Information Fusion 96 (2023) 117–129

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion

Sen Jia, Zhichao Min, Xiyou Fu *

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China
Guangdong-Hong Kong-Macau Joint Laboratory for Smart Cities, Shenzhen University, Shenzhen, 518060, China
Key Laboratory for Geo-Environmental Monitoring of Coastal Zone, Ministry of Natural Resources, Shenzhen University, Shenzhen, 518060, China

ARTICLE INFO

Keywords:
Hyperspectral image (HSI)
Multispectral image (MSI)
Transformer
Pre-training
Spectral multi-head self-attention
Image fusion

ABSTRACT

Fusing hyperspectral images (HSIs) and multispectral images (MSIs) is an economic and feasible way to obtain images with both high spectral resolution and spatial resolution. Due to the limited receptive field of convolution kernels, fusion methods based on convolutional neural networks (CNNs) fail to take advantage of the global relationship in a feature map. In this paper, to exploit the powerful capability of Transformer to extract global information from the whole feature map for fusion, we propose a novel Multiscale Spatial–spectral Transformer Network (MSST-Net). The proposed network is a two-branch network that integrates the self-attention mechanism of the Transformer to extract spectral features from HSI and spatial features from MSI, respectively. Before feature extraction, cross-modality concatenations are performed to achieve cross-modality information interaction between the two branches. Then, we propose a spectral Transformer (SpeT) to extract spectral features and introduce multiscale band/patch embeddings to obtain multiscale features through SpeTs and spatial Transformers (SpaTs). To further improve the network's performance and generalization, we proposed a self-supervised pre-training strategy, in which a masked bands autoencoder (MBAE) and a masked patches autoencoder (MPAE) are specially designed for self-supervised pre-training of the SpeTs and SpaTs. Extensive experiments on simulated and real datasets illustrate that the proposed network can achieve better performance when compared to other state-of-the-art fusion methods. The code of MSST-Net will be available at http://www.jiasen.tech/papers/ for the sake of reproducibility.

1. Introduction

Hyperspectral image (HSI) contains hundreds of continuous narrow spectral bands from visible wavelengths to near-infrared wavelengths, which greatly benefits the precise identification of the composed materials of the ground objects [1]. In view of their capability to accurately characterize the attribute information of objects, HSIs play an important role in a wide range of tasks including dynamic monitoring of the environment [2], land cover classification [3], precision agriculture [4], and anomaly detection [5].

However, the high spectral resolution of HSIs generally comes with a compromise of its spatial resolution due to the limitations of the imaging platform. Since the bandwidth of the electromagnetic wave scattered into the instantaneous field of view is narrow, the spatial resolution of HSIs has to be lower to increase the signal-to-noise ratio. The low spatial resolution of HSIs hinders the potential applications of HSIs in many areas [6]. Normally, low-resolution hyperspectral images (LR-HSIs) and high-resolution multispectral images (HR-MSIs) can be acquired with different sensors, respectively. Therefore, obtaining high-resolution hyperspectral images (HR-HSIs) at the algorithmic level is a necessary and promising research direction [7].

Generally, the observed LR-HSI and HR-MSI can be viewed as the spatial and spectral degradation versions of an underlying HR-HSI, respectively. Thus, the underlying HR-HSI can be reconstructed using the observed images. There are approximately two ways to reconstruct the underlying HR-HSI, i.e., single LR-HSI super-resolution and fusion of LR-HSI with HR-MSI or panchromatic images. Since image reconstruction is an ill-posed problem, it is very tricky to restore HR-HSI from only a single LR-HSI. The reconstruction could be easier and more meaningful with the introduction of HR-MSI or panchromatic images as an auxiliary help. Therefore, it has become a popular research field to develop high-performance algorithms to effectively fuse the spatial and spectral information from HR-MSI and LR-HSI, respectively, to achieve complementary feature fusion, and improve the spatial and spectral resolution of the images.

* Corresponding author.
E-mail addresses: senjia@szu.edu.cn (S. Jia), minzhichao2020@email.szu.edu.cn (Z. Min), fuxy0623@szu.edu.cn (X. Fu).

https://doi.org/10.1016/j.inffus.2023.03.011
Received 19 November 2022; Received in revised form 9 March 2023; Accepted 15 March 2023
Available online 21 March 2023

Many machine learning based methods have been proposed to fuse LR-HSI and HR-MSI, such as matrix factorization based methods [8–10] and tensor factorization based methods [11,12]. Most of them rely on hand-crafted priors, which are time-consuming with limited representation ability. With the rapid development of deep learning, methods based on convolutional neural networks (CNNs) show impressive performance in the fusion of LR-HSI and HR-MSI. Considering the different characteristics of LR-HSI and HR-MSI, Shen et al. [13] formulated the fusion problem into a spectral optimization problem and a spatial optimization problem by using matrix decomposition. In order to solve the difficult problem of cross-modality information fusion, Zhang et al. [14] proposed a straightforward physical model, which includes a spatial edge loss and a spectral edge loss for the spatial and spectral restorations. The impressive performance of CNN based fusion methods boils down to the powerful inferential capability of CNNs. However, the small receptive field of CNNs makes them fail to capture global features effectively, which, to a certain extent, limits the fusion performance of CNN-based methods.

The strong ability to capture long-distance dependencies makes Transformer an appropriate method to extract global features from images [15], which has been proved in many Transformer-based methods, such as Vision Transformer (ViT) [16]. Since MSI contains rich spatial details and HSI has a high correlation in spectral dimension, it will be promising to design a Transformer-based network to excavate these features to reconstruct the HR-HSI. However, ViT divides the image into patches with a fixed spatial size and uses the patches as tokens to calculate self-attention to reduce high computational complexity, which not only limits its ability to characterize spatial features at different scales but also limits its ability to extract spectral features of hyperspectral images. In addition, ViT lacks certain desirable properties inherently built into the CNN architecture that make CNNs uniquely suited to solve vision tasks, e.g., locality and translation invariance. Thus, the training of the Transformer usually requires much more training data to obtain a competitive result. Motivated by the reasons above, we propose a novel Multiscale Spatial–spectral Transformer Network (MSST-Net) for MSI and HSI fusion. The proposed network is a two-branch network that integrates the self-attention mechanism of the Transformer to extract spectral features of HSI and spatial features of MSI, respectively. Moreover, to address the problem that the Transformer is difficult to train on a small dataset, we proposed a self-supervised pre-training strategy based on the idea of masked autoencoders (MAE) [17]. The main contributions of this paper are described in detail as follows:

1. We propose a Multiscale Spatial–spectral Transformer Network (MSST-Net) for hyperspectral and multispectral image fusion. The MSST-Net extracts deep spectral and spatial features using multiscale spectral Transformers (SpeTs) and spatial Transformers (SpaTs), respectively, and then combines the extracted multiscale spectral and spatial features with shallow features, using a long skip connection, to reconstruct HR-HSI. Before feature extraction, cross-modality concatenations are performed to achieve cross-modality information interaction between the two branches.
2. The spectral Transformer is proposed to better capture spectral features from HSIs. In the spectral Transformer, spectral multi-head self-attention is designed to effectively obtain spectral features from HSIs by calculating the self-attention in the spectral domain and dividing the multi-head in the spatial domain.
3. To overcome the limitation of the Transformer on extracting detailed information, we introduce multiscale band/patch embeddings to extract multiscale spectral/spatial features from the observed images. The final features are then obtained by fusing the multiscale features using learnable weights, which enhance the abundance of the extracted features.
4. We propose a self-supervised pre-training strategy, in which a masked bands autoencoder (MBAE) and a masked patches autoencoder (MPAE) are specially designed for self-supervised pre-training of the SpeTs and SpaTs. The pre-trained SpeTs and SpaTs are then loaded into the proposed network for end-to-end fine-tuning to improve the performance and generalization of the network.

The remainder of this article is organized as follows. Section 2 gives the related works of representative MSI and HSI fusion methods and the applications of Transformers in hyperspectral images. In Section 3, we describe the proposed MSST-Net in detail. The experimental results on four datasets are presented and analyzed in Section 4. Finally, we provide a conclusion in Section 5.

2. Related work

In this section, we first introduce some existing representative MSI and HSI fusion methods. Then, we provide an overview of the applications of Transformers in hyperspectral images.

2.1. Hyperspectral and multispectral image fusion

The existing HSI and MSI fusion methods can be approximately divided into matrix factorization based methods, tensor factorization based methods, deep learning based methods, and pan-sharpening methods extended for the fusion of HSIs and MSIs. Matrix factorization based methods fuse the images by decomposing the target HR-HSI into several matrices. Based on the estimation method of the spectral basis and coefficients, matrix factorization based methods can be mainly divided into three classes [18]. The first class obtains the spectral basis and the coefficients only from the observed HSI and the observed MSI, respectively. For example, Kawakami et al. [19] first proposed the sparse matrix factorization (SMF) method, in which the spectral dictionary was estimated with the sparse dictionary learning method, and the sparse coefficients were obtained by sparse coding algorithm. Akhtar et al. [20] learned the spectral dictionary from the LR-HSI and estimated the coefficients from the high-resolution MSI by Bayesian sparse coding. The second class firstly estimates the spectral basis from the observed HSI, and then calculates coefficients from both two images, such as [21–23]. To reduce computation time, Wei et al. [8] proposed a fast multiband image fusion algorithm (FUSE) by solving a Sylvester equation. The third class formulates the fusion task based on the coupled matrix decomposition and then alternatively updates the spectral basis and coefficients, rather than using the fixed dictionary. For example, Yokoya et al. [24] used a coupled nonnegative matrix factorization (CNMF) to solve the fusion problem of spectral unmixing. Lanaras et al. [25] obtained spectral basis and coefficients via the proximal alternating linearized minimization by imposing several priors on spectral unmixing.

Tensor factorization-based methods are a kind of method that treat the images as a tensor to preserve the spatial and spectral structure of the images rather than reshaping them into matrices. Tucker decomposition and Canonical polyadic (CP) decomposition are two widely used decompositions used in the fusion of HSIs and MSIs. Li et al. [11] proposed a coupled sparse Tucker decomposition (CSTF) scheme for HSI-MSI fusion, which estimates the core tensor and dictionary of each mode via proximal alternating optimization. Dian et al. [26] proposed a nonlocal sparse tensor factorization approach (NLSTF_SMBF) for the fusion of HSI and MSI in a semi-blind manner. Prévost et al. [27] made use of the truncated SVD to obtain the dictionaries of three modes to reduce the computational burden. Kanatsoulis et al. [28] factored the HR-HSI using CP decomposition and estimate each factor matrice via solving the least squares equation. Xu et al. [29] further proposed a non-local CP decomposition for HSI-MSI fusion. In addition to Tucker decomposition and CP decomposition, many other TR methods have

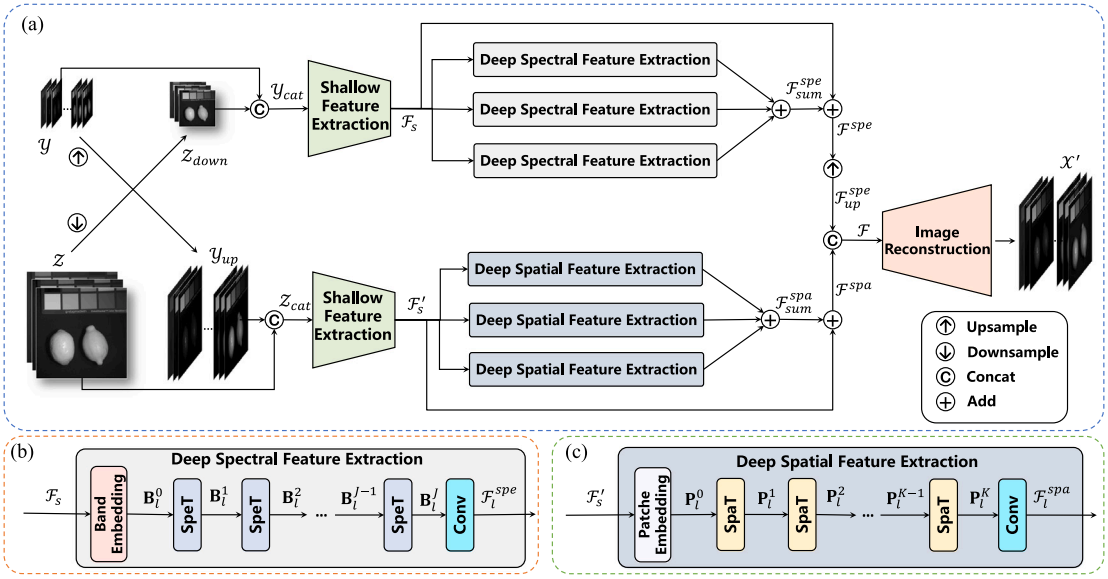

Fig. 1. The overall architecture diagram of our proposed multiscale spatial–spectral Transformer network.

also been actively studied, such as low tensor-train rank regularized HSI-MSI fusion (LTTR) [30], and coupled tensor ring factorization (CTRF) method [31].

However, most traditional methods are based on the modeling of image priors, which is usually sensitive to parameter selection. With the development of deep learning, people realize that all the parameters can be learned from training data via deep learning networks without imposing any assumptions on the images [32]. Therefore, Dian et al. [33] proposed a deep HSI sharpening method (DHSIS) for the fusion of hyperspectral and multispectral images, which directly learns the image priors via a deep residual network instead of using hand-crafted image priors. In addition, Palsson et al. [34] noticed the importance of HSI in the spectral dimension and proposed a method by training a 3-D CNN for learning filters used to fuse the MSI and HSI. Furthermore, to reduce the computational complexity of the 3-D CNN, they used principal component analysis (PCA) [35] for dimensionality reduction before fusing. Zheng et al. [36] proposed an edge-conditioned feature transform network (EC-FTN) to maintain low-level structure information such as sharp edges. However, most learning based methods are supervised. Thus they need an extensive training set, which is not easy to obtain in real life. So Qu et al. [37] proposed an unsupervised sparse Dirichlet-Net (uSDN), which uses an unsupervised encoder–decoder architecture to extract the spatial information and spectral information of two modalities with different dimensions. Since the priors of high-dimensional HSIs can be highly complicated and the degeneration is often unknown, Zhang et al. [38] proposed a semi-supervised network. They pre-trained the fusion module in a supervised manner and learned the adaptation module in an unsupervised manner. Liu et al. [39] designed a model-inspired deep network for HSI super-resolution in an unsupervised manner and an additional unsupervised network to estimate the point spread function and spectral response function.

Over the past two decades, component substitution based methods [40], multiresolution analysis based methods [41], and sparse representation based methods [42] have been developed to enhance the spatial resolution of multispectral images. These pansharpening methods can be generalized to fuse HSIs and MSIs after some modifications. For example, assigning to each hyperspectral band, whose enhancement is separately performed, a single channel of the multispectral data [43], or dividing the spectrum of hyperspectral data into several regions and fusing hyperspectral and multispectral images in each region using conventional pan-sharpening techniques. Recently, Selva et al. [44] proposed a framework, called *hypersharpening*, that utilizes a weighted combination of all the multispectral bands for the spatial improvement of each hyperspectral band, achieving significantly better fusion results than simply selecting a band from multispectral images.

2.2. Applications of transformer in hyperspectral images

Although various CNN-based fusion methods have been derived to fuse hyperspectral and multispectral images, the limited receptive field of CNNs implies that they are not good at extracting global information from the image. However, since HSI is highly correlated in spectral dimension, obtaining global features of the spectra is crucial for improving the fusion performance. Since 2020, thanks to the self-attention mechanism to obtain long-range information, Transformer has begun to shine in the CV field: image classification (ViT) [16], target detection (DETR) [45], semantic segmentation (SETR) [46], image generation (GANsformer) [47], etc. In the field of HSI processing, Transformer has proved its advantages in processing sequential data. For example, Hong et al. [48] proposed a backbone network (SpectralFormer) for HSI classification, which is capable of learning spectrally local sequence information from neighboring bands of HSIs, yielding group-wise spectral embeddings. He et al. [49] proposed a spatial–spectral Transformer (SST) classification network, which used a well-designed CNN to extract the spatial features, and a modified Transformer to capture sequential spectra relationships. Selen and Esra [50] proposed a spectral-swin transformer (SpectralSWIN) classification network which makes use of a swin-spectral module to process the spatial and spectral features concurrently. Transformers have also been gradually applied in the field of HSI reconstruction. For example, Cai et al. [51] proposed the first Transformer-based HSI reconstruction method, which uses the feature map of each spectral channel as a token to calculate self-attention. Bandara and Patel [52] used the self-attention mechanism of the Transformer to transfer high-resolution textural to low-resolution features for pan-sharpening. Wang et al. [53] proposed a convolution and contextual Transformer (CCoT) block to simultaneously utilize the inductive bias ability of convolution and the powerful modeling ability of Transformers to restore the reconstruction details. Recently, a novel Transformer-based fusion network, namely Fusformer [54], was proposed to reconstruct HR-HSI from an LR-HSI and an HR-MSI, achieving state-of-the-art fusion performance.

3. Multiscale Spatial–spectral Transformer

3.1. The network architecture

Our network is a two-branch structure with two shallow feature extraction modules, two kinds of deep feature extraction modules, and an image reconstruction module. The overall architecture of our proposed model is shown in Fig. 1(a). The shallow feature extraction module contains a convolution layer to extract shallow features. Two kinds of deep feature extraction modules are the spectral and spatial feature extraction modules. The image reconstruction module contains two convolution layers and a Gaussian Error Linear Unit (GELU) activation function between the convolution layers.

Let $\mathcal{Y} \in \mathbb{R}^{h\times w\times S}$ denote the observed LR-HSI, where h, w, and S are the numbers of rows, columns, and spectral bands in the LR-HSI. Let $\mathcal{Z} \in \mathbb{R}^{H\times W\times s}$ denote the observed HR-MSI, where H, W, and s are the numbers of rows, columns, and spectral bands in the HR-MSI. First, we upsample the LR-HSI and downsample the HR-MSI to obtain $\mathcal{Y}_{up} \in \mathbb{R}^{H\times W\times S}$ and $\mathcal{Z}_{down} \in \mathbb{R}^{h\times w\times s}$ using the bilinear interpolation method considering the trade-off between the performance and the processing speed, which can be written as

$$\mathcal{Y}_{up} = \mathrm{Up}(\mathcal{Y}), \tag{1}$$

$$\mathcal{Z}_{down} = \mathrm{Down}(\mathcal{Z}), \tag{2}$$

where $\mathrm{Up}(\cdot)$ and $\mathrm{Down}(\cdot)$ denote the bilinear interpolation upsampling and downsampling functions, respectively. Then, we concatenate $\mathcal{Y}$ and $\mathcal{Z}_{down}$ to get $\mathcal{Y}_{cat} \in \mathbb{R}^{h\times w\times(S+s)}$, and concatenate $\mathcal{Z}$ and $\mathcal{Y}_{up}$ to get $\mathcal{Z}_{cat} \in \mathbb{R}^{H\times W\times(s+S)}$, which can be written as

$$\mathcal{Y}_{cat} = \mathrm{Concat}(\mathcal{Y}, \mathcal{Z}_{down}), \mathcal{Z}_{cat} = \mathrm{Concat}(\mathcal{Z}, \mathcal{Y}_{up}), \tag{3}$$

where $\mathrm{Concat}(\cdot)$ represents the concatenation in the channel dimension. The cross-modality concatenations help to achieve cross-modality information interaction between the two branches.

Since convolution is a simple yet effective way to map the image to a higher dimensional feature space, we adopt a 2-D convolution with kernel size = 3, channel number $C = 64$, and strides = 1 to form a residual network with blocks = 5 to extract shallow features $\mathcal{F}_s \in \mathbb{R}^{h\times w\times C}$ and $\mathcal{F}'_s \in \mathbb{R}^{H\times W\times C}$ as

$$\mathcal{F}_s = \mathrm{SFE}(\mathcal{Y}_{cat}), \quad \mathcal{F}'_s = \mathrm{SFE}(\mathcal{Z}_{cat}), \tag{4}$$

where $\mathrm{SFE}(\cdot)$ denotes the shallow feature extraction module. Then, three deep spectral and spatial feature extraction modules are employed in the network to extract multiscale deep features, respectively. That is, we set the number of deep spectral and spatial feature extraction modules L to 3 in this paper. Each deep spectral feature extraction module contains a band embedding layer, J Spectral Transformers (SpeT), and a convolution layer used in the shallow feature extraction module (Fig. 1(b)). Each deep spatial feature extraction module contains a patch embedding layer, K Spatial Transformers (SpaT), and a convolution layer used in the shallow feature extraction module (Fig. 1(c)). The J and K are both set to 5. Before $\mathcal{F}_s$ is fed into SpeT, we first obtain the band embedding of $\mathcal{F}_s$. For the lth deep spectral feature extraction modules, the band embedding is denoted as $\mathbf{B}_l^0 \in \mathbb{R}^{c\times D_{spe}}$, where $c = 16\times 2^{l-1}$ is the number of channels, D_{spe} is set to 32. Similarly, before $\mathcal{F}'_s$ is fed into SpaT, we first obtain the patch embedding of $\mathcal{F}'_s$. The patch embedding of the lth deep spatial feature extraction module is denoted as $\mathbf{P}_l^0 \in \mathbb{R}^{N\times D_{spa}}$, where $p = 8\times 2^{l-1}$ is the size of each patch, $N = \frac{H\times W}{p^2}$ is the number of patches, and D_{spa} is set to 256. In order to retain positional information, we need to add position embeddings $\mathbf{B}_l^{pos} \in \mathbb{R}^{c\times D_{spe}}$ and $\mathbf{P}_l^{pos} \in \mathbb{R}^{N\times D_{spa}}$, obtained as in [16], to the band embedding and patch embedding, respectively. Next, we extract spectral and spatial intermediate features $\mathbf{B}_l^j (j = 1, \ldots, J)$ and $\mathbf{P}_l^k (k = 1, \ldots, K)$ using SpeTs and SpaTs, respectively, as follows.

$$\mathbf{B}_l^j = \mathrm{SpeT}_j(\mathbf{B}_l^{j-1}), \quad j = 1 \ldots J, \tag{5}$$

$$\mathbf{P}_l^k = \mathrm{SpaT}_k(\mathbf{P}_l^{k-1}), \quad k = 1 \ldots K, \tag{6}$$

where $\mathrm{SpeT}_j(\cdot)$ and $\mathrm{SpaT}_k(\cdot)$ represent the jth SpeT and kth SpaT, respectively. Finally, we reshape $\mathbf{B}_l^j$ and $\mathbf{P}_l^k$ into 3-D matrice and feed them into the convolution layer to obtain the output deep features $\mathcal{F}_l^{spe} \in \mathbb{R}^{h\times w\times C}$ and $\mathcal{F}_l^{spa} \in \mathbb{R}^{H\times W\times C}$. Then, the extracted multiscale features are aggregated using learnable weights to obtain $\mathcal{F}_{sum}^{spe} \in \mathbb{R}^{h\times w\times C}$ and $\mathcal{F}_{sum}^{spa} \in \mathbb{R}^{H\times W\times C}$.

The low-level information in an image can easily be extracted by the shallow feature extraction modules. The deep feature extraction modules focus on extracting high-level information to represent semantic content. So we aggregate shallow and deep features to obtain spectral feature $\mathcal{F}^{spe} \in \mathbb{R}^{h\times w\times C}$ and spatial feature $\mathcal{F}^{spa} \in \mathbb{R}^{H\times W\times C}$ using a long skip connection, which can help the network to retain low-level information and high-level information at the same time. To ensure the feature size is the same before reconstructing the image, we use a sub-pixel convolution layer [55] to upsample the spectral feature $\mathcal{F}^{spe}$ to obtain $\mathcal{F}_{up}^{spe} \in \mathbb{R}^{H\times W\times C}$, and concatenate $\mathcal{F}_{up}^{spe}$ and $\mathcal{F}^{spa}$ to get spatial–spectral feature $\mathcal{F} \in \mathbb{R}^{H\times W\times 2C}$. Then, spatial–spectral feature $\mathcal{F}$ is fed into the image reconstruction module to obtain the estimated HR-HSI $\mathcal{X}' \in \mathbb{R}^{H\times W\times S}$. The procedures can be expressed as

$$\mathcal{F}_{up}^{spe} = \mathrm{SpConv}(\mathcal{F}_{sum}^{spe} + \mathcal{F}_s), \tag{7}$$

$$\mathcal{F}^{spa} = \mathcal{F}_{sum}^{spa} + \mathcal{F}'_s, \tag{8}$$

$$\mathcal{F} = \mathrm{Concat}(\mathcal{F}_{up}^{spe}, \mathcal{F}^{spa}), \tag{9}$$

$$\mathcal{X}' = \mathrm{IR}(\mathcal{F}), \tag{10}$$

where $\mathrm{SpConv}(\cdot)$ denotes the sub-pixel convolution layer, and $\mathrm{IR}(\cdot)$ denotes the function of the image reconstruction module. Finally, we optimize the parameters of the network by minimizing the ℓ_1 pixel loss

$$\mathcal{L} = \left\| \mathcal{X}' - \mathcal{X} \right\|_1, \tag{11}$$

where $\mathcal{X} \in \mathbb{R}^{H\times W\times S}$ is ground truth HR-HSI.

3.2. Spectral multi-head self-attention

In the spectral Transformer, the spectral self-attention is specially designed by calculating self-attention in the spectral dimension to obtain the correlation among the spectra, as shown in Fig. 2. First, we input $\mathbf{B}_l^j$ to obtain query matrix $\mathbf{Q} \in \mathbb{R}^{hw\times D_{spe}}$, key matrix $\mathbf{K} \in \mathbb{R}^{hw\times D_{spe}}$ and value matrix $\mathbf{V} \in \mathbb{R}^{hw\times D_{spe}}$ by a trainable linear projection as

$$\mathbf{Q} = \mathbf{B}_l^j \mathbf{W}_\mathbf{Q}, \mathbf{K} = \mathbf{B}_l^j \mathbf{W}_\mathbf{K}, \mathbf{V} = \mathbf{B}_l^j \mathbf{W}_\mathbf{V}, \tag{12}$$

where $\mathbf{W}_\mathbf{Q}$, $\mathbf{W}_\mathbf{K}$, and $\mathbf{W}_\mathbf{V} \in \mathbb{R}^{D_{spe}\times D_{spe}}$ are learnable projection matrices. Then, the scaled dot-product attention function with the query, key, and value matrices as input is defined as

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{V}\left(\mathrm{softmax}\left(\frac{\mathbf{K}^T\mathbf{Q}}{\sqrt{D_{spe}}}\right)\right), \tag{13}$$

where $\mathrm{Attention}(\cdot)$ is the scaled dot-product attention function. The multi-head attention mechanism, which is also used in ViT, is employed to enhance the feature extraction ability of the network. However, unlike ViT splits $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ in the spectral dimension, we split them into M^2 heads in the spatial domain, where M is set to 2 for all the datasets. The function of spectral multi-head self-attention (SpeMSA) is formulated as:

$$\mathbf{head}_m = \mathrm{Attention}(\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m), \quad m = 1 \ldots M^2, \tag{14}$$

$$\mathrm{SpeMSA}(\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m) = \mathrm{Concate}_{m=1}^{M^2}\left(\mathbf{head}_m\right)\mathbf{W}, \tag{15}$$

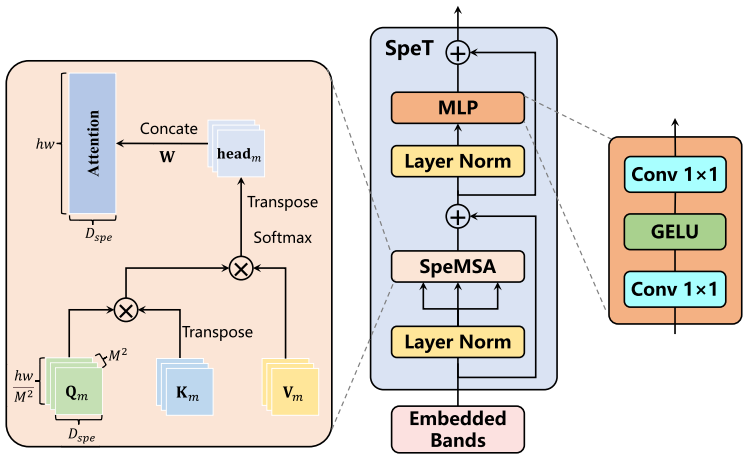

Fig. 2. The architecture diagram of the proposed spectral Transformer.

where $\mathbf{Q}_m$, $\mathbf{K}_m$, $\mathbf{V}_m$ are the mth query matrix, key matrix, and value matrix obtained by a trainable linear projection, respectively. SpeMSA(·) is the spectral multi-head self-attention function, $\mathbf{head}_m \in \mathbb{R}^{h_d w_d \times D_{spe}}$ represents the mth head and $\mathbf{W} \in \mathbb{R}^{D_{spe} \times D_{spe}}$ is a projection matrix with learnable parameters.

3.3. Multiscale embeddings

3.3.1. Multiscale band embeddings

The SpeT is proposed to extract the strong correlation from the HSI. To enhance the ability of the SpeT on excavating features, we introduce multiscale band embeddings to extract multiscale spectral features, which are then fused to enhance the details of the final extracted features, as shown in Fig. 1(a). To achieve this goal, we use a convolution operator with kernel size = 3 and channel number c to output band embeddings with different scales. Then, we feed the band embeddings with different scales into the SpeTs to extract multiscale spectral features. Finally, by aggregating the extracted multiscale spectral features using learnable weights [56,57], we can obtain $\mathcal{F}_{sum}^{spe}$ as follows,

$$\mathcal{F}_{sum}^{spe} = \sum_{l=1}^{L} \left(w_l^{spe} \mathcal{F}_l^{spe} \right) \text{ s.t.}, \sum_{l=1}^{L} w_l^{spe} = 1, w_l^{spe} > 0, \tag{16}$$

where w_l^{spe} represents the learnable weights of spectral features extracted by the lth deep spectral feature extraction.

3.3.2. Multiscale patch embeddings

We use SpaTs to extract spatial features from the HR-MSIs. The SpaT shares the same structure as ViT, which splits an image into patches with only a fixed size and provides the sequence of linear embeddings of these patches. Since the fineness of the features obtained under patch embeddings of different scales can be quite different, we introduce multiscale patch embeddings to enrich the extracted features, as shown in Fig. 1(a). To this end, we first split an image into patches with different sizes p and provide the sequence of linear embeddings of these patches, as done in [16]. Then, we input the patch embeddings with different scales into the SpaTs to capture multiscale deep spatial features. Lastly, we aggregate the extracted multiscale deep spatial features using learnable weights [56,57], which can be expressed as

$$\mathcal{F}_{sum}^{spa} = \sum_{l=1}^{L} \left(w_l^{spa} \mathcal{F}_l^{spa} \right) \text{ s.t.}, \sum_{l=1}^{L} w_l^{spa} = 1, w_l^{spa} > 0, \tag{17}$$

where w_l^{spa} represents the learnable weights of spatial features extracted by the lth deep spatial feature extraction.

3.4. Self-supervised pre-training

Training the network from scratch requires a huge amount of time and data. Therefore, we hope the network can be trained with a better initialization. In other words, we need a pre-training network that can quickly obtain better results when performing similar tasks next time.

Generally, pre-training learning can be divided into supervised learning and unsupervised learning. Self-supervised learning is an intermediate form between supervised and unsupervised learning. It uses unlabeled data sets in the pre-training stage and has produced promising results in various applications [17,58]. Therefore, we pre-trained our model in a self-supervised learning manner.

3.4.1. Masked patches autoencoder

To do self-supervised pretraining, He et al. [17] proposed an MAE with an asymmetric encoder–decoder structure, in which the decoder adopts ViT [16] to reconstruct the random lost patches from the unmask parts of the input image. Following the idea of ViT, the MAE encoder first encodes the patches through linear projection, followed by the addition of position information, and feeds them into a stack of continuous Transformer blocks. Nevertheless, unlike standard ViT, the encoder of MAE only needs to run on visible patches, which enables MAE to train a very large encoder.

MAE has a simple structure with robust scalability and good generalization. The representation learned by MAE can be well generalized to the downstream tasks. In this paper, we propose a masked patches autoencoder (MPAE) by removing the class token of MAE. Besides, we set MPAE as a symmetric encoder–decoder structure, and the masked patches are set as the learnable tokens. Finally, we keep the patch embeddings (Fig. 3) that have learned the spatial features of HR-MSI via the encoder for subsequent fine-tuning.

Since MPAE pre-trains the network by reconstructing the randomly masked patches from the unmasked patches of the input HR-MSIs. The patch masking ratios of HR-MSIs at the pre-training stage will have a significant influence on the final fusion performance. Fig. 4 presents the reconstructed PSNR values as a function of masking ratios of HR-MSIs using the CAVE dataset with a downsampling ratio of 8. From Fig. 4, we can see that the best PSNR value appears when the masking ratio is 50%. Thus, we set the masking ratios of HR-MSIs to 50% for all the datasets.

3.4.2. Masked bands autoencoder

In MPAE, the slicing of the patches is performed in the spatial dimension, so the encoder of MPAE can effectively obtain the spatial representation of HR-MSI. To effectively obtain the spectral features from LR-HSI, we further propose a masked spectral band autoencoder (MBAE), as shown in Fig. 5. Similarly, the proposed MBAE has a

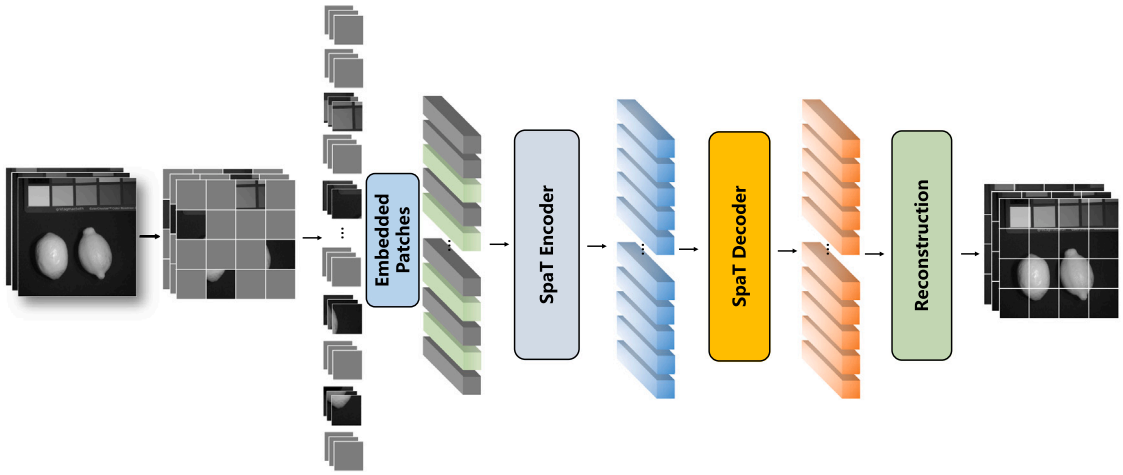

Fig. 3. The architecture diagram of the masked patches autoencoder.

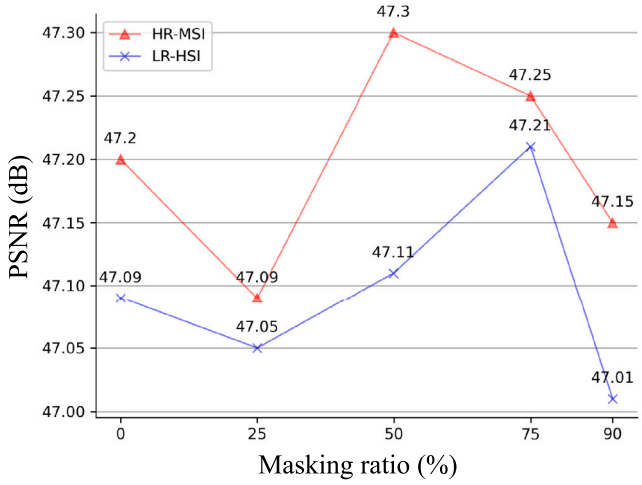

Fig. 4. Reconstructed PSNR values as a function of the masking ratios of HR-MSIS and LR-HSIs on the CAVE dataset with a downsampling ratio equal to 8.

symmetrical encoder–decoder structure. However, different from MPAE which divides the image into patches in the spatial domain, the proposed MBAE extracts spectral features by operating the data across the spectral domain to better extract spectral features. More specifically, we randomly mask some spectral bands of the input images and then use the decoder to reconstruct the masked spectral bands.

The encoder of MBAE adopts the spectral Transformer, whose inputs are unmasked spectral bands and masked spectral bands that are set as the tokens. We first embed spectral bands by a linear projection. Then, the embedded bands with added positional embeddings are fed into a series of Transformer layers to learn the global spectral features. At the decoding stage, a series of Transformer layers of the decoder are used to reconstruct the representation of each masked spectral band.

Similarly, MBAE pre-trains the network by reconstructing the randomly masked bands from the unmasked bands of the input HR-MSIs. To investigate the influence of the band masking ratios of LR-HSIs on the final fusion performance, we also present in Fig. 4 the reconstructed PSNR values with different masking ratios of LR-HSIs using the CAVE dataset with a downsampling ratio of 8. It can be observed from Fig. 4 that the best PSNR value appears at the masking ratio of 75%. Thus, we set the masking ratios of the LR-HSIs to 75% for all the datasets.

3.5. Fine-tuning

The purpose of fine-tuning is to apply the pre-trained model to the subsequent image fusion task. We first updated the parameters of our network using the pre-trained encoders, and then end-to-end fine-tuned the whole network. Fine-tuning with patches larger than pre-training is usually more beneficial [59,60]. In order to further improve the ability of the pre-trained Transformer encoder to extract the spectral features of LR-HSI and the spatial features of HR-MSI, we used patches with larger sizes than pre-training for the end-to-end fine-tuning.

4. Experiments

4.1. Datasets

In this paper, we employ two benchmark datasets and two remote sensing datasets for evaluation. The Columbia computer vision laboratory (CAVE) dataset [61], the Harvard dataset [62], and the Washington DC Mall (WDCM) dataset [63] are used for simulations. The Yellow River Estuary (YRE) dataset [64] is a full resolution dataset used for full resolution experiments.

The CAVE dataset contains 32 indoor HSIs. Each HSI has a dimension of 512 × 512 pixels with 31 spectral bands. The images were acquired at 10 nm wavelength intervals in the range from 400 nm to 700 nm. We use the first 22 HSIs for training, five HSIs for validation, and the last five HSIs for testing.

The Harvard dataset includes 50 HSIs of both indoor and outdoor scenes, featuring a diversity of objects in daylight illumination. Each HSI has 31 spectral bands whose wavelengths range from 420 nm to 720 nm. The size of each HSI in this dataset is 1040 × 1392 pixels. We use the first 34 HSIs for training, eight HSIs for validation, and the last eight HSIs for testing.

The WDCM dataset is a remote sensing HSI captured by the Hydice sensor over the National Mall in Washington, DC, in 1995. It has 191 bands covering the wavelength range from 400 nm to 2400 nm. Each band of HSI contains 1280 × 307 pixels with a spatial resolution of 2.5 m. We cropped two sub-images with a size of 128 × 128 pixels on the bottom left corner for validation and testing, respectively. The rest of the dataset is used for training.

The YRE dataset is a full resolution dataset. It contains a remote sensing HSI captured by the advanced hyperspectral imager (AHSI) aboard the GaoFen-5 satellite and a remote sensing MSI captured by the multispectral imager (MSI) aboard the Sentinel-2 A satellite. The HSI contains 280 bands with wavelength ranges covering 400 nm to 2500 nm. Each band has a size of 1400 × 1400 pixels with a spatial resolution of 30 m. The MSI contains four bands with wavelength ranges from 430 nm to 680 nm. Each band has 4200 × 4200 pixels in size with a spatial resolution of 10 m.

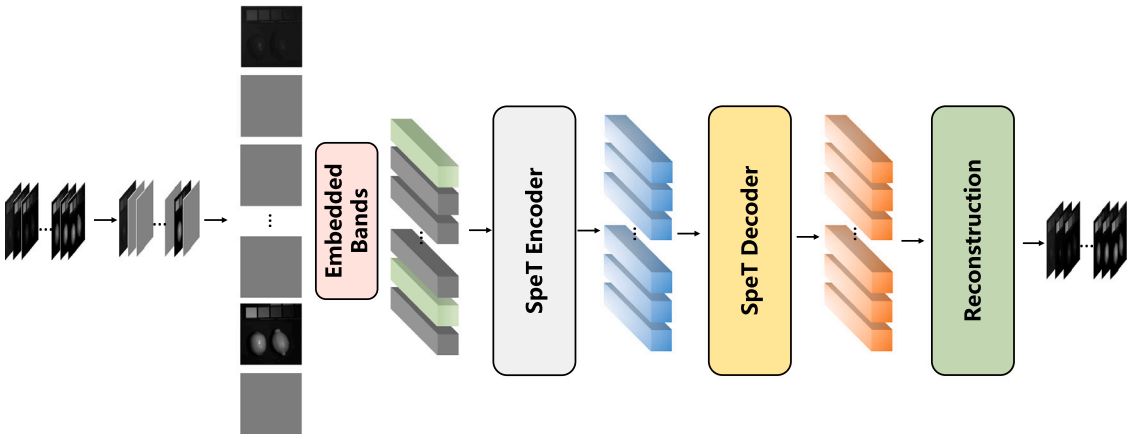

Fig. 5. The architecture diagram of the masked bands autoencoder.

4.2. Experimental settings

For the CAVE, Harvard, and WDCM datasets, we need to simulate LR-HSIs and HR-MSIs from HR-HSIs. We followed Wald's protocol to simulate the observed LR-HSIs and HR-MSIs [14,37,63,65]. First, we applied a Gaussian filter to the HR-HSI of the CAVE, Harvard, and WDCM datasets, respectively, to generate the blurred HSIs. Then, we generated the LR-HSIs by downsampling the blurred HSIs with ratios of 4 and 8, respectively, to simulate different spatial resolutions. For the CAVE and Harvard datasets, the HR-MSIs with three bands were generated using the given spectral response matrix of Nikon D700 [14,33,37]. For the WDCM dataset, the HR-MSI with ten bands was generated using the spectral response matrix of the Sentinel-2 A instrument [63]. As to the real dataset, YRE, we generated the training samples after downsampling the observed HSI and MSI with a factor of 3, as done in [66,67]. The original HSI is regarded as the ground truth. After training, we fused the original HSI and MSI to estimate the HR-HSI using the trained model.

We first input HR-MSIs into MPAE and LR-HSIs into MBAE for self-supervised pre-training and then loaded the parameters of the two pre-trained encoders into the proposed network for end-to-end fine-tuning. Since the generalization of the network is of crucial importance for deep learning based methods, we pre-trained the network using only the CAVE dataset, and then fine-tuned the network using the CAVE dataset and the Harvard dataset, respectively. Thus, the experiments on the Harvard dataset can be viewed as a test for the network's generalization. The optimizer used was Adaptive Moment Estimation (AdamW). The learning rates in the pre-training stage and the end-to-end fine-tuning stage were set to 1.0e−3 and 1.0e−4, respectively. Batch size and epoch were set to 32 and 5000. In addition, in the pre-training, the HR-MSIs of the training set were cropped into patches of 128×128 pixels in size, so the LR-HSIs of the training set were cropped into patches of size $\frac{128}{r} \times \frac{128}{r}$ pixels, where r is the downsampling ratio. In the end-to-end fine-tuning, the HR-MSIs were cropped into patches of size 192×192 pixels, and the LR-HSIs were cropped into patches of size $\frac{192}{r} \times \frac{192}{r}$ pixels.

To effectively evaluate the performance of the proposed method, we introduce seven state-of-the-art fusion methods for comparison, including two traditional methods, i.e., FUSE [8] and CNMF [24], four CNN-based methods, i.e., DBIN [66], MHF-Net [32], UAL [38], and SSR-NET [14], and a newly proposed Transformer-based method, Fusformer [54]. The parameters in different compared methods were set based on either authors' codes or suggestions in the reference articles. The two traditional methods were tested in MATLAB (R2013a) on Windows Server 2012 with two Intel Xeon E5-2650 processors and 128-GB RAM, the deep learning based methods were tested by Pytorch 1.10.0 on Python 3.7 using a GPU of NVIDIA A40.

4.3. Evaluation metrics

Five popular indexes are used in this paper to fully evaluate the quality of the reconstructed HR-HSI at reduced resolution. They are given in detail in the following.

- Peak signal-to-noise ratio (PSNR): PSNR is an objective evaluation index used to evaluate the noise level or image distortion. The higher the value of PSNR, the less distortion and the better quality of the estimated image. Its calculation formula is as follows:

$$\mathrm{PSNR}(\mathcal{X}, \mathcal{X}') = 10 \lg \left(\frac{\max\left(\mathbf{X}_k\right)^2}{\frac{1}{HW} \left\| \mathbf{X}_k - \mathbf{X}'_k \right\|_2^2} \right), \tag{18}$$

  where $\max(\cdot)$ is a function that returns the maximum value, $\mathcal{X}'$ is the estimated HR-HSI, $\mathcal{X}$ is the ground truth HR-HSI, $\mathbf{X}_k$ and $\mathbf{X}'_k$ denote the kth band of the reference HR-HSI and the estimated HR-HSI, respectively.
- Spectral angle mapper (SAM): SAM is a metric for estimating the spectral quality of an image. It is obtained by computing the averaged spectral angle over the entire spatial domain. The lower the value of SAM, the less spectral distortion. The optimal value is 0.

$$\mathrm{SAM}(\mathcal{X}, \mathcal{X}') = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \arccos \left(\frac{\mathbf{X}^T(i,j)\mathbf{X}'(i,j)}{\|\mathbf{X}(i,j)\|_2 \, \|\mathbf{X}'(i,j)\|_2} \right), \tag{19}$$

  where H and W are the numbers of rows and columns in the ground truth HR-HSI, $\mathbf{X}(i,j)$ and $\mathbf{X}'(i,j)$ represent the pixel vector of the reference HR-HSI and the estimated HR-HSI at position (i,j).
- Structural similarity index metric (SSIM): SSIM is used to evaluate the level of similarity between two images. The higher the value of SSIM, the better the spatial structure preservation.

$$\mathrm{SSIM}(\mathcal{X}, \mathcal{X}') = \frac{1}{S} \sum_{k}^{S} \frac{\left(2\mu_{\mathbf{X}_k}\mu_{\mathbf{X}'_k} + C_1\right)\left(2\sigma_{\mathbf{X}_k\mathbf{X}'_k} + C_2\right)}{\left(\mu_{\mathbf{X}_k}^2 + \mu_{\mathbf{X}'_k}^2 + C_1\right)\left(\sigma_{\mathbf{X}_k}^2 + \sigma_{\mathbf{X}'_k}^2 + C_2\right)}, \tag{20}$$

  where S is the number of spectral bands in the ground truth HR-HSI, C_1 and C_2 are constants, $\sigma_{\mathbf{X}_k\mathbf{X}'_k}$ denotes the covariance matrix between $\mathbf{X}_k$ and $\mathbf{X}'_k$, $\mu_{\mathbf{X}_k}$ and $\mu_{\mathbf{X}'_k}$ represent the average values, $\sigma_{\mathbf{X}_k}$ and $\sigma_{\mathbf{X}'_k}$ are the standard deviation of $\mathbf{X}_k$ and $\mathbf{X}'_k$, respectively.
- Erreur relative globale adimensionnelle de synthèse (ERGAS): ERGAS is specially designed to evaluate the overall quality of the fused images. The lower the value of ERGAS, the better the fusion

Table 1
Experimental results of different fusion methods on the CAVE, Harvard, and WDCM datasets with different downsampling ratios.

Ratio	Method	CAVE					Harvard					WDCM				
		PSNR	SAM	SSIM	ERGAS	RMSE	PSNR	SAM	SSIM	ERGAS	RMSE	PSNR	SAM	SSIM	ERGAS	RMSE
8	FUSE	37.53	2.93	0.9927	1.3648	0.0128	37.64	3.24	0.9932	1.2363	0.0122	31.16	4.53	0.9854	1.2754	0.0211
	CNMF	38.97	2.71	0.9942	1.2663	0.0117	41.11	2.74	0.9970	0.9248	0.0053	32.23	3.88	0.9896	1.1514	0.0191
	DBIN	42.86	1.98	0.9980	0.8078	0.0067	42.25	2.59	0.9983	0.7192	0.0047	35.25	3.38	0.9953	0.9668	0.0171
	MHF-Net	43.19	1.93	0.9980	0.7635	0.0063	42.56	2.46	0.9980	0.7234	0.0046	36.92	2.76	0.9967	0.7983	0.0141
	UAL	44.49	1.75	0.9986	0.6881	0.0056	43.87	2.11	0.9986	0.6153	0.0039	38.48	2.31	0.9977	0.6666	0.0118
	SSR-NET	44.26	1.79	0.9985	0.6962	0.0058	43.59	2.09	0.9986	0.6125	0.0034	37.83	2.49	0.9973	0.7188	0.0127
	Fusformer	46.61	1.35	0.9990	0.5341	0.0044	44.31	2.04	0.9987	0.5973	0.0036	38.93	2.19	0.9979	0.6328	0.0112
	MSST-Net	**47.44**	**1.22**	**0.9991**	**0.4841**	**0.0040**	**45.39**	**1.79**	**0.9990**	**0.5212**	**0.0031**	**40.41**	**1.85**	**0.9986**	**0.5337**	**0.0095**
4	FUSE	38.41	2.79	0.9932	2.6104	0.0119	38.13	2.98	0.9952	2.2253	0.0079	32.04	4.01	0.9889	2.3754	0.0198
	CNMF	39.78	2.35	0.9955	2.2672	0.0101	42.47	2.49	0.9980	1.4953	0.0046	33.12	3.57	0.9910	2.2005	0.0178
	DBIN	43.92	1.68	0.9984	1.4041	0.0061	43.48	2.19	0.9985	1.2815	0.0040	38.86	2.78	0.9972	1.6068	0.0142
	MHF-Net	44.85	1.58	0.9985	1.2983	0.0056	43.95	2.02	0.9987	1.2110	0.0038	37.66	2.53	0.9978	1.4651	0.0129
	UAL	45.98	1.41	0.9983	1.1201	0.0046	44.65	1.95	0.9988	1.1350	0.0033	39.04	2.10	0.9980	1.2496	0.0111
	SSR-NET	45.24	1.55	0.9987	1.2443	0.0053	44.31	1.91	0.9988	1.1527	0.0032	38.80	2.22	0.9979	1.2853	0.0114
	Fusformer	47.43	1.53	0.9990	0.9635	0.0041	45.06	1.82	0.9990	1.0684	0.0033	39.89	1.95	0.9983	1.1341	0.0100
	MSST-Net	**48.37**	**1.09**	**0.9994**	**0.8624**	**0.0036**	**46.00**	**1.65**	**0.9992**	**0.9599**	**0.0029**	**41.73**	**1.59**	**0.9989**	**0.9172**	**0.0081**

result.

$$\mathrm{ERGAS}(\mathcal{X},\mathcal{X}') = \frac{100}{r}\sqrt{\frac{1}{S}\sum_{k=1}^{S}\frac{\left\|\mathbf{X}_k-\mathbf{X}'_k\right\|_2^2}{\mu^2\left(\mathbf{X}_k\right)}}, \tag{21}$$

where r is the downsampling ratio and $\mu(\cdot)$ denotes the mean value.

- Root mean squared error (RMSE): RMSE is used to represent the difference between $\mathcal{X}$ and $\mathcal{X}'$. Smaller RMSE means smaller reconstruction errors and better reconstruction quality.

$$\mathrm{RMSE}(\mathcal{X},\mathcal{X}') = \sqrt{\frac{\sum_{k=1}^{S}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(\mathbf{X}_k(i,j)-\mathbf{X}'_k(i,j)\right)^2}{HWS}}, \tag{22}$$

where $\mathbf{X}_k(i,j)$ and $\mathbf{X}'_k(i,j)$ denote the element value at position (i,j) in the kth band of the reference HR-HSI and the estimated HR-HSI, respectively.

For the full resolution experiments, since a reference image at full resolution is unavailable, we adopt two commonly used quality indexes without reference to evaluate the fusion performance at full resolution quantitatively [68]. Detailed descriptions of the two indexes are given in the following.

- Quality with No Reference (QNR): QNR is the product of one's complements of the spatial and spectral distortion indices [69]. It is calculated as:

$$\mathrm{QNR} \triangleq \left(1-D_\lambda\right)^\alpha \cdot \left(1-D_s\right)^\beta, \tag{23}$$

where D_λ and D_s denote the spectral distortion index and spatial distortion index, respectively, as defined in [69]. α and β are two real-valued exponents that attribute the relevance of spectral and spatial distortions to the overall quality. The two exponents jointly determine the non-linearity of response in the interval [0,1]. The highest value of QNR is one and is obtained when the spectral and spatial distortions are both zero.

- Hybrid Quality with No Reference (HQNR): HQNR is a unique quality index that combines the spatial distortion of the QNR protocol and the spectral distortion of Khan's protocol [70]. The calculation formula of HQNR is:

$$\mathrm{HQNR} \triangleq \left(1-D_\lambda^{(K)}\right) \cdot \left(1-D_s\right), \tag{24}$$

where $D_\lambda^{(K)}$ is the spectral distortion of Khan's protocol, as defined in [70].

4.4. Experimental results

Table 1 shows the average quantitative results on the CAVE, Harvard, and WDCM datasets. The optimal results are marked in bold for clarity. From Table 1, we can clearly see that the two traditional methods, FUSE and CNMF fail to produce desired results as deep learning based methods on all three datasets. All the deep learning based methods obtain comparable fusion performance on the three datasets. When the downsampling ratio is 8, Fusformer achieves the second-best fusion performance on the CAVE and WDCM datasets, and SSR-NET ranks second on the Harvard dataset. When the downsampling ratio is 4, UAL, SST-NET, and Fusformer rank second on the CAVE, Harvard, and WDCM datasets, respectively. Based on the superior ability of the Transformer to capture long-term information, Fusformer can produce better or more competitive fusion results in comparison with other CNN-based methods on all three datasets. Compared with Fusformer, Our method significantly outperforms all the competitors on three datasets with different downsampling ratios. Different from Fusformer, which uses a Transformer to extract spatial and spectral features simultaneously and obtain features at a single scale, the proposed MSST-Net uses multiscale spectral and spatial Transformers to extract features from different modalities, considering the high spectral correlation of the HSI and rich spatial information of the MSI. The superior fusion performance of the proposed method can be jointly attributed to the cross-modality concatenations, the elaborated multiscale spectral and spatial Transformers, and the self-supervised pre-training strategy.

To further evaluate the reconstruction performance of each band, we show the PSNR values of each reconstructed band for the three tested datasets in Fig. 6. It can be clearly observed from Fig. 6 that the proposed method produces significantly higher PSNR values than that of the compared methods at almost all the bands, suggesting that the proposed method achieves better overall reconstruction quality than other methods on all the datasets.

In order to evaluate the quality of the fused images visually, we present some bands and their corresponding error maps of the fusion results in Figs. 7 to 9. The error maps are obtained by calculating the absolute difference between the ground truth HR-HSI and the estimated HR-HSI. Fig. 7 shows the 21st estimated bands and their corresponding error maps of the CAVE dataset with different downsampling ratios. Obviously, the two traditional methods, i.e., FUSE and CNMF, fail to produce competitive results at this band, which can be easily observed from the estimated bands and the error maps. Deep learning based methods can produce estimated bands with very similar visual effects, but the fewer residuals left in error maps of the proposed method as indicated in the red boxes suggest that the proposed method can retain more spatial details in the estimated bands. The 26th estimated band and its corresponding error maps of the Harvard dataset with different downsampling ratios are shown in Fig. 8. It can be observed from Fig. 8 that all the methods produce competitive results at different downsampling ratios. However, the proposed method can produce results with higher quality, since there are fewer residuals remaining

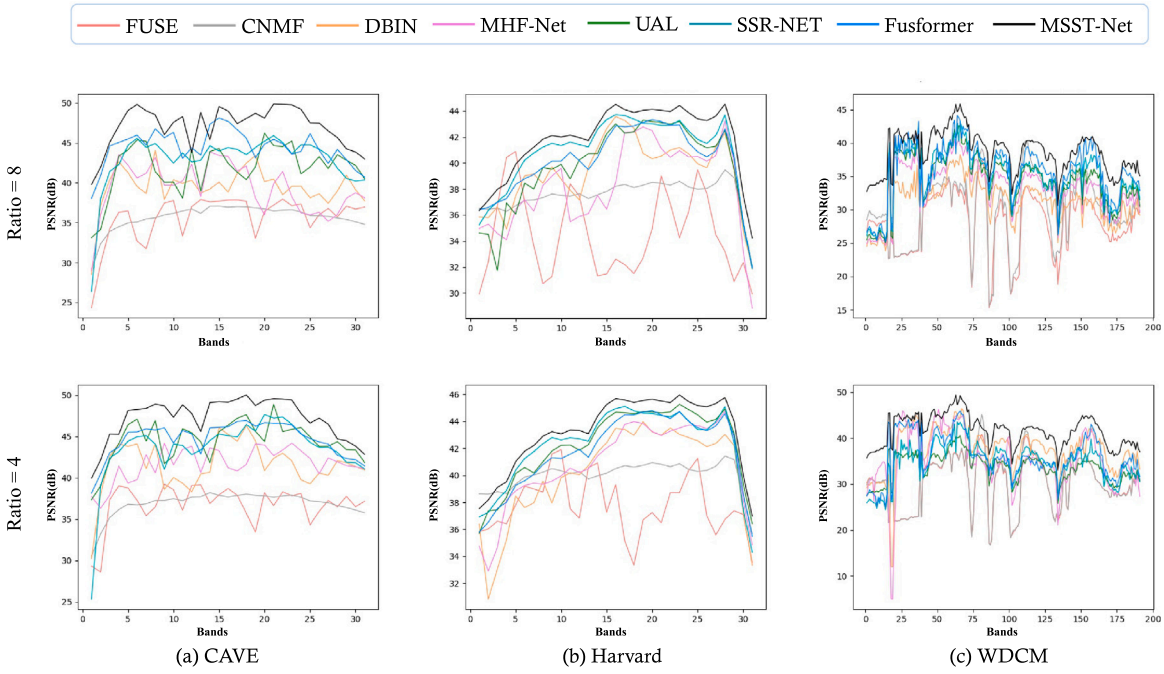

Fig. 6. The band-wise PSNR values for the fake and real beer's image of the CAVE dataset, the imgb6 image of the Harvard dataset, and the WDCM dataset with downsampling ratios equal to 8 and 4.

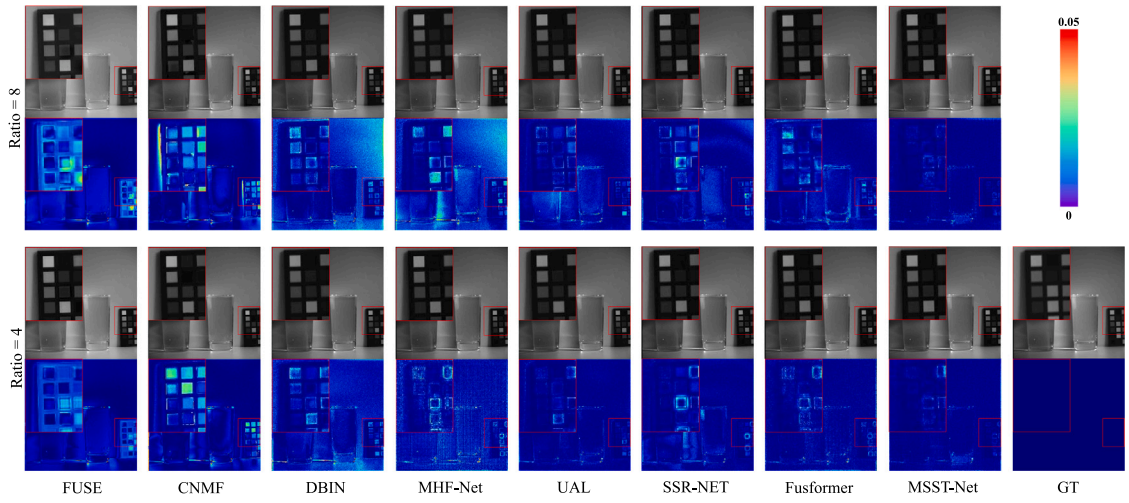

Fig. 7. The 21st band of the fusion images and the reconstruction error maps on the fake and real beer's image of the CAVE dataset with downsampling ratios equal to 8 and 4.

in the error maps of the proposed method, in particular in the areas marked by the red boxes. We present the 14th estimated band and its corresponding error maps of the WDCM dataset in Fig. 9. In this band, all the methods fail to obtain satisfying results except for the proposed method. One of the challenges in image reconstruction is to recover the texture features of the image. The proposed method can better recover spatial textures compared with other methods, which can be obviously observed in the areas of rich textures, for example, the areas marked by the red boxes.

Note that the results of the Harvard dataset were obtained by pre-training the network on the CAVE dataset first, and then fine-tuning the network using the Harvard dataset, which can be viewed as a test of the proposed network's generalization. The proposed network achieves the best fusion performance on the Harvard dataset for all the metrics, demonstrating the good generalization of the proposed network. We believe that the excellent generalization of the proposed network comes mainly from the well-designed self-supervised pre-trained strategy.

To test the performance of the proposed method on the full resolution dataset, we use a sub-image of 576 × 576 pixels in size cropped from the YRE dataset for experiments. The reconstructed HR-HSIs generated by different methods are presented as pseudo-color images in Fig. 10. It can be seen that although the traditional methods can obtain HR-HSIs with high spatial resolutions, they will induce severe spectral distortion. In contrast, the methods based on deep learning can well preserve spectral information while improving spatial resolution. In particular, the reconstructed HR-HSI obtained by our proposed method has a better visualization effect compared with other deep learning based methods. We utilized QNR and HQNR to evaluate the fusion performance quantitatively [68]. The evaluation results are presented in Fig. 11. It can be easily observed that the proposed method achieves the best scores for both indexes, indicating the better reconstruction performance of the proposed method compared with its competitors.

To analyze the computational burden, we present the training time, testing time, floating point operations (FLOPs), and the number of

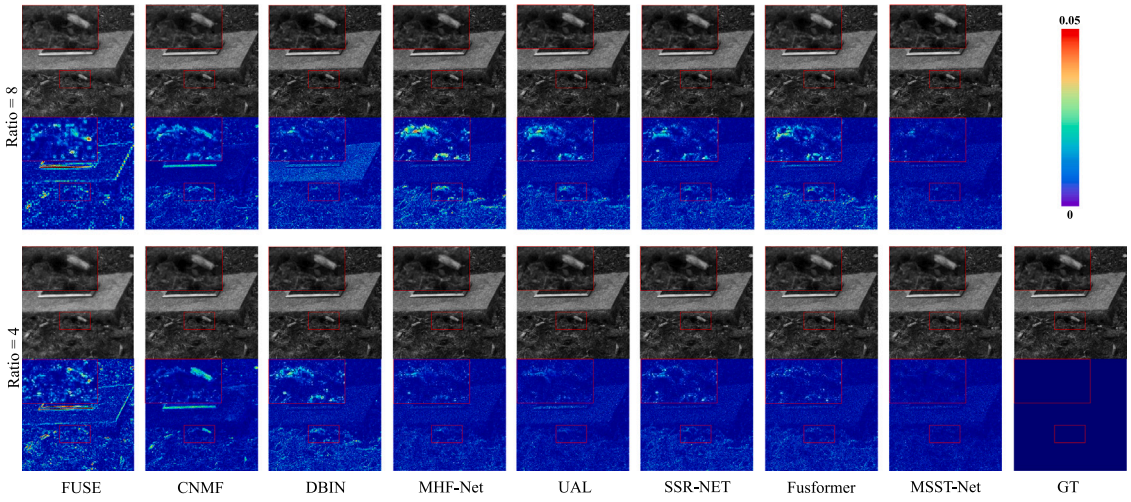

Fig. 8. The 26th band of the fusion images and the reconstruction error maps on the imgb6 image of the Harvard dataset with downsampling ratios equal to 8 and 4. .

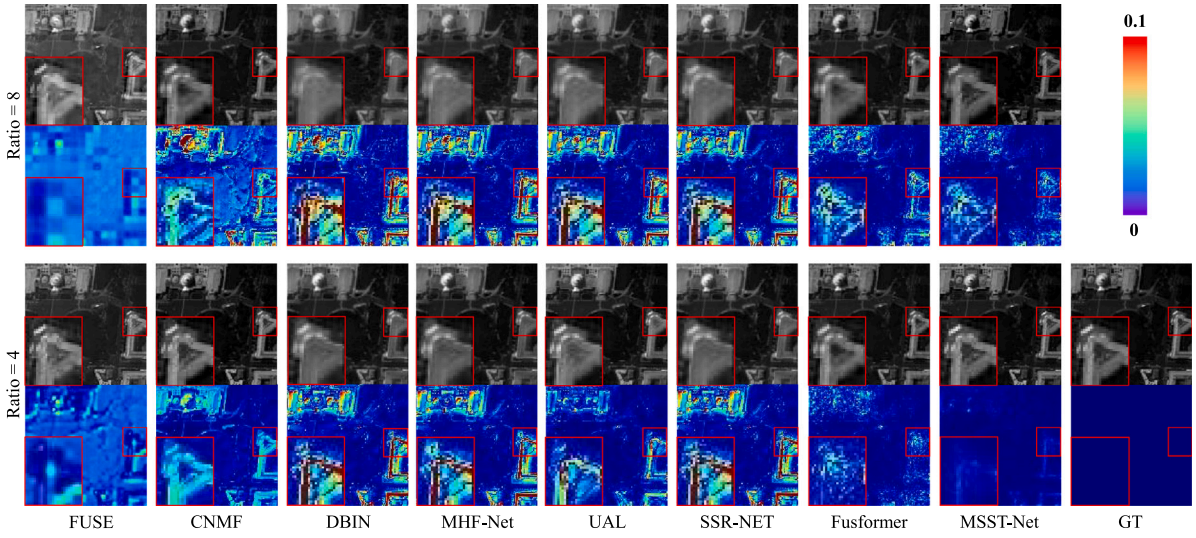

Fig. 9. The 14th band of the fusion images and the reconstruction error maps on the WDCM dataset with downsampling ratios equal to 8 and 4.

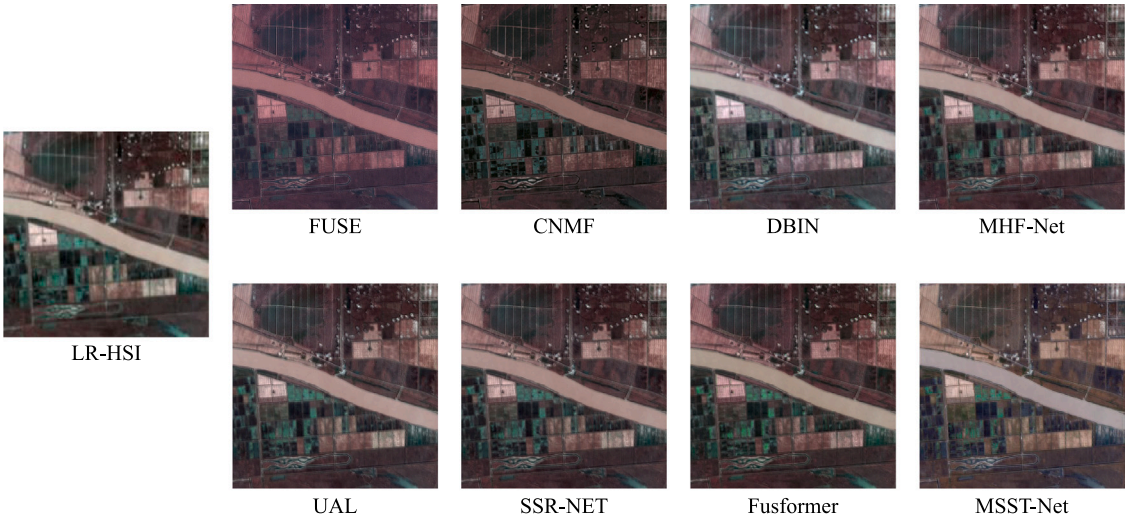

Fig. 10. The pseudo-color images (R-66, G-36, B-31) of the LR-HSI and the reconstructed HR-HSIs of the YRE dataset.

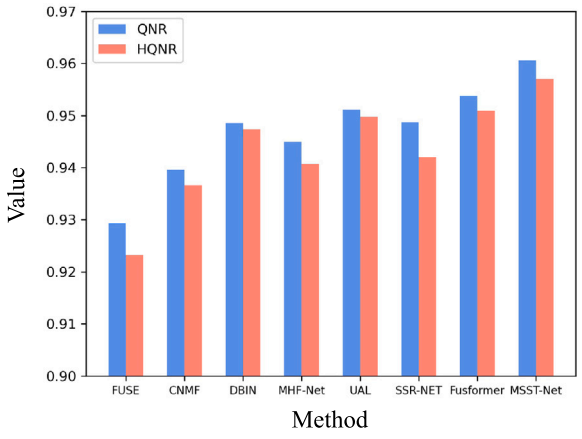

Fig. 11. No reference indexes for fusion result of the YRE dataset.

Table 2
Training time, testing time, FLOPs, and number of parameters of different fusion methods on the CAVE dataset with a downsampling ratio equal to 8.

Method	CAVE			
	Training time (s)	Testing time (s)	FLOPs (G)	Parameters (M)
FUSE	/	34.08	/	/
CNMF	/	544.32	/	/
DBIN	1.68×10^5	2.12	130.37	1.43
MHF-Net	7.37×10^4	5.17	22.54	3.63
UAL	9.56×10^4	2.41	213.30	7.10
SSR-NET	6.45×10^4	3.87	0.43	0.03
Fusformer	1.29×10^5	4.51	1.83	0.11
MSST-Net	1.10×10^5	4.06	188.72	34.40

parameters of different fusion methods on the CAVE dataset in Table 2. According to the experimental results, it can be easily seen that the testing time of traditional methods is much longer than that of deep learning based methods. Since the computation of the Transformer is time-consuming, our proposed method is not the fastest in terms of training time and testing time compared with CNN-based methods. However, our method takes less time compared to Fusformer, which takes pixel-wise tokens as input. As for FLOPs and the number of parameters, our method has relatively larger FLOPs and parameters, which is most likely due to the use of convolution with large-scale transpose in the multiscale patch embeddings.

4.5. Ablation studies

To study the influence of the main modules in our proposed network, we conducted a series of ablation studies on the CAVE dataset with a downsampling ratio of 8 by taking the result without pre-training as the baseline (the penultimate column of Table 3). To verify the effect of the cross-modality concatenations, we conducted an experiment without cross-modality concatenations. The results are shown in the first column of Table 3. It can be seen that the PSNR without cross-modality concatenations decreased by 1.87 dB compared to the baseline (the penultimate column of Table 3), and all the other metrics decreased as well, which indicates that the cross-modality information interaction between the dual branches could help to improve the fusion performance.

SpeTs and SpaTs are introduced to extract spectral and spatial information due to their strong ability to capture long-distance dependencies. In the SpeTs and SpaTs, spectral and spatial multi-head self-attentions are specially designed to capture more detailed spectral and spatial features, respectively. We first performed an ablation experiment to verify the effectiveness of the SpeTs by replacing the SpeTs with the SpaTs. The results are shown in the second column of Table 3. It can be clearly seen that the results are significantly worse than the baseline (the penultimate column of Table 3) in all metrics, which indicates that the spectral Transformers are more powerful in extracting spectral features. At the same time, SpaTs were replaced with SpeTs to verify the effectiveness of SpaTs in extracting spatial information. It can also be clearly seen from the third column of Table 3 that the results are significantly worse than the baseline (penultimate column of Table 3) in all metrics, which indicates that the proposed network would be weaker in extracting spatial features without SpaTs.

Multiscale band embeddings are proposed to extract the multiscale spectral features, which will be fused using learnable weights to enhance the abundance of the extracted features. We can see from Table 3 that the results of single-scale band embedding (the fourth column) are significantly worse than the results of multiscale band embeddings (the penultimate column of Table 3), which proves the effectiveness of the extracted multiscale spectral features in improving the fusion performance. Fig. 12(a)–(c) show the spatial features of an image obtained with patch embeddings at different scales, i.e., $8 \times 8, 16\times 16$, and 32×32, for qualitative assessment. It is obvious that spatial features obtained with different patch embeddings are quite different, which motivates us to use multiscale patch embeddings to extract spatial features. Fig. 12(d) and 12(e) present the fused features using fixed weights and with learnable weights, respectively. It can be easily observed that the fused feature with learnable weights contains more details than the fused feature with fixed weights. We show the quantitative results of single-scale patch embedding with a patch size of 16×16 in the fifth column of Table 3. All the metrics of single-scale path embedding decreased, to some extent, compared with the results of multiscale patch embeddings (the penultimate column of Table 3), demonstrating the effectiveness of the proposed multiscale patch embeddings. The above results demonstrate the effectiveness of extracted multiscale features in improving fusion performance.

To verify the effectiveness of the proposed pre-training strategy, we first pre-trained the SpeTs and SpaTs and then loaded the pre-trained SpeTs and SpaTs into the proposed network for fine-tuning. The experimental results are given in the final columns of Table 3. We can see that the results with self-supervised pre-training and fine-tuning are significantly improved compared with baseline metrics (the penultimate column of Table 3), demonstrating the effectiveness of the proposed pre-training strategy.

In summary, we propose a multiscale spatial–spectral Transformer Network for the fusion of hyperspectral and multispectral images. The proposed network extracts spectral and spatial information from HSIs and MSIs using SpeTs and SpaTs, respectively, in which spectral and spatial multi-head self-attentions are specially designed to obtain the strong ability to capture long-distance dependencies. Considering the limitation of existing Transformers on excavating detailed information at different levels of granularity, multiscale band/patch embeddings are proposed to take full advantage of the high spectral correlation of HSIs and rich spatial textures of the MSIs. To further improve the fusion performance and generalization of the network, an MPAE, which is specially designed to randomly mask the bands of the HSI, as well as an MPAE, are employed for the self-supervised pre-training of the SpeTs and SpaTs. The above reasons jointly contribute to the excellent fusion performance of the proposed network for hyperspectral and multispectral image fusion, which has been fully validated using a series of ablation experiments.

5. Conclusion

This paper proposes a Multiscale Spatial–spectral Transformer Network (MSST-Net) to address the HR-MSI and LR-HSI fusion task. Our MSST-Net mainly contains two shallow feature extraction modules, two kinds of deep feature extraction modules, and an image reconstruction module. The deep spectral feature extraction module contains a series of spectral Transformers working on the spectral dimension to extract

Table 3
Ablation experimental results on the CAVE dataset with a downsampling ratio equal to 8.

Cross-modality concatenations	✗	✓	✓	✓	✓	✓	✓
Spectral Transformers	✓	✗	✓	✓	✓	✓	✓
Spatial Transformers	✓	✓	✗	✓	✓	✓	✓
Multiscale band embeddings	✓	✓	✓	✗	✓	✓	✓
Multiscale patch embeddings	✓	✓	✓	✓	✗	✓	✓
Pre-training	✗	✗	✗	✗	✗	✗	✓
PSNR	45.05	45.89	46.37	46.29	46.03	46.92	**47.44**
SAM	1.64	1.43	1.37	1.38	1.41	1.32	**1.22**
SSIM	0.9985	0.9987	0.9990	0.9989	0.9988	0.9990	**0.9991**
ERGAS	0.6572	0.5965	0.5552	0.5604	0.5851	0.5344	**0.4841**
RMSE	0.0055	0.0049	0.0048	0.0046	0.0048	0.0045	**0.0040**

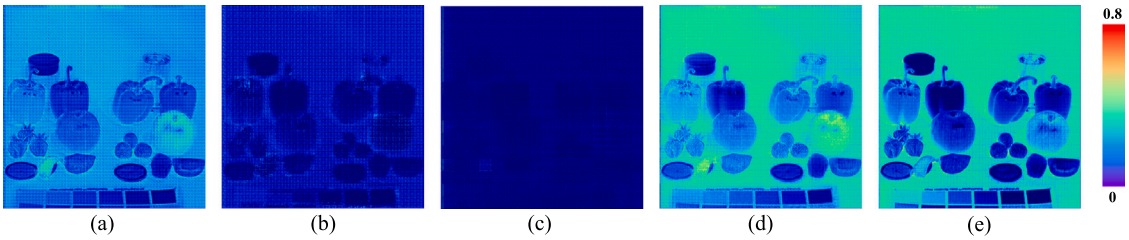

Fig. 12. Spatial feature maps of patch embeddings at different scales. (a) 8 × 8. (b) 16 × 16. (c) 32 × 32. (d) Fused feature map with fixed weights. (e) Fused feature map with learnable weights.

the spectral features of LR-HSI. The deep spatial extraction modules with different scale patch embeddings are proposed to obtain multiscale spatial features of HR-MSI. Furthermore, we propose a self-supervised pre-training strategy to further improve the fusion performance. Two autoencoders, i.e., MPAE and MBAE, are designed for self-supervised pre-training of the spatial and spectral Transformers, respectively. Extensive experiments were performed on three simulated datasets and one real dataset. The experimental results suggest that our model can achieve excellent performance compared with other state-of-the-art methods.

However, the proposed network only focuses on the fusion of well-registered image pairs. If there are large misalignments between the images, our proposed network may be unworkable. Therefore, in future research, we will focus on extending the network to be able to handle both well-registered and unregistered cases by introducing some regularizations to constrain the representations of the two modalities.

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.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62271327, 42271336, 41971300), in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022A1515011290, 2022A1515110076), and in part by the Shenzhen Science and Technology Program (Grant No. RCJC20221008092731042, JCYJ20220818100206015, KQTD20200909113951005).

References

[1] L. Zhuang, M.K. Ng, X. Fu, J.M. Bioucas-Dias, Hy-demosaicing: Hyperspectral blind reconstruction from spectral subsampling, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.
[2] J. Bian, A. Li, Z. Zhang, W. Zhao, G. Lei, G. Yin, H. Jin, J. Tan, C. Huang, Monitoring fractional green vegetation cover dynamics over a seasonally inundated alpine wetland using dense time series HJ-1A/B constellation images and an adaptive endmember selection LSMM model, Remote Sens. Environ. 197 (2017) 98–114.
[3] S. Jia, L. Shen, J. Zhu, Q. Li, A 3-D gabor phase-based coding and matching framework for hyperspectral imagery classification, IEEE Trans. Cybern. 48 (4) (2018) 1176–1188.
[4] J. Zhao, Y. Zhong, X. Hu, L. Wei, L. Zhang, A robust spectral-spatial approach to identifying heterogeneous crops using remote sensing imagery with high spectral and spatial resolutions, Remote Sens. Environ. 239 (2020) 111605.
[5] X. Fu, S. Jia, L. Zhuang, M. Xu, J. Zhou, Q. Li, Hyperspectral anomaly detection via deep plug-and-play denoising CNN regularization, IEEE Trans. Geosci. Remote Sens. 59 (11) (2021) 9553–9568.
[6] L. Zhuang, X. Fu, M.K. Ng, J.M. Bioucas-Dias, Hyperspectral image denoising based on global and nonlocal low-rank factorizations, IEEE Trans. Geosci. Remote Sens. 59 (12) (2021) 10438–10454.
[7] H. Ghassemian, A review of remote sensing image fusion methods, Inf. Fusion 32 (2016) 75–89.
[8] Q. Wei, N. Dobigeon, J.-Y. Tourneret, Fast fusion of multi-band images based on solving a sylvester equation, IEEE Trans. Image Process. 24 (11) (2015) 4109–4121.
[9] R. Dian, S. Li, L. Fang, Q. Wei, Multispectral and hyperspectral image fusion with spatial-spectral sparse representation, Inf. Fusion 49 (2019) 262–270.
[10] X. Fu, S. Jia, M. Xu, J. Zhou, Q. Li, Fusion of hyperspectral and multispectral images accounting for localized inter-image changes, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–18.
[11] S. Li, R. Dian, L. Fang, J.M. Bioucas-Dias, Fusing hyperspectral and multispectral images via coupled sparse tensor factorization, IEEE Trans. Image Process. 27 (8) (2018) 4118–4130.
[12] R. Dian, L. Fang, S. Li, Hyperspectral image super-resolution via non-local sparse tensor factorization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5344–5353.
[13] D. Shen, J. Liu, Z. Xiao, J. Yang, L. Xiao, A twice optimizing net with matrix decomposition for hyperspectral and multispectral image fusion, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13 (2020) 4095–4110.
[14] X. Zhang, W. Huang, Q. Wang, X. Li, SSR-NET: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion, IEEE Trans. Geosci. Remote Sens. 59 (7) (2020) 5953–5965.
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, 2021, arXiv preprint arXiv:2111.06377.

[18] R. Dian, S. Li, B. Sun, A. Guo, Recent advances and new guidelines on hyperspectral and multispectral image fusion, Inf. Fusion 69 (2021) 40–51.

[19] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, K. Ikeuchi, High-resolution hyperspectral imaging via matrix factorization, in: CVPR 2011, 2011, pp. 2329–2336.

[20] N. Akhtar, F. Shafait, A. Mian, Bayesian sparse representation for hyperspectral image super resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 3631–3640.

[21] M. Simões, J. Bioucas-Dias, L.B. Almeida, J. Chanussot, A convex formulation for hyperspectral image superresolution via subspace-based regularization, IEEE Trans. Geosci. Remote Sens. 53 (6) (2015) 3373–3388.

[22] R. Dian, S. Li, X. Kang, Regularizing hyperspectral and multispectral image fusion by CNN denoiser, IEEE Trans. Neural Netw. Learn. Syst. 32 (3) (2021) 1124–1135.

[23] R. Dian, S. Li, Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization, IEEE Trans. Image Process. 28 (10) (2019) 5135–5146.

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

[25] C. Lanaras, E. Baltsavias, K. Schindler, Hyperspectral super-resolution by coupled spectral unmixing, in: 2015 IEEE International Conference on Computer Vision, ICCV, 2015, pp. 3586–3594.

[26] R. Dian, S. Li, L. Fang, T. Lu, J.M. Bioucas-Dias, Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion, IEEE Trans. Cybern. 50 (10) (2020) 4469–4480.

[27] C. Prévost, K. Usevich, P. Comon, D. Brie, Hyperspectral super-resolution with coupled tucker approximation: Recoverability and SVD-based algorithms, IEEE Trans. Signal Process. 68 (2020) 931–946.

[28] C.I. Kanatsoulis, X. Fu, N.D. Sidiropoulos, W.-K. Ma, Hyperspectral super-resolution: A coupled tensor factorization approach, IEEE Trans. Signal Process. 66 (24) (2018) 6503–6517.

[29] Y. Xu, Z. Wu, J. Chanussot, P. Comon, Z. Wei, Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion, IEEE Trans. Geosci. Remote Sens. 58 (1) (2020) 348–362.

[30] R. Dian, S. Li, L. Fang, Learning a low tensor-train rank representation for hyperspectral image super-resolution, IEEE Trans. Neural Netw. Learn. Syst. 30 (9) (2019) 2672–2683.

[31] W. He, Y. Chen, N. Yokoya, C. Li, Q. Zhao, Hyperspectral super-resolution via coupled tensor ring factorization, Pattern Recognit. 122 (2022) 108280.

[32] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, Z. Xu, Multispectral and hyperspectral image fusion by MS/HS fusion net, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1585–1594.

[33] R. Dian, S. Li, A. Guo, L. Fang, Deep hyperspectral image sharpening, IEEE Trans. Neural Netw. Learn. Syst. 29 (11) (2018) 5345–5355.

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

[35] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.

[36] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, Y. Shi, J. Chanussot, Edge-conditioned feature transform network for hyperspectral and multispectral image fusion, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.

[37] Y. Qu, H. Qi, C. Kwan, Unsupervised sparse dirichlet-net for hyperspectral image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2511–2520.

[38] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, L. Shao, Unsupervised adaptation learning for hyperspectral imagery super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3073–3082.

[39] J. Liu, Z. Wu, L. Xiao, X.-J. Wu, Model inspired autoencoder for unsupervised hyperspectral image super-resolution, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–12.

[40] B. Aiazzi, S. Baronti, M. Selva, Improving component substitution pansharpening through multivariate regression of MS ++Pan data, IEEE Trans. Geosci. Remote Sens. 45 (10) (2007) 3230–3239.

[41] J.G. Liu, Smoothing filter-based Intensity Modulation: A spectral preserve image fusion technique for improving spatial details, Int. J. Remote Sens. 21 (18) (2000) 3461–3472.

[42] X.X. Zhu, R. Bamler, A sparse image fusion algorithm with application to pan-sharpening, IEEE Trans. Geosci. Remote Sens. 51 (5) (2013) 2827–2836.

[43] D. Picone, R. Restaino, G. Vivone, P. Addesso, M. Dalla Mura, J. Chanussot, Band assignment approaches for hyperspectral sharpening, IEEE Geosci. Remote Sens. Lett. 14 (5) (2017) 739–743.

[44] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, S. Baronti, Hyper-sharpening: A first approach on SIM-GA data, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (6) (2015) 3008–3024.

[45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, 2020, pp. 213–229.

[46] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

[47] D.A. Hudson, L. Zitnick, Generative adversarial transformers, in: International Conference on Machine Learning, 2021, pp. 4487–4499.

[48] D. Hong, Z. Han, J. Yao, L. Gao, B. Zhang, A. Plaza, J. Chanussot, SpectralFormer: Rethinking hyperspectral image classification with transformers, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–15.

[49] X. He, Y. Chen, Z. Lin, Spatial-spectral transformer for hyperspectral image classification, Remote Sens. 13 (3) (2021) 498.

[50] A. Selen, T.-G. Esra, SpectralSWIN: a spectral-swin transformer network for hyperspectral image classification, Int. J. Remote Sens. 43 (11) (2022) 4025–4044.

[51] Y. Cai, J. Lin, X. Hu, H. Wang, X. Yuan, Y. Zhang, R. Timofte, L. Van Gool, Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17502–17511.

[52] W.G.C. Bandara, V.M. Patel, HyperTransformer: A textural and spectral feature fusion transformer for pansharpening, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1767–1777.

[53] L. Wang, Z. Wu, Y. Zhong, X. Yuan, Snapshot spectral compressive imaging reconstruction using convolution and contextual transformer, Photon. Res. 10 (8) (2022) 1848–1858.

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

[55] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.

[56] M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10781–10790.

[57] X. Wu, T.-Z. Huang, L.-J. Deng, T.-J. Zhang, Dynamic cross feature fusion for remote sensing pansharpening, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 14667–14676.

[58] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[59] H. Touvron, A. Vedaldi, M. Douze, H. Jégou, Fixing the train-test resolution discrepancy, Adv. Neural Inf. Process. Syst. 32 (2019).

[60] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, N. Houlsby, Big transfer (bit): General visual representation learning, in: European Conference on Computer Vision, 2020, pp. 491–507.

[61] F. Yasuma, T. Mitsunaga, D. Iso, S.K. Nayar, Generalized assorted pixel camera: postcapture control of resolution, dynamic range, and spectrum, IEEE Trans. Image Process. 19 (9) (2010) 2241–2253.

[62] A. Chakrabarti, T. Zickler, Statistics of real-world hyperspectral images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2011, pp. 193–200.

[63] N. Yokoya, C. Grohnfeldt, J. Chanussot, Hyperspectral and multispectral data fusion: A comparative review of the recent literature, IEEE Geosci. Remote Sens. M 5 (2) (2017) 29–56.

[64] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, J. Huang, A locally optimized model for hyperspectral and multispectral images fusion, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.

[65] L. Wald, T. Ranchin, M. Mangolini, Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images, Photogramm. Eng. Remote Sens. 63 (6) (1997) 691–699.

[66] W. Wang, W. Zeng, Y. Huang, X. Ding, J. Paisley, Deep blind hyperspectral image fusion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4150–4159.

[67] T. Huang, W. Dong, J. Wu, L. Li, X. Li, G. Shi, Deep hyperspectral image fusion network with iterative spatio-spectral regularization, IEEE Trans. Comput. Imaging 8 (2022) 201–214.

[68] G. Vivone, Multispectral and hyperspectral image fusion in remote sensing: A survey, Inf. Fusion 89 (2023) 405–417.

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

[70] B. Aiazzi, L. Alparone, S. Baronti, R. Carlà, A. Garzelli, L. Santurri, Full-scale assessment of pansharpening methods and data products, in: L. Bruzzone (Ed.), Image and Signal Processing for Remote Sensing XX, Vol. 9244, 9244, International Society for Optics and Photonics, SPIE, 2014, 924402.