Spectral–Spatial Transformer for Hyperspectral Image Sharpening

Lihui Chen, *Member, IEEE*, Gemine Vivone, *Senior Member, IEEE*, Jiayi Qin, Jocelyn Chanussot, *Fellow, IEEE*, and Xiaomin Yang, *Member, IEEE*

***Abstract*— Convolutional neural networks (CNNs) have recently achieved outstanding performance for hyperspectral (HS) and multispectral (MS) image fusion. However, CNNs cannot explore the long-range dependence for HS and MS image fusion because of their local receptive fields. To overcome this limitation, a transformer is proposed to leverage the long-range dependence from the network inputs. Because of the ability of long-range modeling, the transformer overcomes the sole CNN on many tasks, whereas its use for HS and MS image fusion is still unexplored. In this article, we propose a spectral–spatial transformer (SST) to show the potentiality of transformers for HS and MS image fusion. We devise first two branches to extract spectral and spatial features in the HS and MS images by SST blocks, which can explore the spectral and spatial long-range dependence, respectively. Afterward, spectral and spatial features are fused feeding the result back to spectral and spatial branches for information interaction. Finally, the high-resolution (HR) HS image is reconstructed by dense links from all the fused features to make full use of them. The experimental analysis demonstrates the high performance of the proposed approach compared with some state-of-the-art (SOTA) methods.**

***Index Terms*— Deep learning (DL), hyperspectral (HS) imaging, image fusion, multispectral (MS) imaging, remote sensing, transformer.**

I. Introduction

DUE to some physical limitations in imaging sensors, hyperspectral (HS) images with abundant spectral information always get a low spatial resolution. On the other hand, multispectral (MS) images with limited spectral information can achieve a high spatial resolution. However, high-resolution (HR) HS (HRHS) images are preferred for many downstream tasks, such as classification [1] and assessing grain yield [2], due to their rich spectral and spatial information. Therefore, to obtain HRMS images, researchers studied methods to fuse low-resolution (LR) HS (LRHS) and HRMS images.

Deep-learning (DL) methods are one of the most effective ways to address the HS and MS image fusion task. Many deep convolutional neural networks (CNNs) have been proposed for HS and MS image fusion because of their strong learning ability. Although CNNs can achieve state-of-the-art (SOTA) performance for HS and MS image fusion, its inherent limitation about local receptive fields makes it hard to capture long-range dependence. Nevertheless, the long-range dependence is a significant cue for HS and MS image fusion in twofold. On the one hand, the spatial long-range dependence enables the network to reconstruct details and structures by pixel similarity, which is relevant for remote sensing images, where similarities among different areas of the image, showing a similar landscape (e.g., similar buildings and lands), can easily be found. On the other hand, we can explore the spectral long-range dependence to maintain the spectral features of HS images, in particular, when we deal with HS images with a relevant number of spectral bands.

Some common ways to capture the long-term dependence is to improve the receptive fields by larger kernel size, pooling, deeper networks, or dilated convolution. However, these methods still capture local features in each layer. The use of kernels with large sizes and deep networks leads to considerable parameters. Pooling always suffers from information loss, as for dilated convolution that can also lead to checkerboard artifacts. To alleviate the above limitations, attention and transformer [4], [5] have been proposed to capture the long-range dependence. Constructed by multihead attention, the transformer model is proposed to fully leverage the long-range dependence of its inputs. Benefiting from its strong capability of modeling long-range dependence, transformer achieves SOTA performance on many tasks, see, e.g., machine translation [4], image recognition [6], HS image classification [7], [8], [9], [10], image restoration [11], spectral reconstruction [12], and image super-resolution [13], [14]. However, their use is still unexplored for HS and MS image fusion. Although two recent reports [15], [16] available online adopted transformer for HS and MS image fusion, the transformer

Manuscript received 22 March 2022; revised 7 December 2022 and 20 March 2023; accepted 17 July 2023. This work was supported in part by the China Postdoctoral Science Foundation under Grant 2023M730425, in part by the Fundamental Research Funds for the Centeral Universities under Project 2023CDJXY-037, in part by the China Scholarship Council under Grant 202006240191, in part by the Science Foundation of Sichuan Science and Technology Department under Grant 2021YFH0119, and in part by Sichuan University under Grant 2020SCUNG205. *(Corresponding author: Xiaomin Yang.)*

Lihui Chen is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: lihui.chen@cqu.edu.cn).

Gemine Vivone is with the National Research Council, Institute of Methodologies for Environmental Analysis (CNR-IMAA), 85050 Tito Scalo, Italy, and also with the National Biodiversity Future Center (NBFC), 90133 Palermo, Italy (e-mail: gemine.vivone@imaa.cnr.it).

Jiayi Qin and Xiaomin Yang are with the College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China (e-mail: qinjiayi@stu.scu.edu.cn; arielyang@scu.edu.cn).

Jocelyn Chanussot is with Inria, CNRS, Grenoble INP, LJK, Université Grenoble Alpes, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn.chanussot@grenoble-inp.fr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3297319.

Digital Object Identifier 10.1109/TNNLS.2023.3297319

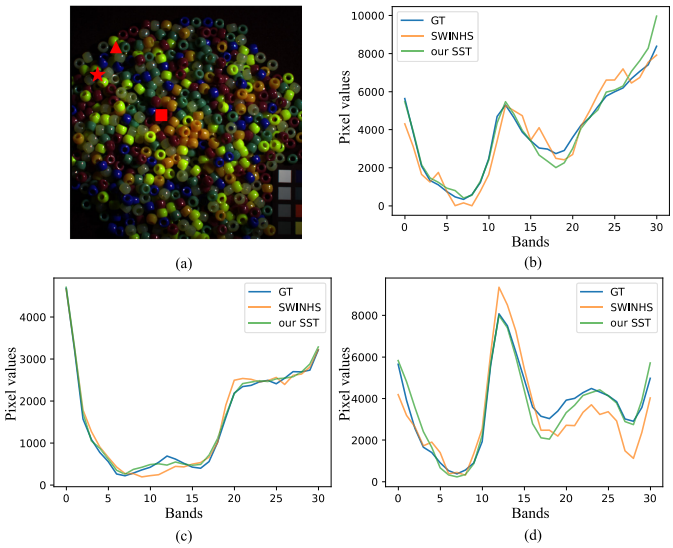


Fig. 1. Comparison of the spectral signatures between the proposed network (our SST), which considers both spectral and spatial dependence, and the baseline (SWINHS) relied upon the SWIN transformer block [3] and only considering the spatial dependence. In (a), the sampling points are depicted, marked by the red star, triangle, and square related to (b)–(d), respectively. GT stands for the ground truth.

blocks were directly adopted from [3] and [17], without any dedicated design for HS and MS image fusion. Besides, both of them only explored spatial transformer blocks (SpaTs), thus neglecting the spectral dependence. Unlike the transformer blocks in [3], [11], and [16], we devise the spectral and spatial transformer (SST) blocks to build two branches in the network dedicated to the spectral and spatial information, respectively. As shown in Fig. 1, we can observe that the proposed SST achieves better spectral preservation than the one built by considering only SpaTs.

Specifically, we propose an SST for HS and MS image fusion, in which two branches are dedicated to extracting spectral and spatial features, named spectral and spatial branches, respectively. The spectral branch built by spectral transformer blocks (SpeTs) enhances the spectral features by exploring channel correlation representing the spectral dependence, while the spatial branch built by SpaTs enhances the spatial features by exploring pixel correlation in the spatial domain. Afterward, the enhanced spectral and spatial features are fused by a simple fusion module. The fused features will further transfer and fed back into the spectral and spatial branches for information interaction. Finally, to fully use the fused features, these latter are connected to the final layer by dense links to reconstruct the HRHS image (the so-called dense reconstruction).

The contributions of this work can be summed up as follows.

1) To exploit the long-range spectral and spatial dependencies in HS and MS images, we propose an SST for HS and MS image fusion. In the spectral transformer branch, spectral features and long-range spectral dependency are explored, which contribute to spectral preservation. In the spatial transformer, spatial features and long-range spatial dependency are considered, which contribute to spatial detail enhancement.
2) To reduce the computational burden for attention, an adaptive calculation based on the associative rule is proposed for both spectral and spatial attention. In the spectral attention, pooled spectral features are calculated for further computational reduction. Because of these solutions, the proposed SST can reduce the computation and simultaneously capture global dependency.
3) To fully fuse spectral and spatial information, each spatial–spectral transformer block pair is followed by a fusion module to aggregate spatial and spectral features, and these fused features are densely connected to reconstruct the fused product.
4) To achieve an extremely lightweight transformer [fewer than 1M parameters and 20G float point operations (FLOPs)] for HS and MS image fusion, a spectral transformer based on the sole SpeT is proposed.

This article is organized as follows. Section II briefly reviews HS and MS image fusion. Afterward, Section III presents the proposed SST. Experimental results, even including the presentation of the exploited data and benchmark for comparison, are shown in Section IV. Finally, the conclusions are drawn in Section V.

II. Related Works

In this section, we focus on reviewing the HS and MS image fusion, with a special focus on DL-based methods, and a brief review of long-range exploration in DL, such as self-attention, nonlocal networks, and transformers. For a more comprehensive review about HS and MS image fusion, interested readers can refer to [1], [18], [19], [20], [21], and [22]. Readers interested in transformers can refer to [23], [24], [25], and [26] for more details.

A. Classical Methods for HS and MS Image Fusion

Most early HS and MS image fusion methods originate from pansharpening [19], [27], including component substitution (CS) and multiresolution analysis (MRA)-based methods. Chen et al. [28] introduced spectral-coverage-based band assignment to group HS and MS bands and adopted Gram–Schmidt adaptive (GSA) [29] to pansharpen HS bands in each group. According to the criterion of minimizing a spectral distortion, Picone et al. [30] presented a spectral angle mapper (SAM)-based [31] band assignment method for adapting pansharpening methods to HS and MS image fusion. Rather than using a band assignment method, Selva et al. [32] synthesized, for each HS band, a related HR band from MS images by using a regression framework. Afterward, the generalized Laplacian pyramid (GLP) with a matched modulation transfer function (MTF) [33] approach is used to fuse the synthesized HR band and the corresponding HS band. Although these methods are efficient and easy to implement, they often encounter spectral and/or spatial distortions. Zhang et al. [34] utilized convolutional sparse decomposition for the fusion of MS and panchromatic images.

In matrix factorization methods, LRHS and HRMS images are always factorized into spectral bases and a matrix of

coefficients [20]. Dong et al. [35] used an over-completed dictionary to construct the bases obtaining the coefficients by sparse coding. Simoes et al. [36] learned the bases in a low-rank subspace by vertex component analysis (VCA). Yokoya et al. [37] proposed a coupled nonnegative matrix factorization (CNMF) to alternatively update the bases and coefficients. Recently, some tensor factorization methods have also been proposed. Xu et al. [38] explored nonlocal coupled tensor decomposition for HS and MS image fusion. Dian et al. [39] presented a low tensor-train rank-based method to fuse HS and MS images. Li et al. [40] introduced a coupled sparse tensor decomposition for HS and MS image fusion. Prévost et al. [41] used the truncated singular value decomposition (SVD) to obtain the dictionaries of image tensors efficiently reducing the computational cost by reliable core tensor estimation for HS and MS image fusion.

B. DL Methods for HS and MS Image Fusion

Similar to classical methods, most DL-based pansharpening methods can also be applied to LRHS and HRMS image fusion by easily changing the input and output channels for the first and the last convolutional layers. Masi et al. [43] proposed the first CNN, named PNN, for pansharpening inspired by Dong et al. [44]. To boost the performance of flat CNNs with shallow networks for pansharpening, Wei et al. [45] introduced residual learning to build a deep network. For superior spectral and spatial preservation, PanNet [46] received the high-passed MS image and panchromatic images rather than the original images. Besides, DL has also been proposed specially to fuse LRHS and HRMS images. We can roughly divide DL methods into two categories, i.e., supervised DL methods and unsupervised DL methods.

For supervised DL methods, Dian et al. [47] proposed to reconstruct the HRMS image from an initial HRMS acquired by a model-based method. Xie et al. [48] incorporated the low-rank prior into the deep CNN and proposed an interpretable network for HS and MS image fusion. Han et al. [49] presented a multiscale deep CNN to progressively increase the size of the features of the HS image, finally obtaining the reconstructed HRMS image. Fu et al. [50] proposed a multiscale detail network for MS image sharpening. Xie et al. [51] extracted a deep prior by injecting high frequencies with a deep residual mapping into a transformed HS image adding a further constraint in a Sylvester equation. Zheng et al. [52] acquired edge maps by applying the Sobel operator to features extracted from the RGB band of the HRMS images using a pretrained model, and then utilized a feature fusion and a transform network to fuse features extracted from HS and MS images conditioned on the acquired edge maps to avoid loss of sharp edges in deep networks. Formulating the HRMS image as the multiplication of a subspace and spatial coefficients, Dian et al. [53] applied SVD to obtain the subspace and achieve the coefficients via the maximum posterior criterion regularized by a CNN denoiser. Hu et al. [54] introduced a spatial–spectral attention CNN, named HSRnet, for HS and MS image fusion. Guided by an observation model for HS images, Dong et al. [55] unfolded an iteration reconstruction and denoising for HS and MS image fusion by deep networks.

In addition to the above-mentioned supervised DL methods, there are also some works exploring unsupervised learning for HS and MS image fusion. Qu et al. [56] proposed the first unsupervised DL method for HS and MS image fusion by an encoder–decoder architecture, into which the sparse Dirichlet distribution is incorporated to force the physical constraints for spatial coefficients, i.e., sum-to-one and nonnegative coefficients. Wang et al. [57] proposed a deep blind iterative fusion network to estimate the observation model and to alternatively fuse HS and MS images. Uezato et al. [58] presented a guided deep decoder for paired image fusion, which could also be trained by unsupervised learning without training data. The encoder exploited the multiscale features from MS images, and then generated the HRHS image by decoder with the guidance of the multiscale features from the encoder. Yao et al. [59] proposed a coupled unmixing network with cross attention for unsupervised HS and MS image fusion. They unmixed the HS and MS images into spectral bases and coefficients by deep networks learning the spectral response function and point spread function by two simple layers based on the assumption that the degraded LRHS and HRMS images from the reconstructed HRHS image are consistent with the LRHS and HRMS data in the input. Zheng et al. [60] proposed a network to adaptively learn the spectral response functions. Taking middle results from self-supervised learning, Wei et al. [61] proposed to recurrently refine the reconstructed HRMS image for unsupervised HS and MS image fusion. Formulating the degradation by a convolutional layer and full connection for LRHS and HRMS images with zero-mean Gaussian prior, respectively, Zhang et al. [62], [63] introduced an image-specific network optimized by a loss function for unsupervised blind HS and MS image fusion. Diao et al. [64] proposed an unsupervised GAN to fuse MS and panchromatic images.

C. Long-Range Dependence in DL

Long-range dependence is an important cue for many signal processing tasks, such as image deraining [65], human-skeleton motion prediction [66], and activity recognition [67]. However, the early CNNs rarely explored the long-range dependence due to the inherited limitation of local receptive fields until nonlocal networks [5], graph convolutional networks (GCNs) [65], attention [4], and transformer [3], [4] were proposed. Although GCNs and nonlocal networks can explore long-range dependence, our SST differs from them in four aspects. First, the architectures of the overall network and the building blocks are different among the proposed network, [5] and [65], i.e., GCN or nonlocal blocks versus transformer blocks. Second, although GCN (with graph built by all pixels) and nonlocal module can be seen as attention, they only have one head attention, while the transformer is built by multihead attention and feed-forward networks (FFNs). Vaswani et al. [4] demonstrated that multihead attention can capture the long-range dependence from multiple subspaces, while GCN and nonlocal modules capture the long-range dependence from the entire feature space. In our experiments, we will show the performance comparison between one-head and multihead attention, noting that

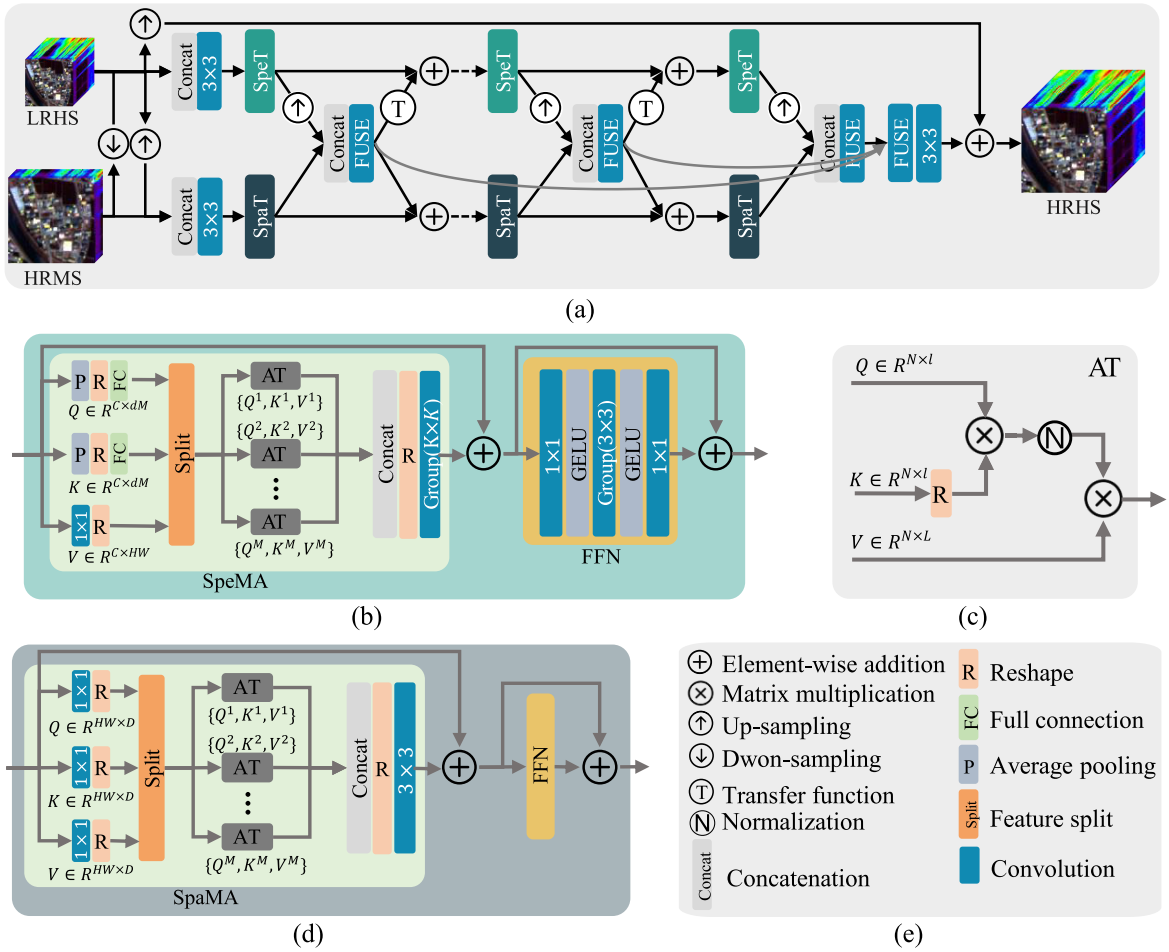


Fig. 2. (a) Architecture of the proposed SST. (b) Proposed SpeT. (c) Attention module (AT) in SpeMA and SpaMA. (d) Proposed SpaT. (e) Legend. "GELU" is the nonlinear activation function [42], "Group" refers to a group convolution, "$(K \times K)$" refers to the kernel size of the convolution, and "FUSE" is achieved by a (1×1) convolution.

multihead attention achieves better results for HS and MS image fusion. Besides, the FFN is also important for the fusion of features from multihead attention, as discussed in Section IV-D2. Third, softmax is used to avoid numerical instabilities and normalize features in [65], while we use rectified linear unit (ReLU) and inner production to this aim. Compared with softmax, ReLU and inner production have smaller computational complexity. Finally, the purpose of the spectral transformer is to explore spectral dependency in HS images, while the purpose of channel GCN in [65] is to explore feature channel dependence for image deraining. Noteworthy, it is hard to explore spectral dependency in [65], since there is a little spectral dependency for RGB images in the deraining task.

To explore long-range dependence in image fusion, recently, Meng et al. [68] proposed a visual transformer for MS and panchromatic image fusion. Zhou et al. [69] combined the transformer and the invertible neural network for pansharpening. Two transformer-based HS and MS image fusion methods have also been reported. Hu et al. [15] integrated spatial transformer into an encoder–decoder architecture for HS and MS image fusion. Ma et al. [16] took the transformer as a prior for HS and MS image fusion. Unlike their works, which apply the existing transformer blocks, we design transformer blocks dedicated to the HS and MS image fusion problem. Moreover, our SST considers both spatial and spectral transformers, while the abovementioned works only consider a spatial transformer. Finally, we adopt the inner production for normalization to obtain the attended features rather than softmax, and we further use a switcher to adaptively choose a calculation order for reducing the computation; see (13).

III. Methodology

This section is devoted to the description of the adopted methodology. The overview of the architecture of the proposed SST will be presented first. Afterward, the SST blocks will be detailed.

A. Architecture of SST

As shown in Fig. 2, our SST has two branches, where the upper branch, named the spectral branch, consists of several SpeTs, while the bottom branch, named the spatial branch, is obtained by several SpaTs. The dual-branch architecture is motivated by the philosophy of divide and conquer, which means that the spectral transformer branch is dedicated to exploring spectral information, while the spatial transformer

is devoted to the exploration of spatial information. In the middle of the two branches, a simple fusion module (i.e., the concatenation followed by a pointwise convolution in Fig. 2) after each pair of SpeT and SpaT is utilized to fuse spectral and spatial features as well as establishing information exchange between them.

Given the LRHS image, $\mathcal{X} \in \mathbb{R}^{C\times h\times w}$ (with height, w, width, h, and number of bands, C), and the HRMS image, $\mathcal{Y} \in \mathbb{R}^{c\times H\times W}$ (with height, W, width, H, and number of bands, c), a convolution in each branch is used first to extract the initial spectral or spatial features. For the spectral branch, the LRHS image and the downsampled HRMS image are adopted as input. Therefore, the initial spectral features, $\mathcal{S}_0$, are obtained by

$$\mathcal{S}_0 = f_s([\downarrow(\mathcal{Y}), \mathcal{X}]) \tag{1}$$

where $f_s(\cdot)$ denotes the convolution operation. The filter number can be regarded as the number of the spectra bases, so that each channel in $\mathcal{S}_0$ can be seen as the spectral feature for a specific spectral basis. $\downarrow(\cdot)$ is the downsampling operator, and $[\cdot]$ represents the concatenation operation. For the spatial branch, we argue that spatial features should derive from both the LRHS and the HRMS image, since the low-frequent information and the high-frequent information in the fused production are from the LRHS image and the HRMS image, respectively. Therefore, we concatenate the interpolated LRHS image with the HRMS as input for the spatial branch. Specifically, the initial spatial features, $\mathcal{H}_0$, are extracted as follows:

$$\mathcal{H}_0 = f_h([\mathcal{Y}, \uparrow(\mathcal{X})]) \tag{2}$$

where $f_h(\cdot)$ refers to the convolution operation and $\uparrow(\cdot)$ denotes the upsampling operator.

Afterward, the spectral and spatial features are separately fed to SpeT and SpaT. To increase the capacity of the network, we sequentially stack L SpeTs or SpaTs in these two branches. More specifically, each SpeT is used to enhance the input spectral features by exploring the spectral dependency, while each SpaT does similar work for spatial features by exploring spatial dependency. Focusing on the lth SpeT and SpaT, we have

$$\mathcal{S}'_l = f^l_{\text{spet}}(\mathcal{S}_{l-1}) \tag{3}$$

$$\mathcal{H}'_l = f^l_{\text{spat}}(H_{l-1}) \tag{4}$$

where $f^l_{\text{spet}}(\cdot)$ and $f^l_{\text{spat}}(\cdot)$ denote the functions of the lth ($1 \le l \in \mathbb{Z} \le L$) SpeT and the lth SpaT in the spectral and spatial branches, respectively; $\mathcal{S}'_l$ and $\mathcal{H}'_l$ are the outputs of SpeT and SpaT, respectively; and S_{l-1} and $\mathcal{H}_{l-1}$ are the inputs of the SpeT and SpaT, respectively, which can be obtained by (1) and (2) or (6) and (7).

The enhanced spectral and spatial features are then fused by the fusion module between the two branches. Thus, we have

$$\mathcal{F}_l = f^l_{\text{pw}}\left(\left[\mathcal{S}'_l, \uparrow\left(\mathcal{H}'_l\right)\right]\right) \tag{5}$$

where $f^l_{\text{pw}}(\cdot)$ denotes the pointwise convolution for fusion. To enable information interaction between spectral and spatial features, the fused features are taken as residual features to be added back to the spectral and the spatial branches. Since the fused features have a different size from that of spectral features, a transfer function achieved by downsampling followed by pointwise convolution is applied to change the size before adding it back to the spectral branch. The transfer function can also play the role of transferring features from the spectral–spatial domain to the spectral domain. Hence, the input features for the next SpeT and SpaT are obtained as follows:

$$\mathcal{S}_l = \mathcal{S}'_l + f^l_t(\downarrow(\mathcal{F}_l)) \tag{6}$$

$$\mathcal{H}_l = \mathcal{H}'_l + \mathcal{F}_l \tag{7}$$

where $\downarrow(\cdot)$ denotes the downsampling operation and $f^l_t(\cdot)$ refers to the pointwise convolution for the transferring. Since the last fused features are no longer fed back to the spectral and spatial branches, the index l in (6) and (7) belongs to $\{1, 2, \ldots, L-1\}$.

Finally, the fused image, $\mathcal{Z}$, is obtained by adding the interpolated LRHS image and the residual image reconstructed from the fused features obtained by each fusion module, $\mathcal{F}_L$. Thus, we have

$$\mathcal{Z} = \uparrow(\mathcal{X}) + f_{\text{rec}}(f_f([\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_L])) \tag{8}$$

where $f_f(\cdot)$ is a pointwise convolution to fuse all the fused features and $f_{\text{rec}}(\cdot)$ denotes the convolution related to residual image reconstruction.

B. Spectral Transformer Block

As shown in Fig. 2(c), the SpeT consists of a spectral multihead attention (SpeMA) module and an FFN. In SpeT, spectral features are enhanced first by the spectral attention representing the channel dependency. Then, the enhanced spectral features are fed into the FFN to fuse them. Similar to most transformer blocks [3], [11], residual learning is adopted for both SpeMA and FFN. Therefore, the formulation of the lth SpeT is as follows:

$$\mathcal{S}^{\text{ma}}_{l-1} = \mathcal{S}_{l-1} + f_{\text{spema}}(\mathcal{S}_{l-1}) \tag{9}$$

$$\mathcal{S}'_l = \mathcal{S}^{\text{ma}}_{l-1} + f_{\text{ffn1},l}\left(\mathcal{S}^{\text{ma}}_{l-1}\right) \tag{10}$$

where $\mathcal{S}_{l-1}$ and $\mathcal{S}'_l$ are the input and the output of the SpeT, $\mathcal{S}^{\text{ma}}_{l-1}$ is the output of the SpeMA, $f_{\text{spema}}(\cdot)$ is the SpeMA function, and $f_{\text{ffn1},l}(\cdot)$ denotes the FFN function in the SpeT module consisting of a pointwise, a group, and a pointwise convolution, as shown in Fig. 2(c).

In SpeMA, spectral features are transferred in query, key, and value. Differently from the way of general transformer blocks to obtain the query, the key, and the value, we obtain the query and the key by a sequential combination of pooling, reshaping, and full connection, while obtaining the value by a pointwise convolution and reshaping. The pooling in generating query and key is used for two reasons: 1) we think the pooled matrix is enough to represent the spectral features for a channel in the query and key, as shown in the experiments in Section IV-D3; this is also corroborated by Hu et al. [54], where the global average pooling can capture the spectral dependence for HS and MS image fusion and 2) the pooling operation can reduce the dimension of the features for calculating the channel dependency, thereby reducing

the computational burden for SpeMA. Besides, the ReLU [70] is applied to the query and the key for nonnegative values. Specifically, the query, $\mathbf{Q} \in \mathbb{R}^{C\times MP^2}$ (where M is the head number of attention and P is the output size of pooling), the key, $\mathbf{K} \in \mathbb{R}^{C\times MP^2}$, and the value, $\mathbf{V} \in \mathbb{R}^{C\times hw}$, can be formulated as follows:

$$\begin{aligned}\mathbf{Q} &= \delta(f_{\text{fc}}(R(P(\mathcal{S}))))\\ \mathbf{K} &= \delta(f_{\text{fc}}(R(P(\mathcal{S}))))\\ \mathbf{V} &= R(f_{\text{pw}}(\mathcal{S}))\end{aligned} \tag{11}$$

where $P(\cdot)$ represents the average pooling operation, $R(\cdot)$ refers to the reshape operation, $f_{\text{fc}}(\cdot)$ denotes the full connection, and $\delta(\cdot)$ is the ReLU function.

Afterward, the query, key, and value are split into M parts, and each part of them is fed into a head of an attention module, as shown in Fig. 2(b). Unlike the classical self-attention in [4], the proposed SST utilizes ReLU in (11) and inner production in (12) to obtain the normalized similarity matrix instead of a softmax layer. Thus, our SST has two advantages in calculating self-attention. First, the softmax layer with complicated exponent arithmetic is replaced by the simple ReLU, thereby reducing the computational complexity. Second, the normalization of the similarity matrix by inner production, instead of softmax, enables the associative rule to hold in the formula, thus reducing again the computational complexity of the classical self-attention. Specifically, the attention module can be formulated as follows:

$$\mathbf{U}_c^m = \frac{\sum_{j=1}^{C} \mathbf{Q}_c^m \left(\mathbf{K}_j^m\right)^T \mathbf{V}_j^m}{\sum_{j=1}^{C} \mathbf{Q}_c^m \left(\mathbf{K}_j^m\right)^T + \epsilon} \tag{12}$$

where $\mathbf{Q}_c^m \in \mathbb{R}^{1\times P^2}$ is the query of the cth band fed into the mth head of attention, $\mathbf{K}_j^m \in \mathbb{R}^{1\times P^2}$ and $\mathbf{V}_j^m \in \mathbb{R}^{1\times (hw/M)}$ are the key and value of the jth band fed into the mth head of attention, $\mathbf{U}_i^m$ is the reweighted values of the query, i.e., the cth band output by the mth head of attention, and $\cdot^T$ is the transpose operator. Rather than using the softmax function to normalize, we use $\sum_{j=1}^{C} \mathbf{Q}_c^m (\mathbf{K}_j^m)^T + \epsilon$ to normalize the numerator; thus, we can leverage the associative rule in (12) to reduce the computation complexity and memory-consuming. ϵ equal to 10^{-7} is used to avoid a zero denominator. Furthermore, the reweighted spectral features from all the heads are concatenated and reshaped to the size of the input spectral features, i.e., $C \times h \times w$. To adaptively choose the least computations and memory consuming for (12), we switch the computation order to obtain the reweighted features according to the input size as follows[1]:

$$\mathbf{U}^m = \begin{cases} \dfrac{\mathbf{Q}^m (\mathbf{K}^m)^T \mathbf{V}^m}{\text{sum}(\mathbf{Q}^m (\mathbf{K}^m)^T, \text{dim} = 1) + \epsilon}, & \text{if } \textbf{Condition-1} \\ \dfrac{\mathbf{Q}^m ((\mathbf{K}^m)^T \mathbf{V}^m)}{\mathbf{Q}^m \text{sum}((\mathbf{K}^m)^T, \text{dim} = 1) + \epsilon}, & \text{otherwise} \end{cases} \tag{13}$$

[1]Noteworthy, this equation is formulated considering all the channels simultaneously instead of a specific cth channel in (12), and it is formulated using the PyTorch style with broadcasting of shape.

where $\text{sum}(\cdot, \text{dim} = 1)$ refers to the sum of the items in the matrix along the row axis. **Condition-1** is as follows:

$$\begin{aligned} &CMP^2C + CChw < MP^2Chw + CMP^2hw \\ &\iff MP^2C + Chw < 2MP^2hw \end{aligned} \tag{14}$$

used to judge if the computations of the first equation in (13) are smaller than the second one.

Finally, a group convolution is utilized to fuse and project the reshaped spectral features; i.e., the output of the SpeMA is obtained by

$$\mathcal{S}^{\text{ma}} = f_{\text{gc},K}(\mathcal{U}_{\text{spe}}, C) \tag{15}$$

where the values of $\mathcal{U}_{\text{spe}} \in \mathbf{R}^{C\times h\times w}$ are the reshaped spectral features and $f_{\text{gc},K}(\cdot, C)$ denotes the function of the group convolution with C groups to fuse spectral features in a neighborhood of $K \times K$.

C. Spatial Transformer Block

The architecture of the SpaT shown in Fig. 2(d) is similar to the SpeT containing again two modules in the residual structure, i.e., the spatial multihead attention (SpaMA) and the FFN. The use of SpaMA instead of SpeMA is the only difference between SpaT and SpeT. SpaMA differs from SpeMA, because SpeMA relied upon the band (spectral) dependency to reweigh spectral features, while SpaMA explores the spatial dependency to reweigh spatial features.

The query, $\mathbf{Q} \in \mathbb{R}^{HW\times D}$ (where D is the number of channels of the spatial features), the key, $\mathbf{K} \in \mathbb{R}^{HW\times D}$, and the value, $\mathbf{V} \in \mathbb{R}^{HW\times D}$, in SpaMA are generated by a pointwise convolution and a reshape operation. Similar to SpeMA, the ReLU is applied to the query and the key. Then, the query, the key, and the value are split into M heads of attention. On the other hand, the formula differs from (12), since its purpose is to use the spatial similarity to reweigh the MS features, while the purpose of (12) is to utilize channel similarity to reweigh the input HS features. Each attention module can be formulated as follows:

$$\mathbf{U}_i^m = \frac{\sum_{j=1}^{HW} \mathbf{Q}_i^m \left(\mathbf{K}_j^m\right)^T \mathbf{V}_j^m}{\sum_{j=1}^{HW} \mathbf{Q}_i^m \left(\mathbf{K}_j^m\right)^T + \epsilon} \tag{16}$$

where $\mathbf{Q}_i^m \in \mathbb{R}^{1\times D}$ is the query at the ith pixel fed into the mth head of attention, $\mathbf{K}_j^m \in \mathbb{R}^{1\times D}$ and $\mathbf{V}_j^m \in \mathbb{R}^{1\times D}$ are the key and the value at the jth pixel fed into the mth head attention, and $\mathbf{U}_i^m$ is the reweighted value at the query, i.e., the ith pixel output by the mth head of attention. Afterward, the outputs of the M heads of attention are concatenated and reshaped to the input the features' size (i.e., $D \times H \times W$). Similar to spectral attention, a switcher equation, such as (13), is used in the spatial attention to achieve the lowest computational burden for (16). Finally, a convolutional layer is used to fuse the features by all the heads of attention. Consequently, the output of the SpaMA is formulated as follows:

$$\mathcal{H}^{\text{ma}} = f_c(\mathcal{U}_{\text{spa}}) \tag{17}$$

where the values of $\mathcal{U}_{\text{spa}} \in \mathbb{R}^{D\times H\times W}$ are the reshaped spatial features and $f_c(\cdot)$ denotes the convolution.

IV. Experiments

This section is devoted to the presentation of the experimental results. Details about how to implement the proposed approach and the training phase will be presented first. Afterward, the datasets will be described. Finally, the performance in comparison with SOTA methods and an ablation study will be provided to readers.

A. Network and Training Details

1) Networks Details: Similar to most networks [55], the ℓ_1 norm is adopted as the loss function to train the network. According to the experiments in Section IV-D3, we empirically set L in (5) to 14, P for generating spectral query and key in (11) to 4, K in (15) to 9, M for the head numbers in (12) and (16) to 8, and both C in (12) and D for spatial features in (16) to 120.

2) Training Details: All the networks in this article are trained using the same training set with 10^5 iterations for a fair comparison. The batch size and the LRHS patch size are set to 8 and 18×18, respectively. The Adam [71] optimizer with default setting and initial learning rate of 0.0002 is used for training. The learning rate is halved every 2×10^4. The official (available online) code is used for the compared networks tuning their parameters to get the best results.

B. Datasets

Three datasets are adopted in this article, including a natural image dataset, i.e., CAVE [72], one synthetic (widely used) remote sensing image set, i.e., Chikusei [73], and one real remote sensing image set, i.e., Hyperion/ALI.

1) The HS images in CAVE have 31 bands with a wavelength range from 400 to 700 nm. Similar to [54], 20 images are randomly used as the training set, and 11 images are used as the test set. We generated the LRHS image by a Gaussian filter with a support ($2\times s - 1$) and standard deviation of 1.5, where $s = 4$ is the scale ratio between HRHS and LRHS. The HRMS images are synthesized by the relative spectral response function of the Nikon-D700. For the training set, we cropped the LRHS images into patches with a size of 32×32 without overlapping them; 20% of training samples have been adopted randomly as the validation set.
2) We select 120 bands for HS images cropping them to reach the size of 2400×2200 pixels in our Chikusei experiments. We split first the datasets into training and test sets without any overlap and with a ratio of 8:2. Then, 20% of the training area is chosen as the validation set. Finally, we split HRHS images for the training, the validation, and the test sets into 96×96 patches. To generate the LRHS image, we exploit the MTFs of the Hyperion sensor imposing a scale ratio of 3. To synthesize the HRMS images, we applied the estimated spectral response coefficients (obtained by the nonnegative least square from the HS and MS pairs acquired by Hyperion/ALI) to the HRHS images. The spectral response coefficients are estimated using images captured over the area of Shenzheng, China. The synthetic HRMS images have nine bands.
3) For the Hyperion/ALI dataset, we collect first real HS and MS image pairs from the United States Geological Survey (USGS) website.[2] Then, we remove noisy bands retaining 120 bands for HS images. Afterward, we cropped the overlapped areas between HS and MS images registering them by Vivone et al. [74] and Guizar-Sicairos et al. [75]. Finally, we acquired seven HS/MS image pairs from different areas: 1) three areas in China, i.e., Suzhou, Zaozhuang, and Beijing; 2) Melbourne, Australia; 3) Paris, France; 4) Layaffete, USA; and 5) London, U.K. We adopted the Melbourne and Paris datasets as validation and test sets, respectively, and the remaining datasets as the training set. Similar to Chikusei, the MTFs of the Hyperion sensor imposing a scale ratio of 3 are exploited to generate the LRHS images. For the training set, we divided HRHS images into 96×96 patches. For validation and test sets, we cropped HRHS images into 189×189 patches.

For all the abovementioned training sets, we adopted random cropping, rotation, and flipping to increase the number of training samples. Specifically, the samples and image sizes for each dataset are shown in Table I. The reported results in all the tables in this article always refer to an average outcome on all the test/validation samples.

C. Comparison With SOTA Methods

1) Benchmark: In this section, we compared our SST and SPE (a lightweight network consisting of the only proposed spectral branch) with seven classical methods and three state-of-art DL-based techniques to show the effectiveness of the SST. Six metrics, i.e., the erreur relative globale adimensionnelle de synthése (ERGAS) [76], the SAM [31], the Q2n index [77], the peak signal-to-noise ratio (PSNR), the structure similarity index (SSIM), and the root-mean-square error (RMSE), are used for measuring the similarity with the GT. Ideal values are 0 for ERGAS, SAM, and RMSE; 1 for Q2n and SSIM; and infinity for PSNR. The adopted benchmark is as follows.

1) Classical methods
   a) *Bicubic:* Bicubic interpolation method applied to LRHS images.
   b) *GSA:* The GSA [29] method for HS and MS image fusion [1].
   c) *Smoothing filter-based intensity modulation (SFIM):* SFIM for HS and MS image fusion [1], [78].
   d) *GLPHS:* GLP with MTF as in [1] and [32].
   e) *CNMF:* CNMF for HS and MS image fusion [37].
   f) *HySure:* HS image super-resolution via subspace regularization [36].
   g) *FUSE:* Fast fusion for multiband images by solving a Sylvester equation [79].

[2] https://earthexplorer.usgs.gov/

TABLE I

SIZE OF THE GT AND THE NUMBERS OF SAMPLES FOR THE DATASETS USED IN THIS ARTICLE. (·/· BANDS) REFERS TO THE NUMBER OF BANDS FOR HS AND MS IMAGES, RESPECTIVELY

Dataset	CAVE (31/3 bands)			Chikusei (120/9 bands)			Hyperion/ALI (120/9 bands)		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
GT Size	128×128	128×128	512×512	96×96	96×96	96×96	96×96	189×189	189×189
Numbers	450	25	11	352	88	110	304	14	14

TABLE II

AVERAGE RESULTS ON THE CAVE TEST SET. THE BEST RESULTS ARE IN BOLDFACE, AND THE SECOND BEST RESULTS ARE UNDERLINED

Methods	SAM	ERGAS	Q2n	PSNR	SSIM	RMSE
Bicubic	5.0611	12.2321	0.7047	31.07	0.6562	1985.15
GSA	6.9656	4.0809	0.8457	40.60	0.8306	652.47
SFIMHS	6.0673	9.1776	0.8382	24.74	0.8066	5758.29
GLPHS	3.5839	2.2664	0.9351	45.70	0.9005	369.56
CNMF	4.1447	2.2045	0.9278	45.30	0.8824	396.15
HySure	6.8759	3.7103	0.8841	40.89	0.8152	656.89
FUSE	3.4393	3.0074	0.9339	40.11	0.8776	704.31
MHFnet	3.6422	1.8208	0.9456	47.40	0.9204	326.65
HSRnet	**2.3263**	1.2431	**0.9608**	**50.95**	0.9524	**211.38**
MogCDN	2.5694	1.3463	0.9576	50.05	0.9495	241.02
Our SPE	3.0702	1.6424	0.9535	48.25	0.9083	293.10
Our SST	2.4239	**1.2195**	**0.9608**	50.87	**0.9544**	221.31

2) DL-based methods

a) *MHFnet:* An interpretable MS and HS image fusion network [48].

b) *HSRnet:* A spatial–spectral attention network for HS and MS image fusion [54].

c) *MoGDCN:* A model-guided network for HS and MS image fusion [55].

The results of classical methods are obtained by the toolbox as proposed in [1]. The results for DL-based methods are obtained by the official codes (available online) for the CAVE dataset, and by changing inputs and outputs for Chikusei and Hyperion/ALI datasets, since their number of HS bands is different from that of the CAVE dataset.

2) Results on CAVE: To demonstrate the effectiveness of the proposed SST on natural images, we quantitatively and qualitatively evaluate all the methods on the CAVE dataset.

The quantitative results are shown in Table II, from which we can observe that HSRnet and our SST achieve better results. Overall, the DL-based methods have much better performance than classical methods. Noteworthy, although our method does not achieve the best result on the SAM, PSNR, and RMSE metric, this represents just a partial index measuring only the spectral preservation and average errors. Moreover, PSNR is derived from RMSE. Instead, the two overall quality indexes, i.e., ERGAS and Q2n, testify the goodness of our approach showing the best performance in both cases. The best SSIM result also shows that our SST has a better structure reconstruction.

It is worth to be noted that the reason why our SST does not achieve better results in a clear way with respect to the other DL-based methods in this test case is that CAVE data do not have so many similarity structures with a reduced number of spectral bands. Hence, the spatial and spectral transformers in our SST cannot leverage the spatial and spectral dependencies in a sufficient way to create a better gap with respect to the other DL-based methods. Instead, remote sensing images, such as Chikusei and Hyperion/ALI, have these kinds of features. Thus, the advantages on remotely sensed images will be much more clear (see Section IV-C3).

To show the perceptual quality, we present a visual comparison among all the methods in Figs. 3 and 4. For Fig. 3, the bottom two rows show the residual image between the

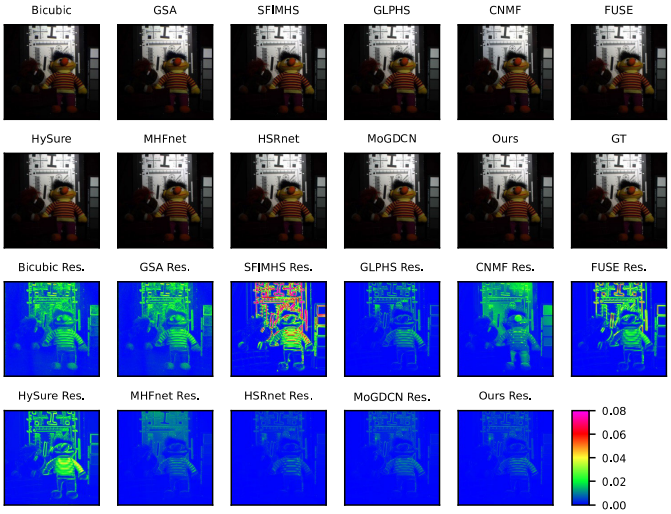


Fig. 3. Visual comparison for "chart_and_stuffed_toy_ms" in the CAVE dataset. The color images are composed by taking (29, 19, 9) bands as RGB bands. "Res." means the residual between the corresponding result and the GT averaged along spectral bands. Ours stands for the proposed SST.

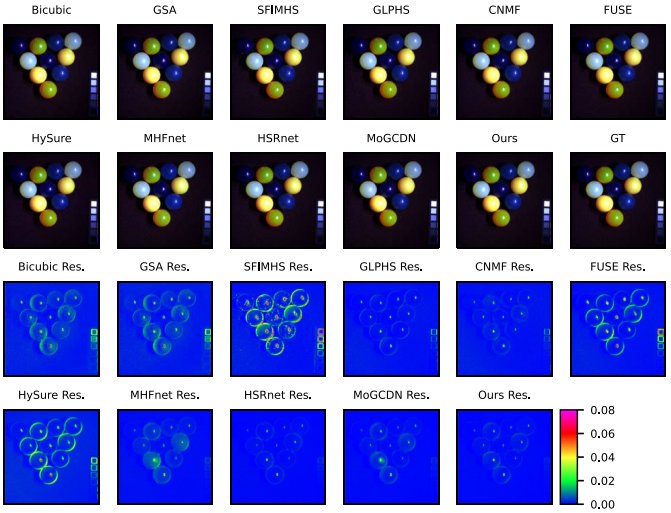


Fig. 4. Visual comparison for "superballs_ms" in the CAVE dataset. The color images are composed by taking (29, 19, 9) bands as RGB bands. "Res." means the residual between the corresponding result and the GT averaged along spectral bands. Ours stands for the proposed SST.

TABLE III

AVERAGE RESULTS ON THE CHIKUSEI TEST SET. THE BEST RESULTS ARE IN BOLDFACE, AND THE SECOND-BEST RESULTS ARE UNDERLINED

Methods	SAM	ERGAS	Q2n	PSNR	SSIM	RMSE
Bicubic	3.0109	7.8242	0.7547	25.72	0.4106	221.46
GSA	1.9415	2.7177	0.9749	35.45	0.8678	72.91
SFIMHS	2.3073	5.3879	0.9224	29.33	0.7839	149.72
GLPHS	2.4202	5.2968	0.9205	29.53	0.7841	142.37
FUSE	2.9081	5.3082	0.9190	33.06	0.8380	95.84
CNMF	2.0127	4.0063	0.9551	32.93	0.8679	94.74
HySure	2.5181	5.4804	0.9301	29.21	0.7167	147.70
MHFnet	0.7689	0.8515	0.9955	46.20	0.9390	20.44
HSRnet	0.7578	1.0817	0.9936	45.25	0.9230	25.51
MoGCDN	0.8147	0.9731	0.9945	43.54	0.9207	27.79
Our SPE	0.7011	**0.5223**	0.9970	46.68	0.9390	19.66
Ours SST	**0.5885**	0.5250	**0.9980**	**48.57**	**0.9520**	**15.88**

TABLE IV

AVERAGE RESULTS ON THE HYPERION/ALI TEST SET. BEST RESULTS ARE IN BOLDFACE, AND THE SECOND BEST RESULTS ARE UNDERLINED

Methods	SAM	ERGAS	Q2n	PSNR	SSIM	RMSE
Bicubic	3.4489	5.1974	0.7180	26.83	0.4744	188.02
GSA	3.6577	5.5054	0.7383	25.85	0.5866	214.39
SFIMHS	3.2515	4.8655	0.7789	27.14	0.6020	181.14
GLPHS	3.2194	4.8268	0.7791	27.26	0.6034	178.63
CNMF	3.2852	5.1566	0.7563	26.93	0.5729	187.24
HySure	4.5658	6.8258	0.6899	24.88	0.4780	245.67
FUSE	8.9813	9.3637	0.5230	19.89	0.3639	489.69
MHFnet	2.7215	4.6685	0.8413	29.40	0.6554	139.64
HSRnet	2.5708	3.8348	0.8665	29.52	0.7282	137.83
MOGCDN	2.5862	3.8893	0.8583	29.39	0.6831	140.10
Our SPE	2.5998	**2.9186**	0.8554	29.39	0.6891	140.36
Our SST	**1.9923**	2.9620	**0.9163**	**31.76**	**0.7941**	**106.58**

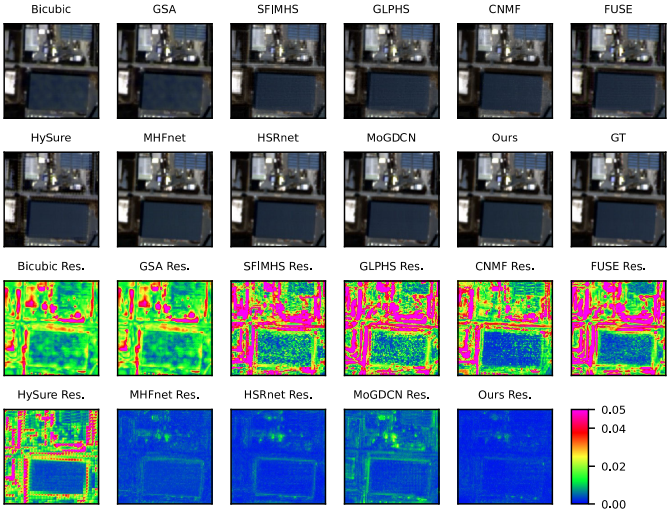


Fig. 5. Visual comparison considering a close-up of a Chikusei test image. The color images are composed by taking (29, 19, 9) bands as RGB bands. "Res." means the residual between the corresponding result and the GT averaged along spectral bands. Ours stands for the proposed SST.

corresponding result and the ground truth (GT). This is obtained by averaging over the spectral bands the band-by-band residuals. From these images, we can easily find that DL methods show fewer distortions than classical methods. Among DL-based methods, MHFnet achieves worse results than the others. It is hard to assess which one is the best among HSRnet, MoGDCN, and our SST, since the visual results are very close. In Fig. 4, we can find similar results to those of Fig. 3, but HSRnet achieves fewer errors than the other methods, while our method gets the second-fewest errors. Overall, the visual effectiveness is consistent with the quantitative results, where DL methods achieve better results than classical methods, showing quite similar outcomes for HSRnet, MoGDCN, and the proposed approach.

3) Results on Chikusei: To demonstrate the effectiveness of the proposed method on remote sensing images, we exploited the widely used Chikusei dataset.

The quantitative results are reported in Table III. We can easily find out that our approaches (i.e., SST and SPE) achieve the best results. For this dataset, DL-based methods achieve much better results than classical images. This is because the test and the training areas are quite close to each other, i.e., located in the Chikusei, Japan. Therefore, DL methods can fully exploit the learned features to reconstruct HRMS images. MHFnet achieves the second best result after our approaches. It is not surprising that a model, consisting of a huge amount of parameters (see Section IV-C5), achieves good results on very close images. Despite that, the gap among MHFnet, HSRnet, and MoGDCN is quite small. On the other hand, our SST gets much better performance than the comparative methods because of the superiority of capturing spatial and spectral dependencies using the spatial and spectral transformers.

The visual comparison is shown in Fig. 5. Obviously, DL methods achieve better results than classical methods. Classical methods show many distortions on smooth areas (e.g., roofs). About DL methods, MoGDCN gets more distortions compared with MHFnet, HSRnet, and our SST. Although MHFnet and HSRnet get great fidelity in reproducing many areas, they show some distortions near some edges. Our SST obtains the closest result with respect to the GT.

4) Results on Hyperion/ALI: To demonstrate the effectiveness of the proposed method on real HS/MS image pairs and the generalization of the results to different areas, we exploit the Hyperion/ALI dataset.

The quantitative results are reported in Table IV. Our SST achieves the best results followed by HSRnet. Although the DL-based methods still outperform the classical ones, the gap between them is reduced in this case pointing out that the potentiality of DL models is not sufficiently explored for real cases, yet. On the other hand, the gap in performance between our SST and the other DL methods increases, demonstrating the effectiveness of our approach in real test cases. This dataset is the most important one in this article, because it includes some different types of area (e.g., buildings and arable lands), showing a big variability between test and training sets, thus simulating an operating environment.

A visual comparison is provided in Fig. 6. Consistently with the quantitative results, some classical methods, such as SFIMHS, GLPHS, and CNMF, obtain slightly worse or even comparable results with respect to MHFnet. Instead, outcomes obtained by HSRnet and MoGDCN have lower distortions than those of MHFnet. Finally, there is an evident visual gap (see the residual images) between all the compared approaches and our SST.

TABLE V

NUMBER OF PARAMETERS AND THE RUNNING TIMES FOR ALL THE METHODS. "/" MEANS THAT THERE ARE NO LEARNABLE PARAMETERS. "C" REFERS TO A METHOD RUNNING ON CPU, AND "G" REFERS TO A METHOD RUNNING ON GPU. M INDICATES A MILLION. IT IS WORTH TO BE REMARKED THAT MHFNET AND HSRNET USE TENSORFLOW; INSTEAD, MOGDCN AND OUR METHODS EXPLOIT PYTORCH

Methods	Bicubic	GSA	SFIMHS	GLPHS	FUSE	CNMF	HySure	MHFnet	HSRnet	MoGCDN	Our SPE	Our SST
#Params./FLOPs	/	/	/	/	/	/	/	12.84M/198.94G	1.85M/13.40G	11.79M/474.30G	0.9735M/16.90G	4.87M/70.92G
Time(s)	0.0024(C)	0.0736(C)	0.0405 (C)	0.3378(C)	0.0475(C)	0.8360(C)	2.8768(C)	0.1370 (G)	0.2734(G)	0.0360(G)	0.0086(G)	0.0305(G)

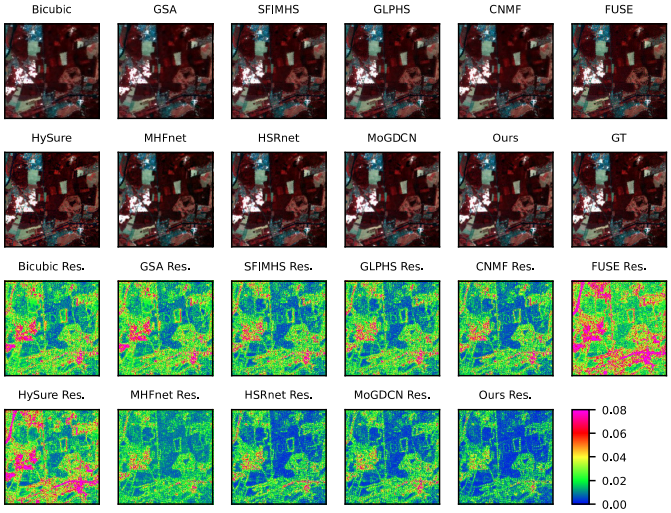


Fig. 6. Visual comparison using a close-up of the Paris test case of the Hyperion/ALI dataset. The color images are composed by taking (29, 19, 9) bands as RGB bands. "Res." means the residual between the corresponding result and the GT averaged along spectral bands. Ours stands for the proposed SST.

TABLE VI

RESULTS OF SIX DIFFERENT NETWORKS EVALUATED ON THE CAVE VALIDATION SET. THE BEST RESULTS ARE IN BOLDFACE

Networks	ERGAS	SAM	Q2n	#Params.
SWINHS	1.5491	2.3280	0.9609	0.7716
SPA	1.3994	2.2930	0.9568	0.8844
SPE	1.9509	2.7199	0.9488	**0.2715**
Dual-T	1.4729	2.3704	0.9592	1.2205
Dual-T-F	1.3318	2.2823	0.9604	1.3508
SST	**1.3080**	**2.1413**	**0.9633**	1.3740

5) Model Efficiency: To show the efficiency of the proposed SST, we compare the number of parameters (#Params.) and the running times for all the compared approaches. The parameters and running times are evaluated using the Chikusei dataset with a desktop equipped an Intel I7-12700K CPU and an NVIDIA RTX-3090 GPU. The size of the HRHS image is $96 \times 96 \times 120$ for this test. As reported in Table V, although our SST is not the most lightweight model for DL-based methods, the running time is the shortest one because of the linear computational complexity attention used in the transformer block. Moreover, if a lightweight network is required, one can adopt SPE built only by the spectral branch. SPE has much fewer parameters and a reduced running time than the other compared DL-based methods while achieving SOTA performance.

D. Ablation Study

1) Investigation on Architecture: To validate the effectiveness of each module in the proposed network, we conduct an ablation study to compare five variants of the SST reported in Table VI with a baseline network (called SWINHS) built by using SWIN transformer blocks. To save training time, we reduced L to 4 and the number of iterations to 2×10^4 for all the networks in Table VI. Noteworthy, the Q2n index is less sensitive to changes in the network architecture and also in varying hyperparameters (see Section IV-D3). Therefore, we will mainly consider ERGAS and SAM metrics in this ablation study. The meaning of each compared network is as follows.

1) *SWINHS:* The network by simply replacing the transformer blocks in SPA (see the following) by the ones from SWIN [3].
2) *SPA:* The network using only the spatial branch.
3) *SPE:* The network using only the spectral branch.
4) *Dual-T:* The network with both spectral and spatial branches but only one fusion module at the end of the two branches.

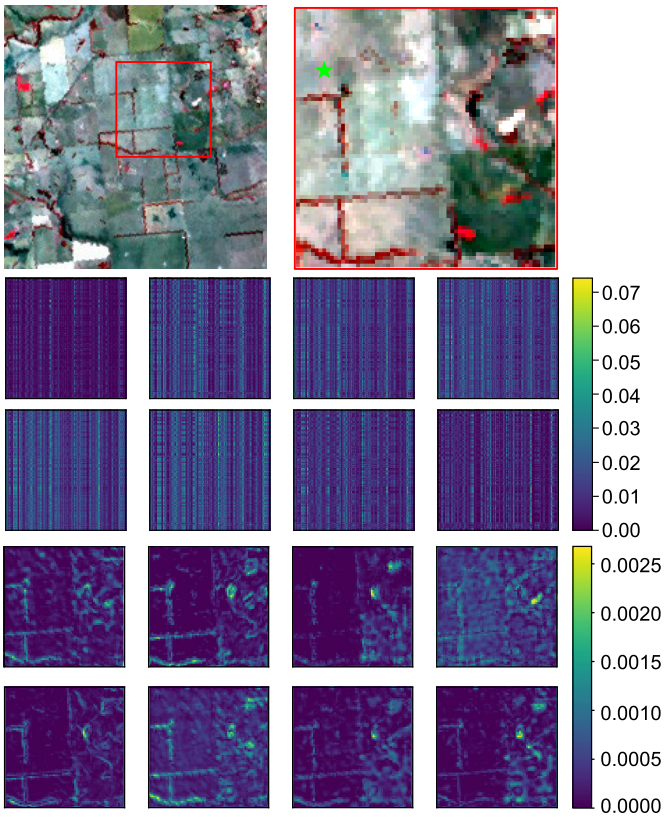


Fig. 7. Visualization of spectral and spatial dependence for a sample in the Hyperion/ALI validation set. The first row shows the test samples and the pixel, marked with a green star in the right image, for the visualization of the spatial dependence. The second and third rows show the spectral attention map (120×120) obtained by eight heads in the 13th SpeT for our SST. The last two rows show the spatial dependence of the marked pixel obtained by eight heads in the 13th SpaT for our SST.

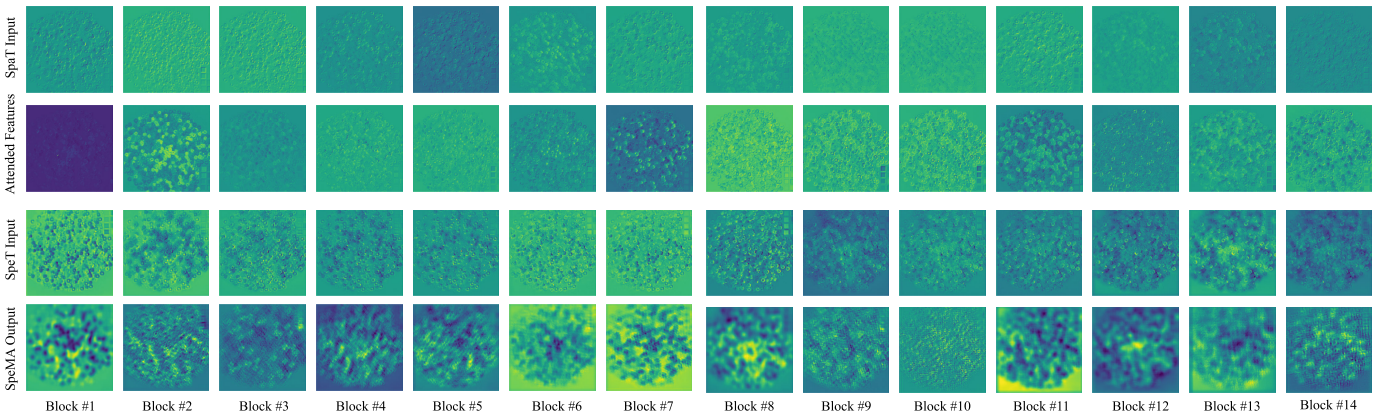


Fig. 8. Visualization of input spatial features (the first row) versus spatial attended features (the second row) and input spectral features (the third row) versus SpeMA output features (the last row).

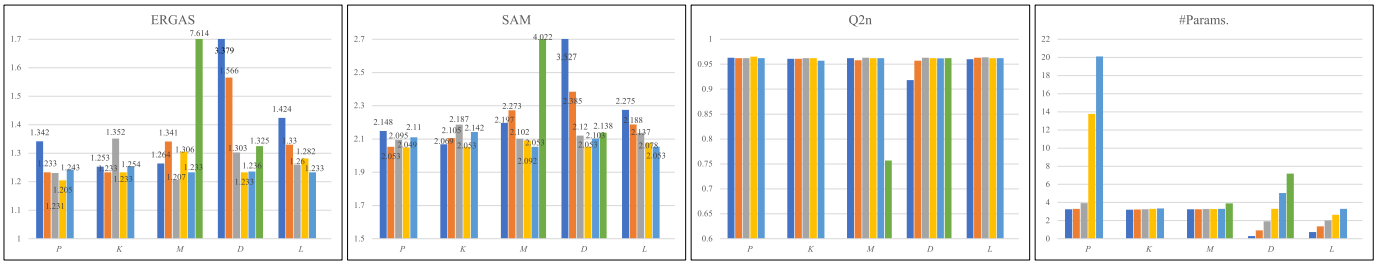


Fig. 9. (a) ERGAS, (b) SAM, (c) Q2n, and (d) #Params (unit of million, M) versus different hyperparameters for our SST. The performance is evaluated on the CAVE validation set. For each group of hyperparameters, the value increases from left to right; specifically, we have $P = \{1, 4, 8, 16, 18\}$, $K = \{3, 5, 7, 9, 11\}$, $M = \{1, 2, 4, 6, 8, 120\}$, $D = \{30, 60, 90, 120, 150, 180\}$, and $L = \{2, 4, 6, 8, 10\}$. To save the training time, we reduced the number of iterations to 2×10^4 for the networks in this figure.

5) *Dual-T-F:* The network with dual branches and a fusion module after each pair of SST blocks.
6) *SST:* The proposed approach adding the dense reconstruction to the Dual-T-F to make full use of all the fused features.

a) Difference between our SST and SWIN blocks: The proposed transformer blocks differ from those of SWIN for three main aspects: 1) SWIN transformer blocks only consider spatial attention, while our SST not only considers spatial attention in SpaTs but also considers the spectral attention, which can enhance the spectral features for HS and MS image fusion; 2) the FFN in SWIN transformer blocks consists of full connections, while the one in our transformer is constructed by convolution layers, which can receive context pixels that cannot be achieved by full connections; and 3) the softmax is used to normalize the attention map in SWIN transformer blocks, while the inner production is used to normalize attention in our transformer blocks. Recalling the switcher formula (13), our attention module can adaptively select a calculation order for reducing the computation. Benefiting from that, our transformer blocks globally acquire the attention map and do not have to divide features into widows to save computational resources.

b) Effectiveness of SPA: Comparing SPA with SWINHS, we can easily find that SPA achieves better results, despite the use of a few more parameters. This shows that the SpaTs in SPA (considering the global attention and the context reception) overcome the local window attention and the point reception of full connections for HS and MS image fusion.

c) Effectiveness of SPE: Although SPE achieves the worst results among the different configurations, it has much fewer parameters than the other methods and also has much fewer computations than the others because of the pooling in (13). This is indeed a good choice to achieve a lightweight transformer model with satisfactory results, as shown in Section IV-C.

d) Effectiveness of dual-branch fusion: Comparing Dual-T with SPA and SPE, we can see that Dual-T achieves worse results than SPA. Although both the spatial and spectral enhanced features are explored in Dual-T, the fusion, just at the end of the two branches, is not enough. Besides, there is no information interaction between spectral and spatial enhanced features. Therefore, we added a fusion module following each pair of SST blocks in the middle of the two branches, achieving the Dual-T-F. This latter obtains better results than both SPA and SPE showing the effectiveness of dual-branch fusion for SST.

e) Effectiveness of dense reconstruction: Although spatial and spectral features are fused by the middle fusion module, the fused features used to reconstruct the final HRMS images are still only from the final fusion module. To exploit the fused features in a better way, we use dense links from all the previous fusion modules to the final one achieving the proposed SST. It can easily be seen that, because of the use of dense reconstruction, the performance is further improved.

2) Analysis on Multihead Attention: To get some insights on the multihead attention, we show the spectral attention and

spatial attention maps for a sampling pixel for the spectral and spatial multihead attention; see Fig. 7. Observing the spectral attention maps for eight heads, we can see that each head focuses on some specific spectral bases (as already mentioned), since there are some vertical lines in these attention maps, which is unexpected, because the attention map should be symmetrical and have high values on its diagonal. We think it is because the multihead attention forces each head to explore different spectral bases in different heads. This is also the reason why a fused layer and an FFN are important for a transformer; thus, the features enhanced by the different spectral bases can be fused. Having a look at the spatial attention maps, we can also find that the different heads explore different dependencies for the same pixel, which is also consistent with the statement in [4], demonstrating that different attention heads explore features in different subspaces.

Besides, we show the average features before and after the processing of the multihead attention as Fig. 8. Comparing the first and second rows for SpaT, we can note that the attended features by SpaMA have clearer edges, such as the results in Blocks 2, 7, and 13, which indicate that SpaMA enhances the spatial features. For SpeMA, we can remark that its output contains low frequency of the input, such as results in Blocks 1, 7, 8, and 11, which can show that our SpeT has the potentiality to capture, explore, and exploit spectral features.

3) Investigation on Key Hyperparameters: There are five key hyperparameters for our SST including the pooling size to get the query and the key in SpeT, P, in (12); the number of heads for multihead attention, M, in (12) and (16); the number of channels for spatial features, D, in (16); and the number of transformer blocks in each branch, L, in (3) and (4). The influence of the kernel size used to fuse the spectral features in SpeT, K, is also considered. We conducted five groups of experiments to study the effects of each hyperparameter. The results are shown in Fig. 9. We can observe that the Q2n is insensitive to the different hyperparameters. Therefore, we mainly analyze the results based on ERGAS, SAM, and learnable parameter variations.

a) Performance versus P: We studied the effect of varying P in $\{1, 4, 8, 16, 18\}$ fixing the rest of the parameters as $K = 9$, $L = 10$, $M = 8$, and $D = 120$. Observing the bars of P for ERGAS and SAM, see Fig. 9(a) and (b), we can find that although a slight performance gain is achieved with the increase in the output size of the pooling, the learnable parameters are sharply increased. This is due to the fact that the parameters for the full connections used to generate the query and the key are quadratic with respect to the pooling size. Besides, we can find that $P = 4$ achieves a good performance, because 4×4 pixels are enough to represent the query and the key for spectral features, as discussed in Section III-B.

b) Performance versus K: We studied the effect of varying K in $\{1, 3, 5, 7, 9, 11\}$ fixing the rest of the parameters as $P = 4$, $L = 10$, $M = 8$, and $D = 120$. Having a look at the ERGAS and SAM results, we can find that $K = 9$ is the best setting.

c) Performance versus M: Intuitively, more heads enable more attention maps for different groups of features, which can enrich the features. As demonstrated by Vaswani et al. [4], multiple heads can explore the features in different subspaces. Thus, it is interesting to study how the number of heads impacts the performance for fusing HS and MS images. Therefore, we studied different numbers of heads in the multihead attention. Moreover, we added two special cases, i.e., only a head, $M = 1$, and each channel has a head, $M = D$. According to the results shown in Fig. 9, we can find that the performance increases for the SAM index from $M = 2$ to $M = 8$, while the performance for the ERGAS is up and down. For the two special cases, $M = 1$ only achieves slightly better results than $M = 2$; instead, the performance for $M = 120$ dramatically decreases. In our opinion, this is because the spatial attention map for each head is obtained by the similarity of a single pixel rather than a set of pixels when $M = D$. Therefore, the attention map for each head only considers the intensity similarity and cannot acquire similarity in other ways as done when a set of pixels is used. Overall, our experiments show that the performance can be improved in an appropriate range for the SAM metric and the performance visibly decreases when each channel is a head.

d) Performance versus D: We increased D from 30 to 180 with a step of 30. As shown in Fig. 9, the performance increases up to $D = 120$, and then decreases. Thus, $D = 120$ is set for the proposed approach.

e) Performance versus L: We increased L from 2 to 10 with a step of 2. It is obvious that the performance is improved by increasing L. Despite that, we do not choose a very big value for L, because we found that our SST cannot obtain a clear performance improvement when $L > 14$, and the training becomes unstable when L is too large. Therefore, $L = 14$ is set for the proposed approach.

V. Conclusion

In this article, we proposed an SST for HS and MS image fusion. We showed the SST blocks to extract spectral and spatial features. Experimental results demonstrated that our method achieves better results than all comparative methods in all the test cases highlighting the potentiality of the use of the transformer for HS and MS image fusion. The effectiveness of each module in our SST has been studied in the ablation study. Besides, a specific analysis about multihead attention for HS and MS image fusion has been presented together with some considerations from a computational point of view. Although the proposed SST is a lightweight transformer architecture, its computational burden is still higher than lightweight CNNs. Besides, as in almost all the DL solutions, the proposed SST shows a reduced generalization ability. Hence, future developments will be related to the study of more lightweight and generalized transformers for HS and MS image fusion.

Acknowledgment

Lihui Chen performed this work while he was at the National Research Council, Institute of Methodologies for Environmental Analysis (CNR-IMAA), Tito Scalo, Italy.

REFERENCES

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

[2] M. D. Raya-Sereno et al., "High-resolution airborne hyperspectral imagery for assessing yield, biomass, grain N concentration, and N output in spring wheat," *Remote Sens.*, vol. 13, no. 7, p. 1373, Apr. 2021.

[3] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[4] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[5] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[6] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[7] X. He and Y. Chen, "Optimized input for CNN-based hyperspectral image classification using spatial transformer network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1884–1888, Dec. 2019.

[8] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[9] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715.

[10] X. He, Y. Chen, and Z. Lin, "Spatial–spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, p. 498, Jan. 2021.

[11] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van G., and R. Timofte, "Swinir: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) Workshops*, Oct. 2021, pp. 1833–1844.

[12] Y. Cai et al., "Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction," 2021, *arXiv:2111.07910*.

[13] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.

[14] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615611.

[15] J.-F. Hu, T.-Z. Huang, and L.-J. Deng, "Fusformer: A transformer-based fusion approach for hyperspectral image super-resolution," 2021, *arXiv:2109.02079*.

[16] Q. Ma, J. Jiang, X. Liu, and J. Ma, "Learning a 3D-CNN and transformer prior for hyperspectral image super-resolution," 2021, *arXiv:2111.13923*.

[17] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–9.

[18] L. Loncan et al., "Hyperspectral pansharpening: A review," *IEEE Geosci. Remote Sens. Mag. Replaces Newsletter*, vol. 3, no. 3, pp. 27–46, Sep. 2015.

[19] G. Vivone et al., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag. Replaces Newsletter*, vol. 9, no. 1, pp. 53–81, Mar. 2021.

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

[21] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[22] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, vol. 89, pp. 405–417, Jan. 2023.

[23] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2020, *arXiv:2009.06732*.

[24] K. Han et al., "A survey on visual transformer," 2020, *arXiv:2012.12556*.

[25] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surveys*, vol. 54, no. 10s, pp. 1–41, Jan. 2022.

[26] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021, *arXiv:2106.04554*.

[27] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[28] Z. Chen, H. Pu, B. Wang, and G.-M. Jiang, "Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1418–1422, Aug. 2014.

[29] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS +pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[30] D. Picone, R. Restaino, G. Vivone, P. Addesso, M. Dalla Mura, and J. Chanussot, "Band assignment approaches for hyperspectral sharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 739–743, May 2017.

[31] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. JPL Summ. Annu. JPL Airborne Geosci. Workshop, AVIRIS Workshop*, vol. 1, 1992, pp. 1–3.

[32] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: A first approach on SIM-GA data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, Jun. 2015.

[33] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

[34] K. Zhang, F. Zhang, Z. Feng, J. Sun, and Q. Wu, "Fusion of panchromatic and multispectral images using multiscale convolution sparse decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 426–439, 2021.

[35] W. Dong et al., "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.

[36] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.

[37] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[38] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 348–362, Jan. 2020.

[39] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019.

[40] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

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

[42] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[43] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.

[44] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[45] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[46] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1753–1761.

[47] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.

[48] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022.

[49] X.-H. Han, Y. Zheng, and Y.-W. Chen, "Multi-level and multi-scale spatial and spectral fusion CNN for hyperspectral image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–10.

[50] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2090–2104, May 2021.

[51] W. Xie, J. Lei, Y. Cui, Y. Li, and Q. Du, "Hyperspectral pansharpening with deep priors," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1529–1543, May 2020.

[52] Y. Zheng et al., "Edge-conditioned feature transform network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5513315.

[53] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by CNN denoiser," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1124–1135, Mar. 2021.

[54] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

[55] W. Dong, C. Zhou, F. Wu, J. Wu, G. Shi, and X. Li, "Model-guided deep hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 5754–5768, 2021.

[56] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-net for hyperspectral image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2511–2520.

[57] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, "Deep blind hyperspectral image fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4150–4159.

[58] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 87–102.

[59] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 208–224.

[60] K. Zheng et al., "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, Mar. 2021.

[61] W. Wei, J. Nie, L. Zhang, and Y. Zhang, "Unsupervised recurrent hyperspectral imagery super-resolution using pixel-aware refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5500315.

[62] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3070–3079.

[63] L. Zhang, J. Nie, W. Wei, Y. Li, and Y. Zhang, "Deep blind hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2388–2400, Jun. 2021.

[64] W. Diao, F. Zhang, J. Sun, Y. Xing, K. Zhang, and L. Bruzzone, "ZeRGAN: Zero-reference GAN for fusion of multispectral and panchromatic images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 4, 2022, doi: 10.1109/TNNLS.2021.3137373.

[65] X. Fu, Q. Qi, Z.-J. Zha, X. Ding, F. Wu, and J. Paisley, "Successive graph convolutional network for image de-raining," *Int. J. Comput. Vis.*, vol. 129, no. 5, pp. 1691–1711, May 2021.

[66] X. Shu, L. Zhang, G.-J. Qi, W. Liu, and J. Tang, "Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3300–3315, Jun. 2022.

[67] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5281–5292, Aug. 2022.

[68] X. Meng, N. Wang, F. Shao, and S. Li, "Vision transformer for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5409011.

[69] M. Zhou, X. Fu, J. Huang, F. Zhao, A. Liu, and R. Wang, "Effective pan-sharpening with transformer and invertible neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5406815.

[70] A. Fred Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[72] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Post-capture control of resolution, dynamic range and spectrum," Dept. Comput. Sci., Columbia Univ., New York, NY, USA, Tech. Rep., 2008.

[73] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep., 2016.

[74] G. Vivone, M. Dalla Mura, A. Garzelli, and F. Pacifici, "A benchmarking protocol for pansharpening: Dataset, preprocessing, and quality assessment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6102–6118, 2021.

[75] M. Guizar-Sicairos, S. T. Thurman, and J. R. Fienup, "Efficient subpixel image registration algorithms," *Opt. Lett.*, vol. 33, no. 2, pp. 156–158, Jan. 2008.

[76] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris: Presses des MINES, 2002.

[77] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.

[78] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, Jan. 2000.

[79] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4109–4121, Nov. 2015.

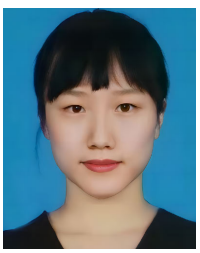
Lihui Chen (Member, IEEE) received the B.Sc. degree in electronics and information engineering and the Ph.D. degree in information and communication engineering from Sichuan University, Chengdu, Sichuan, China, in 2018 and 2022, respectively.

He visited the National Research Council, Institute of Methodologies for Environmental Analysis (CNR-IMAA), Tito Scalo, Italy, from August 2021 to August 2022. He is currently an Assistant Researcher with Chongqing University, Chongqing, China. He is interested in image processing, remote sensing, and deep learning.

Gemine Vivone (Senior Member, IEEE) received the B.Sc. (summa cum laude), the M.Sc. (summa cum laude), and the Ph.D. degrees in information engineering from the University of Salerno, Salerno, Italy, in 2008, 2011, and 2014, respectively.

In 2014, he joined the North Atlantic Treaty Organization (NATO), Science and Technology Organization (STO), Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy, as a Scientist. He was a Visiting Professor with the Grenoble Institute of Technology (INPG), Grenoble, France. In 2019, he was an Assistant Professor with the University of Salerno. He is currently a Researcher with the National Research Council, Institute of Methodologies for Environmental Analysis (CNR-IMAA), Tito Scalo, Italy, and also with the National Biodiversity Future Center (NBFC), Palermo, Italy. His main research interests include statistical signal processing, detection of remotely sensed images, data fusion, and tracking algorithms.

Dr. Vivone is an Editorial Board Member of Multidisciplinary Digital Publishing Institute (MDPI) Remote Sensing, MDPI Sensors, and MDPI Encyclopedia. He received the IEEE Geoscience and Remote Sensing Society (GRSS) Early Career Award in 2021, the Symposium Best Paper Award at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2015, and the Best Reviewer Award of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING in 2017. He is the Cochair of the IEEE Image Analysis and Data Fusion Technical Committee. He is an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL). He served as a guest associate editor for several special issues.



Jiayi Qin received the B.Sc. degree in information security and the M.Sc. degree in electronics and information engineering from Sichuan University, Chengdu, China, in 2020 and 2023, respectively.

She is interested in image processing, image super-resolution, and deep learning.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA; the KTH Royal Institute of Technology, Stockholm, Sweden; and the National University of Singapore (NUS), Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. He holds the AXA Chair in remote sensing and is an Adjunct Professor with the Aerospace Information research Institute, Chinese Academy of Sciences, Beijing, China. He has coauthored over 250 papers in international journals, gathering more than 31 500 citations, with H-index = 78. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008 and the Institut Universitaire de France from 2012 to 2017. He is a fellow of the European Laboratory for Learning and Intelligent Systems (ELLIS) and the Asia-Pacific Artificial Intelligence Association. He was the Founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE Geoscience and Remote Sensing Society (GRSS) Chapter Excellence Award. He has received multiple outstanding paper awards. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS). He was the Chair from 2009 to 2011 and the Cochair of the Geoscience and Remote Sensing (GRS) Data Fusion Technical Committee from 2005 to 2008. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. From 2017 to 2019, he was the Vice President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia. Since 2018, he has been a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters). He was an Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. In 2014, he served as a Guest Editor for the *IEEE Signal Processing Magazine*. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE.

Xiaomin Yang (Member, IEEE) received the B.Sc. and Ph.D. degrees in communication and information system from Sichuan University, Chengdu, China, in 2002 and 2007, respectively.

She was a Post-Doctoral Researcher with The University of Adelaide, Adelaide, SA, Australia, for one year. She is currently a Professor with the College of Electronics and Information Engineering, Sichuan University. Her research interests include image processing and pattern recognition.