Hybrid FusionNet: A Hybrid Feature Fusion Framework for Multisource High-Resolution Remote Sensing Image Classification

Yongjie Zheng, *Student Member, IEEE*, Sicong Liu, *Senior Member, IEEE*, Hao Chen, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*— With the increasing number of high-resolution (HR) images captured by various platforms, integrating spectral and spatial properties of data across different HR image types, such as multispectral (MS), hyperspectral (HS), and multitemporal (MT) images, remains a challenging task for object classification. This article proposes a novel hybrid framework named hybrid FusionNet (HFN) that jointly exploits 2-D–3-D convolutional neural networks (CNNs) and a transformer encoder to address a complex classification problem. By incorporating 2-D and 3-D convolutional layers, the proposed HFN generates rich multidimensional hybrid features, including spectral, spatial, and temporal features. These features are then fed into a transformer encoder to learn global saliency and discriminative information, enabling the identification of spatially irregular and spectrally similar objects. The hybrid architecture efficiently captures local intricate spectral-spatial-temporal contextual features through convolutional layers. Then, it learns global long-range dependencies and the spectral dimension through the transformer encoder, thus effectively reducing spectral-spatial mutations, distortions, and variations of ground objects. Experimental results from an high-resolution multispectral (HR-MS) dataset, an high-resolution hyperspectral (HR-HS) dataset, and an high-resolution multitemporal (HR-MT) dataset covering complex urban scenarios confirm the effectiveness of the proposed approach compared to the main state-of-the-art methods. Notably, the proposed HFN can achieve satisfactory classification performance even with limited training samples. The source code will be made available at https://github.com/MissYongjie/Hybrid-FusionNet.**

***Index Terms*— Convolutional neural networks (CNNs), feature fusion, high resolution (HR), image classification, remote sensing, transformer.**

I. Introduction

NOWADAYS, the increasing availability of multisource remote sensing images with much finer resolutions (e.g., high-resolution multispectral (HR-MS), high-resolution hyperspectral (HR-HS), and high-resolution multitemporal (HR-MT) images) allows a more precise classification of land surface objects on the Earth [1], [2], [3]. The intricate details inherent in high-resolution (HR) images are particularly valuable for applications such as urban monitoring, disaster monitoring, and precision agriculture. However, most of the current HR satellite images only have limited spectral bands and the high spatial resolution amplifies the intraclass variation and interclass similarity [4]. The HR-HS images acquired by aerial remote sensing platforms such as unmanned aerial vehicles (UAVs) or satellites have redundant spectral information. Thus, it is still difficult to adaptively extract spatially discriminative features without any intermediate operations (e.g., handcraft feature extraction) [5]. Considering these challenges with HR images, traditional classification techniques face limitations in handling the complexities associated with multiscale objects, such as small object identification, particularly in the case of few-shot learning [3], [6]. Therefore, it is necessary to develop more advanced methodologies that can achieve a precise and accurate classification of HR remote sensing data [6], [7], [8].

In the past decade, deep learning (DL)-based methods have been the hot topic in the field of image processing. The remarkable capabilities of DL-based models in automatically learning discriminative features from raw data have gained large attention in remote sensing image classification, particularly for HS images [9], [10], [11]. Extensive research efforts have been devoted to exploring the effectiveness of DL-based techniques, such as those based on convolutional neural network (CNN)-based with 1-D, 2-D, and 3-D convolutional layers, and Transformers [12], [13]. As the most classical and common DL-based model, CNNs exhibit a powerful ability to capture hierarchical deep features by employing convolutional filters and pooling operations [14]. Numerous CNN-based architectures have been adapted and optimized for remote sensing image classification. For instance, Hu et al. [15] applied a 1-D-CNN to the extraction of hierarchical spectral features for HS image classification. This technique only used spectral information without considering spatial information. Liu et al. [16] proposed a novel approach to HR remote sensing image classification based on a 2-D-CNN framework. By combining low-level and high-level spectral-spatial representations, this method effectively captures detailed spatial

Manuscript received 21 August 2023; revised 31 December 2023; accepted 8 January 2024. Date of publication 12 January 2024; date of current version 29 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42071324 and Grant 42241130, in part by the Shanghai Rising-Star Program under Grant 21QA1409100, in part by the Research Project of Tongji Architectural Design (Group) Company Ltd. under Grant 2023J-JB11, and in part by the China Scholarship Council. *(Corresponding authors: Lorenzo Bruzzone; Sicong Liu.)*

Yongjie Zheng and Lorenzo Bruzzone are with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it; yongjie.zheng@unitn.it).

Sicong Liu is with the College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China (e-mail: sicong.liu@tongji.edu.cn).

Hao Chen is with the Institute of Geodesy and Geoinformation Science, Technische Universität Berlin, 10553 Berlin, Germany (e-mail: hao.chen.2@campus.tu-berlin.de).

Digital Object Identifier 10.1109/TGRS.2024.3352812

information and high-level semantics. From the perspective of 3-D-CNN and dilated convolution, Ye et al. [17] designed a distinctive approach to HS image classification, which effectively extracts multiscale spatial and spectral features, enhances the discriminative power of features and improves the accuracy of HS image classification. Such models are effective in extracting spectral and spatial features. However, problems and challenges still exist due to the specific feature learning operations focused on a single type of HR images (e.g., HR-HS images). This limits the modeling of the interaction among multidimensional information, and the possible resulting redundant features may involve performance with limited training samples [18].

Inspired by the strong capability of adaptively recalibrating the nonlinear interdependence of deep features, the integration of attention mechanisms and CNN-based frameworks is widely employed in the field of image classification [19] and [20]. For instance, in [13], [21], [22], [23], [24], [25], [26], [27], [28], several spectral-spatial attention networks have been proposed that effectively integrate both spectral and spatial attention mechanisms. An attention-based adaptive spectral-spatial Kernel ResNet (A^2S^2K-ResNet) was introduced in [25]. It employs improved 3-D ResBlocks and an efficient feature recalibration (EFR) mechanism to effectively extract discriminative spectral-spatial features for HS image classification. Also, a triple-path spectral-spatial network with interleave-attention (TP-Net) was proposed for HS image classification [26]. It efficiently integrates spectral and spatial information through a hybrid branch, while an interleave-attention mechanism enhances feature interaction across branches. In [27], a multiscale and cross-level attention learning (MCAL) network was proposed to explore complex local land cover structures at different scales in HS images. MCAL incorporates a multiscale feature extraction module to capture local spatial context and a cross-level feature fusion module for hierarchical feature integration using attention mechanisms. In addition, the spectral attention module (SAM) enhances spectral information. Considering the challenge of few-shot learning, a 3-D multihead self-attention spectral-spatial feature fusion network (3DMHSA-SSFFN) was proposed in [28]. It incorporates step-by-step feature extraction (SBSFE) blocks with a 3-D multihead self-attention (3DMHSA) module, which effectively enhances the classification stability by extracting and correlating different spectral-spatial features.

More recently, the transformer model that uses the self-attention mechanism proposed by Vaswani et al. [29] also represents an essential milestone in the use of the attention mechanism and its variants [30], [31], [32] have been successfully applied to HS remote sensing image classification. For instance, Hong et al. [33] proposed a new Transformer-based model named SpectralFormer (SF). By utilizing the power of self-attention and global context modeling, it effectively captures intrinsic spectral information to support the HS image classification. A novel approach to HS image classification that integrates spectral and spatial information has been proposed in [34]. This method effectively captures the intricate patterns and dependencies within HS data by tokenizing spectral features and leveraging a Transformer-based architecture, thus improving classification accuracy and robustness. The technique in [35] incorporates group-aware attention mechanisms and a hierarchical transformer architecture, effectively capturing both local and global spatial dependencies, allowing more accurate and context-aware HS image classification. However, there is a challenge related to the training of such models that are characterized by a large number of parameters and poorly exploited limited training samples. Huang et al. [36] proposed a spectral-spatial masked transformer (SS-MTr) for HR-HS image classification, using a two-stage training strategy. In the first stage, SS-MTr pretrains a vanilla transformer with local inductive bias using masked HS image inputs. In the second stage, the pretrained transformer is fine-tuned with a fully connected (FC) layer for the final classification. Recent studies have demonstrated the considerable advantages of Transformer-based frameworks in remote sensing image classification. These frameworks offer a more comprehensive feature learning mechanism, showing promising results for HS image classification. However, further research is needed to explore the optimal utilization of Transformer-based models to reduce both the labeled sample requirements and the time consumption in HR image classification.

Despite the above-mentioned significant advancements in DL-based frameworks, there are still several challenges in HR image classification. The main ones can be summarized as follows.

1) *Multisource and Multitemporal HR Images:* The existing methods in the literature mainly target a single type of HR images (e.g., HS images), with feature extraction and fusion models specifically designed on their properties. Consequently, these network architectures are not able to jointly process different types of HR images.
2) *Ineffective Modeling of Hybrid Feature Information:* One of the key challenges in HR image classification is the effective modeling and fusion of spectral, spatial, and temporal information present in different objects in HR-MS, HR-HS, and HR-MT images. DL-based networks often struggle to comprehensively capture the complex interactions and dependencies among these different types of information, resulting in suboptimal classification performance.
3) *Limited Stability in Few-Shot Learning:* DL-based methods often require a large amount of labeled data for effective training. In the context of HR image classification, this requirement is seldom satisfied. Consequently, the stability of the models trained with limited sample sizes remains an area that requires further improvement, highlighting the need for improved strategies to deal with few-shot learning.

To address the aforementioned problems, in this article, we propose a novel framework for HR (including MS, HS, and MT) image classification, which consists of three main steps: data augmentation based on 3-D-CNN, hybrid feature interaction based on 2-D-CNN and triplet (TRP) attention mechanism, and global information discrimination based on transformer encoder. The proposed framework

combines the strengths of multidimensional features (spectral, spatial, and even temporal), enables end-to-end learning, and incorporates few-shot learning capabilities. The main contributions of this work can be summarized as follows.

1) *Adaptive Feature Prioritization:* The proposed Hybrid FusionNet (HFN) can model spectral-spatial-temporal feature interactions due to its adaptive feature prioritization capability. This adaptive functionality facilitates the extraction of significant and diverse features to support multisource and multistream classification tasks.
2) *Hierarchical Hybrid Feature Fusion:* The proposed approach fully exploits the information content of HR images by utilizing both 2-D and 3-D convolutional layers accompanied by a transformer encoder. Hierarchical hybrid feature extraction and fusion steps maintain spectral, spatial, and temporal consistency. This results in an end-to-end process that achieves improved classification performance with remarkable accuracy.
3) *Few-Shot Learning:* The proposed method, which contains a hybrid structure and a data augmentation strategy, exhibits high robustness and flexibility, achieving remarkable stability and generalization even with a small number of training samples.

The article provides an extensive experimental analysis on three kinds of HR images (MS, HS, and MT) to validate the effectiveness of the proposed HFN. The proposed method is compared against state-of-the-art techniques in various complex scenarios. Results show its superiority in multidimensional feature learning and classification performance.

The remainder of this article is structured as follows. Section II presents in detail the proposed HFN approach. Section III shows the experimental results and provides the quantitative and qualitative analysis. Finally, Section IV concludes the article, offering insights into future research directions.

II. Proposed Hybrid FusionNet: HFN

Considering the requirements of HR image classification, the proposed HFN framework is designed to address various types of classification tasks, including MS, HS, and MT tasks. The framework comprises several integrated components to achieve accurate and effective classification results. In this section, we first provide a description of individual components, and then introduce the whole HFN framework.

A. Convolutional Neural Networks

In the past decades, CNNs have played a major role in remote sensing image classification due to its strong deep feature exploration ability [37], [38]. Regarding the dimension of the convolutional layer, CNNs can be categorized into three types: 1-D, 2-D, and 3-D. In this article, we mainly utilize the 2-D and 3-D convolutional layers to properly extract multidimensional information to address the hybrid HR data tasks.

1) 2-D-CNN: 2-D-CNN employs 2-D convolutions, scanning across the height and width of an image to capture spatial context information [39]. Given an input tensor $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ (where C is the number of channels (channel dimension), and $H \times W$ represents the height and width (spatial dimension), respectively), the mathematical formulation of the 2-D convolution can be expressed as

$$\boldsymbol{X}_{i,j}^{l} = f_{\delta}\left(\omega^{l} * \boldsymbol{X}_{i,j}^{l-1} + b^{l}\right) \tag{1}$$

where $\boldsymbol{X}_{i,j}^{l}$ represents the output feature map at pixel spatial location (i, j) in the lth 2-D convolution layer, $\boldsymbol{X}_{i,j}^{l-1}$ is the input feature map of the lth layer, $*$ is the convolution operation, ω and b indicate the weight vector and the bias term, respectively, and f_{δ} represents the Rectified Linear Unit (ReLU) activation function.
2) 3-D-CNN: 3-D-CNN incorporates both the spatial dimension (height and width) and the temporal dimension (depth), and applies 3-D filters to collaborative capture spectral-spatial-temporal information that slides across the input data cube [40]. For HS data, each pixel in the data cube provides spectral-spatial information. The 3-D filters during convolution slide over these cubes, capturing spectral-spatial patterns. For MT data, each frame corresponds to the spectral-spatial data at a specific time instance. Stacking these frames along the depth dimension creates the 3-D data cube, thus the 3-D filters can capture spectral-spatial-temporal patterns. The mathematical formalization of the 3-D convolution can be expressed as

$$\boldsymbol{X}_{i,j,r}^{l} = f_{\delta}\left(\omega^{l} * \boldsymbol{X}_{i,j,r}^{l-1} + b^{l}\right) \tag{2}$$

where $\boldsymbol{X}_{i,j,r}^{l}$ represents the output feature map at pixel spatial location (i, j) and spectral or temporal location r in the lth 3-D convolution layer.

B. Transformer Encoder

Considering the strong ability of global feature generation, many Transformer-based models have been applied to remote sensing image classification, especially for classifying the HS image. Unlike CNN-based frameworks, which mainly rely on local receptive fields and might lose global information, Transformers exploit self-attention mechanisms that allow for global receptive fields, they can better capture long-range dependencies and contextual information in the remote sensing images.

By exploiting the self-attention mechanism, the transformer encoder can effectively model the complex relationships within the image data, enabling accurate and robust classification of HR remote sensing images [41], [42]. The architecture of the transformer encoder used in this article is shown in Fig. 1. It is composed of three parts: input embeddings, multihead self-attention mechanisms, and feed-forward neural networks, where the key characteristic of the transformer encoder is the multihead self-attention mechanisms. It involves three representing parameters and then obtains the global long-distance relationship. The core formalization of a self-attention mechanism is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{3}$$

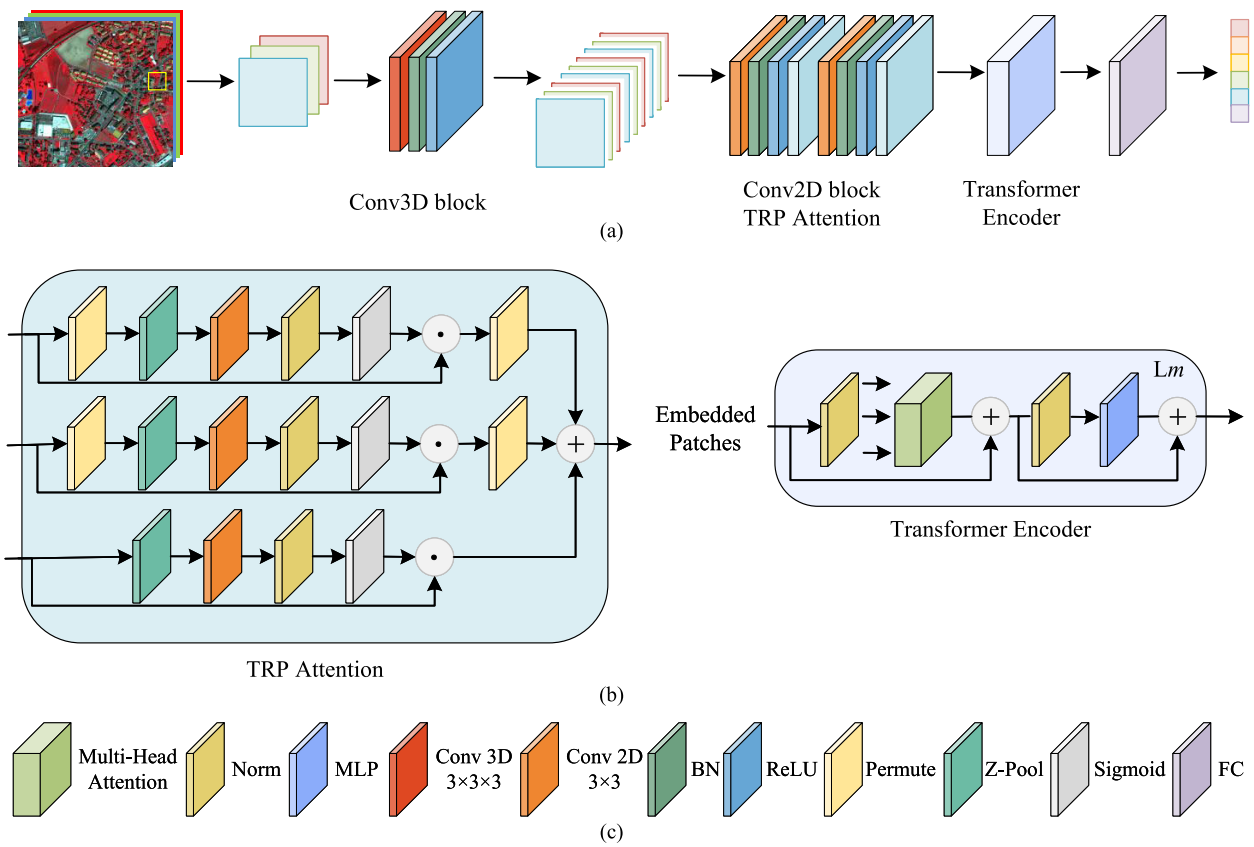


Fig. 1. Overview of proposed HFN, which comprises three connected steps. (a) Local data augmentation. (b) Local feature interaction. (c) Global information discrimination.

where V, K, and Q are the value, key, and query abbreviations, respectively. d_k represents the dimension of Q and K [43].

Let the multiple attention heads (MHA) be indexed by a, which can then be expressed as

$$\mathrm{MHA}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_a)W^O \quad (4)$$

where W^O represents a final projection matrix owned by the whole multiheaded attention head [42]. Each head is expressed as

$$\mathrm{head}_a = \mathrm{Attention}\left(QW_a^Q, KW_a^K, VW_a^V\right) \quad (5)$$

where W_a^Q, W_a^K, and W_a^V are projection matrices owned by individual attention head a.

C. TRP Attention

Attention mechanisms in DL-based models effectively handle the intricate relationships between different parts of an image by selectively focusing on the most relevant and significant features. One lightweight but effective mechanism is TRP attention [44]. Its structure is shown in Fig. 1, which consists of three branches, each responsible for capturing cross-dimension between the spatial dimensions and channel dimension of the input. Given an input feature tensor $\boldsymbol{X}$ with shape ($H \times W \times C$), each branch is responsible for aggregating cross-dimensional interactive features between either the spatial dimension H or W and the channel dimension C [45], [46]. In the last branch, the channels of the input tensor are reduced to two using Z-pooling, followed by a standard convolution layer and batch normalization (BN). Attention weights are generated by passing the output through a sigmoid activation layer, which is then applied to the input tensor. The TRP tensor is created by averaging the outputs of the three branches.

The Z-pool layer is used to efficiently reduce the dimensionality of the C-dimensional tensor to just 2-D. This reduction is achieved by concatenating the average pooled (AvgPool) features and the maximum pooled (MaxPool) features along this dimension. As a result, the layer retains a comprehensive representation of the original tensor, while significantly reducing its depth, thus making subsequent computations more efficient

$$Z\text{-pool}(\boldsymbol{X}) = \mathrm{Concat}(\mathrm{MaxPool}(\boldsymbol{X}), \mathrm{AvgPool}(\boldsymbol{X})) \quad (6)$$

$$\boldsymbol{y} = \mathrm{TRP}(\boldsymbol{X}) = \frac{1}{3}(\overline{\hat{\boldsymbol{X}}_1\omega_1} + \overline{\hat{\boldsymbol{X}}_2\omega_2} + \hat{\boldsymbol{X}}_3\omega_3) \quad (7)$$

where ω_1, ω_2, and ω_3 represent the three cross-dimensional attention weights computed in TRP attention; $\hat{\boldsymbol{X}}_1$, $\hat{\boldsymbol{X}}_2$, and $\hat{\boldsymbol{X}}_3$ represent $\boldsymbol{X}$ rotate 90° anticlockwise along the W-, H-, and C-axis, respectively [46].

D. Hybrid FusionNet

The fine architecture of the proposed HFN is shown in Fig. 1. It consists of three parts: 1) Conv3D block for local data augmentation; 2) Conv2D block with TRP attention for local feature interaction; and 3) transformer block for global information discrimination.

TABLE I
NETWORK AND PARAMETER SETTINGS OF THE PROPOSED HFN

Layers	Output shape	Parameters
Conv3d	[1, 64, 4, 7, 7]	432
ReLU	[1, 64, 4, 7, 7]	0
BatchNorm3d	[1, 64, 4, 7, 7]	32
Conv2d	[1, 256, 7, 7]	92.16M
ReLU	[1, 256, 7, 7]	0
BatchNorm2d	[1, 256, 7, 7]	6400
TRP	[1, 256, 7, 7]	200
Conv2d	[1, 256, 7, 7]	92.16M
ReLU	[1, 256, 7, 7]	0
BatchNorm2d	[1, 256, 7, 7]	6400
TRP	[1, 256, 7, 7]	200
Linear Embedding	[1, 49, 64]	51.216k
Transformer Encoder	[49, 1, 64]	5088
Linear Head	[1, 7]	272
Total Parameters: 185.194840M		

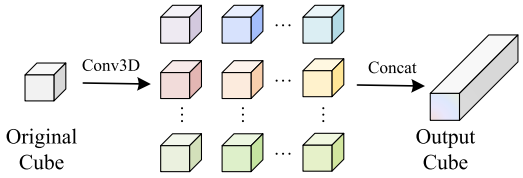


Fig. 2. Illustration of the local data augmentation based on the Conv3D block.

1) Local Data Augmentation (Conv3D): The Conv3D block enhances local data within HR images by applying 3-D convolutions. It expands and enriches the original multidimensional features, capturing spectral correlations across time or spectral dimensions and expressing spatial context information for improved classification performance.

To improve the rich expression of original features of HR images, the Conv3D block is first employed for data augmentation. As shown in Fig. 1 and Table I, this block includes three layers: a 3-D convolution layer, a ReLU layer, and a BN layer. First, the main convolution layer is used to introduce additional dimensions to the convolutional operations, simultaneously allowing for the augmentation of spectral or even spectral-temporal channels (see Fig. 2). Assuming the input patch $\boldsymbol{I^0} \in \mathbb{R}^{1\times C\times w\times w}$, a $3 \times 3 \times 3$ convolution layer with 64 kernels is first employed to transform $\boldsymbol{I^0}$ into $\boldsymbol{I^1} \in \mathbb{R}^{64\times C\times w\times w}$ for augmenting the original features. Next, the ReLU and BN layers follow a convolutional layer to add nonlinearity, prevent vanishing gradients, normalize activations, and improve the convergence of this network. This stage can be formalized as

$$\boldsymbol{I}^1 = F^{3d}_{\delta,\mathrm{BN}}(\boldsymbol{I}^0) = \mathrm{BN}(f_\delta(\omega^{3d} * (\boldsymbol{I}^0) + b^{3d})). \tag{8}$$

2) Local Feature Interaction (Conv2D + TRP): This block combines the Conv2D module with the TRP attention mechanism module. The Conv2D extracts detailed spatial context information, while TRP attention captures spatial-spectral relationships. This fusion enhances spatial-spectral joint features, promoting effective feature interaction.

From the perspective of local hybrid spectral-spatial feature interaction, two Conv2D blocks with the TRP modules are proposed further to capture effective detailed spatial-spectral contextual information without any pooling operation. Specifically, the Conv2D block consists of a conv2D layer with a kernel size of 3×3, a BN layer, and a ReLU layer, where the 2-D convolution operation is mainly applied to extract the detailed spatial contextual information inherent in HR images.

In addition to the Conv2D block, the TRP module is designed to capture significant spectral-spatial information by leveraging the interdependencies between different spatial locations. Considering the relationships among nearby pixels can effectively enhance the model's ability to capture the interaction between spectral-spatial features along with the spectral and spatial dimensions.

The output tensor $\boldsymbol{I}^1$ of the Conv3D block is first reshaped as $\boldsymbol{I}^1 \in \mathbb{R}^{(64\times C)\times w\times w}$. Then, the output tensor of one Conv2D block with a TRP module can be formalized as

$$\begin{aligned}\boldsymbol{I}^2 &= \mathrm{TRP}\big(F^{2d}_{\delta,\mathrm{BN}}(\boldsymbol{I}^1)\big)\\ &= \mathrm{TRP}(\mathrm{BN}(f_\delta(\omega^{2d} * (\boldsymbol{I}^1) + b^{2d}))).\end{aligned} \tag{9}$$

3) Global Information Discrimination (Transformer Encoder): The transformer encoder block globally fuses spectral-spatial information, capturing long-range dependencies across different channels. This enables HFN to comprehend discriminative information, reducing spatial-spectral variations in land objects and improving recognition accuracy.

The acquired fusion features are given as input to a transformer encoder block to generate global discriminative information from the perspective of spectral or temporal channels. The transformer encoder block is well-known for its ability to capture long-range dependencies. In the context of global spectral information learning, the transformer encoder block is utilized to analyze and discriminate the spectral consistency. It consists of several self-attention layers, which allow the model to attend to different parts of the multidimensional features and discriminatively extract the meaningful relationships along with the spectral dimension. By dynamically weighting the spectral information, the transformer encoder block can emphasize discriminative features and suppress irrelevant or noisy components.

From the perspective of the spectral channels, here the output features of the previous stage are reshaped as $\boldsymbol{I}^2 \in \mathbb{R}^{(w\times w)\times 64\times C}$. From Table I, one can see that, the $\boldsymbol{I}^2$ is first given as input to the linear embedding layer, then a transformer encoder is used to project the input data into a higher dimensional space. This allows the transformer encoder to effectively capture complex patterns and relationships in the data. Afterward, three transformer encoder layers are used to further process the data and fuse higher level hybrid features for the final task. Notably, the parameter a was set to 8 in this article. Thus, the output tensor $\boldsymbol{I}^3 \in \mathbb{R}^{(w\times w)\times 64}$ passing the MHA in the proposed HFN can be then expressed as

$$\begin{aligned}\boldsymbol{I}^3 &= \mathrm{MHA}\big(\boldsymbol{I}^2 W_q + \boldsymbol{I}^2 W_k + \boldsymbol{I}^2 W_v\big)\\ &= \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_8) W^O\end{aligned} \tag{10}$$

TABLE II
NUMBER OF SAMPLES FOR EACH CLASS IN DIFFERENT DATASETS

No.	ZH1		ZH2		ZH17		UP		SH	
	Classes	Samples	Classes	Samples	Classes	Samples	Classes	Samples	Classes	Samples
C1	Roads	120532	Roads	86551	Roads	154786	Asphalt	6631	Building	13400
C2	Buildings	251222	Buildings	154164	Buildings	150627	Meadows	18649	Trees	13954
C3	Trees	79153	Trees	81033	Trees	111072	Gravel	2099	Water	4318
C4	Grass	311833	Grass	7201	Grass	129125	Trees	3064	Roads	1358
C5	Bare Soil	72429			Bare Soil	10619	Painted metal sheets	1345	Grass	2582
C6	Railways	16043			Water	9040	Bare soil	5029		
C7	Swimming Pools	5070			Swimming Pools	6052	Bitumen	1330		
C8							Self-blocking bricks	3682		
C9							Shadows	947		

where W_q, W_k, and W_v are the corresponding matrices to transform the $\boldsymbol{I}^2$ vector into the Q, K, and V vectors, respectively.

After three sequential transformer encoder blocks, the acquired discriminative information is directly input to a FC layer for the final classification. The FC layer takes the output $\boldsymbol{I}^3$ from the transformer encoder as its input and performs a linear transformation, which acts as the classifier, making use of the learned representations from the transformer encoder to classify the input data into the appropriate classes. The final classification result is obtained as

$$\boldsymbol{Y} = \mathrm{FC}(\boldsymbol{I}^3) \tag{11}$$

where $\boldsymbol{Y} \in \mathbb{R}^{1\times c}$, and c is the number of classes.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Description of Datasets

In this section, three HR-MS subdatasets, one HR-HS dataset, and one HR-MT dataset from three benchmark datasets are first introduced. Then, experimental settings, results, and ablation studies are described.

To evaluate the effectiveness of the proposed HFN, we choose three benchmark datasets, which include three types of HR data: HR-MS, HR-HS, and HR-MT. Three HR-MS subdatasets, i.e., Zurich 1 (ZH1), Zurich 2 (ZH2), and Zurich 17 (ZH17), were acquired by the QuickBird (QB) satellite over the urban area of Zurich, Switzerland. One HR-HS dataset, i.e., PaviaU (UP), was acquired by the reflective optics systems imaging spectrometer (ROSIS) sensor over the Pavia University, Northern Italy. One HR-MT dataset, i.e., Shanghai (SH), was acquired by the Gaofen (GF) satellite over the urban area of SH. Details of these three datasets are given as follows.

1) Zurich (ZH) Dataset: The ZH dataset includes four spectral channels with a spatial resolution of 0.62 m after the pan sharpening operation. The false-color images and the corresponding ground truth maps of three HR-MS datasets (ZH1, ZH2, and ZH17) are shown in Figs. 4–6, respectively. Table II shows the number of samples and classes of these three datasets, respectively.

2) PaviaU (UP) Dataset: The UP dataset captures high spatial resolution (1.3 m) and high spectral resolution (103 bands after removing some noisy bands) information. The false-color image and the ground truth map are shown in Fig. 7(a) and (b), respectively. As shown in Table II, the UP dataset includes a total of nine classes.

3) SH Dataset: The SH dataset includes four MT images acquired from different satellites [GF-1B (T1), GF-1 (T2), GF-1D (T3), and GF-6 (T4)] in SH on April 8, April 12, April 15, and April 18, 2019, respectively. These sensors include HR cameras and wide-format cameras. Each image consists of 2 m panchromatic images and 8 m MS images with four spectral bands. Therefore, we first used the Gram–Schmidt pan sharpening fusion algorithm to complete the fusion processing of panchromatic and MS data to obtain 2 m HR-MS data, and then performed registration and cropping of different images. False-color composite images of different times and the ground truth map can be seen in Fig. 8(a)–(e), respectively. As shown in Table II, there are five land-cover classes in this scenario.

B. Experimental Settings

To evaluate the effectiveness of the proposed HFN approach, we performed a comparative analysis against eight state-of-the-art methods, namely, A^2S^2K-ResNet [25], SF [33], shallow-to-deep feature fusion network (SDF^2N) [16], spectral–spatial feature tokenization transformer (SSFTT) [34], group-aware hierarchical transformer (GAHT) [35], TP-Net [26], 3DMHSA-SSFFN [28], and supervised contrastive SS-MTr (SC-SS-MTr) [36]. Note that the publicly available source code 3DMHSA-SSFFN was not found, and therefore it was reimplemented based on the information provided in the original paper. To ensure a fair comparison, all methods were implemented using PyTorch and executed on an NVIDIA RTX A6000 GPU with 48 685 MB of available memory.

The main parameters that influence on the classification accuracy were set as follows. The input patch size was set as 7×7 (experiments with different patch sizes and network structures can be found in Section III-C), the batch size was set to 128, and the number of epochs was equal to 300. In our method, we utilized the Adam optimizer with a fixed learning rate of 1e-4 and did not employ any learning rate scheduler. For the compared networks, while maintaining the network parameters as consistent as possible with those suggested in the related papers, we adjusted the patch size to 7×7, the number of epochs to 300, and the batch size to 128 to ensure a fair comparison. Considering the limited availability of the

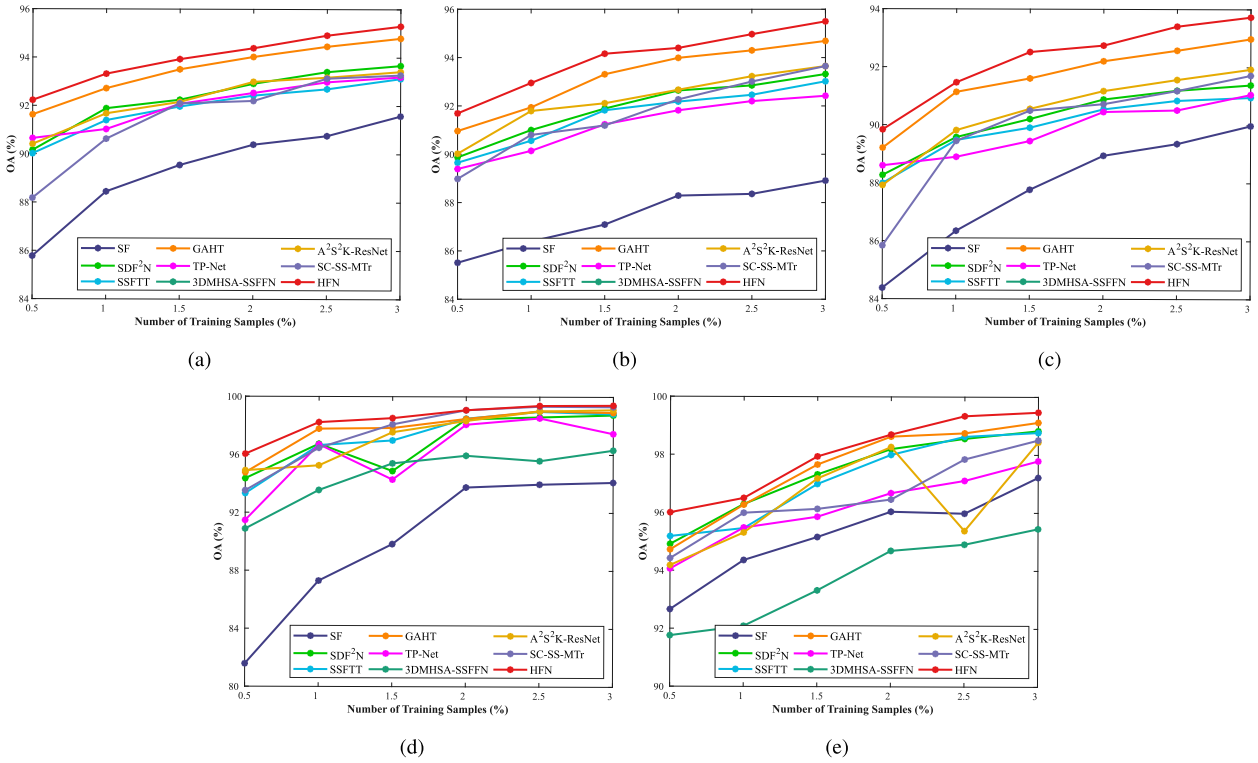


Fig. 3. OA obtained by different methods under different numbers of randomly selected training samples on (a) ZH1 dataset, (b) ZH2 dataset, (c) ZH17 dataset, (d) UP dataset, and (e) SH dataset.

TABLE III
AVERAGE ACCURACIES ON FIVE REPEATED TRIALS ON FIVE CONSIDERED DATASETS FOR DIFFERENT SIZES OF INPUT PATCH WITH 0.5% TRAINING SAMPLES

	3×3	5×5	7×7	9×9	11×11
ZH1	87.58 ±0.26	90.11 ± 0.27	92.10 ±**0.15**	93.90 ±**0.15**	**94.90** ±0.17
ZH2	86.58 ±0.16	88.91 ±0.20	91.70 ±**0.15**	92.28 ±0.32	**93.32** ±0.23
ZH17	85.71 ±0.16	87.71 ±0.30	89.70 ±0.17	91.68 ±0.20	**92.10** ±**0.15**
UP	92.57 ±1.50	94.62 ±**0.36**	96.32 ±0.51	96.68 ±0.48	**97.79** ±0.60
SH	94.72 ±1.26	94.88 ±1.37	95.58 ±1.14	95.74 ±0.96	**95.84** ±**0.79**

TABLE IV
AVERAGE OA ON FIVE REPEATED TRIALS OBTAINED IN ABLATION STUDIES OF DIFFERENT FUSION STEPS WITH 0.5% TRAINING SAMPLES

Strategies	OA(%)				
	ZH1	ZH2	ZH17	UP	SH
step 1	88.74 ±0.09	87.76 ±0.36	87.18 ±0.36	89.85 ±0.67	94.77 ±1.36
step 2	89.11 ±0.25	88.04 ±0.18	87.50 ±0.24	94.24 ±0.77	95.31 ±1.27
step 3	88.69 ±0.20	85.61 ±0.31	86.80 ±0.38	92.94 ±0.60	94.03 ±0.86
step 1&2	91.24 ±0.28	90.60 ±0.28	88.76 ±0.38	92.93 ±1.17	95.00 ±1.35
step 1&3	91.97 ±0.17	91.43 ±0.34	89.54 ±0.21	95.39 ±1.03	95.04 ±1.55
step 1&2&3	**92.10** ±**0.15**	**91.70** ±**0.15**	**89.70** ±**0.17**	**96.32** ±**0.51**	**95.58** ±**1.14**

minimum window size of the SC-SS-MTr method, and we set it to 9 × 9. In order to avoid gradient explosion, the number of epochs of the SC-SS-MTr method were set to 100, 100, and 300 on ZH, SH, and UP datasets, respectively.

To quantitatively evaluate the classification performance of the compared methods, different quality indices, i.e., overall accuracy (OA), average accuracy (AA), class accuracy (CA), Kappa, and time cost are employed.

C. Ablation Studies

1) Impact of Patch Size: Different patch sizes have distinct effects on the acquired information for image classification. Smaller patch sizes (such as 3 × 3 or 5 × 5) can usually capture more local information, while larger patch size (such as 9 × 9 or 11 × 11) contains more spatial context information. In general, the larger patch size of the input map will conduct higher classification accuracy, also with more time-consuming computations. To effectively identify which size is suitable for the experiments, we consider five different patch sizes, i.e., 3 × 3, 5 × 5, 7 × 7, 9 × 9, and 11 × 11. From Table III, one can note that, as expected, as the patch size increases, the classification accuracy generally improves across all datasets. This result confirms that larger patch sizes capture more spatial information, allowing the model to better distinguish between different land-cover classes. However, it is important to consider several factors when selecting the appropriate patch size, i.e., the spatial resolution of the datasets, the suitable

TABLE V
CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS WITH 0.5% TRAINING SAMPLES ON THE ZH1 DATASET

Classes	A^2S^2K-ResNet 2021 [28]	SF 2022 [36]	SDF^2N 2022 [18]	SSFTT 2022 [37]	GAHT 2022 [38]	TP-Net 2022 [29]	3DMHSA-SSFFN 2023 [31]	SC-SS-MTr 2023 [39]	HFN
C1	84.74	77.60	83.01	83.18	86.69	85.41	55.10	81.73	87.48
C2	90.89	85.32	90.21	89.77	92.05	92.02	73.60	87.83	92.29
C3	76.71	61.73	78.24	78.07	79.39	74.92	50.78	68.51	77.72
C4	95.69	96.14	94.94	94.97	95.21	95.78	95.64	96.89	96.10
C5	99.31	97.98	98.80	99.16	99.32	99.11	83.89	99.53	99.32
C6	67.67	22.55	68.34	69.34	79.68	70.42	14.57	7.86	83.06
C7	91.98	94.56	94.39	93.50	94.92	94.13	86.64	86.87	96.31
OA	90.74 ±0.19	85.94 ±0.24	90.16 ±0.34	90.09 ±0.10	91.68 ±0.20	90.91 ±0.28	76.76 ±0.15	87.97 ±0.16	92.10 ±0.15
AA	86.71 ±0.63	76.55 ±0.50	86.85 ±1.48	86.86 ±0.36	89.61 ±0.44	86.19 ±0.37	65.75 ±1.02	75.60 ±0.95	90.33 ±0.36
Kappa	87.55 ±0.24	80.95 ±0.30	86.78 ±0.48	86.69 ±0.13	88.83 ±0.25	87.73 ±0.36	68.41 ±0.11	83.70 ±0.23	89.38 ±0.20
Time (s)	2146.29 ±81.90	473.43 ±1.45	165.27 ±1.16	161.51 ±8.76	286.32 ±3.00	389.47 ±1.70	636.97 ±0.82	345.22 ±19.5	298.42 ±5.01

computational time, the model stability, and the performance of other literature methods. Therefore, by considering all these valuables, a patch size of 7×7 is chosen as the benchmark size for the experiments. The size can provide a good balance between accuracy and model stability.

2) Impact of Sequential Structure: To show the impact of different network structures, a detailed ablation study (see Table IV) is made based on different combinations of the three fusion steps in the HFN approach on the five datasets. As shown in Table IV, when only a single step (step 1, step 2, or step 3) is considered (see rows 1–3), all five datasets show relatively poor performance due to insufficient multidimensional feature learning. In contrast, combining all three steps leads to the highest classification accuracy and the lowest deviation on ZH1 (92.10% ± 0.15%), ZH2 (91.70% ± 0.15%), and ZH17 (89.70% ± 0.17%) three datasets, and UP (96.32% ± 0.51%) and SH (95.58% ± 1.14%) datasets, demonstrating the effectiveness of the proposed sequential structure in improving the performance of the HFN network.

D. Results and Discussion

1) Quantitative Analysis: The quantitative classification results of the nine considered DL-based methods on the five datasets are shown in Fig. 3 and Tables V–IX.

Fig. 3(a)–(e) illustrates the OA values obtained by different methods for six percentages (0.5%, 1%, 1.5%, 2%, 2.5%, and 3%) of training samples extracted from the available labeled data on ZH1, ZH2, ZH17, UP, and SH datasets, respectively. The results indicate that the proposed HFN approach, represented by the red curves, achieves the highest OA values on all datasets with largest improvements in the first three HR-MS datasets. Among the eight comparison methods, GAHT exhibits the most favorable performance, whereas SF presents the lowest accuracy due to the insufficient number of training samples. Several other methods (A^2S^2K-ResNet, SDF^2N, SSFTT, TP-Net, 3DMHSA-SSFFN, and SC-SS-MTr) achieve intermediate accuracies. Note that A^2S^2K-TP-Net and ResNet on the UP and SH datasets display significant fluctuates in performance when varying training sample sizes.

To further evaluate the performance of these nine DL-based methods, we report the average OA values with 0.5% of training samples on the five datasets in Tables V–IX.

For the ZH1 dataset (see Table V), the proposed HFN method performed consistent well for most of the classes, achieving the highest average OA (92.10%), AA (90.33%), and Kappa (89.38%) values with minimal fluctuations. For the specific class C6 (railways), the accuracies varied significantly in the different methods ranging from 7.86% to 83.06%. The HFN method achieved the highest accuracy of 83.06%, while GAHT had the second highest accuracy of 79.68%. This approved the ability of the proposed HFN to effectively capture and preserve the details of the irregular and narrow objects. With regard to the computational time of the nine methods, the SSFTT method shows the lowest value with 161.51 s, whereas the A^2S^2K-ResNet took the highest time of 2146.29 s. The proposed method consumed almost 298.42 s.

For the ZH2 dataset (see Table VI), it can be observed that the HFN method shows consistently robust performance across all classes, especially for C1 (roads) with the highest CA value (88.02%). In general, it also achieved the highest average OA value (91.70%). Note that the HFN method did not achieve the highest accuracy on the C4 (grass) class (64.02%), but this value is close to that of the highest accuracy produced by the GAHT method (65.47%), while the SC-SS-MTr method had the lowest accuracy of only 0.31%. In addition, the best accuracy was obtained by the proposed method with a low computation cost (116.16 s).

Similar to the previous two HR-MS datasets, Table VII also demonstrates the superiority of the HPN method on the ZH17 dataset, which achieving the highest CA on roads (89.63%), buildings (86.09%), water (93.93%), and swimming pools (96.64%). Overall, it also shows the highest average OA (89.70% ± 0.17), AA (90.63% ± 0.55), and Kappa (86.58% ± 0.21) values with minimal fluctuations. Although

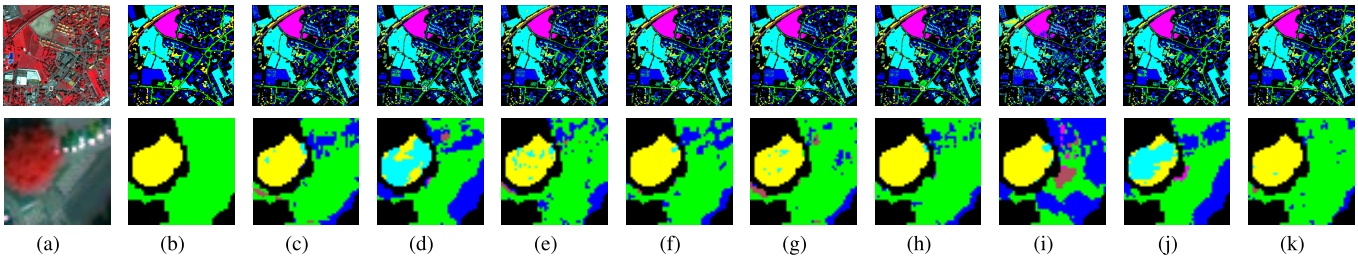


Fig. 4. Classification maps and zoom on a specific region obtained by different methods on the ZH1 dataset. (a) False-color composite image (RGB: bands 4, 3, and 2). (b) Ground reference map. (c) A^2S^2K-ResNet (90.41%). (d) SF (85.78%). (e) SDF^2N (90.16%). (f) SSFTT (90.02%). (g) GAHT (91.63%). (h) TP-Net (90.65%). (i) 3DMHSA-SSFFN (76.75%). (j) SC-SS-MTr (88.09%). (k) Proposed HFN (92.23%).

TABLE VI

CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS WITH 0.5% TRAINING SAMPLES ON THE ZH2 DATASET

Classes	A^2S^2K-ResNet 2021 [28]	SF 2022 [36]	SDF^2N 2022 [18]	SSFTT 2022 [37]	GAHT 2022 [38]	TP-Net 2022 [29]	3DMHSA-SSFFN 2023 [31]	SC-SS-MTr 2023 [39]	HFN
C1	85.50	79.78	84.80	83.45	87.44	85.10	78.37	72.74	88.02
C2	90.24	86.07	90.45	90.79	90.93	90.66	77.31	92.29	92.19
C3	97.02	97.02	96.62	96.40	96.82	96.69	97.89	98.66	97.14
C4	57.34	18.65	62.84	58.55	65.47	59.99	11.66	0.31	64.02
OA	89.94 ±0.38	85.64 ±0.15	89.88 ±0.2	89.54 ±0.13	90.91 ±0.2	90.01 ±0.32	81.22 ±0.39	86.70 ±4.17	91.70 ±0.15
AA	82.52 ±1.04	70.38 ±2.82	83.68 ±1.64	82.30 ±2.01	85.16 ±0.48	83.11 ±1.44	66.31 ±1.98	66.00 ±4.24	85.34 ±1.26
Kappa	84.53 ±0.59	77.86 ±0.24	84.42 ±0.27	83.88 ±0.19	86.04 ±0.3	84.62 ±0.53	71.36 ±0.55	79.02 ±7.07	87.23 ±0.24
Time (s)	743.10 ±0.5	182.67 ±0.94	66.73 ±0.59	61.67 ±0.99	111.78 ±0.52	148.63 ±1.53	244.55 ±0.94	124.17 ±1.29	116.16 ±0.77

TABLE VII

CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS WITH 0.5% TRAINING SAMPLES ON THE ZH17 DATASET

Classes	A^2S^2K-ResNet 2021 [28]	SF 2022 [36]	SDF^2N 2022 [18]	SSFTT 2022 [37]	GAHT 2022 [38]	TP-Net 2022 [29]	3DMHSA-SSFFN 2023 [31]	SC-SS-MTr 2023 [39]	HFN
C1	87.41	86.93	88.75	89.03	88.93	89.27	87.59	88.68	89.63
C2	84.54	78.32	83.05	82.91	84.58	84.68	67.06	81.71	86.09
C3	91.48	86.47	91.07	92.02	91.63	88.86	87.27	90.01	91.63
C4	91.67	89.60	91.77	90.03	92.29	92.95	91.63	94.15	92.15
C5	77.25	72.41	75.21	79.53	84.45	78.34	51.90	12.37	84.33
C6	88.64	79.30	86.24	89.75	92.67	85.63	86.30	42.09	93.93
C7	91.96	96.22	94.97	95.36	94.97	95.12	96.63	72.20	96.64
OA	88.29 ±0.27	84.88 ±0.53	88.16 ±0.16	88.12 ±0.22	89.11 ±0.25	88.61 ±0.19	82.44 ±0.85	86.01 ±0.48	89.70 ±0.17
AA	87.56 ±0.71	84.18 ±1.46	87.30 ±0.81	88.37 ±0.33	89.93 ±0.54	87.84 ±0.92	81.20 ±1.58	68.74 ±4.03	90.63 ±0.55
Kappa	84.74 ±0.34	80.31 ±0.69	84.56 ±0.21	84.53 ±0.28	85.82 ±0.32	85.15 ±0.25	77.10 ±1.11	81.68 ±0.64	86.58 ±0.21
Time (s)	1289.10 ±0.63	314.93 ±1.06	113.93 ±0.59	108.94 ±6.15	195.88 ±1.43	263.09 ±2.59	424.42 ±0.38	232.26 ±10.22	200.97 ±3.67

the time consumption of the proposed HFN (200.97 ± 3.67 s) is not the smallest, it is relatively low and efficient enough for classification tasks.

On the HR-HS dataset [UP dataset (see Table VIII)], the HFN method also shows the highest accuracy with an OA value of 96.32% ± 0.51, outperforming all other methods. Notably, while the A^2S^2K-TP-Net, SDF^2N, and GAHT methods also perform well, the HFN method provided much higher OA, AA, and kappa values than them, indicating its superior classification performance. However, in our proposed method, the data augmentation operation based on Conv3D increases the time consumption for a high number of channels such as HS. This time consumption can be decreased by adjusting the convolution kernel parameters.

TABLE VIII

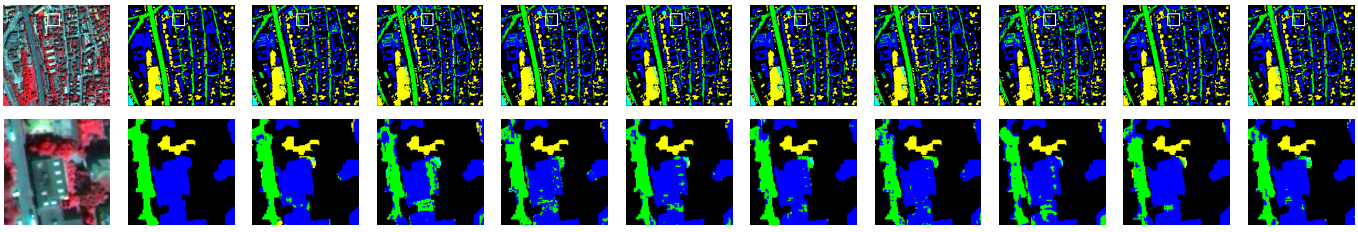
CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS WITH 0.5% TRAINING SAMPLES ON THE UP DATASET

Classes	A^2S^2K-ResNet 2021 [28]	SF 2022 [36]	SDF^2N 2022 [18]	SSFTT 2022 [37]	GAHT 2022 [38]	TP-Net 2022 [29]	3DMHSA-SSFFN 2023 [31]	SC-SS-MTr 2023 [39]	HFN
C1	92.23	80.92	95.47	95.31	94.71	94.27	91.03	98.39	95.82
C2	97.82	94.74	98.04	97.96	98.32	97.34	95.03	99.53	98.88
C3	71.03	51.42	73.50	65.44	80.40	61.12	54.36	20.92	82.89
C4	90.63	82.67	95.85	97.07	96.40	91.72	86.71	89.48	94.88
C5	99.90	99.97	99.82	99.61	99.98	99.11	99.77	99.96	99.97
C6	93.09	58.06	93.29	93.13	92.65	80.17	65.70	97.00	96.59
C7	92.43	48.88	77.62	75.00	84.18	65.15	33.64	73.16	88.44
C8	92.41	66.32	87.01	85.87	88.09	81.10	82.81	97.33	93.31
C9	99.79	95.07	99.77	99.83	99.66	98.44	95.13	85.07	99.74
OA	94.05 ±0.79	81.59 ±0.99	94.23 ±0.44	93.66 ±0.92	94.84 ±0.78	90.35 ±0.73	85.56 ±1.18	93.16 ±1.16	96.32 ±0.51
AA	92.15 ±1.69	75.34 ±1.93	91.15 ±1.34	89.91 ±2.24	92.71 ±1.67	85.38 ±0.68	78.24 ±3.08	84.54 ±4.10	94.50 ±1.31
Kappa	92.09 1.07	75.23 ±1.30	92.35 ±0.60	91.59 ±1.21	93.15 ±1.04	87.11 ±0.95	80.63 ±1.65	90.88 ±1.57	95.11 ±0.67
Time (s)	104.96 ±0.44	86.58 ±0.4	17.2 ±0.95	15.23 ±0.13	22.05 ±0.06	29.27 ±0.12	376.55 ±0.93	58.54 ±0.97	85.63 ±2.82

TABLE IX

CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS WITH 0.5% TRAINING SAMPLES ON THE SH DATASET

Classes	A^2S^2K-ResNet 2021 [28]	SF 2022 [36]	SDF^2N 2022 [18]	SSFTT 2022 [37]	GAHT 2022 [38]	TP-Net 2022 [29]	3DMHSA-SSFFN 2023 [31]	SC-SS-MTr 2023 [39]	HFN
C1	95.79	95.25	96.03	93.47	95.61	95.88	97.38	98.22	97.14
C2	98.47	97.86	98.81	98.38	97.75	99.03	96.00	99.26	99.21
C3	99.11	98.83	99.68	99.34	99.43	99.04	98.43	99.20	99.74
C4	39.09	28.48	38.85	45.64	39.06	37.69	16.16	4.94	40.03
C5	87.53	76.64	87.65	89.91	91.05	84.24	67.37	85.28	90.20
OA	94.48 ±1.22	92.82 ±1.12	94.78 ±0.87	94.02 ±1.76	94.42 ±1.27	94.43 ±1.44	91.70 ±0.06	94.25 ±0.37	95.58 ±1.14
AA	79.41 ±1.24	84.21 ±3.08	85.35 ±0.98	84.58 ±0.79	83.18 ±2.02	75.07 ±1.47	84.00 ±1.52	77.38 ±0.91	85.26 ±1.86
Kappa	89.39 ±1.77	92.31 ±1.61	91.29 ±1.25	91.83 ±2.46	91.79 ±1.80	87.65 ±2.09	91.88 ±0.09	91.44 ±0.52	93.49 ±1.68
Time (s)	74.46 ±0.2	11.25 ±0.75	8.8 ±0.77	17.39 ±0.06	22.87 ±0.41	33.63 ±0.39	163.28 ±0.24	20.03 ±0.75	40.69 ±1.6



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 5. Classification maps and zoom on a specific region obtained by different methods on the ZH2 dataset. (a) False-color composite image (RGB: bands 4, 3, and 2). (b) Ground reference map. (c) A^2S^2K-ResNet (90.02%). (d) SF (85.51%). (e) SDF^2N (89.87%). (f) SSFTT (89.65%). (g) GAHT (90.96%). (h) TP-Net (89.39%). (i) 3DMHSA-SSFFN (81.34%). (j) SC-SS-MTr (89.25%). (k) Proposed HFN (91.69%).

On the short-term HR-MT dataset [SH dataset (see Table IX)], one can see that among all the methods, the proposed HFN still has the highest OA value of 95.58% with a relatively low computational cost of 40.69 s. This indicates that the HFN method outperforms the reference methods also on the short-term HR-MT dataset.

2) Qualitative Analysis: Fig. 4 provides the classification maps (Fig. 4 row 1) and zoom on a specific region (Fig. 4

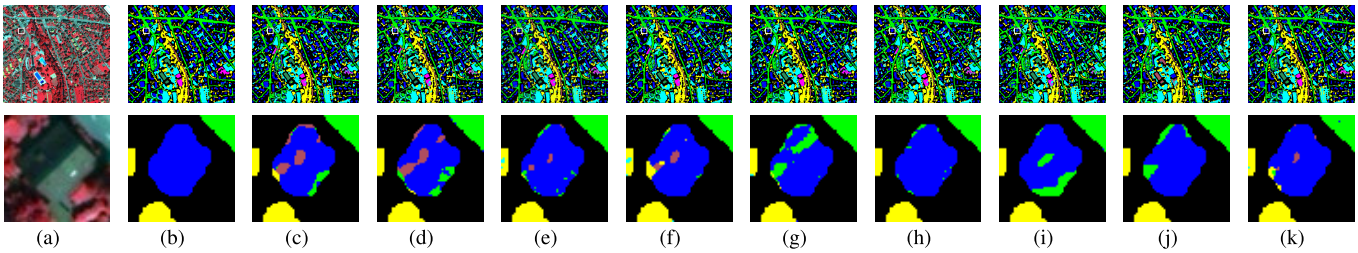


Fig. 6. Classification maps and zoom on a specific region obtained by different methods on the ZH17 dataset. (a) False-color composite image (RGB: bands 4, 3, and 2). (b) Ground reference map. (c) A^2S^2K-ResNet (87.92%). (d) SF (84.38%). (e) SDF^2N (88.27%). (f) SSFTT (87.99%). (g) GAHT (89.21%). (h) TP-Net (88.60%). (i) 3DMHSA-SSFFN (82.60%). (j) SC-SS-MTr (85.95%). (k) Proposed HFN (89.84%).

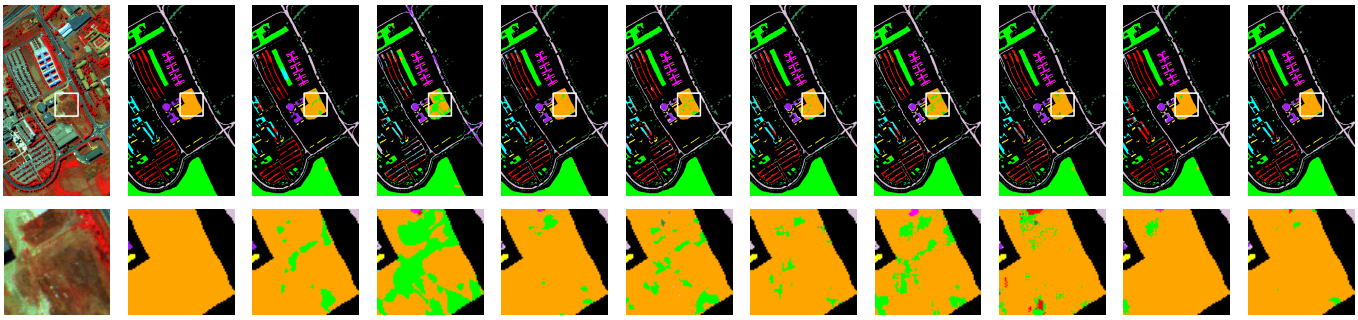


Fig. 7. Classification maps and zoom on a specific region obtained by different methods on the UP dataset. (a) False-color composite image (RGB: bands 90, 50, and 10). (b) Ground reference map. (c) A^2S^2K-ResNet (94.90%). (d) SF (81.58%). (e) SDF^2N (94.35%). (f) SSFTT (93.32%). (g) GAHT (94.76%). (h) TP-Net (91.47%). (i) 3DMHSA-SSFFN (90.88%). (j) SC-SS-MTr (93.52%). (k) Proposed HFN (96.04%).

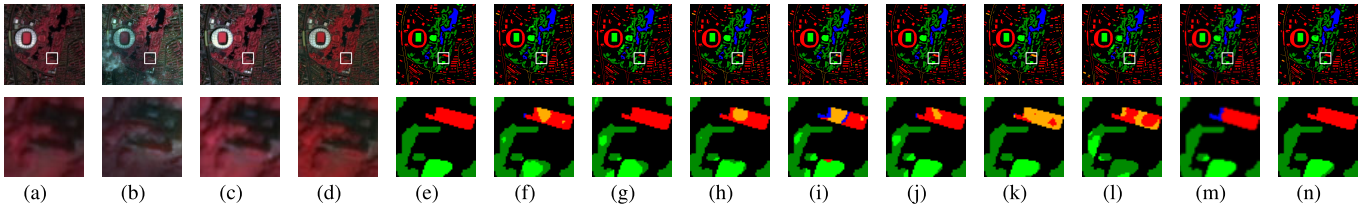


Fig. 8. Classification maps and zoom on a specific region obtained by different methods on the SH dataset. (a)–(d) False-color composite images T1–T4 (RGB: bands 4, 3, and 2). (e) Ground reference map. (f) A^2S^2K-ResNet (94.18%). (g) SF (92.66%). (h) SDF^2N (94.91%). (i) SSFTT (95.18%). (j) GAHT (94.72%). (k) TP-Net (94.06%). (l) 3DMHSA-SSFFN (91.76%). (m) SC-SS-MTr (94.42%). (n) Proposed HFN (96.00%).

row 2) of different methods with 0.5% training samples on the ZH1 dataset. From the zoomed classification maps, the proposed approach presents fewer noisy pixels and better spatial continuity, particularly in spectrally similar classes, such as roads (bright green) and buildings (blue), trees (yellow) and grass (cyan).

Similar to the ZH1 dataset, the zoomed classification maps of the other HR-MS datasets [ZH2 (Fig. 5) and ZH17 (Fig. 6)] obtained by the eight reference methods, one can see that there are many misclassified pixels in the roads (bright green) and buildings (blue) classes. In contrast, the proposed HFN approach better models the external edges and internal homogeneity of these two classes, thus reducing the classification errors and achieving the best classification performance with the highest OA accuracies of 91.69% and 89.84%, respectively.

Fig. 7 provides the classification maps (Fig. 7 row 1) and zoom on a specific region (Fig. 7 row 2) obtained by different methods with 0.5% training samples on the UP dataset. From the comparison of zoomed classification maps, we can see that the map obtained by the HFN method contains more regular and correctly classified objects (see Fig. 7). The proposed method effectively reduces misclassification errors, especially for spatially irregular objects [e.g., bare soil (orange) and meadows (bright green)].

Fig. 8 presents the classification maps (Fig. 8 row 1) and zoom on a specific region (Fig. 8 row 2) obtained by different methods with 0.5% training samples on the SH dataset. The zoomed classification maps show that the HFN method is effective in reducing misclassification errors and better distinguishing spectrally similar classes [e.g., buildings (red) and roads (orange), trees (dark green) and grass (bright green)].

IV. Conclusion

In this article, a novel adaptive hybrid feature extraction and fusion network (HFN) based on CNNs and transformer models has been proposed for multisource HR remote sensing image classification. The proposed HFN introduces a novel strategy to solve complex classification problems by effectively fusing multidimensional information (spectral-spatial or spectral-spatial-temporal) and leveraging the power of 2-D–3-D CNN and transformer encoder models. The proposed technique demonstrates remarkable performance in

identifying spatially irregular and spectrally similar objects. Specifically, the proposed HFN is composed of three core multidimensional feature learning steps: 1) a local data augmentation, which is used to increment the dimension of raw HR data, especially for MS images; 2) a local feature fusion step, which utilizes 2-D convolutional layers with a TRP mechanism for integrating multiscale spatial context information and further interacting with spectral information; and 3) a global information discrimination step, which introduces the transformer encoder for capturing global long dependencies in the hybrid fusion features obtained from the previous steps. Compared with eight popular state-of-the-art reference methods, experimental results obtained on real HR-MS, HR-HS, and HR-MS datasets confirmed the superior performance of the proposed HFN. Notably, by introducing data augmentation strategies and local-to-global feature interaction mechanisms, it effectively alleviates the inaccurate identification of the inherent spectral complexity and spatial variability present in some spatially complex objects and spectrally similar objects. In addition, the proposed HFN approach shows the highest accuracy with the limited training samples, and maintains excellent model stability and generalization.

Although the current model has demonstrated remarkable performance in multisource HR remote sensing image classification tasks, its efficiency in HR-HS classification is reduced, with increased time consumption. For future development, we plan to explore more lightweight spectral-spectral-temporal feature augmentation and fusion techniques to address the issue of time consumption. In addition, we will explore and incorporate more efficient and stable pretrained generative networks (e.g., generative adversarial network (GAN) or diffusion models) for the classification of large complex scenes in HR satellite images. Moreover, we plan to extend the application of the method to real-world large-scale urban land use and land cover (LULC) classification tasks, where it can provide more efficient and accurate results.

ACKNOWLEDGMENT

The author Yongjie Zheng thanks the China Scholarship Council (CSC) for the financial support of her Ph.D. study at the University of Trento. The authors thank all the researchers and institutions involved for providing these benchmark HR datasets. In particular, the authors would like to express their gratitude to Dr. Michele Volpi from the Swiss Federal Institute of Technology Zurich for providing the QB images, and to Prof. Paolo Gamba from the Telecommunications and Remote Sensing Laboratory, Pavia University for providing the HS image, and also to the China Center for Resources Satellite Data and Application for providing the GF images.

REFERENCES

[1] Y. Tao, M. Xu, F. Zhang, B. Du, and L. Zhang, "Unsupervised-restricted deconvolutional neural network for very high resolution remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 6805–6823, Dec. 2017.

[2] Y. Zheng, S. Liu, Q. Du, H. Zhao, X. Tong, and M. Dalponte, "A novel multitemporal deep fusion network (MDFN) for short-term multitemporal HR images classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10691–10704, Oct. 2021.

[3] S. Liu, Y. Zheng, Q. Du, A. Samat, X. Tong, and M. Dalponte, "A novel feature fusion approach for VHR remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 464–473, Dec. 2021.

[4] J. R. Bergado, C. Persello, and A. Stein, "Recurrent multiresolution convolutional networks for VHR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6361–6374, Nov. 2018.

[5] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial–spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Sep. 2019.

[6] C. Persello and A. Stein, "Deep fully convolutional networks for the detection of informal settlements in VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2325–2329, Dec. 2017.

[7] S. Dong, Y. Zhuang, Z. Yang, L. Pang, H. Chen, and T. Long, "Land cover classification from VHR optical remote sensing images by feature ensemble deep learning network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1396–1400, Aug. 2020.

[8] C. Shi, L. Fang, Z. Lv, and H. Shen, "Improved generative adversarial networks for VHR remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, Sep. 2022.

[9] W. Li and Q. Du, "A survey on representation-based classification and detection in hyperspectral remote sensing imagery," *Pattern Recognit. Lett.*, vol. 83, pp. 115–123, Nov. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865515003207

[10] D. Datta, P. K. Mallick, A. K. Bhoi, M. F. Ijaz, J. Shafi, and J. Choi, "Hyperspectral image classification: Potentials, challenges, and future directions," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–36, Apr. 2022.

[11] H. Pan, X. Zhao, H. Ge, M. Liu, and C. Shi, "Hyperspectral image classification based on multiscale hybrid networks and attention mechanisms," *Remote Sens.*, vol. 15, no. 11, p. 2720, May 2023.

[12] L. Yang et al., "FusionNet: A convolution-transformer fusion network for hyperspectral image classification," *Remote Sens.*, vol. 14, no. 16, p. 4066, Aug. 2022.

[13] X. Mei et al., "Spectral–spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 8, p. 963, Apr. 2019.

[14] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[15] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jul. 2015.

[16] S. Liu et al., "A shallow-to-deep feature fusion network for VHR remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, May 2022, Art. no. 5410213.

[17] Z. Ye, J. Wang, and L. Bai, "Multi-scale spatial–spectral feature extraction based on dilated convolution for hyperspectral image classification," in *Proc. 6th Int. Conf. Image Graph. Process.* New York, NY, USA: Association for Computing Machinery, Jan. 2023, pp. 97–103.

[18] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[19] C. Pu, H. Huang, and L. Yang, "An attention-driven convolutional neural network-based multi-level spectral–spatial feature learning for hyperspectral image classification," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115663.

[20] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[21] Y. Zhan, K. Wu, and Y. Dong, "Enhanced spectral–spatial residual attention network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7171–7186, 2022.

[22] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[23] Y. Cui, J. Xia, Z. Wang, S. Gao, and L. Wang, "Lightweight spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5510114.

[24] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, Nov. 2020.

[25] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.

[26] Z. Deng et al., "A triple-path spectral–spatial network with interleave-attention for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5906–5923, 2022.

[27] F. Xu, G. Zhang, C. Song, H. Wang, and S. Mei, "Multiscale and cross-level attention learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, Art. no. 5501615.

[28] Q. Zhou, S. Zhou, F. Shen, J. Yin, and D. Xu, "Hyperspectral image classification based on 3-D multihead self-attention spectral–spatial feature fusion network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1072–1084, 2023.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[30] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[31] Z. Liu et al., "Video Swin Transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3202–3211.

[32] Z. Liu et al., "Swin Transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12009–12019.

[33] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art. no. 5518615.

[34] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5522214.

[35] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5539014.

[36] L. Huang, Y. Chen, and X. He, "Spectral–spatial masked transformer with supervised and contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023, Art. no. 5508718.

[37] W. Lv and X. Wang, "Overview of hyperspectral image classification," *J. Sensors*, vol. 2020, pp. 1–13, Jul. 2020.

[38] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2485–2501, 2020.

[39] H. Fırat, M. E. Asker, M. İ. Bayindir, and D. Hanbay, "Spatial–spectral classification of hyperspectral remote sensing images using 3D CNN based LeNet-5 architecture," *Infr. Phys. Technol.*, vol. 127, Dec. 2022, Art. no. 104470.

[40] D.-H. Jung, J. D. Kim, H.-Y. Kim, T. S. Lee, H. S. Kim, and S. H. Park, "A hyperspectral data 3D convolutional neural network classification model for diagnosis of gray mold disease in strawberry leaves," *Frontiers Plant Sci.*, vol. 13, Mar. 2022, Art. no. 837020.

[41] J. Bai, J. Lu, Z. Xiao, Z. Chen, and L. Jiao, "Generative adversarial networks based on transformer encoder and convolution block for hyperspectral image classification," *Remote Sens.*, vol. 14, no. 14, p. 3426, Jul. 2022.

[42] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, "MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art. no. 5224212.

[43] D. Wang, J. Zhang, B. Du, L. Zhang, and D. Tao, "DCN-T: Dual context network with transformer for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 2536–2551, 2023.

[44] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3138–3147.

[45] X. Yang et al., "Fast multi-shadow tracking for video-SAR using triplet attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art. no. 5224212.

[46] Y.-C. Chen, K.-T. Lai, D. Liu, and M.-S. Chen, "TAGNet: Triplet-attention graph networks for hashtag recommendation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1148–1159, Mar. 2022.

Yongjie Zheng (Student Member, IEEE) received the B.S. degree in remote sensing science and technology from Henan Polytechnic University, Jiaozuo, China, in 2018, and the M.S. degree in photogrammetry and remote sensing from Tongji University, Shanghai, China, in 2021. She is currently pursuing the Ph.D. degree in remote sensing image analysis with the University of Trento, Trento, Italy.

Her research interests include deep learning, feature extraction, feature fusion, and remote sensing image classification and change detection.

Sicong Liu (Senior Member, IEEE) received the B.Sc. degree in geographical information system and the M.E. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2009 and 2011, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, 2015.

He is currently an Associate Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. His research interests include multitemporal data analysis, change detection, multi-spectral/hyperspectral remote sensing in Earth observation and planetary exploration.

Dr. Liu serves as the Program Committee Member for the SPIE Remote Sensing Symposium: Image and Signal Processing for Remote Sensing (since 2020). He was the winner (ranked as the third place) of Paper Contest of the 2014 IEEE GRSS Data Fusion Contest. He is the Technical Co-Chair of the Tenth International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp 2019). He served as the Session Chair for many international conferences, such as International Geoscience and Remote Sensing Symposium (since 2017). He is an Associated Editor and a Guest Editor for several international journals.

Hao Chen was born in 1994. He received the M.Sc. degree in geographic information system from Tongji University, Shanghai, China, in 2020. He is currently pursuing the Ph.D. degree in surveying and mapping engineering with the Technical University of Berlin, Berlin, Germany.

From July 2020 to January 2021, he was a Research Assistant with the College of Surveying and Geo-Informatics, Tongji University. His research interests include spatial data processing, planetary mapping (mainly for lunar surface and small bodies), and deep learning.

Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (M.S.) degree in electronic engineering (summa cum laude) and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is the Founder and the Director of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among the others, he is currently the Principal Investigator of the Radar for Icy Moon Exploration (RIME) instrument in the framework of the JUpiter ICy moons Explorer (JUICE) mission of the European Space Agency (ESA) and the Science Lead for the High Resolution Land Cover project in the framework of the Climate Change Initiative of ESA. He is the author (or coauthor) of more than 360 scientific publications in referred international journals, more than 350 papers in conference proceedings, and 22 book chapters. He is the editor/coeditor of 18 books/conference proceedings and one scientific book. His papers are highly cited, as proven from the total number of citations (more than 49 000) and the value of the H-index (104) (source: Google Scholar). He was invited as keynote speaker at more than 40 international conferences and workshops. His research interests are in the areas of remote sensing, radar and SAR, signal processing, machine learning, and pattern recognition.

Dr. Bruzzonee has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) since 2009, where he has been the Vice-President for Professional Activities since 2019. He ranked the First Place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, July 1998. Since then, he was a recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards, and the 2019 WHISPER Outstanding Paper Award. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He was a Guest Coeditor of many Special Issues of international journals. He is the Co-Founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) Series and is currently a member of the Permanent Steering Committee of this series of workshops. He has been the Founder of the *IEEE Geoscience and Remote Sensing Magazine* for which he has been the Editor-in-Chief from 2013 to 2017. Currently, he is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He was a Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society from 2012 to 2016.