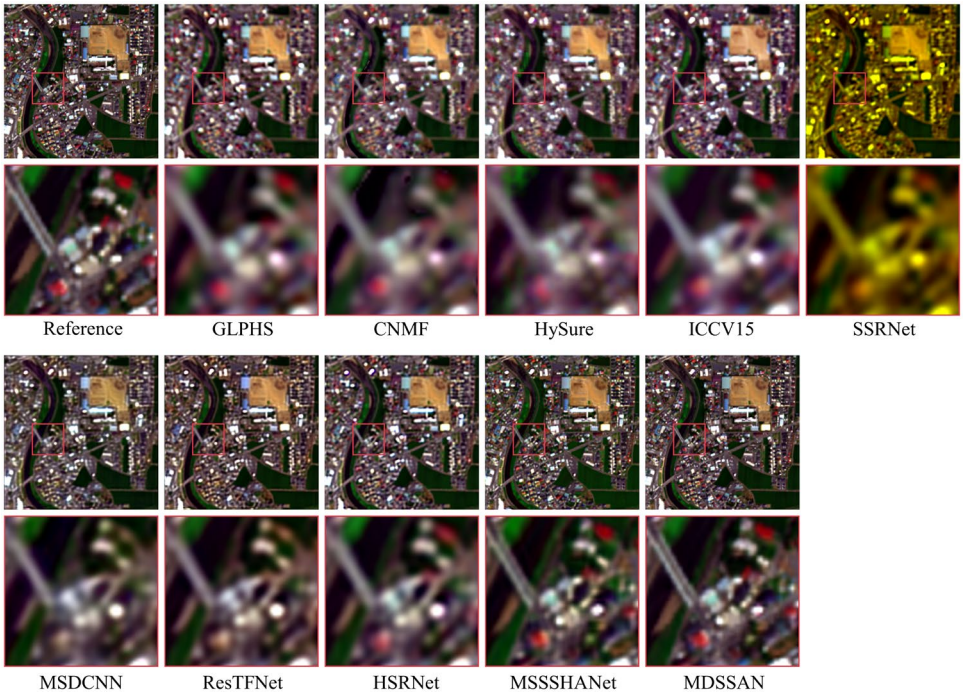

FIGURE 5 | RGB composite image for each method on the 20th sample of the Chikusei dataset. In the second row are the zoom-in images of the red box in the first-row images.

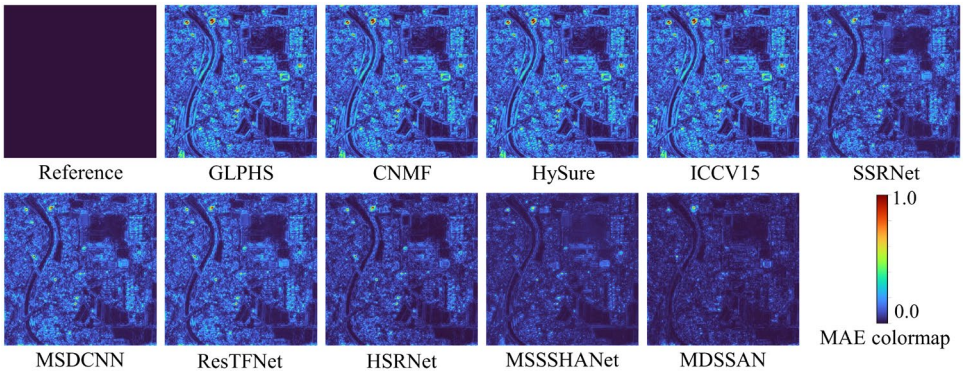

FIGURE 6 | MAE for each method on the 20th sample of the Chikusei dataset.

ResTFNet, HSRNet, and MSSSHANet demonstrated improved spatial restoration of fine objects and edges, although their retention of spatial information remains inferior to that of MDSSAN. Specifically, MSDCNN and ResTFNet produced spectral distortions in the visual recovery of certain features such as buildings. Overall, a comprehensive comparison of the fused images indicates that MDSSAN achieves a superior balance between spectral and spatial fidelity, recovering as much spatial detail as possible while maintaining better spectral integrity. The MAE statistics for each method are illustrated in Figure 6. Although MSSSHANet demonstrates competent spatial reconstruction capabilities, it exhibits significant spectral distortion artifacts in smooth regions. Among all the compared methods, MDSSAN achieves the lowest MAE, indicating a greater number of pixels with lower spectral errors. It is noteworthy that the Chikusei dataset not only offers higher spatial resolution but also comprises a larger number of samples, with pixel ratios among the four datasets, such as Pavia Centre, Chikusei, MDAS, and Daxing, being 1:10:2:5. The relatively larger volume of data are especially beneficial for training the proposed method, which utilizes high-complexity self-attention mechanisms. As a result, the performance enhancements over traditional CNN-based methods are particularly evident when applied to the Chikusei dataset.

TABLE 6 | The metrics for the fused images of the comparative methods on the MDAS dataset.

Method	PSNR (dB ↑)	CC (↑)	SAM (↓)	ERGAS (↓)
GLPHS	26.0063	0.9199	4.9323	6.4168
CNMF	27.3573	0.9322	4.7090	6.0960
HySure	24.8286	0.8904	7.2408	8.9265
ICCV15	27.4876	0.9465	3.9679	5.8597
SSRNet	30.5889	0.9494	3.6768	4.3277
MSDCNN	31.9571	0.9666	2.8341	3.8694
ResTFNet	32.0395	0.9665	2.9766	3.9260
HSRNet	30.9226	0.9604	3.0092	3.8892
MSSSHANet	31.8750	0.9683	2.9899	3.5040
MDSSAN	**32.9074**	**0.9746**	**2.4613**	**3.3299**
Improvement	2.7%	0.7%	13.2%	4.9%

Note: The best values are highlighted in bold.

3.4.3 | Experiment on the MDAS Dataset

The experimental results of each method on the MDAS dataset for various quantitative metrics are presented in Table 6. The proposed MDSSAN achieved optimal fused images across all four metrics, with improvements of 0.7%, 2.7%, 13.2%, and 4.9% in CC, PSNR, SAM, and ERGAS, respectively. In the false color composite images (Figure 7), the outcomes of all the comparative methods seem quite similar in overall visual effect due to the low resolution of the MDAS dataset. However, in terms of local details, the four traditional methods, along with SSRNet, presented a loss of spatial detail, whereas MSDCNN, ResTFNet, HSRNet, and MSSSHANet demonstrated poorer quality in detail recovery. In comparison, the MDSSAN method produced results that were closest to the reference image. When evaluating spectral reconstruction quality (Figure 8), MDSSAN achieved the smallest MAE.

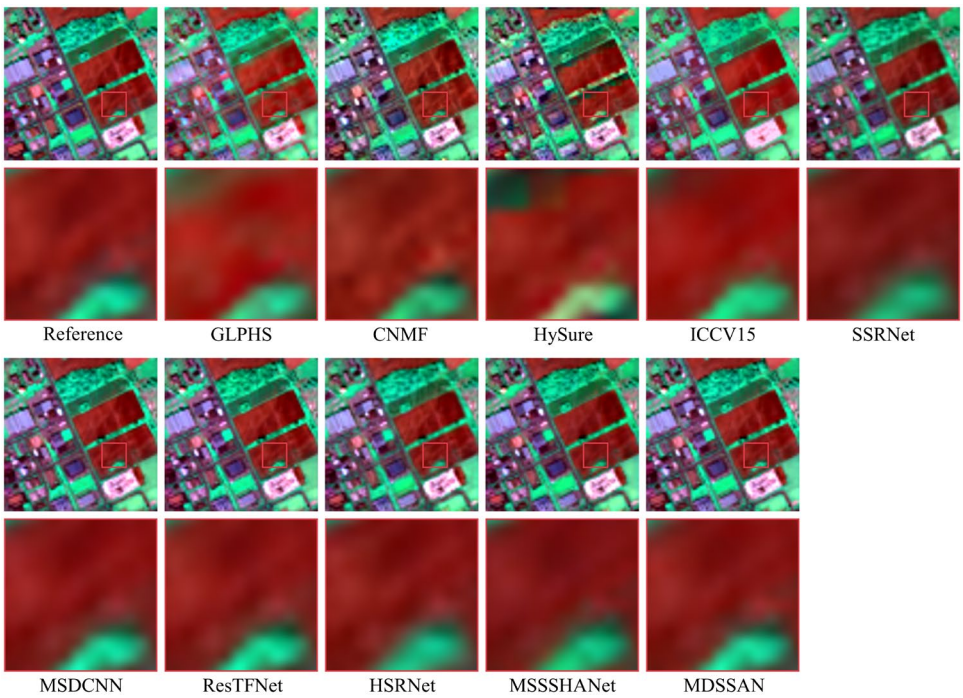

FIGURE 7 | False color (R-200, G-100, B-34) composite image for each method on the 13th sample of the MDAS dataset. In the second row are the zoom-in images of the red box in the first-row images.

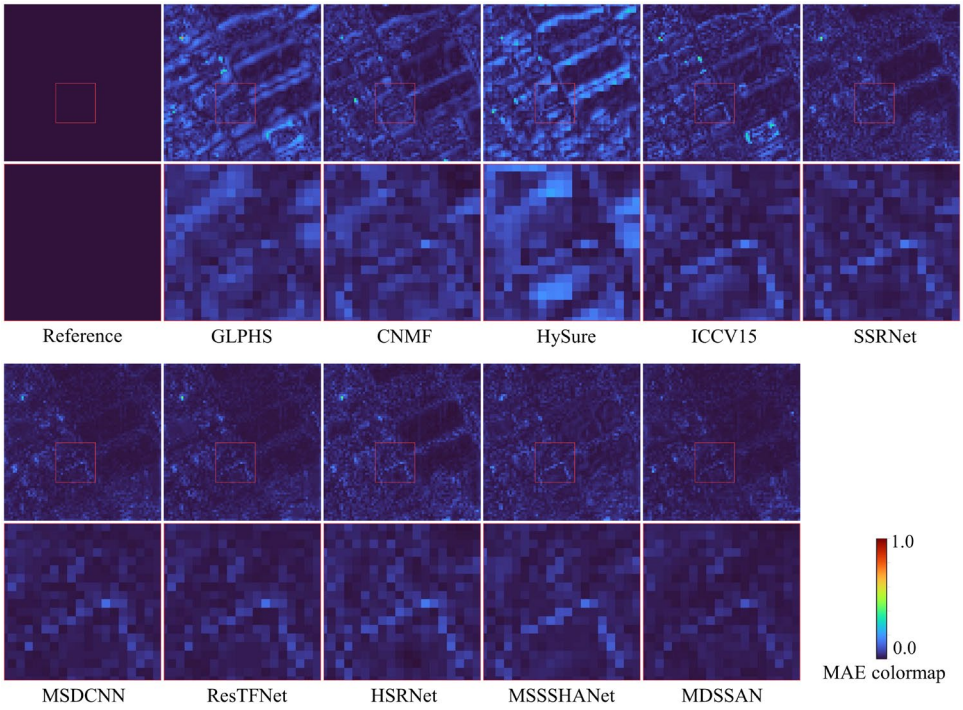

FIGURE 8 | MAE for each method on the 13th sample of the MDAS dataset.

3.4.4 | Experiment on the Daxing Dataset

The quantitative metrics of the resultant fusion image of all the methods with the Daxing dataset are presented in Table 7. Notably, the displayed sample in this subsection exhibits an alignment error of approximately 7 pixels. The proposed MDSSAN achieved optimal results across all the assessment metrics, with improvements of 0.3%, 1.2%, 5.5%, and 3.6% in CC, PSNR, SAM, and ERGAS, respectively. Figure 9 displays the reference image, RGB, and 4× upsampled LR-HSI for the HSI sample. It is evident that significant pixel offsets and luminance differences exist between the RGB and HSI, compounded by the presence of moving targets in the RGB image. Such substantial modal differences are common in practical image fusion applications. Consequently, all traditional methods inadvertently incorporate the spatial features of moving targets in the RGB image, leading to erroneous fusion results. Due to their sensitivity to alignment errors, these methods produced noticeable edge blurring and bilateral artifacts in feature reconstruction. Although the deep learning approaches effectively mitigated the misestimation of moving target features and demonstrated significant advantages in feature extraction and fusion, SSRNet suffered from severe spectral distortion due to the large pixel offsets and luminance differences. Meanwhile, MSDCNN, ResTFNet, and HSRNet experienced considerable blurring and deformation in spatial detail recovery; consequently, ResTFNet, HSRNet, and MSSSHANet failed to eliminate the effects of moving targets (Figure 9, red box). This indicates that existing residual convolutional fusion networks with relatively small receptive fields fail to capture features when an obvious pixel shift occurs, thereby limiting their fusion performance. Moreover, NGSTNet failed to fully exploit the long-range dependency capturing capability of self-attention mechanisms, owing to its trade-off between model performance and computational efficiency (e.g., employing 8×8 windowed attention). In contrast, MDSSAN achieves significantly broader attention coverage—with self-attention window size expanded to 64×64 through quadruple sampling—demonstrating clear performance superiority, resulting in visual effects closest to the reference image. Additionally, MDSSAN achieved optimal results in spectral reconstruction (Figure 10). This demonstrated that MDSSAN enlarges the receptive field of the residual network through the SSFM, effectively utilizes features at different depth levels, and significantly improves the adaptability of the residual fusion network in real-world scenarios.

TABLE 7 | The metrics for the fused images of the comparative methods on the Daxing validation dataset.

Method	PSNR (dB ↑)	CC (↑)	SAM (↓)	ERGAS (↓)
GLPHS	30.7236	0.9210	1.8953	2.2552
CNMF	27.9259	0.9198	1.4649	2.5547
HySure	24.7493	0.8390	3.3340	3.8269
ICCV15	28.6941	0.8676	2.2120	2.8432
SSRNet	28.5468	0.7951	18.9608	17.5041
MSDCNN	32.4626	0.9277	2.0778	2.6265
ResTFNet	32.2146	0.9233	1.9455	1.7368
HSRNet	34.6925	0.9563	1.1772	1.3758
NGSTNet	35.4519	0.9607	1.1265	1.2145
MSSSHANet	35.4786	0.9622	1.1650	1.2520
MDSSAN	**35.9006**	**0.9657**	**1.0644**	**1.1699**
Improvement	1.2%	0.3%	5.5%	3.6%

Note: The best values are highlighted in bold.

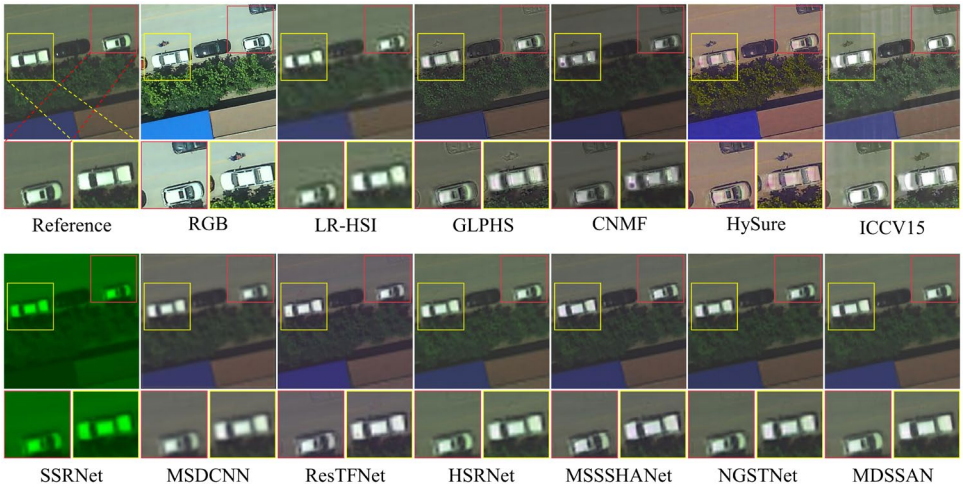

FIGURE 9 | RGB composite image for each method on the 23rd sample of the Daxing dataset. In the first row are the zoom-in images of the red box in the first-row images.

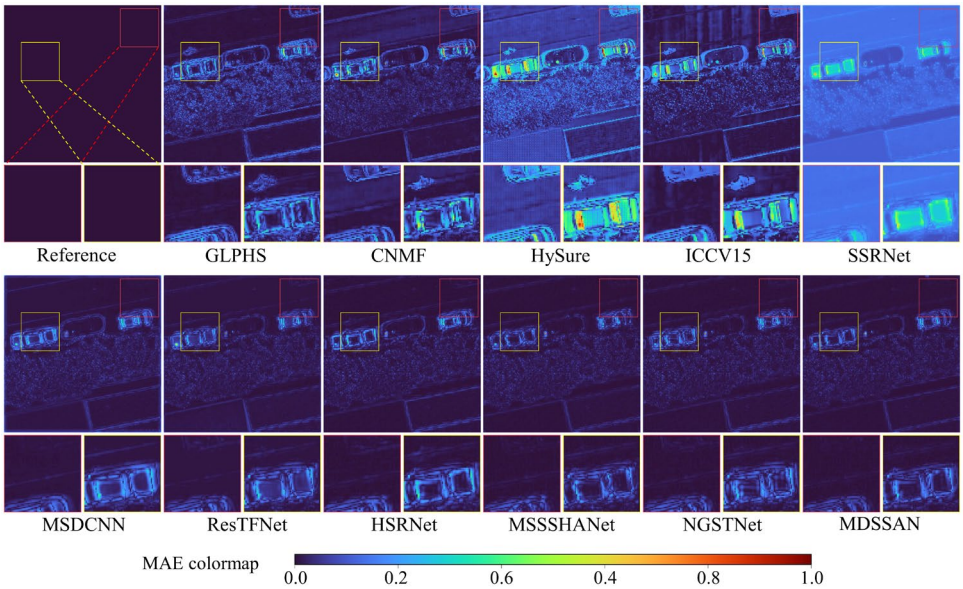

FIGURE 10 | MAE for each method on the 23rd sample of the Daxing dataset.

To further evaluate generalization performance, we conducted inference on the Daxing test set using the well-trained deep learning-based fusion methods. As shown in Table 8, all deep learning methods exhibited a decrease in performance metrics on the test set compared to the validation set (Table 7). Nevertheless, the proposed MSSDAN still achieved the best results across all four spatial–spectral evaluation metrics. Furthermore, as illustrated in Figure 11, our method demonstrated the smallest overall spectral reconstruction differences for vegetation, impervious surfaces, and bare soil, indicating its strong generalization capability.

TABLE 8 | The metrics for the fused images of the comparative models on the Daxing test set.

Model	PSNR (dB ↑)	CC (↑)	SAM (↓)	ERGAS (↓)
SSRNet	26.0010	0.7188	28.7898	12.0935
MSDCNN	30.4685	0.9003	2.0811	1.7128
ResTFNet	32.4595	0.9374	1.8478	1.4340
HSRNet	33.9748	0.9576	1.2278	1.3184
NGSTNet	34.7451	0.9632	1.1713	1.2074
MSSSHANet	34.6476	0.9606	1.2174	1.1821
MDSSAN	**35.1218**	**0.9667**	**1.1094**	**1.1374**

Note: The best values are highlighted in bold.

4 | Discussion

4.1 | Ablation Experiments

To elucidate the contribution of each module of the proposed method to the fusion performance, we set the number of RFEM layers in the feature extraction branch to one. Ablation experiments were then conducted on the proposed SSFM, edge spectral loss L_{edge}, and the applied RM, as summarized in Table 9. The baseline configuration comprises a single layer of RFEMs

without any of the SSFM, RM, or L_{edge}. The ablation of SSFM was achieved by replacing the spatial–spectral fusion step with simple concatenation and convolution operations. When the L_{edge} is first appended to the baseline, the CC remains almost unchanged while SAM improves markedly, demonstrating that the proposed edge spectral supervision is indispensable for spectral fidelity. After further integrating the RM, both spatial and spectral metrics exhibit modest gains, attributable to channel-wise and pixel-wise weighting and reuse of spatial–spectral features, which refines the fused features. Finally, the introduction of the SSFM yields significant improvements in PSNR and SAM, confirming that its stepwise sampling and self-attention enlarge the spatial receptive field, allowing the network to leverage more contextual pixels for accurate spectral and spatial reconstruction.

4.2 | Robustness Analysis

To further demonstrate the robustness of the proposed method, simulation experiments were conducted with the Pavia Centre dataset. Pixel offsets between the reference image and the HR-MSI were introduced by rotating the reference image clockwise around its center. This section evaluates the fusion performance of various deep learning methods at rotation angles of 1°, 2°, 3°, 4°, and 5° (see Figure 12). As the rotation angle increases, the spatial and spectral reconstruction quality of all methods generally declines, indicating that the residual learning-based fusion methods with local receptive fields are sensitive to alignment errors. In contrast, the proposed model, which can aggregate similar features from other pixel locations—including the misaligned pixels—demonstrates superior robustness compared to the comparative residual-based fusion methods.

4.3 | Sensitivity Analysis

To illustrate the impact of varying depth levels on fusion results, a sensitivity analysis was conducted on the number of RFEM layers (Table 10). The results demonstrate that model

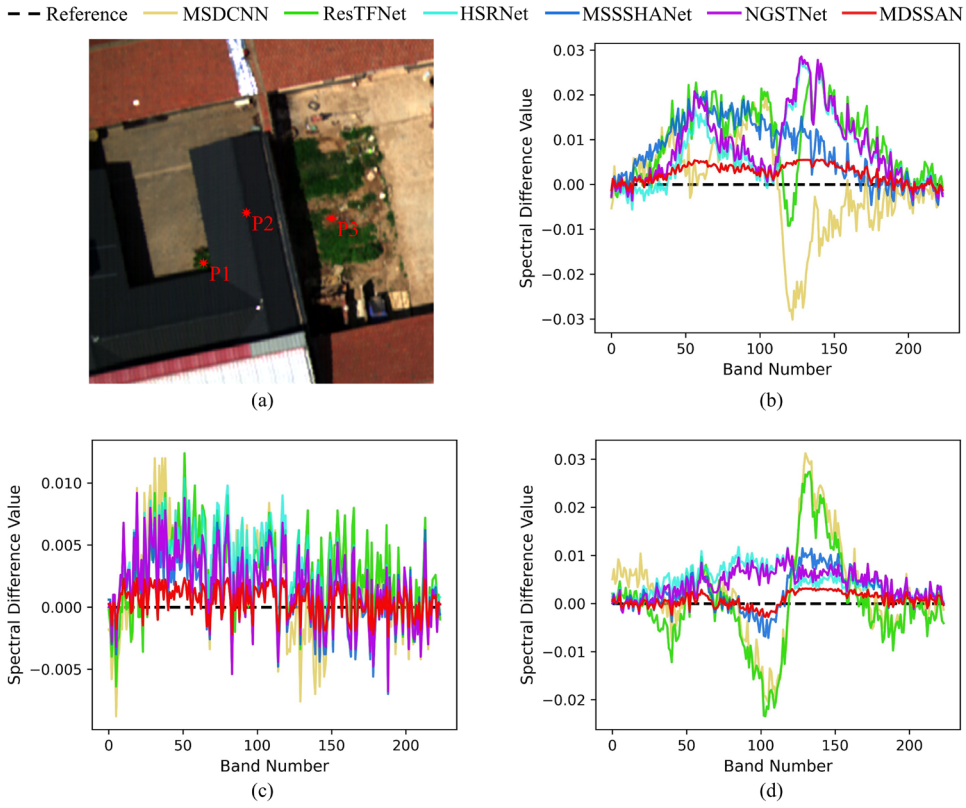

FIGURE 11 | (a) RGB composite image and selected points p1 at position (84, 167), p2 at position (117, 129), and p3 at position (180, 135) on the 4th sample of the Daxing test set; (b) p1-vegetation, (c) p2-roof, and (d) p3-soil spectral difference curves of comparative models.

TABLE 9 | Ablation study for the modules and edge spectral loss with the Pavia Centre dataset.

Model	PSNR (dB ↑)	CC (↑)	SAM (↓)	ERGAS (↓)
Baseline	32.2560	0.9653	6.5539	4.2359
Baseline + L_{edge}	32.6593	0.9658	6.3744	4.0661
Baseline + RM + L_{edge}	32.9158	0.9686	6.3042	3.9392
Baseline + SSFM + RM + L_{edge}	33.2867	0.9692	5.5653	3.7981

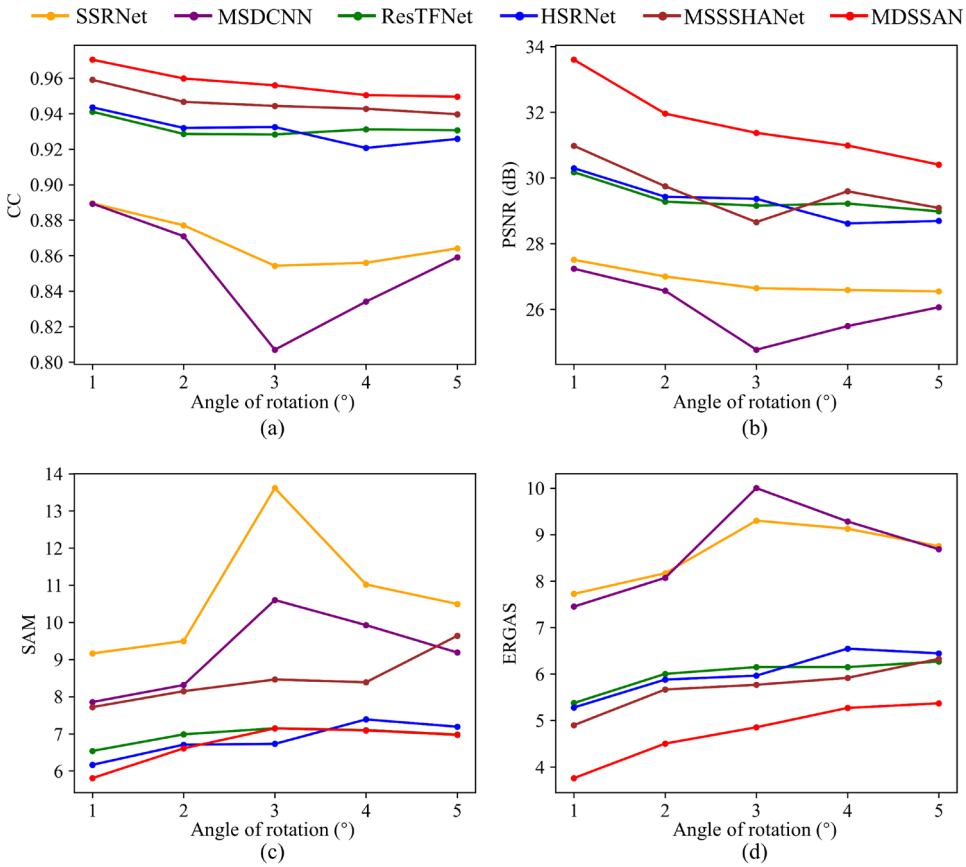

FIGURE 12 | Variations in CC (a), PSNR (b), SAM (c), and ERGAS (d) values for the fused results of comparative CNN methods when rotated at different angles with the Pavia Centre dataset.

TABLE 10 | Sensitivity of the proposed method to the number of RFEM and SSFM with the Pavia Centre dataset.

Layers	PSNR (↑)	CC (↑)	SAM (↓)	ERGAS (↓)
1	33.2867	0.9692	5.5653	3.7981
2	33.5388	0.9707	5.4660	3.7025
3	**34.4189**	**0.9747**	**5.3325**	**3.4094**
4	34.3910	0.9736	5.3544	3.4506
5	34.2479	0.9733	5.3630	3.4821

Note: The best values are highlighted in bold.

TABLE 11 | Model complexity analysis for each model in single-sample inference with the MDAS dataset.

Model	Params (MB)	Inference time (ms)	FLOPs (G)	Graphic memory (MB)
SSRNet	6.03	42.80	8.20	68.31
MSDCNN	9.85	79.89	13.38	74.81
ResTFNet	9.33	39.28	2.97	86.47
HSRNet	19.69	37.04	4.61	75.57
MSSSHANet	3.90	68.43	5.28	**70.60**
MDSSAN	**1.39**	**27.76**	**1.25**	70.97

Note: The best values are highlighted in bold.

performance continues to improve with an increasing number of RFEMs, reaching an optimum at three layers. Beyond this point, fusion performance declines, which may be attributed to difficulties in global spatial–spectral feature aggregation when training deeper SSFMs. Empirically, three RFEM layers yield optimal fusion images in this study.

4.4 | Complexity Analysis

The network parameters, inference time, number of floating-point operations (FLOPs), and GPU memory usage of each deep learning method were further analysed (Table 11). For these experiments, the number of RFEM layers in the proposed MDSSAN was set to one. All tests were conducted on sample pairs of sizes 72×72×4 and 24×24×242 using the same device (12th Gen Intel Core i7-12700@2.10 GHz). The proposed model showcased a significant advantage in both parameter count and FLOPs. Although a pixel-dense self-attention mechanism is employed to enlarge the receptive field of the fused features, the sampling block ensures that the process does not introduce excessive computational overhead. As a result, MDSSAN can considerably amplify the performance of residual learning-based convolutional fusion networks while maintaining a small parameter footprint, thereby improving the practicality of such efficient methods in hyperspectral image fusion. However, it should be noted that the self-attention mechanism, which scales quadratically with the spatial dimension of

the input, may lead to reduced inference speed and increased GPU memory usage for larger input images.

5 | Conclusion

This article proposes a novel multiple depth-level spatial–spectral aggregation network for multispectral and hyperspectral image fusion tasks. The proposed spatial–spectral feature fusion module significantly expands the receptive field of fused features, resulting in enhanced detail texture transfer and spectral reconstruction. Furthermore, the edge spectral loss function directs the model's attention to high-frequency edges in the input images, further boosting fusion performance. Overall, the developed MDSSAN network effectively leverages the complementary strengths of multispectral and hyperspectral data, significantly improving spatial–spectral feature fusion quality within residual learning-based networks. Extensive experiments across diverse datasets have demonstrated MDSSAN's robust performance and efficiency, characterized by a small parameter count, underscoring its potential for hyperspectral image fusion applications.

In addition, the experiments suggest that appropriately increasing the receptive field of the residual fusion network can augment its adaptability to alignment errors by enabling the model to consider a broader range of pixel features. However, both alignment errors and the geometric scale of key targets influence fusion outcomes. Therefore, future research should focus on elucidating the dependency between alignment errors and the scale of reconstructed objects to ensure that fusion algorithms can effectively reconstruct most scene objects.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Aharon, M., M. Elad, and A. Bruckstein. 2006. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation." *IEEE Transactions on Signal Processing* 54, no. 11: 4311–4322. https://doi.org/10.1109/TSP.2006.881199.

Akhtar, N., F. Shafait, and A. Mian. 2014. "Sparse Spatio-Spectral Representation for Hyperspectral Image Super-Resolution." *European Conference on Computer Vision*, Cham, 63–78. https://doi.org/10.1007/978-3-319-10584-0_5.

Borsoi, R. A., C. Prevost, K. Usevich, D. Brie, J. C. M. Bermudez, and C. Richard. 2021. "Coupled Tensor Decomposition for Hyperspectral and Multispectral Image Fusion With Inter-Image Variability." *IEEE Journal of Selected Topics in Signal Processing* 15, no. 3: 702–717. https://doi.org/10.1109/JSTSP.2021.3054338.

Burt, P., and E. Adelson. 1983. "The Laplacian Pyramid as a Compact Image Code." *IEEE Transactions on Communications* 31, no. 4: 532–540. https://doi.org/10.1109/TCOM.1983.1095851.

Carper, W. J., T. M. Lillesand, and R. W. Kiefer. 1990. "The Use of Intensity-Hue-Saturation Transformations for Merging Spot Panchromatic and Multispectral Image Data." *Photogrammetric Engineering and Remote Sensing* 56, no. 4: 459–467.

Chang, Y., L. Yan, X. L. Zhao, H. Fang, Z. Zhang, and S. Zhong. 2020. "Weighted Low-Rank Tensor Recovery for Hyperspectral Image Restoration." *IEEE Transactions on Cybernetics* 50, no. 11: 4558–4572. https://doi.org/10.1109/TCYB.2020.2983102.

Chavez, P. S., and A. Y. Kwarteng. 1989. "Extracting Spectral Contrast in Landsat Thematic Mapper Image Data Using Selective Principal Component Analysis." *Photogrammetric Engineering and Remote Sensing* 55, no. 3: 339–348.

Churchill, S., C. Randell, D. Power, and E. Gill. 2004. "Data Fusion: Remote Sensing for Target Detection and Tracking." *IEEE International Geoscience and Remote Sensing Symposium*, 20–24 Sept. 2004, 1–612. https://doi.org/10.1109/IGARSS.2004.1369101.

Demirel, H., and G. Anbarjafari. 2011. "Image Resolution Enhancement by Using Discrete and Stationary Wavelet Decomposition." *IEEE Transactions on Image Processing* 20, no. 5: 1458–1460. https://doi.org/10.1109/TIP.2010.2087767.

Deng, S. Q., L. J. Deng, X. Wu, R. Ran, D. Hong, and G. Vivone. 2023. "PSRT: Pyramid Shuffle-and-Reshuffle Transformer for Multispectral and Hyperspectral Image Fusion." *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–15. https://doi.org/10.1109/TGRS.2023.3244750.

Dian, R., L. Fang, and S. Li. 2017. "Hyperspectral Image Super-Resolution via Non-Local Sparse Tensor Factorization." *IEEE Conference on Computer Vision and Pattern Recognition*, 21–26 July 2017, 3862–3871. https://doi.org/10.1109/CVPR.2017.411.

Ding, X., X. Zhang, J. Han, and G. Ding. 2022. "Scaling up Your Kernels to 31×31: Revisiting Large Kernel Design in CNNs." *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18–24 June 2022, 11953–11965. https://doi.org/10.1109/CVPR52688.2022.01166.

Dong, W., F. Fu, G. Shi, et al. 2016. "Hyperspectral Image Super-Resolution via Non-Negative Structured Sparse Representation." *IEEE Transactions on Image Processing* 25, no. 5: 2337–2352. https://doi.org/10.1109/TIP.2016.2542360.

Ehlers, M. 1991. "Multisensor Image Fusion Techniques in Remote Sensing." *ISPRS Journal of Photogrammetry and Remote Sensing* 46, no. 1: 19–30. https://doi.org/10.1016/0924-2716(91)90003-E.

Feng, Z., X. Zhang, B. Zhou, M. Ren, and X. Chen. 2024. "NGST-Net: A N-Gram Based Swin Transformer Network for Improving Multispectral and Hyperspectral Image Fusion." *International Journal of Digital Earth* 17, no. 1: 2359574. https://doi.org/10.1080/17538947.2024.2359574.

Hong, D., N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu. 2019. "Learnable Manifold Alignment (Lema): A Semi-Supervised Cross-Modality Learning Framework for Land Cover and Land Use Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 147: 193–205. https://doi.org/10.1016/j.isprsjprs.2018.10.006.

Hu, J., R. Liu, D. Hong, et al. 2023. "MDAS: A New Multimodal Benchmark Dataset for Remote Sensing." *Earth System Science Data* 15, no. 1: 113–131. https://doi.org/10.5194/essd-15-113-2023.

Hu, J. F., T. Z. Huang, L. J. Deng, H. X. Dou, D. Hong, and G. Vivone. 2022. "Fusformer: A Transformer-Based Fusion Network for Hyperspectral Image Super-Resolution." *IEEE Geoscience and Remote Sensing Letters* 19: 1–5. https://doi.org/10.1109/LGRS.2022.3194257.

Hu, J. F., T. Z. Huang, L. J. Deng, T. X. Jiang, G. Vivone, and J. Chanussot. 2022. "Hyperspectral Image Super-Resolution via Deep Spatiospectral Attention Convolutional Neural Networks." *IEEE Transactions on*

Neural Networks and Learning Systems 33, no. 12: 7251–7265. https://doi.org/10.1109/TNNLS.2021.3084682.

Jia, S., Z. Min, and X. Fu. 2023. "Multiscale Spatial–Spectral Transformer Network for Hyperspectral and Multispectral Image Fusion." *Information Fusion* 96: 117–129. https://doi.org/10.1016/j.inffus.2023.03.011.

Li, S., R. Dian, L. Fang, and J. M. Bioucas-Dias. 2018. "Fusing Hyperspectral and Multispectral Images via Coupled Sparse Tensor Factorization." *IEEE Transactions on Image Processing* 27, no. 8: 4118–4130. https://doi.org/10.1109/TIP.2018.2836307.

Lin, T. Y., P. Goyal, R. Girshick, K. He, and P. Dollár. 2017. "Focal Loss for Dense Object Detection." *IEEE International Conference on Computer Vision*, 2017, 2999–3007. https://doi.org/10.1109/ICCV.2017.324.

Liu, N., L. Li, W. Li, R. Tao, J. E. Fowler, and J. Chanussot. 2021. "Hyperspectral Restoration and Fusion With Multispectral Imagery via Low-Rank Tensor-Approximation." *IEEE Transactions on Geoscience and Remote Sensing* 59, no. 9: 7817–7830. https://doi.org/10.1109/TGRS.2020.3049014.

Liu, Q., X. Meng, F. Shao, and S. Li. 2023. "Supervised-Unsupervised Combined Deep Convolutional Neural Networks for High-Fidelity Pansharpening." *Information Fusion* 89: 292–304. https://doi.org/10.1016/j.inffus.2022.08.018.

Liu, X., Q. Liu, and Y. Wang. 2020. "Remote Sensing Image Fusion Based on Two-Stream Fusion Network." *Information Fusion* 55: 1–15. https://doi.org/10.1016/j.inffus.2019.07.010.

Mallat, S. G. 1989. "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, no. 7: 674–693. https://doi.org/10.1109/34.192463.

Masi, G., D. Cozzolino, L. Verdoliva, and G. Scarpa. 2016. "Pansharpening by Convolutional Neural Networks." *Remote Sensing* 8, no. 7: 22. https://doi.org/10.3390/rs8070594.

Meng, X., N. Wang, F. Shao, and S. Li. 2022. "Vision Transformer for Pansharpening." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–11. https://doi.org/10.1109/TGRS.2022.3168465.

Nascimento, J. M. P., and J. M. B. Dias. 2005. "Vertex Component Analysis: A Fast Algorithm to Unmix Hyperspectral Data." *IEEE Transactions on Geoscience and Remote Sensing* 43, no. 4: 898–910. https://doi.org/10.1109/TGRS.2005.844293.

Nezhad, Z. H., A. Karami, R. Heylen, and P. Scheunders. 2016. "Fusion of Hyperspectral and Multispectral Images Using Spectral Unmixing and Sparse Coding." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9, no. 6: 2377–2389. https://doi.org/10.1109/JSTARS.2016.2528339.

Plaza, A., J. A. Benediktsson, J. W. Boardman, et al. 2009. "Recent Advances in Techniques for Hyperspectral Image Processing." *Remote Sensing of Environment* 113: S110–S122. https://doi.org/10.1016/j.rse.2007.07.028.

Prévost, C., K. Usevich, P. Comon, and D. Brie. 2020. "Hyperspectral Super-Resolution With Coupled Tucker Approximation: Recoverability and SVD-Based Algorithms." *IEEE Transactions on Signal Processing* 68: 931–946. https://doi.org/10.1109/TSP.2020.2965305.

Rahmani, S., M. Strait, D. Merkurjev, M. Moeller, and T. Wittman. 2010. "An Adaptive IHS Pan-Sharpening Method." *IEEE Geoscience and Remote Sensing Letters* 7, no. 4: 746–750. https://doi.org/10.1109/LGRS.2010.2046715.

Selva, M., B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti. 2015. "Hyper-Sharpening: A First Approach on SIM-GA Data." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8, no. 6: 3008–3024. https://doi.org/10.1109/JSTARS.2015.2440092.

Shah, V. P., N. H. Younan, and R. L. King. 2008. "An Efficient Pan-Sharpening Method via a Combined Adaptive PCA Approach and Contourlets." *IEEE Transactions on Geoscience and Remote Sensing* 46, no. 5: 1323–1335. https://doi.org/10.1109/TGRS.2008.916211.

Shettigara, V. K. 1992. "A Generalized Component Substitution Technique for Spatial Enhancement of Multispectral Images Using a Higher Resolution Data Set." *Photogrammetric Engineering and Remote Sensing* 58, no. 5: 561–567.

Simoes, M., J. Bioucas-Dias, L. B. Almeida, and J. Chanussot. 2015. "A Convex Formulation for Hyperspectral Image Superresolution via Subspace-Based Regularization." *IEEE Transactions on Geoscience and Remote Sensing* 53, no. 6: 3373–3388. https://doi.org/10.1109/TGRS.2014.2375320.

Wald, L. 2000. "Quality of High Resolution Synthesised Images: Is There a Simple Criterion?" *Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images*, 01-26 2000 Sophia Antipolis, France, 99–103.

Wald, L., T. Ranchin, and M. Mangolini. 1997. "Fusion of Satellite Images of Different Spatial Resolutions: Assessing the Quality of Resulting Images." *Photogrammetric Engineering and Remote Sensing* 63, no. 6: 691–699.

Wang, N., X. Meng, and F. Shao. 2022. "Convolution-Embedded Vision Transformer With Elastic Positional Encoding for Pansharpening." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–9. https://doi.org/10.1109/TGRS.2022.3227405.

Wei, Y., Q. Yuan, X. Meng, H. Shen, L. Zhang, and M. Ng. 2017. "Multi-Scale-and-Depth Convolutional Neural Network for Remote Sensed Imagery Pan-Sharpening." *IEEE International Geoscience and Remote Sensing Symposium*, 23–28 July 2017, 3413–3416. https://doi.org/10.1109/IGARSS.2017.8127731.

Xiao, J., J. Li, Q. Yuan, and L. Zhang. 2022. "A Dual-UNet With Multistage Details Injection for Hyperspectral Image Fusion." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–13. https://doi.org/10.1109/TGRS.2021.3101848.

Xu, Y., Z. Wu, J. Chanussot, P. Comon, and Z. Wei. 2020. "Nonlocal Coupled Tensor CP Decomposition for Hyperspectral and Multispectral Image Fusion." *IEEE Transactions on Geoscience and Remote Sensing* 58, no. 1: 348–362. https://doi.org/10.1109/TGRS.2019.2936486.

Yokoya, N., and A. Iwasaki. 2016. "Airborne Hyperspectral Data Over Chikusei."

Yokoya, N., T. Yairi, and A. Iwasaki. 2012. "Coupled Nonnegative Matrix Factorization Unmixing for Hyperspectral and Multispectral Data Fusion." *IEEE Transactions on Geoscience and Remote Sensing* 50, no. 2: 528–537. https://doi.org/10.1109/TGRS.2011.2161320.

Zhang, X., M. Chen, F. Liu, S. Li, J. Rao, and X. Song. 2025. "MSSSHANet: Hyperspectral and Multispectral Image Fusion Algorithm Based on Multi-Scale Spatial–Spectral Hybrid Attention Network." *Measurement Science and Technology* 36, no. 3: 035407. https://doi.org/10.1088/1361-6501/adb5af.

Zhang, X., W. Huang, Q. Wang, and X. Li. 2021. "SSR-NET: Spatial–Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion." *IEEE Transactions on Geoscience and Remote Sensing* 59, no. 7: 5953–5965. https://doi.org/10.1109/TGRS.2020.3018732.

Zhao, R., and S. Du. 2022. "Spectral-Spatial Residual Network for Fusing Hyperspectral and Panchromatic Remote Sensing Images." *Remote Sensing* 14, no. 3: 800. https://doi.org/10.3390/rs14030800.

Zhou, B., X. Zhang, X. Chen, M. Ren, and Z. Feng. 2023. "HyperRefiner: A Refined Hyperspectral Pansharpening Network Based on the Autoencoder and Self-Attention." *International Journal of Digital Earth* 16, no. 1: 3268–3294. https://doi.org/10.1080/17538947.2023.2246944.