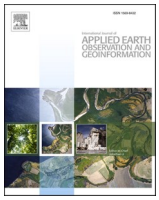
International Journal of Applied Earth Observation and Geoinformation 143 (2025) 104843

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



MHFu-former: A multispectral and hyperspectral image fusion transformer

Xue Wang [a,b,c,d], Songling Yin [a,b,c,d], Xiaojun Xu [e], Yong Mei [f], Yan Huang [g,h], Kun Tan [a,b,c,d,*]

[a] Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China
[b] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
[c] School of Geographic Sciences, East China Normal University, Shanghai 200241, China
[d] School of Geospatial Artificial Intelligence, East China Normal University, Shanghai 200241, China
[e] Shanghai Environmental Monitoring Center, Shanghai 200003, China
[f] Institute of Defense Engineering, AMS, Beijing 100036, China
[g] Geological Exploration Technology Institute of Jiangsu Province, Jiangsu 210000, China
[h] Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology, Jiangsu 210000, China

ARTICLE INFO

Keywords:
Hyperspectral image processing
Satellite remote sensing image
Image fusion
Transformer

ABSTRACT

Hyperspectral images (HSIs) can capture detailed spectral features for object recognition, while multispectral images (MSIs) can provide a high spatial resolution for accurate object location. Deep learning methods have been widely applied in the fusion of hyperspectral and multispectral images, but still face challenges, including the limited capacity to enhance spatial details and preserve spectral information, as well as issues related to spatial scale dependency. In this paper, to solve the above problems and achieve more effective information integration between HSIs and MSIs, we propose a novel multispectral and hyperspectral image fusion transformer (MHFu-former). The proposed MHFu-former consists of two main components: (1) a feature extraction and fusion module, which first extracts deep multi-scale features from the hyperspectral and multispectral imagery and fuses them to form a joint feature map, which is then processed by a dual-branch structure consisting of a Swin transformer module and convolutional module to capture the global context and fine-grained spatial features, respectively; and (2) a spatial-spectral fusion attention mechanism, which adaptively enhances the important spectral information and fuses it with the spatial detail information, significantly boosting the model's sensitivity to the key spectral features while preserving rich spatial details. We conducted comparative experiments on the indoor Cave dataset and the Shanghai and Ganzhou datasets from the ZY1-02D satellite to validate the effectiveness and superiority of the proposed method. Compared to the state-of-the-art methods, the proposed method significantly enhances the fusion performance across multiple key metrics, demonstrating its outstanding ability to process spatial and spectral details.

1. Introduction

Hyperspectral and multispectral imagery are critical in remote sensing, with each offering distinct advantages for various applications. Hyperspectral imaging technology enables the capture of hundreds of continuous and narrow spectral bands, providing rich spectral information that is invaluable for tasks such as detection, classification, and tracking. However, hyperspectral images (HSIs) typically suffer from a low spatial resolution due to the physical constraint of the sensor, limiting their ability to accurately represent fine spatial details. In contrast, multispectral images (MSIs) offer a high spatial resolution but a lower spectral resolution, typically capturing only a few broad spectral bands. As a result, there is growing interest in fusing hyperspectral and multispectral data to combine their respective strengths, resulting in high-resolution HSIs that can provide both detailed spectral and spatial information. This fusion addresses the inherent limitations of the individual sensors, as the current technology cannot simultaneously achieve a high spatial and spectral resolution from a single imaging platform due to constraints such as the sensor capacity, signal-to-noise ratio (SNR), and data transmission limits. The development of effective multispectral

* Corresponding author at: Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China.
E-mail address: tankuncu@gmail.com (K. Tan).

https://doi.org/10.1016/j.jag.2025.104843
Received 26 May 2025; Received in revised form 11 August 2025; Accepted 9 September 2025

and hyperspectral image fusion (MHIF) techniques is therefore crucial to meet the increasing demand for more accurate and detailed remote sensing data in a wide range of applications, such as object detection (Qin et al., 2024), anomaly detection (Wang et al. (2023a), Wang et al. (2023b)), change detection (Yang et al., 2024), and scene classification (Dong et al., 2024).

Generally speaking, MHIF methods are typically either traditional methods or deep learning based methods. The traditional methods optimize the objective function to restore the fused image, while the deep learning based approaches employ a labeled sample learning mechanism for image fusion. MHIF is a specific instance of panchromatic sharpening, and many different methods have been investigated (Tian et al., 2021; Zhu & Bamler, 2012). Selva et al. (2014) employed multivariate linear regression to synthesize HSIs from high-resolution MSIs, which is an approach that has less computational burden, but results in significant spectral distortion when there is a large discrepancy in the spatial resolution. Dong et al. (2021) investigated a more effective component substitution fusion model based on a binary partition tree and image segmentation. However, the above methods are faced with the limitations of linear models and deficiencies in band processing, while also being constrained by the challenge of balancing computational efficiency and fusion quality.

In degradation-based methods for MHIF, the MSIs are treated as a combination of spectral degradation and noise, based on the fused image, while the HSIs are considered as a combination of spatial degradation and noise. Methods based on a low-rank prior consider that the spectral features can be represented in a low-dimensional subspace to learn low-rank spectra from low-resolution images. Both sparse and low-rank representation methods can effectively preserve the spectral characteristics and significantly reduce the redundancy inherent in spectral information. Yokoya et al. (2011) proposed the coupled nonnegative matrix factorization (CNMF) method, which alternately unmixes hyperspectral and multispectral data while incorporating a sensor observation model to generate fused data with a high spatial and spectral resolution. Pansharpening methods are computationally simple but prone to spectral distortion, while matrix factorization and tensor representation methods yield promising results but are computationally intensive and depend on accurate sensor model estimation and prior knowledge.

Recently, due to the powerful feature extraction and representation capabilities of deep learning, it has been extensively applied in the field of remote sensing image fusion. Zhang et al. (2020) proposed an interpretable spatial-spectral reconstruction network (SSR-NET) based on a CNN, integrating cross-mode information insertion, a spatial reconstruction network, and a spectral reconstruction network to enhance the spatial and spectral information recovery under spatial edge loss constraints. Despite its strong performance, the proposed MHFu-former still has limitations. First, the model employs fixed window and slice sizes, which may not be universally optimal across diverse scene types or sensor characteristics. The static configuration could limit generalization performance in highly heterogeneous environments. Future work could explore adaptive or dynamic windowing strategies that adjust to local image complexity, potentially enhancing robustness. Second, the current pipeline relies on pre-processing steps such as spectral alignment and interpolation. These steps, while standard, can introduce subtle artifacts and often require prior knowledge of sensor specifications. A key future direction is the development of end-to-end fusion mechanisms that can learn to align and integrate data directly, thereby improving both autonomy and generalization.

Furthermore, transformer frameworks have also been explored and applied in the fusion of multispectral and hyperspectral imagery, because of their big breakthrough in the field of computer vision. Hu et al. (Hu et al., 2022) proposed a transformer-based architecture (Fusformer), which leverages self-attention to capture the global feature relationships and estimates the spatial residuals to enhance the high-resolution HSI reconstruction while reducing the training complexity. The HyperTransformer model (Bandara & Patel, 2022) addresses the fusion of panchromatic and hyperspectral imagery, and is made up of two independent feature extraction modules: a multi-head feature soft attention module and a spatial-spectral feature fusion module. Deng et al. (Deng, Wu, Ran, & Wen, 2023) introduced the Bidirectional Dilation Transformer (BDT), which integrates dilation spatial self-attention with grouped spectral self-attention to effectively capture multiscale spatial-spectral characteristics. Jia et al. (2023) proposed a Multiscale Spatial-Spectral Transformer Network (MSST-Net) embedding multiscale attention mechanisms to enhance joint feature representation of MSI and HSI. Ma et al. (2024) proposed a dual cross-attention-based reciprocal transformer architecture, enabling bidirectional interaction of spatial and spectral features between HSI and MSI modalities. Wang et al. (2023a), Wang et al. (2023b) developed a Retractable Spatial-Spectral Transformer Network (RSST) with an attention retractable mechanism and a gradient spatial–spectral recovery block to address token interaction limitations and enhance edge detail preservation.

In addition, scholars have explored the integration of model-driven approaches with deep learning to establish interpretable deep fusion networks, thereby enhancing the network's interpretability. Xie et al., (2020) proposed an interpretable multispectral/hyperspectral fusion network, MHF-net, which integrates linear mapping and low-rank priors, while using a proximal gradient algorithm for efficient fusion and robust performance across different sensors. Wang et al. (X. Wang, Borsoi, Richard, & Chen, 2023) combined a lightweight CNN with an iterative optimization method to establish a general imaging model with a super-Laplacian distribution, thereby improving the image fusion accuracy. More recently, Yan et al. (2025) introduced a spatial–spectral unfolding network that embeds an optimization algorithm directly into its architecture. This hybrid approach leverages both prior knowledge and the representation power of deep learning, further advancing the performance and interpretability of hyperspectral–multispectral fusion models.

Numerous data fusion methods for multi-source remote sensing have been developed, with many mature applications and high-performing algorithms in the field of MHIF. However, several limitations and challenges remain to be addressed:

1) Compared to the traditional fusion methods, deep learning based remote sensing image fusion algorithms have advantages in learning spatial-spectral features and obtaining a higher fusion accuracy. Nevertheless, their capabilities in enhancing spatial details and preserving spectral information require further improvement, particularly in complex remote sensing scenes, resulting in unsatisfactory performances in real fusion applications.
2) In MHIF, besides enhancing the spatial features and maintaining spectral consistency, the current deep learning based methods predominantly rely on small-scale public datasets, which are very different from satellite-based image scenarios. Furthermore, many deep learning fusion methods are trained based on the Wald protocol, which causes significant scale dependency. Therefore, more effective feature extraction strategies are vitally needed to capture the spatial-spectral mapping relationships between multispectral and hyperspectral images, to mitigate this dependency.
3) Despite the demonstrated efficacy in hyperspectral and multispectral image fusion, standard Transformer-based architectures are constrained by inherent limitations. The primary challenge stems from the global self-attention mechanism, which imposes a prohibitive computational burden due to its quadratic complexity with respect to input image size. Moreover, standard transformers often lacks a strong inductive bias for local spatial details and struggles to capture the fine-grained textures essential for high-fidelity fusion. These constraints on scalability and spatial precision hinder the practical deployment in large-scale remote sensing scenarios.

To address the current issues of insufficient spatial context details and poor spectral characteristics consistency in image fusion, we propose a novel multispectral and hyperspectral image fusion transformer (MHFu-former). MHFu-former integrates a Swin transformer and a convolutional module within a dual-branch architecture to extract multi-scale spatial-spectral features, thereby effectively handling the spectral disparities while capturing the global contextual correlations and fine-grained spatial details. In addition, we introduce a spatial-spectral fusion attention mechanism that dynamically prioritizes the key spectral bands and integrates multi-level spatial information through global–local dependency modeling, thereby minimizing the spectral distortion and ensuring spectral continuity. The main contributions of the MHFu-former method are summarized as follows:

1) The dual-branch hybrid architecture integrating a Swin transformer and depth-separable convolution enables parallel extraction of the global spectral correlations and fine-grained spatial features, effectively addressing the spectral disparities while preserving the spatial details. The Swin transformer, with its hierarchical structure and shifted window-based self-attention, enables efficient modeling of both long-range dependencies and local context. Compared to standard Transformers, it significantly reduces computational complexity while maintaining high-resolution feature continuity and achieves linear computational complexity. This capability to model long-range dependencies efficiently while maintaining precise local features makes it a uniquely powerful architecture for remote sensing image fusion.
2) The dynamic spatial-spectral fusion attention mechanism adaptively prioritizes the critical spectral bands and hierarchically fuses the multi-level spatial features through global–local dependency modeling, thereby minimizing the spectral distortion and ensuring spectral profile continuity.
3) The end-to-end cascaded refinement framework achieves high-resolution HSI reconstruction via interpretable spatial-spectral decoupling, embedding the cross-modal interactions to bridge the MSI-HSI domain gaps while maintaining radiometric consistency.

2. Previous work

2.1. Convolutional neural networks (CNNS)

CNNs are a type of feed-forward neural network with local connectivity and weight sharing characteristics. The architectural elements of CNNs, namely convolutional layers, pooling layers, and fully connected layers, endow them with a degree of invariance to translation, scaling, and rotation.

In the case of two-dimensional convolution, the mathematical representation can be given as:

$$y_{ij} = \sum_{u=1}^{U} \sum_{v=1}^{V} w_{uv} x_{i-u+1, j-v+1} \tag{1}$$

where U and V represent the size of the convolution kernel in two adjacent layers. i and j denote the position subscript of the output matrix. In deep learning, the input form is generally a three-dimensional tensor X, and the output feature matrix is denoted as Y^p. The convolution process can be represented as:

$$Y^p = f(w^p \otimes X + b^p) \tag{2}$$

where b^p represents the bias, and $f(\cdot)$ represents the activation function of a neuron. The common activation functions include the rectified linear unit (ReLU), sigmoid, and tanh activation functions. The fully connected layer, combined with the activation function, constitutes a continuous matrix multiplication and nonlinear information transformation process:

$$x^l = f\left(w^l x^{l-1} + b^l\right) \tag{3}$$

where l represents the lth layer, w^l and b^l respectively represent the weight matrix and bias from the $(l-1)$th layer to the lth layer, f represents the nonlinear activation function, and x^l is the output of the current layer. With the classification task, the fully connected layer is usually placed at the last layer in a CNN, while it is rarely used as the last layer before the output in image fusion.

2.2. Attention mechanisms

An attention mechanism can effectively extract and represent more important features to tackle the problem of limited computational resources, and is an effective way to address information overload. Attention mechanisms can be categorized into soft attention (Lu, Wang, Zheng, & Li, 2017), hard attention (Feng et al., 2020), and self-attention (Long et al., 2023) mechanisms.

A. Soft and hard attention mechanisms.

Let $\boldsymbol{X} = [x_1, \cdots, x_N]$ represent N sets of inputs, where x represents a set of input information with D dimensional features. To select information from $\boldsymbol{X}$ that is relevant to the target, the attention distribution from all the input information is first computed, followed by the weighted average, based on the attention distribution.

By introducing a query vector q, z denotes the index position of the corresponding information. The probability of selecting the i-th input vector based on q and $\boldsymbol{X}$ is denoted as α_n.

$$\alpha_n = p(z = n|X, q) \tag{4}$$

$$= softmax(s(x_n, q)$$

$$= \frac{exp(s(x_n, q))}{\sum_{j=1}^{N} exp(s(x_n, q))}$$

where α_n is the attention distribution, and $s(x, q)$ is the attention score function, which can be calculated in various ways, including through additive models, dot-product models, scaled dot-product models, and bilinear models. A soft attention mechanism is computed as follows:

$$attn(X, q) = \sum_{N=1}^{n} \alpha_n x_n \tag{5}$$

Unlike the above-mentioned soft attention mechanisms, which consider information from all the input vectors, hard attention mechanisms focus on a single input feature, which can be implemented through the maximum value or stochastic sampling based on the attention distribution. The sampling strategy in hard attention causes a non-differentiable relationship between the loss function and the attention distribution, which precludes the backpropagation for training. A common implementation of hard attention is conducted by selecting the maximum value:

$$attn(X, q) = X_n \tag{6}$$

$$\widehat{n} = argmax(\alpha_n), n = 1, 2 \cdots N \tag{7}$$

B. Self-attention mechanisms.

The encoding–decoding scheme in a typical CNN only explores the local dependencies within the input information. To capture the long-distance dependencies across sequenced inputs such as spectral features, researchers generally enhance the depth of the network layers or employ fully connected layers. However, fully connected layers are not well suited for handling sequences of varying lengths. To tackle this issue, self-attention mechanisms assign weights to the different positions in the feature sequence and employ dynamic connections between features. The self-attention model operates on a query-key-value (QKV) framework, as illustrated in Fig. 1.

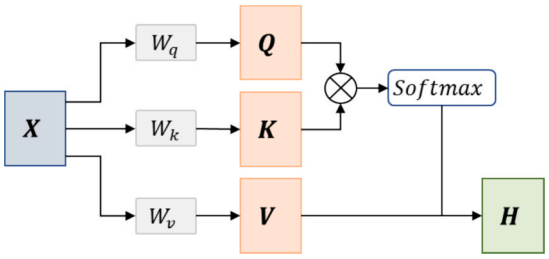


Fig. 1. The self-attention model framework.

Assume that the input is $\boldsymbol{X} = [x_1, \cdots, x_N]$ and the output is $H = [h_1, h_2, \ldots, h_N]$. For each input sequence $\boldsymbol{X}$, linear mapping is applied to transform it into three different feature spaces, yielding the vectors Q, K, and V:

$$Q = W_q X \tag{8}$$

$$K = W_k X \tag{9}$$

$$V = W_v X \tag{10}$$

If the key-value pair attention formula is employed, the output vector h_n is computed as follows:

$$h_n = attn((K, V), q_n)$$

$$= \sum_{j=1}^{N} \alpha_{nj} v_j \tag{11}$$

$$= \sum_{j=1}^{N} softmax(s(k_j, q_n)) v_j$$

If scaled dot-product attention is used for the weight calculation, the output vector H is as shown in the following equation:

$$H = softmax\left(\frac{K^T Q}{\sqrt{D_k}}\right) V \tag{12}$$

where D represents the dimensionality of the input vectors

3. Proposed method

3.1. Multispectral and hyperspectral image fusion formulation

In this work, we utilize lowercase letters to signify scalars, bold letters to represent matrices, and calligraphic letters to denote tensors. Specifically, the HSI with a low spatial resolution is denoted as $\mathbf{X} \in \mathbb{R}^{m \times n \times S}$, and the MSI with a high spatial resolution is denoted as $\mathbf{Y} \in \mathbb{R}^{M \times N \times s}$. The super-resolution (SR) factor is $l = \frac{M}{m} = \frac{N}{n}$, which is 3 for satellite-based remote sensing images. The aim is to estimate a fused image, represented by $\widehat{X} \in \mathbb{R}^{W \times H \times S}$, that encompasses both a high spatial resolution and high spectral resolution. The primary observation of MHIF can be denoted as:

$$\underset{w,b}{\text{argmax}} \| HMformer_{w,b}(\mathrm{X}, \mathrm{Y}) - \widehat{X} \|_1 \tag{13}$$

where $HMformer_{w,b}(.)$ denotes the feature learning of the proposed network with parameters w and b.

3.2. Mhfu-former

The MHFu-former network architecture is depicted in Fig. 2. Firstly, to ensure that the spatial dimensions of the hyperspectral and multispectral images are consistent, an interpolation operation is performed on X to generate $\mathrm{X}^0 \in \mathbb{R}^{M \times N \times S}$. The MHFu-former network employs a dual-branch structure with two kinds of input (multispectral and hyperspectral images) to acquire high spatial-spectral resolution fused imagery through a specified feature fusion strategy. Due to the difference in spectral dimensions between multispectral and hyperspectral images, the input image pairs, X^0 and Y, are first mapped into the feature space using 1×1 convolutional layers, and then processed through the spectral alignment layer. The formula for this is as follows:

$$X_c = Conv\_layer_{HS}(\mathrm{X}^0), Y_c = Conv\_layer_{MS}(Y) \tag{14}$$

where $Conv\_layer_{HS}$ and $Conv\_layer_{MS}$ represent the convolution operations applied to the hyperspectral and multispectral images, respectively. After the spectral alignment, the two features are stacked to generate $\boldsymbol{F}$.

$$\boldsymbol{F} = [X_c; Y_c] \tag{15}$$

After this, the feature $\boldsymbol{F}$ is fed into the Swin transformer module and depthwise convolutional layers. The Swin transformer module is

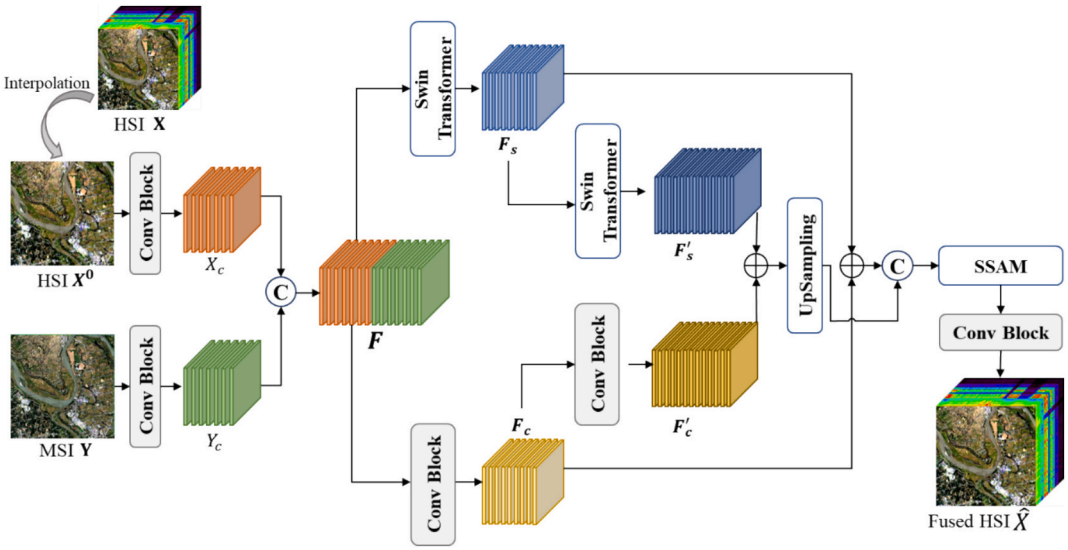


Fig. 2. The MHFu-former network framework.

designed to capture both local and global image features effectively. In this module, the feature map slice size is set to 3×3, based on the image features, and a downsampling rate of 3 is applied to reduce the spatial resolution progressively while retaining the essential spatial and spectral information. The Swin transformer leverages a hierarchical design that operates on progressively smaller patches, enabling efficient processing of large-scale images by incorporating self-attention mechanisms that help capture the long-range dependencies. After two successive Swin transformer operations, two feature maps, $\boldsymbol{F}_s$ and $\boldsymbol{F}_s'$, are generated, each containing increasingly abstract representations of the input image, further refining the feature extraction process.

At the same time, the feature $\boldsymbol{F}$ is passed through the depthwise convolution block to generate feature maps $\boldsymbol{F}_c$ and $\boldsymbol{F}_c'$ at two different scales. The finer-scale feature map $\boldsymbol{F}_c'$ is then fused with the feature map $\boldsymbol{F}_s'$ from the second stage of the Swin transformer module, and the resulting fused feature map undergoes upsampling to recover the spatial resolution. This process ensures that both the fine-grained details and broader contextual information are effectively integrated, allowing the model to leverage multi-scale features for a better performance in tasks such as image reconstruction or segmentation.

$$\boldsymbol{F}' = \left[Upsample\left(\boldsymbol{F}_c' + \boldsymbol{F}_s'\right); \left(\boldsymbol{F}_c + \boldsymbol{F}_s\right)\right] \tag{16}$$

where $[\cdot;\cdot]$ denotes the concatenation operation. *Upsample* is the upsampling operation.

The feature $\boldsymbol{F}'$ is then fed into the spectral attention mechanism module for adaptive feature fusion. Finally, a convolutional layer with a 1×1 kernel size is used to adjust the number of channels and yield the fused imagery with a high spatial-spectral resolution.

3.3. Swin transformer based backbone

In order to achieve more accurate pixel-level prediction and reduce the computational cost of the vision transformer framework, the Swin transformer is introduced as a backbone for the network model. The hierarchical feature mappings are constructed with a linear computational complexity related to the image size. The main difference between the Swin transformer and the original vision transformer is that the self-attention mechanism in the Swin transformer employs shifted window partitioning for computation. The Swin transformer module is depicted in Fig. 3.

After passing the input through a convolutional layer with a kernel size with a fixed stride equal to the window size, the three-dimensional feature map is flattened and linearly transformed into a two-dimensional vector. The embedding and linear projection parts of the module are consistent with the vision transformer architecture. The vector sequence is then input into two continuous Swin transformer modules, with the difference being that the former computes window self-attention scores and the latter computes shifted window self-attention scores.

The window multi-head self-attention (W-MSA) mechanism and the shifted window multi-head self-attention (SW-MSA) mechanism are variants of the multi-head self-attention mechanism. The output feature vector sequence is reshaped to the original size through merging of all the image blocks based on the window size in the W-MSA mechanism, and the self-attention scores for each window are independently calculated, which can significantly reduce the computational cost. However, the feature information between the windows cannot be conveyed, resulting in a smaller receptive field. The input feature $\boldsymbol{F}$ is mapped to queries Q, keys K, and values V through different weight matrices W_Q, W_K, and W_V. In addition, a relative positional encoding B is included. The attention for each head i is computed as follows:

$$\mathrm{head}_i = Attn(Q_i, K_i, V_i) = softmax\left(\frac{Q_i^T Q_i}{\sqrt{D_k}} + B\right) V_i \tag{17}$$

In the multi-head self-attention mechanism, we do not compute a single attention matrix but instead calculate multiple "heads" (i.e., multiple independent attention mechanisms). These heads are then concatenated together and passed through a linear transformation, using the weight matrix W_o to obtain the final multi-head self-attention output:

$$\mathrm{MultiHead}(Q, K, V) = Concat(\mathrm{head}_1, \mathrm{head}_2, \cdots, \mathrm{head}_h)\mathrm{W}_o \tag{18}$$

To facilitate the interaction between different windows, the SW-MSA mechanism is designed to compute new self-attention scores with shifted windows. The implementation achieves the shifted calculation by shifting the feature map itself, as illustrated in Fig. 4. According to the shifting stride, the corresponding number of rows from the top of the feature map are shifted downwards, followed by shifting the corresponding number of columns from the left side to the right. This shifting process effectively introduces connections between the different windows, enabling cross-window information exchange and expanding the receptive field of the network. The window shifting operation can be represented as:

$$F_{wmsa}' = \mathscr{S}(F_{wmsa}) \tag{19}$$

where $\mathscr{S}$ denotes the window shifting operation. F_{wmsa} is the feature obtained through the W-MSA mechanism. Finally, the calculation for the SW-MSA mechanism can be denoted as:

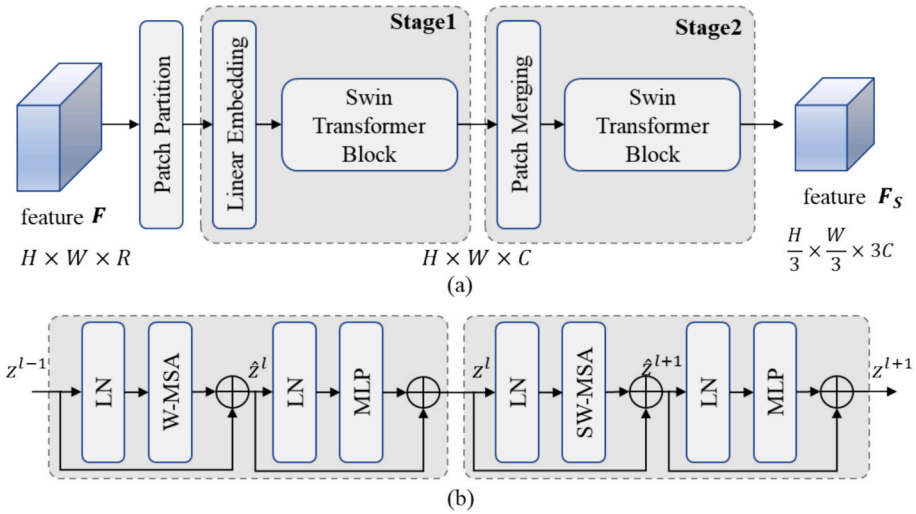


Fig. 3. The Swin transformer module framework. (a) Swin transformer architecture. (b) Two successive Swin transformer blocks.

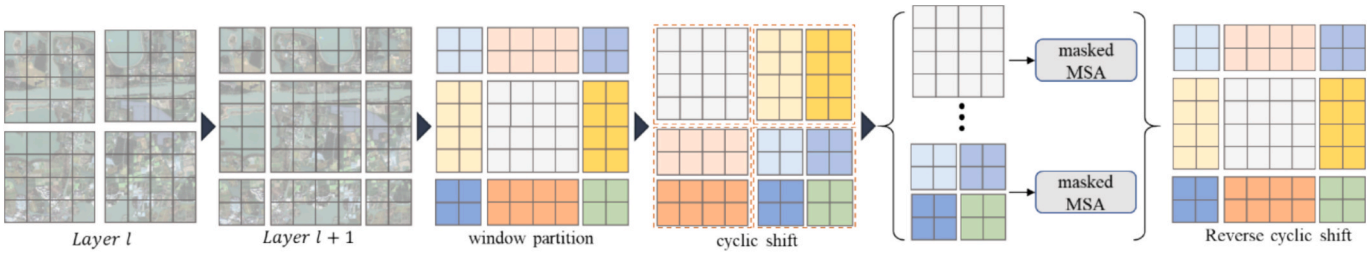


Fig. 4. An illustration of the shifted window approach for computing self-attention in the Swin transformer architecture.

$$F_{swmsa} = \mathscr{S}^{-1}\left(WMSA\left(F'_{wmsa}\right)\right) \tag{20}$$

where $\mathscr{S}^{-1}$ denotes the inverse of the window shifting operation, and *WMSA* represents the window-based multi-head self-attention mechanism.

Moreover, to ensure that the W-MSA and SW-MSA mechanisms have the same number of computational windows during the calculation, a mask matrix is used to compute the self-attention scores within each window. Specifically, the role of the mask matrix is to ensure that the self-attention calculation for each window only focuses on the positions within the current window and does not span the window boundaries. Therefore, the values in the mask matrix are set to -100, so that, during the softmax calculation, the attention scores for positions outside the current window become very small, i.e., close to zero, effectively "masking" these positions. This mechanism ensures that the attention computation is confined to the current window, avoiding interference from external elements, especially during the window shifting operation.

The feature map from the preceding module is first resized to its original dimensions before being flattened through image patch partitioning. Using a downsampling factor of 3, the feature map is divided into 3×3 patches, where pixels at the same location within each patch are grouped to form a new feature map, concatenated along the channel dimension. Finally, a fully connected layer is applied to downsample the feature map. The process of patch merging is illustrated in Fig. 5, where, for instance, a feature map of size $6 \times 6 \times 1$ is transformed into a $2 \times 2 \times 9$ feature map, and then a linear layer further reduces it to a $2 \times 2 \times 3$ feature map, showing that the reduction factor in the spatial dimensions corresponds to the increase factor in the number of channels. This helps to effectively capture both the local and global information between different bands when processing large-scale hyperspectral and multispectral images, enhancing the model's ability to perceive details and variations in the fused image. In addition, through downsampling, the computational cost is reduced, the processing speed is improved, and the sensitivity to important features is maintained.

3.4. Spatial-spectral fusion attention

Due to the significant spatial distribution differences between various land-cover types, especially in multispectral and hyperspectral images, the edges of these land-cover types are often quite blurred, and the traditional CNNs may not effectively capture these details, leading to the loss or excessive smoothing of edge information. Moreover, since each spectral band contains different feature information in the spatial dimension, there is often a lack of sufficient mechanisms to explore the potential correlations between the spectral information. To address these issues, we introduce a spatial-spectral attention mechanism that combines spatial and spectral attention to improve the quality of the feature representation.

Fig. 6 illustrates the spatial-spectral attention mechanism. The spectral attention mechanism operates along the channel dimension, applying max pooling and average pooling operations to process the feature map along the spatial dimension. The pooled results are then processed through a fully connected layer, and after summing, a sigmoid activation function is applied to generate the channel attention map. This process allows the model to focus on the most informative spectral features. The expression for the spectral attention mechanism is as follows:

$$M_C = \sigma(W_C \bullet [MaxPool(F); MinPool(F)]) \tag{21}$$

where W_C is the weight matrix of the fully connected layer, and σ is the sigmoid activation function. M_C is the spectral attention map, and *MaxPool*() and *MinPool*() refer to the max pooling and average pooling operations, respectively.

In contrast, the spatial attention mechanism operates along the channel dimension using max pooling and average pooling operations. These results produce two feature maps, which are then convolved with a 7×7 convolution kernel. The convolutional output is then passed through a sigmoid activation function to generate the spatial attention map. The expression for the spatial attention mechanism is as follows:

$$M_S = \sigma(Comv_{7\times7} \bullet [MaxPool_C(F); MinPool_C(F)]) \tag{22}$$

where $MaxPool_C()$ and $MinPool_C()$ respectively refer to the max pooling and average pooling operations performed along the channel dimension (denoted by c) of the feature map. M_S is the spatial attention map.

Therefore, by fusing the spectral and spatial attention maps, the final weighted feature map is obtained. The fused feature map can be denoted as F_{SSA}, so that:

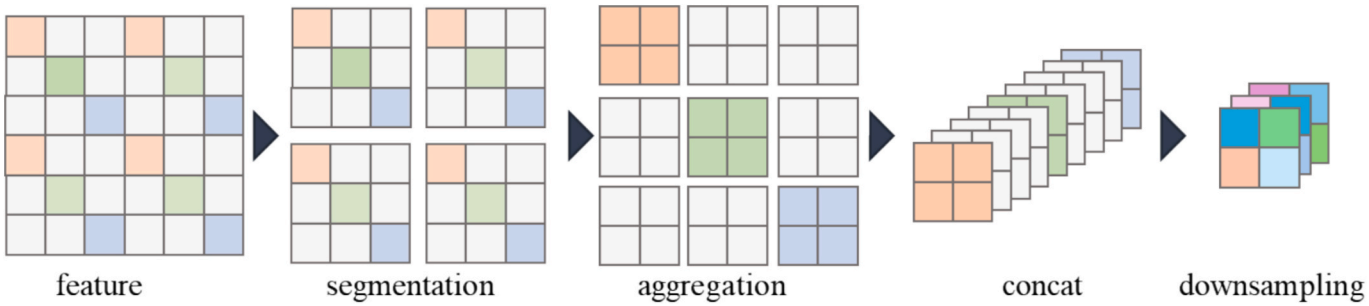


Fig. 5. An illustration of the shifted window approach for computing self-attention in the Swin transformer architecture.

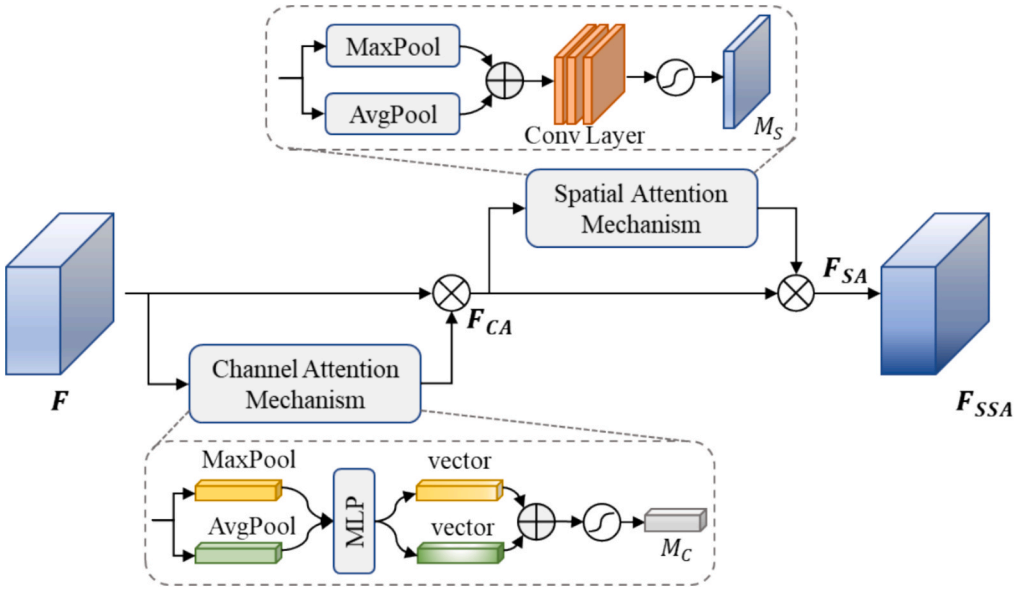


Fig. 6. Structural diagram of the spatial-spectral attention mechanism.

$$F_{SSA} = F \otimes M_C \otimes M_S \tag{23}$$

By fusing the spectral and spatial attention maps, the spatial-spectral fusion attention mechanism further enhances the efficiency of the feature fusion process. The final fused remote sensing imagery is of higher quality because the attention mechanism effectively selects the most relevant feature information and adjusts the contribution of the different channels and spatial positions to the fused image. This improves the accuracy of the spatial and spectral feature extraction, making the model more precise in image fusion tasks.

4. Experiments

4.1. Experimental datasets

The data used for the MHIF experiments in this study were the Cave dataset and satellite-based imagery captured by the Advanced Hyperspectral Imager (AHSI) and Visible Near-Infrared Camera (VNIC) sensors on the ZY-1 02D satellite.

4.1.1. Cave dataset

The Cave dataset (Wang et al. 2019), created by researchers at Harvard University in 2011, is an indoor dataset comprising 32 images that include various materials, foods, paintings etc. Each image has a dimension of 512 × 512 pixels, with a wavelength range of 400–700 nm. The spectral resolution is 10 nm, resulting in 31 spectral bands. The camera model used for this dataset was an Apogee ALTA U260 charge-coupled device (CCD). The original HSIs were used as reference images to obtain the training set. Gaussian noise was added to these original images, followed by downsampling to produce low-resolution HSIs, with the downsampling factor set to 3. Finally, a total of 1300 pairs of image patches were randomly generated as training data with a size of 72 × 72 and 24 × 24. The test images were sized at 480 × 480 × 3 and 160 × 160 × 31, as shown in Fig. 7.

4.1.2. ZY-1 02D satellite dataset

The AHSI hyperspectral camera on the ZY-1 02D satellite has 166 spectral bands, with a wavelength range from 400 to 2500 nm. The resolution is 30 m, with a sensor scanning width of 60 km. The spectral resolution is 10 nm and 20 nm for the visible–near-infrared (VNIR) bands and short-wave infrared (SWIR) bands, respectively. Moreover, the VNIC multispectral camera on the ZY-1 02D satellite has an 8 spectral band, with a wavelength range from 450 to 1047 nm and a spatial resolution of 10 m. In these experiments, we utilized two scenes of data for the training and testing, as shown in Fig. 8. The specific attributes of the ZY-1 02D images are detailed in Table 1.

As illustrated in Fig. 8, the two benchmark datasets contain representative land-cover categories, including bare soil, vegetation, water bodies, and urban structures, which constitute characteristic ground features for HSI fusion studies. The image processing workflow was systematically executed through the following steps. Initial radiometric calibration converted the digital number (DN) values of the multispectral and hyperspectral images into physical radiance units. Subsequent orthorectification eliminated the geometric distortion through rigorous

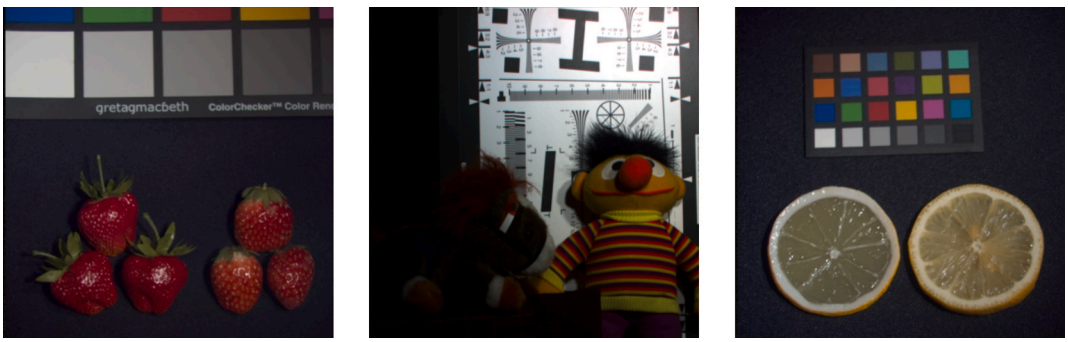
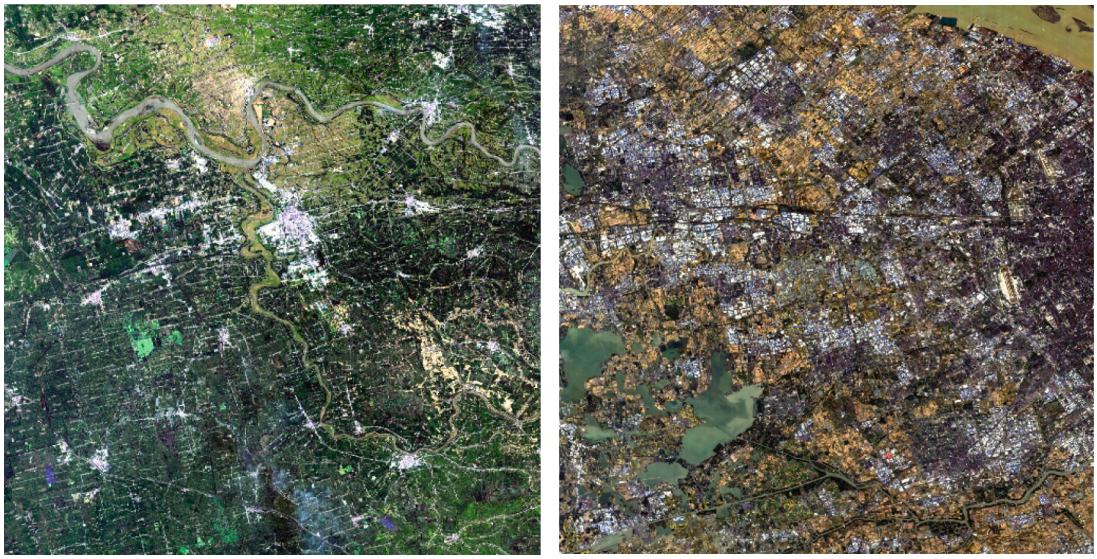


Fig. 7. Examples from the Cave dataset (R-29, G-19, B-9).

Image name	Scene 1	Scene 2
Imaging start time	2020-08-13 11:24:36	2021-11-12 10:55:38
Imaging end time	2020-08-13 11:24:47	2021-11-12 10:55:49
Image size	2000 × 2056	2000 × 2051
Solar azimuth angle	131.237731	166.339721
Solar zenith angle	22.355253	50.074939
Satellite azimuth angle	101.652603	101.4504
Satellite zenith angle	5.6847	6.4723



(a) Scene 1 (b) Scene 2

Fig. 8. Illustration of the AHSI dataset (R-29, G-19, B-9).

Table 1
Attributes of the ZY-1 02D images.

Image name	Scene 1	Scene 2
Imaging start time	2020–08-13 11:24:36	2021–11-12 10:55:38
Imaging end time	2020–08-13 11:24:47	2021–11-12 10:55:49
Image size	2000 × 2056	2000 × 2051
Solar azimuth angle	131.237731	166.339721
Solar zenith angle	22.355253	50.074939
Satellite azimuth angle	101.652603	101.4504
Satellite zenith angle	5.6847	6.4723

geometric correction, achieving standardized spatial resolutions of 30 m for the hyperspectral imagery and 10 m for the multispectral imagery. To ensure optimal data fusion compatibility, precise geometric registration was implemented to rectify the spatial misalignments between the two image modalities. After implementing spectral quality control through removal of noise-corrupted bands and atmospheric absorption bands, the remaining 166 spectral channels were retained for the image fusion processing. For the spatial resolution simulation, a degradation protocol was applied using Gaussian noise injection (5-pixel radius kernel) to generate downsampled image pairs at 90-m (hyperspectral) and 30-m (multispectral) resolutions, with the native 30-m HSIs serving as reference targets for the neural network training.

The dataset construction employed image blocks of 72 × 72 pixels (hyperspectral) and 24 × 24 pixels (multispectral), yielding 725 training blocks from Scene 1 and 730 from Scene 2, which were partitioned into training and validation sets at a strict 9:1 ratio. The final quantitative and qualitative evaluation were performed on full original scenes to assess performance.The test protocols differed between scenes. The Scene 1 evaluation focused on homogeneous land-cover regions (bare

soil, agricultural fields, vegetation stands, and urban structures), while the Scene 2 evaluation emphasized complex spatial configurations with enhanced edge detail preservation analysis. For the simulated-resolution testing, the hyperspectral data cubes were standardized to 160 × 160 × 156, while the multispectral counterparts were maintained at 480 × 480 × 8. In the real-resolution validation scenarios, the dimensional configuration was scaled proportionally to 480 × 480 × 156 (hyperspectral) and 1440 × 1440 × 8 (multispectral), preserving the 3:1 spatial resolution ratio between the multispectral and hyperspectral systems throughout the experimental protocols. The complete dataset will be made publicly available.

4.2. Comparison algorithms

To validate the performance of the proposed deep learning fusion method using multispectral and hyperspectral imagery, five comparison algorithms were selected, including the non-deep learning method of coupled nonnegative matrix factorization (CNMF) (Yokoya et al., 2011) and the deep learning based methods of the spatial-spectral reconstruction network (SSR-NET) (Zhang et al., 2020), a 3-D-convolutional neural network (3DCNN) (Palsson et al., 2017), the two-stream fusion network (TFNet) (Liu, Liu, & Wang, 2020), Fusformer (Hu et al., 2022), MHFNet (Xie et al., 2020), and MSST-Net (Jia et al., 2023).

All the deep learning models employed in the experiments were implemented using the PyTorch deep learning framework. For the pursuit of fair model comparison, no pre-trained models or data augmentation techniques were utilized during the training phase. The Adam optimizer was employed with an initial learning rate of 0.0002. The maximum number of training iterations was set to 150, and the batch size was fixed at 8. The training and validation sets were partitioned in a 9:1 ratio. The mean absolute error (L1 loss) function was selected as the loss function.

$$l_1(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{24}$$

To mitigate overfitting during training and enhance the training efficiency, early stopping was employed after 50 iterations. The training process was terminated when the output loss on the validation set remained unchanged for five consecutive iterations. The model parameters corresponding to the minimum validation loss were utilized for the testing and final fusion. A sliding window approach with overlap was employed during the image fusion at both the simulated and real resolutions. The test images were cropped into overlapping patches of the same size as the training set, with an overlap of 3 pixels. During the final image mosaicking, the overlapping regions between patches were averaged.

4.3. Metrics for evaluating image fusion quality

Remote sensing image fusion performance evaluation can be categorized into two aspects: simulated-resolution evaluation and real-resolution evaluation. In the simulated experiments, following the Wald protocol (Wald, 2000), the high-resolution MSIs acquired by satellites were downsampled proportionally to the low-resolution HSIs, based on their spatial resolution ratio. The fusion algorithms were then applied to the downsampled image pairs, with the original HSI serving as the target reference for the quantitative accuracy assessment of the fused image. For the real-resolution fusion experiments, due to the absence of a true reference image, no-reference image quality assessment metrics were employed.

For the quantitative evaluation metrics for the simulated experiments, we selected the spectral angle mapper (SAM), the structural similarity index (SSIM) (Tian et al., 2021), the peak signal-to-noise ratio (PSNR) (Z. Wang et al., 2004), the relative global dimensional synthesis error (ERGAS) (Renza et al., 2012), and the universal image quality index (Q) (Z. Wang & Bovik, 2002) to assess the algorithm performance. SAM (Yuhas et al., 1992) quantifies the spectral dissimilarity between the fused and reference images by calculating the angle between the spectral vectors of corresponding pixels in both images, with values closer to 0 indicating superior algorithm performance. The SSIM (Tian et al., 2021) measures the similarity between two images, considering luminance, contrast, and structure. An ideal value for this metric is 1, with values closer to 1 indicating higher similarity and accuracy. The PSNR (Z. Wang et al., 2004) measures the fidelity between the fused and reference images as an objective pixel-level evaluation method, with larger values indicating a better image quality and better fusion performance. ERGAS (Renza et al., 2012) is a metric used to evaluate the quality of the fused HSI by quantifying the spectral distortion introduced during the fusion process. It measures the overall composite error, with lower values indicating less error and a higher accuracy. The Q metric (Z. Wang & Bovik, 2002) is a full-reference image quality assessment metric that measures the similarity between the fused image and reference image. This metric is used to measure the overall quality of the fused image, with values closer to 1 indicating a higher quality.

For the quantitative evaluation metrics for the real-resolution fusion experiments, we used the spectral distortion (D_λ) (Alparone et al., 2008), spatial distortion (D_s) (Alparone et al., 2008), and the quality with no reference (QNR) (Alparone et al., 2008) metrics. D_λ is used to measure the degree of spectral distortion, with values closer to 0 indicating a greater spectral similarity and lower error. D_s is employed to measure the spatial distortion between the fused image and the observed image. Values closer to 0 indicate less distortion and a higher fusion accuracy. The QNR metric combines the spectral and spatial distortion to evaluate the overall fusion quality.The ideal value for QNR is 1, with values closer to 1 indicating a better image fusion quality.

5. Results and discussion

5.1. Fusion experiments on simulated-resolution imagery

1) *Cave dataset*: Fig. 9 presents the fusion results of all the algorithms on the indoor Cave dataset under the simulated resolution. The images are true-color composites (R-29, G-19, B-8), with a magnified view of the red region in the lower-left corner. For the Cave dataset with a limited number of bands, all the methods, except Fusformer, can effectively fuse the input synthetic RGB image and simulated low-resolution HSI to generate a high-resolution HSI. The CNMF method effectively preserves the textural features, resulting in clear edges. However, the stripes on the doll's clothing appear orange, whereas they are red in the reference image, indicating spectral distortion in the fusion image. In terms of the spatial detail, except for the proposed MHFu-former method, MSST-Net, SSR-NET, and CNMF, the junction between the doll and the white background in the fused images is blurred. The fusion image of MHFNet shows noticeable artifacts and bright patches at the edges. The Fusformer method fails to capture valid spatial detail features, resulting in a poor visual quality. The proposed MHFu-former method, however, stands out by maintaining superior spectral fidelity and spatial detail retention. It effectively suppresses artifacts, preserves sharp edges, and minimizes color distortion, making it the best-performing method for generating high-quality fused images with both spatial and spectral consistency.

As shown in Table 2, the HSI fusion performance was quantitatively evaluated on the Cave dataset using the eight different methods. Compared to the traditional method of CNMF, the deep learning based methods show significant advantages across all the evaluation metrics. MHFu-former achieves the lowest SAM value, indicating its ability to preserve spectral consistency and minimize spectral distortion. In terms of ERGAS, MHFu-former again achieves the lowest value of 2.8664, demonstrating its superior global error control. Furthermore, MHFu-former reaches a Q-value of 0.9994, further proving its excellent performance in balancing spectral and spatial information. In contrast, the

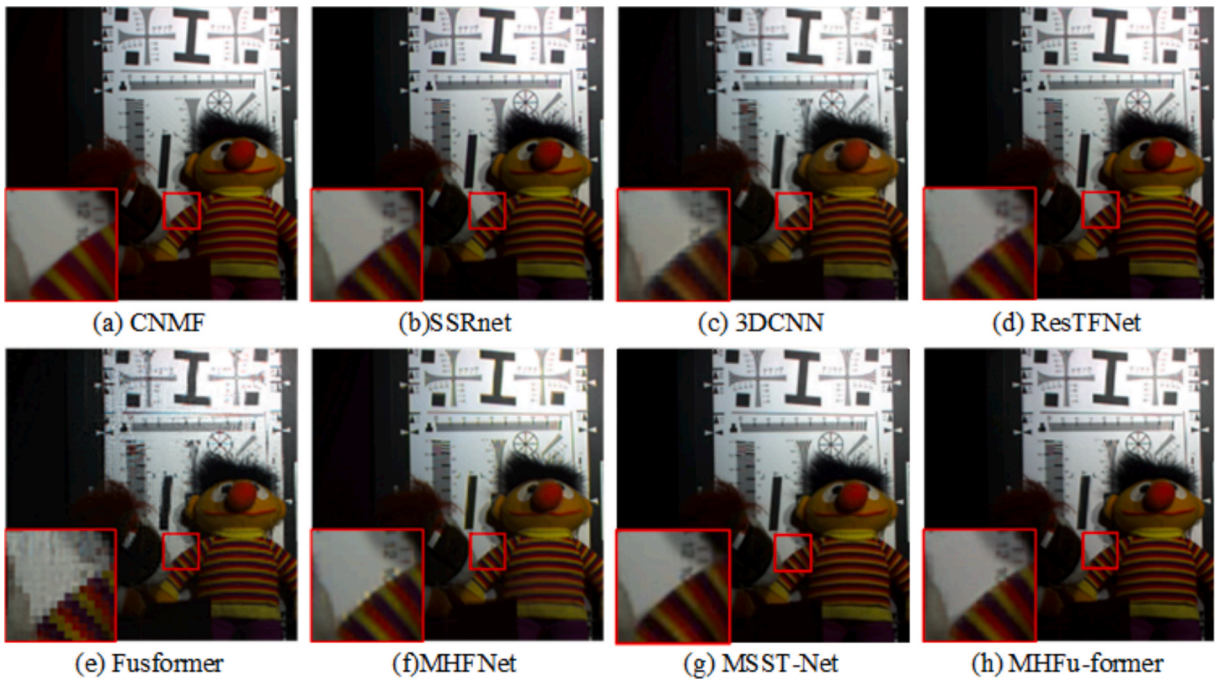


Fig. 9. Fusion images for the simulated-resolution experiment on the Cave dataset.

Table 2
Accuracy assessment for the simulated-resolution experiment on the Cave dataset.

Algorithm	SAM	SSIM	PSNR	ERGAS	Q
CNMF	8.8003	0.8924	19.6132	5.7585	0.9936
SSR-NET	7.2983	0.9813	31.7545	3.0415	0.9983
3DCNN	8.2065	0.9498	26.7734	6.5748	0.9916
TFNet	6.7562	0.9805	31.1767	3.5045	0.9976
Fusformer	8.6516	0.8776	22.7827	6.1559	0.9752
MHFNet	8.8401	0.9488	23.8885	5.7694	0.9938
MSST-Net	5.6576	0.9936	30.2678	3.4387	0.9985
MHFu-former	**5.1269**	**0.9939**	**34.7669**	**2.8664**	**0.9994**

traditional CNMF method exhibits significant spectral distortion, while 3DCNN performs poorly in SAM due to its spatial-spectral feature mapping approach. TFNet, MSST-Net and SSR-NET enhance the fusion quality, to some extent, achieving higher SSIM and PSNR values. Overall, MHFu-former achieves the best fusion image quality, fully demonstrating its effectiveness and robustness in HSI fusion tasks.

2) *ZY1-02D dataset*: Figs. 10 and 11 illustrate the fusion results obtained on simulated-resolution images Scene 1 and Scene 2, respectively. The images are true-color composites (R-29, G-19, B-8), with a magnified view of the red region in the lower-left corner. In the simulated-resolution experiments, HSIs with a spatial resolution of 90 m were fused with the simulated MSIs at a 30-m resolution to generate 30-m resolution fused HSIs, with the original 30-m resolution HSIs serving as reference. As shown in Fig. 10, the land features are relatively continuous and homogeneous, and the visual quality differences between methods are readily observable. In contrast, Scene 2 presents a more complex land-cover distribution with fragmented and patchy regions, posing significant challenges for all the algorithms. Except for CNMF, MSST-Net and MHFu-former, the other deep learning based approaches struggle to effectively capture the spatial details of the input images. SSR-NET generates fused HSIs with rich spatial details and high similarity to the reference images. However, the 3DCNN method exhibits noticeable spectral distortion due to its spatial-spectral feature mapping learned within the projected space after singular value decomposition. This process introduces significant errors during the inverse transformation when generating the fused image. TFNet, which is characterized by extensive consecutive upsampling and downsampling convolution operations, suffers from pronounced checkerboard artifacts in the continuously distributed land-cover types, such as rivers, leading to blurred visual effects. Fusformer demonstrates a suboptimal fusion performance, failing to enhance the visual quality of the HSIs. While MHFNet preserves acceptable edge details, its fusion results still exhibit striping artifacts and brightness inconsistencies, compromising the overall image quality. MSST-Net achieves a relatively balanced performance between spatial and spectral domains, maintaining natural color reproduction and clearer edges than Fusformer and MHFNet. However, slight blurring remains along object boundaries, and fine structural details are not as well preserved as in the MHFu-former results. MHFu-former outperforms the existing methods in spectral fidelity, spatial detail preservation, artifact suppression, and adaptability to complex scenes, demonstrating its effectiveness and robustness in HSI fusion tasks.

Tables 3 and 4 present the accuracy of the fused hyperspectral imagery generated by each method under the simulated-resolution experiments. The experimental results indicate that the MHFu-former method achieves the best performance in the simulated-resolution experiments, particularly in the complex Scene 2, where its SAM, SSIM, PSNR, ERGAS, and Q metric scores outperform the other methods, demonstrating strong spectral fidelity and spatial detail preservation. In contrast, the CNMF method exhibits the most significant spectral distortion, with a much higher SAM value than the other methods, indicating its difficulty in maintaining spectral consistency. MHFu-former performs the best in Scene 1, achieving the highest PSNR of 27.3750 and the lowest ERGAS of 2.1133, highlighting its advantage in image quality. The deep learning based methods, including 3DCNN, TFNet, Fusformer, and MSST-Net generally outperform the traditional method; however, some of them still suffer from checkerboard artifacts and spatial discontinuities. For instance, 3DCNN performs poorly in terms of SAM. The SSIM results suggest that MHFu-former can better preserve the structural information of the images. With Q values approaching 1, these methods achieve an optimal balance between spectral and spatial detail preservation.

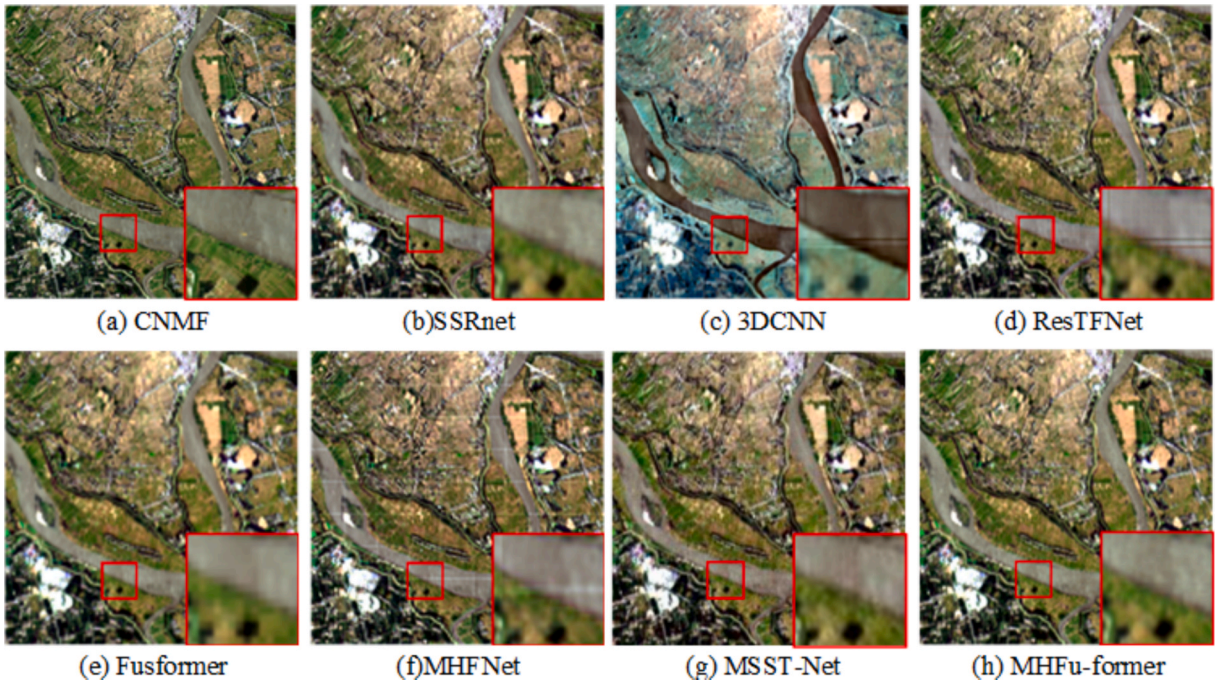


Fig. 10. Fusion images for the simulated-resolution experiment on Scene 1 of the ZY01-02D data.

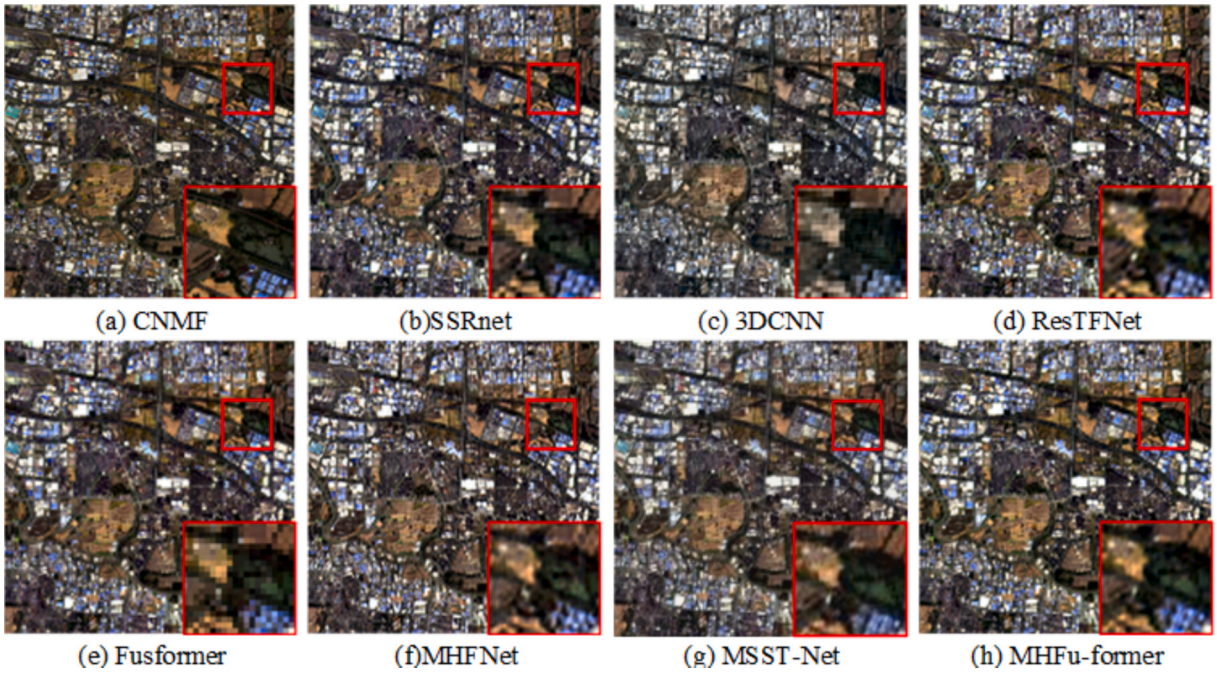


Fig. 11. Fusion images for the simulated-resolution experiment on Scene 2 of the ZY01-02D data.

5.2. Fusion experiments on real-resolution imagery

To assess the efficacy of the proposed MHFu-former network in practical image applications, real-resolution experiments were carried out utilizing the ZY-1 02D imagery. The fusion results of the various algorithms at the real resolution for the Scene 1 and Scene 2 regions are visualized in Figs. 12 and 13, respectively. The accuracy evaluation metrics for each algorithm are tabulated in Table 5. In addition, we specifically analyzed the comparison of the spatial details and generated spectra.

1) *Fusion images*: For the image fusion experiments with the two scenes from the ZY01-02D satellite, Figs. 12 and 13 present the fusion results of the different algorithms at the original resolution. In Scene 1, where the spatial information is relatively uniform and the land cover is more continuous with a broader distribution, CNMF, SSRNet, and MHFu-former significantly enhance the spatial details, outperforming the other methods. However, in the more complex Scene 2, all the methods, except CNMF, show limited effectiveness in enhancing spatial details. Specifically, blurred edges of the fragmented land covers such as buildings, rivers, and bare soil can be observed. In addition, the deep learning based algorithms exhibit a noticeable checkerboard effect caused by the convolution operations, further degrading the spatial

Table 3
Accuracy assessment for the simulated-resolution experiment on Scene 1 of the ZY01-02D data.

Algorithm	SAM	SSIM	PSNR	ERGAS	Q
CNMF	4.1447	0.8004	19.3795	4.9393	0.9788
SSR-NET	2.0773	0.8715	23.5386	3.6105	0.9967
3DCNN	3.7729	0.8875	21.4178	3.5687	0.9904
TFNet	2.0272	0.8379	21.9241	2.2956	0.9972
Fusformer	1.7526	0.9198	24.5390	2.5847	0.9973
MHFNet	4.0446	0.6961	21.3061	2.2698	0.9800
MSST-Net	2.1464	0.9279	22.4988	2.1457	0.9969
MHFu-former	**1.6705**	**0.9320**	**27.3750**	**2.1133**	**0.9975**

Table 4
Accuracy assessment for the simulated-resolution experiment on Scene 2 of the ZY01-02D data.

Algorithm	SAM	SSIM	PSNR	ERGAS	Q
CNMF	6.1931	0.6568	19.2330	9.3805	0.9551
SSR-NET	4.4643	0.6997	22.4057	8.8643	0.9779
3DCNN	4.1776	0.7217	21.9349	6.9435	0.9760
TFNet	4.2214	0.6908	21.8219	13.9206	0.9803
Fusformer	4.1383	0.6704	21.7808	7.6016	0.9836
MHFNet	5.4205	0.6949	20.4965	7.1502	0.9620
MSST-Net	4.0598	0.6977	22.3415	6.8292	0.9851
MHFu-former	**3.7239**	**0.7559**	**23.2131**	**6.7804**	**0.9886**

quality of the fused images. Overall, Fusformer performs the worst in the fused images, failing to effectively enhance the spatial details and leading to a lack of significant improvement in the spatial structures. In contrast, CNMF achieves the most outstanding spatial detail enhancement, with sharp and clearly defined edges in the fused images. However, 3DCNN suffers from significant spectral distortion. Except for MHFu-former, all the deep learning based methods exhibit noticeable blocky artifacts at the stitching edges, forming evenly distributed vertical and horizontal stripes that result in brightness inconsistencies and a loss of spatial details, ultimately lowering the overall fusion quality. MHFu-former demonstrates the best performance in minimizing spatial distortion, and its fused images achieve the highest consistency in spatial feature distribution.

2) *Accuracy evaluation*: The quantitative evaluation of the HSI fusion across the two scenes highlights the trade-off between spectral fidelity, spatial enhancement, and overall fusion quality. As shown in Table 5, the proposed MHFu-former method achieves the best D_λ and D_s values, along with the highest QNR scores, demonstrating its ability to effectively balance spectral preservation and spatial detail enhancement. In Scene 1, MHFu-former outperforms CNMF and TFNet, achieving $D_\lambda = 0.0168$ and $D_s = 0.0385$. CNMF enhances the spatial details but suffers from higher spectral distortion, while the deep learning based 3DCNN method, due to the weaker spectral constraints, exhibits the highest spectral distortion, with $D_\lambda = 0.0736$. In Scene 2, which features complex spatial structures, MHFu-former maintains its advantage, with $D_\lambda = 0.0197$ and $D_s = 0.0216$. The transformer-based methods of Fusformer and 3DCNN show improvements in spectral fidelity, but still exhibit relatively high spatial distortion. The QNR metric further confirms the robustness of MHFu-former, as it outperforms Fusformer and TFNet. In addition, an inverse relationship between the spatial enhancement and spectral fidelity can be observed: CNMF excels in spatial detail preservation but ranks last in spectral accuracy for Scene 2. In summary, MHFu-former effectively mitigates this trade-off through multi-scale fusion and attention mechanisms, demonstrating a superior fusion performance and robustness.

3) *Spatial details*: Fig. 14 illustrates the rectangular distribution of crops and vegetation, highlighting the spatial enhancement features of the different methods. The comparative analysis reveals that TFNet exhibits checkerboard artifacts, particularly manifesting as grid-like distortions in homogeneous regions, which degrade the visual quality and spatial coherence. In contrast, both CNMF and MHFu-former perform well in preserving spatial edge details, especially in delineating the sharp boundaries between vegetation patches and farmland. However, CNMF demonstrates a notable advantage in accurately reconstructing the sub-pixel transitions at vegetation-farmland boundaries, thereby exhibiting a superior performance in maintaining spatial edge details.

4) *Generated spectra*: We selected four typical land-cover types—soil, vegetation, building, and water—and conducted a comparative analysis of the reflectance characteristics of three methods, namely CNMF, TFNet, and MHFu-former, against the reference spectral data, as shown in Fig. 15. The experimental results demonstrate that MHFu-former achieves the best performance across all the land-cover categories,

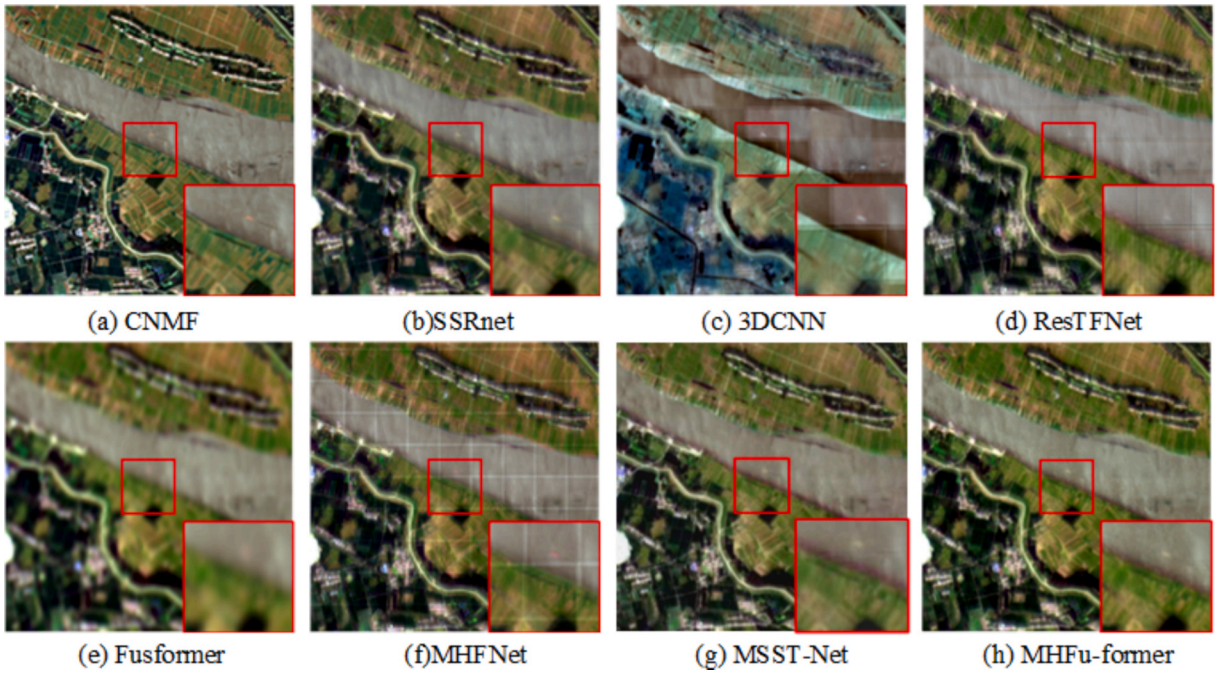


Fig. 12. Fusion images for the real-resolution experiment on Scene 1 of the ZY01-02D data.

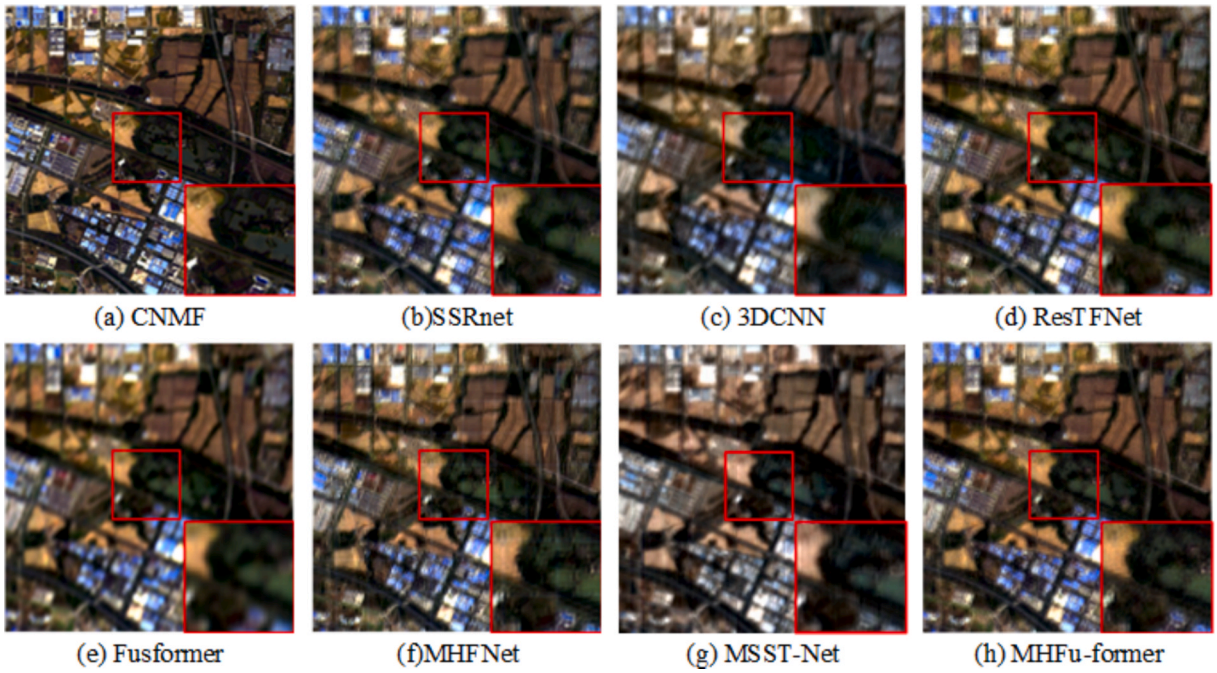


Fig. 13. Fusion images for the real-resolution experiment on Scene 2 of the ZY01-02D data.

Table 5
Accuracy assessment for the real-resolution experiments on the two scenes of ZY01-02D data.

Algorithm	Scene1			Scene2		
	D_λ	D_s	QNR	D_λ	D_s	QNR
CNMF	0.0268	0.0645	0.9541	0.0355	0.0593	0.9525
SSR-NET	0.0310	0.0399	0.9645	0.0283	0.0370	0.9673
3DCNN	0.0736	0.0600	0.9331	0.0294	0.0233	0.9736
TFNet	0.0215	0.0390	0.9697	0.0258	0.0261	0.9740
Fusformer	0.0211	0.0473	0.9657	0.0217	0.0232	0.9775
MHFNet	0.0223	**0.0346**	0.9715	0.0262	0.0688	0.9522
MSST-Net	0.0171	0.0410	0.9695	0.0229	0.0228	0.9659
MHFu-former	**0.0168**	0.0385	**0.9722**	**0.0197**	**0.0216**	**0.9793**

with particularly strong robustness in the building and vegetation categories. For the different land-cover types, MHFu-former achieves the highest spectral fidelity in the soil category, particularly in the near-infrared range of 750–1250 nm, where it exhibits the lowest retrieval error. However, CNMF suffers from spectral inaccuracies around 2250 nm, which can be attributed to the mixing of building materials and the limitations in endmember extraction accuracy. For the water bodies and vegetation, all the methods exhibit relatively small errors in the VNIR range. However, as the wavelength increases, CNMF and TFNet show intensified reflectance fluctuations in the 1500–2500 nm range. In contrast, MHFu-former effectively suppresses noise interference through its hierarchical feature weighting mechanism, thereby improving the overall stability of the spectral retrieval. In summary, MHFu-former demonstrates a superior spectral retrieval performance across the different land-cover categories, with particularly strong robustness in the complex land-cover types such as buildings and water bodies, as well as in the SWIR region.

5.3. Ablation and hyperparameter experiments

1) *Network structure*: To validate the effectiveness of the spectral attention module in the proposed network, ablation studies were conducted on Resource-02D imagery to analyze the impact of hyperparameters in the Swin transformer based architecture. As shown in Table 6, the convolutional layer integrated with the spatial-spectral attention module (CBAM) outperforms the vanilla CNN counterpart across most of the evaluation metrics (i.e., SSIM, PSNR, ERGAS, QNR) in both Scene 1 and Scene 2, despite a slight increase in SAM values. The CBAM, employing a sigmoid activation function for the channel attention, introduces nonlinear mapping to enhance the feature discrimination, which can amplify the spectral deviations in regression-based fusion tasks. Notably, the proposed module achieves superior spectral distortion control (lower D_λ and D_s) and spatial consistency (higher Q), demonstrating its ability to balance spectral fidelity and spatial detail preservation. These results underscore the critical role of adaptive

Fig. 14. Localized magnification of the true-resolution fusion images of the three methods.

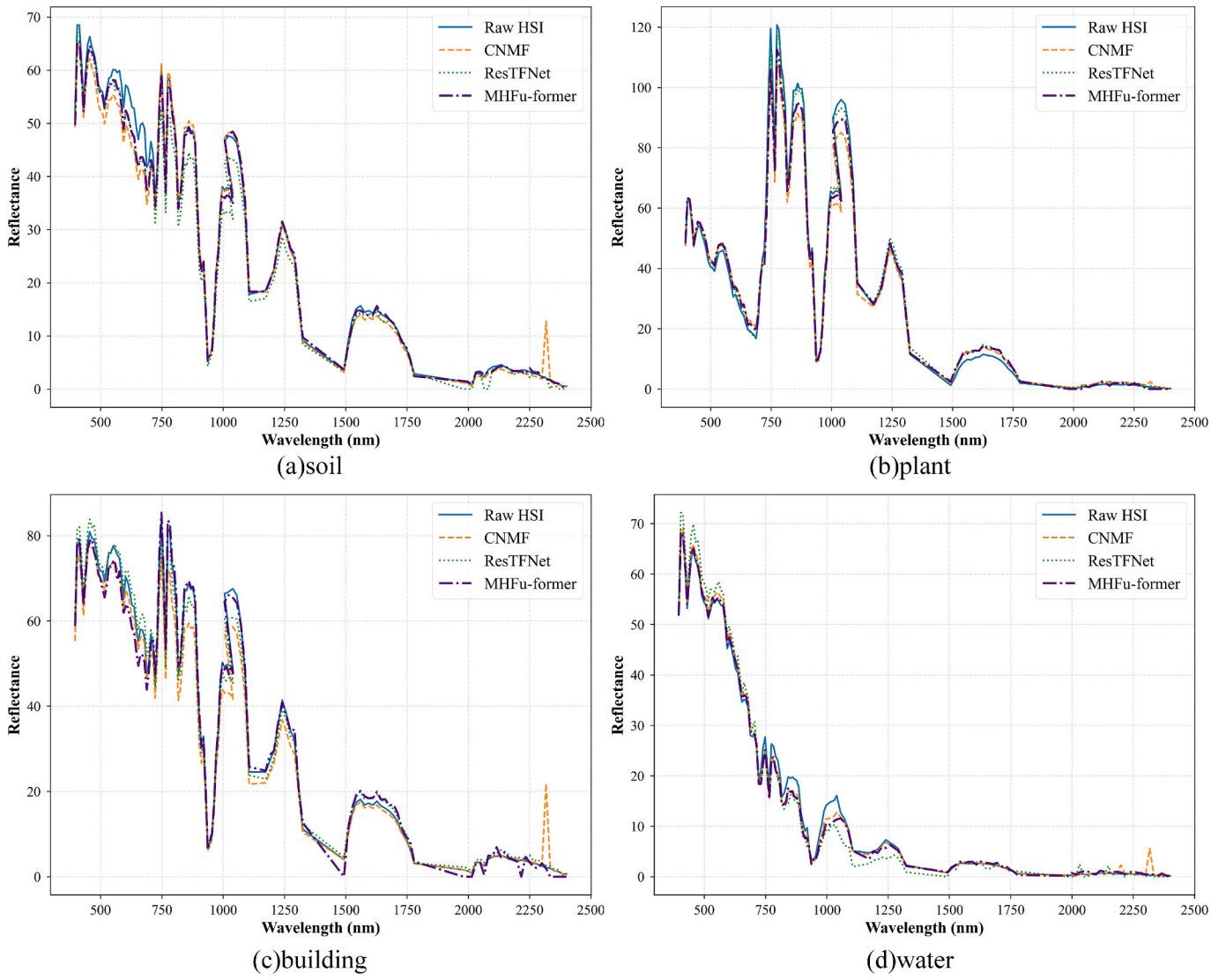


Fig. 15. Comparison of the spectral curves of four ground objects with the different methods.

Table 6
The fusion accuracy for the two scenes with the CBAM.

Data	Model	SAM	SSIM	PSNR	ERGAS	Q	D_λ	D_s	QNR
Scene1	CNN	**1.6580**	0.9260	25.0315	2.4726	0.9965	**0.0140**	0.0451	0.9702
	CBAM	1.6705	**0.9320**	**27.3750**	**2.1133**	**0.9975**	0.0168	**0.0385**	**0.9722**
Scene2	CNN	**3.6223**	0.7493	22.2462	6.9433	0.9819	**0.0097**	0.0430	0.9734
	CBAM	3.7239	**0.7559**	**23.2131**	**6.7804**	**0.9886**	0.0197	**0.0216**	**0.9793**

spatial-spectral attention in mitigating fusion-induced artifacts while highlighting the trade-off between nonlinear feature enhancement and spectral accuracy optimization in deep image fusion frameworks.

2) *Window size*: The hyperparameter selection in the Swin transformer module, particularly the window size for attention computation and the image block size for feature map partitioning, significantly influences the network's fusion performance. The experimental results obtained with fixed 3 × 3 slice sizes and varying window

Table 7
The fusion accuracy for the two scenes with different window sizes.

Data	Window size	SAM	SSIM	PSNR	ERGAS	Q	D_λ	D_s	QNR
Scene 1	3	1.6705	**0.9320**	**27.3750**	**2.1133**	**0.9975**	**0.0168**	**0.0385**	**0.9722**
	5	**1.6572**	0.9299	26.6423	2.2729	0.9973	0.0181	0.0482	0.9667
	7	1.7525	0.9187	26.3676	2.3874	0.9972	0.0192	0.0435	0.9685
Scene 2	3	3.7239	0.7559	23.2131	6.7804	0.9886	0.0197	**0.0216**	**0.9793**
	5	**3.6246**	**0.7621**	**23.3344**	6.4701	0.9888	**0.0184**	0.0252	0.9781
	7	3.6312	0.7539	23.2412	**6.1136**	**0.9893**	0.0185	0.0282	0.9704

sizes of 3 × 3, 5 × 5, and 7 × 7 are listed in Table 7. For Scene 1, the 3 × 3 window configuration achieves the best performance across most of the metrics, including a SSIM of 0.9320, PSNR of 27.3750, ERGAS of 2.1133, and QNR of 0.9722, despite a marginally higher SAM value of 1.6705 when compared to the 5 × 5 window. Conversely, Scene 2 demonstrates superior results with the 5 × 5 window size, attaining balanced improvements in spatial-spectral fidelity, with a PSNR of 23.3344 and ERGAS of 6.4701, along with distortion control metrics of $D_\lambda = 0.0184$ and $D_s = 0.0252$, although with slight trade-offs in QNR at 0.9781 when compared to the smaller windows. The 7 × 7 window exhibits an inconsistent performance across scenes, achieving the lowest ERGAS of 6.1136 in Scene 2, but suboptimal spectral preservation, with a SAM of 3.6312. These findings emphasize the critical need for scene-specific window size adaptation to optimize the balance between global context modeling and local detail preservation in MHIF fusion tasks.

3) *Slice size*: Table 8 presents the experimental results for a 3 × 3 window size paired with slice sizes of 3 × 3 and 4 × 4. As shown in Table 8, the 3 × 3 slice configuration outperforms the 4 × 4 variant across most of the metrics, except for SSIM. For Scene 1, the 3 × 3 slice achieves superior spectral preservation, with a SAM of 1.6705 and higher reconstruction fidelity, yielding a PSNR of 27.3750, ERGAS of 2.1133, and QNR of 0.9722. While the 4 × 4 slice slightly improves the SSIM score to 0.9381, it degrades the spectral accuracy to a SAM of 7.3481, increases the spatial distortion, with $D_\lambda = 0.0240$ and $D_s = 0.0451$, and reduces the overall fusion quality to a QNR of 0.9653. In Scene 2, the 3 × 3 slice maintains a competitive performance, with a SAM of 3.7239, PSNR of 23.2131, ERGAS of 5.7585, robust distortion metrics of $D_\lambda = 0.0184$ and $D_s = 0.0252$, and a QNR of 0.9781. In contrast, the 4 × 4 slice exhibits a higher SSIM of 0.7559 but suffers from spectral degradation, with a SAM of 7.7176, elevated spatial distortion at $D_s = 0.0395$, and a reduced QNR of 0.9674. These results underscore the trade-off between SSIM improvement and spatial-spectral fidelity degradation when increasing the slice size, highlighting the advantage of smaller 3 × 3 slices for a balanced fusion performance across diverse scenes.

4) Computational Complexity:

To further analyze the computational complexity, we report the Floating Point Operations (FLOPs), and testing time of different fusion methods on the CAVE, and ZY-1 02D satellite dataset. As shown in Table 9 and Table 10, traditional methods generally have longer inference time. CNN-based methods strike a balance between computational cost and performance. Transformer-based methods show significantly higher FLOPs and inference time due to the use of self-attention mechanisms. The proposed MHFu-former also involves relatively large FLOPs, primarily because of the integration of Swin Transformer blocks and multiscale patch embeddings.

6. Conclusion

In this paper, a multispectral and hyperspectral image fusion transformer (MHFu-former) has been proposed to reduce spectral distortion, enhance spatial details, and maintain spectral integrity in MHIF. The model integrates a Swin transformer and convolutional module into a two-branch architecture to extract multiscale spatial-spectral features by parallel depth-separable convolution, which efficiently handles spectral disparities while capturing global contextual correlations and fine-grained spatial details. The spatial-spectral fusion attention mechanism dynamically prioritizes the key spectral bands and integrates multilevel spatial information through global–local dependency modeling to minimize spectral distortion and maintain the continuity of the spectral profiles. The end-to-end framework maps raw hyperspectral/multispectral image inputs to a high-resolution HSI through cascading feature extraction, fusion, and refinement, providing interpretable insights for spatial-spectral decoupling and cross-modal interactions. The experimental results obtained on the Cave dataset and ZY0-02D multispectral and hyperspectral images showed that the proposed method can obtain high spatial and spectral resolution fusion images with good spectral consistency and rich spatial details, with a better performance than a classical method and the commonly used deep learning based methods. The application on satellite images proved that the proposed MHFu-former method has a strong spectral preservation ability, without obvious checkerboard effects and uneven brightness at synthetic block edges.

Despite its strong performance, the proposed MHFu-former still has limitations. First, the model employs fixed window and slice sizes, which may not be universally optimal across diverse scene types or sensor characteristics. The static configuration could limit generalization performance in highly heterogeneous environments. Future work could explore adaptive or dynamic windowing strategies that adjust to local image complexity, potentially enhancing robustness. Second, the current pipeline relies on pre-processing steps such as spectral alignment and interpolation. These steps, while standard, can introduce subtle artifacts and often require prior knowledge of sensor specifications. A key future direction is the development of end-to-end fusion mechanisms that can learn to align and integrate data directly, thereby improving both autonomy and generalization.

Table 9
Number of parameters, FLOPs, and testing time of different fusion methods on the CAVE dataset.

Algorithm	FLOPs (G)	Testing time (s)
CNMF	/	12.7119
SSR-NET	2.1521	0.2269
3DCNN	5.9560	0.6632
TFNet	36.4569	0.2589
Fusformer	2506.9419	2.1547
MHFNet	74.5405	0.2187
MSST-Net	15.9957	9.7433
MHFu-former	484.9785	0.6567

Table 10
Number of parameters, FLOPs, and testing time of different fusion methods on the ZY01-02D data.

Algorithm	Scene1		Scene2	
	FLOPs/G	Testing time/s	FLOPs/G	Testing time/s
CNMF	/	1489.02	/	1363.24
SSR-NET	42.0330	6.3848	54.5002	7.1274
3DCNN	6.3144	12.7378	6.3144	16.1532
TFNet	44.1727	6.7358	45.5343	6.9382
Fusformer	2518.2488	25.4826	2520.2660	26.4147
MHFNet	1141.1839	7.3993	1462.2126	8.6224
MSST-Net	67.3353	7.8527	76.7005	20.2607
MHFu-former	485.5501	8.7913	485.6509	10.2591

Table 8
The fusion accuracy for the two scenes with different slice sizes.

Data	Slice size	SAM	SSIM	PSNR	ERGAS	Q	D_λ	D_s	QNR
Scene 1	3	**1.6705**	0.9320	**27.3750**	**2.1133**	**0.9975**	**0.0168**	**0.0385**	**0.9722**
	4	7.3481	**0.9381**	19.9566	3.3204	0.9734	0.0240	0.0451	0.9653
Scene 2	3	**3.7239**	0.7406	**23.2131**	**5.7585**	**0.9915**	**0.0184**	**0.0252**	**0.9781**
	4	7.7176	**0.7559**	18.7752	6.7804	0.9886	0.0256	0.0395	0.9674

CRediT authorship contribution statement

Xue Wang: Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Songling Yin:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Xiaojun Xu:** Validation, Supervision, Formal analysis, Data curation. **Yong Mei:** Software, Resources, Investigation, Formal analysis. **Yan Huang:** Project administration, Investigation, Formal analysis. **Kun Tan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (No. 2024CSJZN01300), Shanghai Municipal Education Commission Science and Technology Project(2024AI02002), National Natural Science Foundation of China (No. 42171335), National Civil Aerospace Project of China (No. D040102), the International Research Center of Big Data for Sustainable Development Goals (No. CBAS2022GSP07), and the Open Foundations of Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology (JSECF2023-10).

Data availability

The data that has been used is confidential.

References

Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., Selva, M., 2008. Multispectral and panchromatic data fusion assessment without reference. Photogramm. Eng. Remote Sens. 74 (2), 193–200.

Bandara, W.G.C., Patel, V.M., 2022. *Hypertransformer: a textural and spectral feature fusion transformer for pansharpening.* Paper Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Deng, S., Wu, X., Ran, R., Wen, R., 2023. Bidirectional dilation transformer for multispectral and hyperspectral image fusion. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-23).

Dong, L., Geng, J., Jiang, W., 2024. Spectral-Spatial Enhancement and Causal Constraint for Hyperspectral image Cross-Scene Classification. IEEE Trans. Geosci. Remote Sens.

Dong, W., Yang, Y., Qu, J., Xiao, S., Du, Q., 2021. Hyperspectral pansharpening via local intensity component and local injection gain estimation. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Feng, J., Feng, X., Chen, J., Cao, X., Zhang, X., Jiao, L., Yu, T., 2020. Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification. Remote Sens. (Basel) 12 (7), 1149.

Hu, J.-F., Huang, T.-Z., Deng, L.-J., Dou, H.-X., Hong, D., Vivone, G., 2022. Fusformer: a transformer-based fusion network for hyperspectral image super-resolution. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Jia, S., Yu, H., Wang, C., Zheng, K., Li, J., Hu, J., 2023. Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion. Inf. Fusion 79, 174–187.

Liu, X., Liu, Q., Wang, Y., 2020. Remote sensing image fusion based on two-stream fusion network. Inf. Fusion 55, 1–15.

Long, Y., Wang, X., Xu, M., Zhang, S., Jiang, S., Jia, S., 2023. Dual self-attention Swin transformer for hyperspectral image super-resolution. IEEE Trans. Geosci. Remote Sens. 61, 1–12.

Lu, X., Wang, B., Zheng, X., Li, X., 2017. Exploring models and data for remote sensing image caption generation. IEEE Trans. Geosci. Remote Sens. 56 (4), 2183–2195.

Ma, J., Zhang, Y., Zhang, W., Fan, H., & Du, Q. (2024). Reciprocal transformer for hyperspectral and multispectral image fusion. IEEE Transactions on Neural Networks and Learning Systems. Advance online publication.

Palsson, F., Sveinsson, J.R., Ulfarsson, M.O., 2017. Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network. IEEE Geosci. Remote Sens. Lett. 14 (5), 639–643.

Qin, H., Xu, T., Liu, P., Xu, J., Li, J., 2024. DMSSN: Distilled mixed Spectral-Spatial Network for Hyperspectral Salient Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*.

Renza, D., Martinez, E., Arquero, A., 2012. A new approach to change detection in multispectral images by means of ERGAS index. IEEE Geosci. Remote Sens. Lett. 10 (1), 76–80.

Selva, M., Aiazzi, B., Butera, F., Chiarantini, L., Baronti, S., 2014. *Hyper-sharpening of hyperspectral data: a first approach.* Paper Presented at the 2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS).

Tian, X., Zhang, W., Chen, Y., Wang, Z., Ma, J., 2021. Hyperfusion: a computational approach for hyperspectral, multispectral, and panchromatic image fusion. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Wald, L. (2000). *Quality of high resolution synthesised images: Is there a simple criterion?* Paper presented at the Third conference" Fusion of Earth data: merging point measurements, raster maps and remotely sensed images".

Wang, W., Zeng, W., Huang, Y., Ding, X., Paisley, J., 2019. *Deep blind hyperspectral image fusion*. Paper Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision.

Wang, X., Borsoi, R.A., Richard, C., Chen, J., 2023a. Deep hyperspectral and multispectral image fusion with inter-image variability. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Wang, Z., Bovik, A.C., 2002. A universal image quality index. IEEE Signal Process Lett. 9 (3), 81–84.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, Z., Wang, X., Tan, K., Han, B., Ding, J., Liu, Z., 2023b. Hyperspectral anomaly detection based on variational background inference and generative adversarial network. Pattern Recogn. 143, 109795.

Xie, Q., Zhou, M., Zhao, Q., Xu, Z., Meng, D., 2020. MHF-Net: an interpretable deep network for multispectral and hyperspectral image fusion. IEEE Trans. Pattern Anal. Mach. Intell. 44 (3), 1457–1473.

Yan, J., Zhang, K., Sun, Q., Ge, C., Wan, W., Sun, J., Zhang, H., 2025. Spatial–spectral unfolding network with mutual guidance for multispectral and hyperspectral image fusion. Pattern Recogn. 161, 111277.

Yang, B., Mao, Y., Liu, L., Fang, L., Liu, X., 2024. Change representation and extraction in stripes: Rethinking unsupervised hyperspectral image change detection with an untrained network. IEEE Trans. Image Process.

Yokoya, N., Yairi, T., Iwasaki, A., 2011. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Trans. Geosci. Remote Sens. 50 (2), 528–537.

Yuhas, R. H., Goetz, A. F., & Boardman, J. W. (1992). *Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm.* Paper presented at the JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop.

Zhang, X., Huang, W., Wang, Q., Li, X., 2020. SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. IEEE Trans. Geosci. Remote Sens. 59 (7), 5953–5965.

Zhu, X.X., Bamler, R., 2012. A sparse image fusion algorithm with application to pan-sharpening. IEEE Trans. Geosci. Remote Sens. 51 (5), 2827–2836.