S^{2}PNet: An Interactive Learning Framework for Addressing Spatial–Spectral Heterogeneity in H^2 Imagery Classification

Shuai Zhang, Yonghua Jiang, *Member, IEEE*, Chengjun Wang, *Member, IEEE*, Meilin Tan, *Member, IEEE*, Bin Du, *Member, IEEE*, and Feng Tian, *Member, IEEE*

***Abstract*—Hyperspectral imagery with high spatial resolution (H^2) imagery can synchronously obtain the spectral and spatial features of objects, thus providing richer information. However, the exacerbated spatial–spectral heterogeneity poses new challenges for classification. In this study, an interactive learning framework was proposed to address the current issues in H^2 imagery classification. Specifically, we propose a spectral–spatial purification network (S^{2}PNet) to improve classification accuracy. First, a multistage spectral purification module is designed to purify noisy information and mitigate spectral heterogeneity, achieving interaction between spectral optimization and classification. Second, a global–local mutual guide module is utilized to realize image–pixel-level feature interaction, thus enhancing the spatial discriminability of extracted features and reducing spatial heterogeneity. Third, the introduction of dual-stream semantic progressive module facilitates shallow-deep feature interaction, reducing the semantic gap in internal network and enabling a smoother information flow. We validated our approach using the public WHU-Hi hyperspectral datasets and large-scale Houston datasets. Experimental results demonstrate that S^{2}PNet achieves the highest classification accuracy across all tests, significantly outperforming state-of-the-art methods.**

***Index Terms*—High spatial resolution (H^2) imagery, Houston datasets, interactive learning, semantic gap, spatial–spectral heterogeneity, WHU-Hi datasets.**

I. Introduction

HYPERSPECTRAL imagery (HSI) is renowned for providing continuous spectral information integrated with spatial data, and this crease rich and crucial data support in various fields, such as geological surveys, environmental monitoring, urban planning, and precision agriculture [1], [2]. With the rapid development of hyperspectral technology in recent years, HSI has evolved toward higher spectral and spatial resolutions [3]. HSI with high spatial resolution (H^2) imagery provides narrower spectral bands and more detailed spatial information. However, this also leads to more severe spatial–spectral heterogeneity, resulting in increased intraclass variance and reduced interclass differences [4], as shown in Fig. 1. This severely affect the ability of the image to distinguish between land cover categories, making HSI classification extremely challenging.

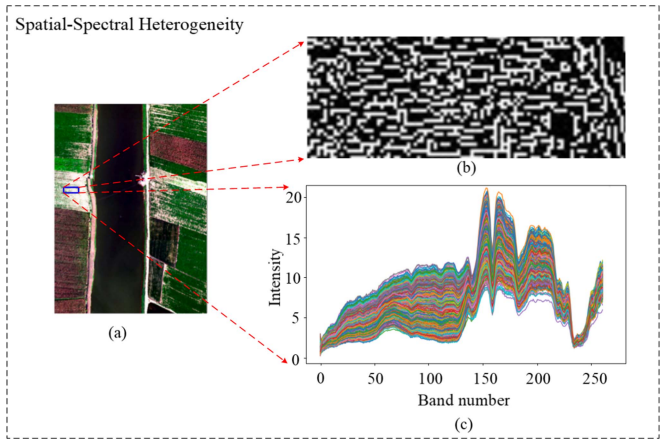


Fig. 1. Spatial–spectral heterogeneity. (a) H^2 Imagery. (b) We used the Sobel operator to extract the gradient map of the selected region, illustrating the heightened spatial heterogeneity in H^2 imagery. (c) Spectral curve demonstrates the spectral heterogeneity within the same category.

Numerous studies have aimed to improve the performance of HSI classification. Traditional methods primarily rely on spectral features, such as kernel-based classifiers [5], support vector machines (SVM) [6], [7], random forests [8], [9], and multinomial logistic regression [10]. Owing to the heightened spectral heterogeneity in H^2 imagery, these methods face significant challenges. Subsequent research revealed that spatial information is more crucial than spectral information for HSI classification [11]. Therefore, methods that integrate both types of information, such as wavelet transformation [12], gray-level co-occurrence matrix [13], Gabor filters [14], [15], and Markov random fields [16], have been proposed to maintain the local consistency of class labels

Received 9 July 2024; revised 26 August 2024; accepted 10 September 2024. Date of publication 20 September 2024; date of current version 21 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 41971412 and in part by the National Key Research and Development Program of China under Grant 2023YFF1303702. *(Corresponding author: Yonghua Jiang).*

Shuai Zhang, Yonghua Jiang, and Feng Tian are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China (e-mail: zhangshuai2023@whu.edu.cn; jiangyh@whu.edu.cn; tian.feng@whu.edu.cn).

Chengjun Wang is with the School of Computer Science and School of Cyberspace Science, Xiangtan University, Xiangtan 430300, China (e-mail: changjunwang@xtu.edu.cn).

Meilin Tan and Bin Du are with the Inner Mongolia Autonomous Region Surveying and Mapping Geographic Information Center, Hohhot 010050, China (e-mail: tanmeilin@whu.edu.cn).

Code is available online at https://github.com/zshandsome/S2PNet

Digital Object Identifier 10.1109/JSTARS.2024.3464758

in pixel neighborhoods. However, the features extracted using aforementioned traditional methods require manual design based on prior knowledge and empirical data, resulting in instability to the classification effect.

Recently, deep learning technology has made significant advances in remote sensing image understanding, particularly in the imaging and analysis of HSI. For example, Li et al. [17] introduced the CasFormer model, which combines deep learning with RGB image fusion, focusing on both spatial and spectral domains to significantly enhance the quality of HSI. Hong et al. [18] proposed a subpixel level hyperspectral superresolution framework, which progressively integrates HS-MS information from the pixel level to the subpixel level, and from the image level to the feature level. Hong et al. [19] designed a domain-adaptive network that effectively preserves the spatial topology of remote sensing images through parallel high- and low-resolution fusion, achieving improved segmentation performance and generalization capability. Notably, deep learning-based methods have gradually become the mainstream approach for HSI classification due to their ability to adaptively extract and integrate advanced spatial–spectral features from images [20]. These methods primarily include stacked sparse autoencoders [21], convolutional neural networks (CNN) [22], [23], recurrent neural networks [24], and deep belief networks [25]. Among these, CNNs and their variants have significantly improved the accuracy of HSI classification when labeled samples are abundant, and they have consequently received widespread attention [26]. Although current deep learning methods have achieved excellent performance, difficulties and challenges remain when dealing with H^2 imagery. Specific limitations include the following.

1) Due to the higher spatial resolution of H^2 imagery, the number of mixed pixels is reduced, and the degree of mixing between different components is significantly lowered, resulting in more pronounced spatial heterogeneity in the images. Specifically, the spectral differences between pixels in H^2 imagery become more distinct, and the boundaries between different materials are clearer. This enhanced spatial heterogeneity imposes greater demands on existing image processing models, requiring them to better understand and integrate both global and local information.
2) H^2 imagery, characterized by narrower spectral bands and reduced mixing degrees, leads to more pronounced spectral differences within the same object, resulting in stronger intraclass spectral heterogeneity. This increased intraclass spectral heterogeneity complicates classification, as even within the same plot, identical objects may exhibit significantly different spectral characteristics. This complexity demands that models more accurately capture spectral variations and effectively differentiate between different object classes.

However, existing methods predominantly use patch-based processing, where the patch size directly affects the model's ability to capture global contextual information. This limitation renders current models inadequate for addressing the enhanced spatial heterogeneity in H^2 imagery. And the significant overlap between adjacent pixel patch regions increases computational redundancy and reduces the model's inference speed. In addition, current approaches to addressing spectral heterogeneity mainly involve band selection to stack dominant bands, with little consideration given to nonlinear relationships or band combinations, resulting in limited effectiveness. Moreover, these methods struggle with end-to-end optimization and cannot adaptively adjust based on a classification performance feedback.

In this study, to address the challenges posed by the spatial–spectral heterogeneity in H^2 imagery classification, we propose a spectral–spatial purification network (S^{2}PNet) based on an interactive learning framework. In S^{2}PNet, we drew inspiration from neural architecture search (NAS) [27] and designed an interactive multistage spectral purification module (MSSP). This module is guided by the classification results and selects most distinguishable bands and their combinations for each category. Unlike NAS, we did not alter the network structure; instead, we implemented this process using learnable convolutional kernel operations, which significantly reduces the computational complexity. In addition, we have discovered a strong underlying connection between convolution and self-attention. Leveraging this relationship, we propose a global–local mutual guide (GLMG) module to facilitate the mutual fusion and complementation of global and local features. It is worth noting that we integrate the parallel concept of grouped convolution into this process, further reducing computational complexity. Furthermore, we re-evaluated the skip connections and designed a dual-stream semantic progressive module (DSSP), which considers local information correlations and the bidirectional long-term dependency to mitigate this semantic gap between encoder and decoder features. The main contributions of this study are summarized as follows.

1) We designed an interactive MSSP that integrates spectral optimization and classification tasks within a unified framework. MSSP allows band selection to be guided by classification results and dynamically evaluates the contribution and relevance of each band to the classification task, thus seeking the optimal spectral combinations for each category, effectively solving the problems associated with spectral heterogeneity and information redundancy.
2) We propose a GLMG, which combines the advantages of convolution and Transformer to enhance the interaction and compensation between image-level and pixel-level features to cope with the challenge of spatial heterogeneity. By frozen kernel shift convolution and the edge masked multihead window self-attention (MWSA) mechanism, more discriminative spatial features are captured, which play a crucial role in solving the spatial heterogeneity problem.
3) We reexamined information interaction across various stages of the network and propose a DSSP that allows information to flow bidirectionally between shallow and deep features, progressively aligning the semantic representations of the shallow features with the deeper ones,

thereby mitigating the semantic discrepancies associated with skip connections and facilitating a more coherent and enriched feature representation.

4) An interactive learning framework is proposed for H^2 imagery classification, which can effectively address the spatial–spectral heterogeneity particularly for limited labeled samples. We conducted extensive experiments on five datasets to demonstrate the superiority of the proposed approach.

The rest of this article is organized as follows. In Section II, we review related studies on HSI classification. In Section III, we provide a detailed description of S^2PNet. An analysis of the experimental results is provided in Section IV, and ablation studies are discussed in Section V. Finally, Section VI concludes this article.

II. Related Work

In this section, we provide a comprehensive review of the CNN-based HSI classification methods and band selection methods.

A. Band Selection

Researchers have employed preprocessing techniques, such as band selection and feature extraction, to address the challenges of spectral variability [13]. Strategies for hyperspectral band selection include methods based on ranking, clustering, searching, embedding, and deep learning [28]. Ranking-based approaches evaluate the importance of each band by using manually selected criteria. For instance, Chang et al. [29] proposed the minimum-variance PCA (MVPCA) method while MVPCA does not account for the band correlation. To address this limitation, Chang et al. [30] introduced a method based on the Kullback–Leibler distance to eliminate redundant bands. Clustering-based methods fully consider the relationships between bands. Sun et al. [31] proposed sparse subspace clustering to obtain the required subset of bands, and Zhai et al. [32] developed Laplacian-regularized low-rank subspace clustering to reduce the representation bias in candidate bands. Search-based and embedding-based methods optimize objective functions to obtain optimal band subsets. Zhang et al. [33] used conflicting indicators, such as information content and redundancy to jointly constrain the search process. In addition, embedding methods, such as REF-SVM, rank feature weights during the training phase to eliminate unimportant features [34]. Recently, deep learning has been widely applied in band selection. Zhan et al. [35] proposed a method combining CNN and distance density. Cai et al. [36] added an attention mechanism to CNN and proposed a unified band selection framework, and Sellami et al. [37] designed a semisupervised low-redundancy criterion to combine semisupervised 3D-CNNs to extract features from the selected bands. Moreover, Feng et al. [38] introduced fixed and adaptive band selection strategies using reinforcement learning, avoiding repeated selection of the same bands. However, the existing band selection methods suffer from two main shortcomings. First, they require manual parameter tuning for evaluation and primarily focus on selecting significant bands, while paying less attention to band combinations. These methods do not screen band information at the initial stage of the model but rather perform equivalent feature selection in the feature space, which still retains noisy band information, making it difficult to effectively address spectral heterogeneity issues. Second, these methods are not coupled with the classification task, thus they cannot optimize the band subset based on feedback from the classification results.

B. CNN-Based HSI Classification

Owing to the fine feature extraction ability of convolution operations, a series of CNN-based methods has been extensively investigated for HSI classification and achieved markable performance. For example, Lee and Kwon [39] introduced a deeper and wider contextual CNN, which uses multiple convolutional layers of different sizes and residual connections to form joint feature maps. Wang et al. [40] presented a fast dense spectral–spatial convolution network (FDSSN) maximizing the network information flow through dense connections, effectively improving training speed and classification accuracy. Zhong et al. [41] developed a spectral–spatial residual network (SSRN), employing consecutive learning modules and utilizing 3D-CNN to consider HSI structural characteristics and extract discriminative spectral–spatial features. Yu et al. [42] leveraged the strengths of both GNN and CNN to propose the graph polarization fusion network, which uses GCN and graph attention networks as feature extraction operators to learn features from large, irregular target regions effectively. However, existing CNN-based methods ignore the different importance of spatial pixels and unequal contributions of spectral bands, leading to inaccurate identification of ground objects with similar local context and spectral characteristics. The tremendous success of attention mechanisms in computer vision has attracted widespread attention in the remote sensing domain. Ma et al. [43] proposed a double-branch multiattention (DBMA) mechanism network, where each branch focuses on extracting either spectral or spatial information independently to avoid interference, thus ensuring the extraction of the most discriminative features. Inspired by DBMA, a double-branch dual-attention (DBDA) mechanism network was proposed to further refine and optimize the extracted spectral–spatial features [44]. Zhu et al. [45] proposed a residual spectral–spatial attention network (RSSAN), incorporating both spectral and spatial attention modules to suppress noisy bands. Roy et al. [46] introduced an adaptive spectral–spatial kernel (A^2S^2KNet) autonomously adjusting receptive field size using adaptive attention kernels in residual blocks. Yu et al. [47] introduced a feedback-attention CNN by incorporating a feedback mechanism into the attention module, thereby enhancing the attention weights with high-level semantic knowledge. However, these attention-based approaches are essentially the enhanced versions of CNN-based methods and suffer from the inherent limitations of local convolution kernels, failing to model long-range dependencies effectively. Transformers, due to their superior capabilities in modeling long-range dependencies and remote information interactions, have shown competitive performance in computer vision tasks

and have attracted interest in the HSI classification task. Gao et al. [48] designed a spatial–spectral vision transformer separately extracting spatial and spectral sequences from HSIs, mapping flattened patches and spectra to the transformer's input vectors. Yang et al. [49] proposed a GCN and Transformer fusion network for spatial–spectral feature extraction, which effectively leverages the contextual information of classified pixels while establishing long-range dependencies in the spectral domain. Hong et al. [50] introduced SpectralFormer, utilizing a pure Transformer to process spectral features. Sun et al. [51] proposed a network comprising 2-D and 3-D convolution layers to preprocess input HSI images, followed by a Gaussian-weighted feature tokenizer to generate input tokens for Transformer blocks. Song et al. [52] proposed a novel bottleneck spatial–spectral transformer (BS2T) to depict the long-range global dependencies of HSI pixels, and on this basis, defined a dual-branch HSI classification framework based on 3D-CNN and BS2T for jointly extracting the local–global features of HSI. Xu et al. [53] proposed the cross spatial–spectral dense transformer for spatial–spectral feature extraction and fusion, utilizing an adaptive dense encoder module and a cross spatial–spectral attention module. Although the vision transformer allows for learning long-range dependencies from a global perspective, it tends to overlook local region features. To address these limitations, Qi et al. [54] proposed a novel method called the global–local 3-D convolutional transformer network, embedding 3-D convolution into a dual-branch transformer to simultaneously capture global–local associations in both spectral and spatial domains. However, these transformer-based approaches are still constrained by the patch size, limiting their ability to effectively capture global information and handle the high spatial heterogeneity in H2 imagery. To mitigate the issues, Zheng et al. [55] first proposed an FPGA framework and a variant of FCN (FreeNet). However, when the sample distribution is imbalanced, FreeNet cannot extract the most discriminative features. Zhu et al. [56] proposed a spectral–spatial-dependent global learning (SSDGL) framework that combined global convolutional long short-term memory and a global joint attention mechanism to capture the long-term dependency of spectral features. Similarly, Yu et al. [57] introduced the cross-level spectral–spatial joint encoding (CLSJE) method. However, existing patch-free methods do not effectively address the spectral heterogeneity issues—caused by narrower spectral bands and higher spatial resolution, leading to negatively impacting classification results. The emergence of large models has significantly accelerated the development of HSI processing. Hong et al. [58] identified a substantial gap in spectral data research and introduced a general remote sensing foundation model called SpectralGPT, offering new approaches for applying spectral information in HSI. Wang et al. [59] introduced a novel sparse sampling attention mechanism and developed HyperSIGMA, effectively addressing spectral and spatial redundancy issues in HSI.

III. Proposed Method

Fig. 2 illustrates the proposed interactive learning framework, in which S²PNet comprises MSSP, GLMG, and DSSP.

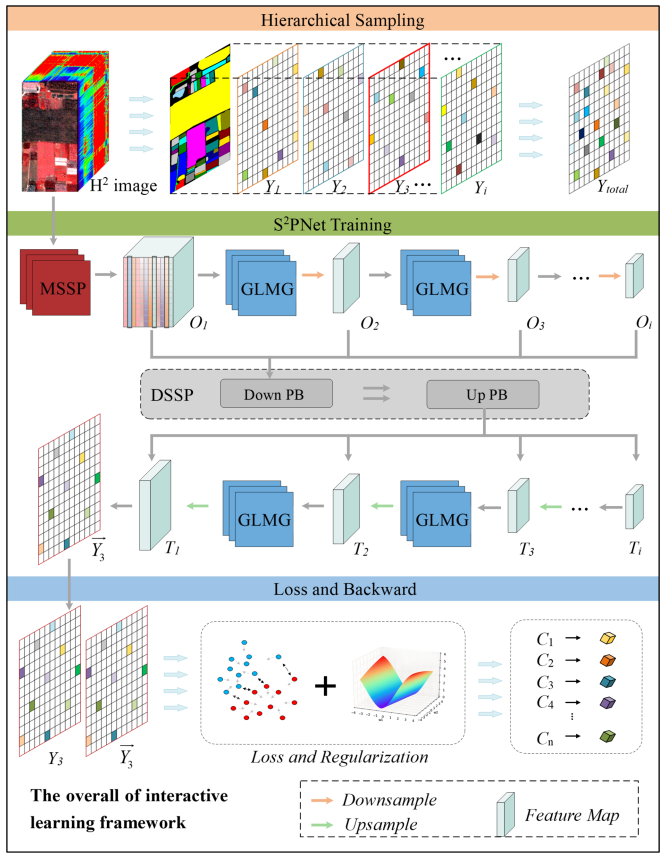


Fig. 2. Overall architecture of the proposed interactive learning framework. Where $Y_i, i \in (1, 2, 3, \ldots)$, represent the labeled pixels obtained from each stratified sampling, and Ytotal represents the total number of sampled labeled pixels. $O_i, i \in (1, 2, 3, \ldots)$, and $T_i, i \in (1, 2, 3, \ldots)$, represent the features at different layers of the encoding and decoding stages, respectively. Different shapes indicate different sizes after pooling layers, with O_1 illustrating the result of band selection and the gray box indicating the selected bands. Finally, C_1–C_n represent each category in the dataset.

Because of the spectral similarity between the phenologically similar crops, H² imagery exhibits significant spectral heterogeneity, which leads to spectral mixing issues. Therefore, at the beginning of the network, MSSP was introduced to purify noisy spectral bands and reduce the spectral variance of the imagery. Improvements in image resolution result in the intensification of spatial heterogeneity. In the encoder section, GLMG is introduced to adaptively aggregate global contextual information and pixel-level local information, leveraging the advantages of convolution and self-attention to establish long-term dependency relationships between features and spatially enhance the discriminability of each pixel. Finally, the semantic gap between the features in the encoding and decoding stages was minimized using DSSP, ensuring a smoother flow of information throughout the network.

A. MSSP Module

An effective approach to addressing spectral heterogeneity is to find the most discriminative combination of bands for each category and remove less distinguishable bands. Fig. 3 shows the average spectral curves for each category in the LongKou

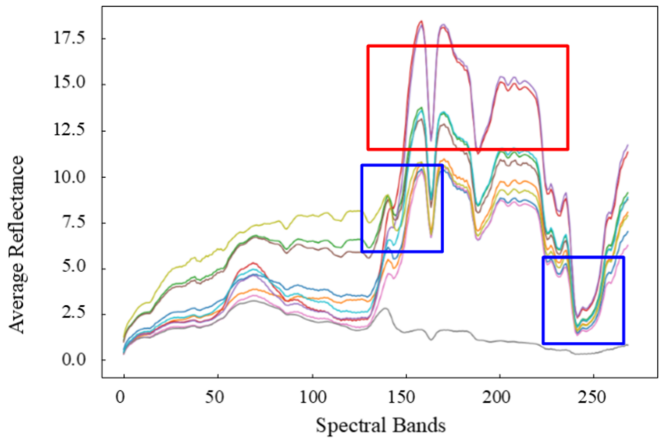


Fig. 3. Average spectral curves for each class in the LongKou dataset, with significant differences in the red-boxed region and severe band mixing in the blue-boxed region.

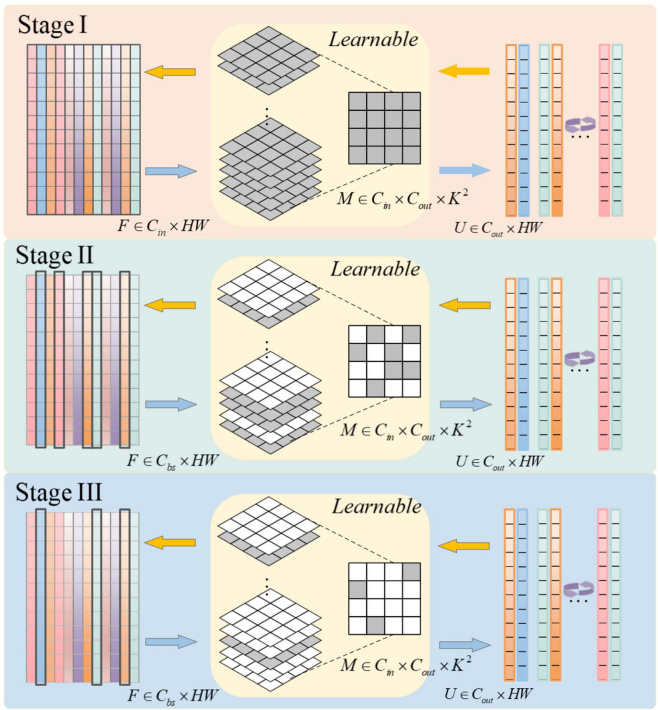


Fig. 4. Structure of MSSP. $F^{C_{\text{in}} \times HW}$ and $U^{C_{\text{out}} \times HW}$ represent the input and output features of MSSP, while M denotes importance matrix. The gray and white squares represent learnable and zeroed convolution weights, respectively.

dataset. As can be clearly seen from the figure, in the region marked by the red box, the spectral curves of different categories exhibit significant differences, indicating high separability. In contrast, in the region marked by the blue box, the spectral bands are more mixed, which can have a negative impact on classification. Therefore, we consider the bands in the blue region as noisy bands that need to be removed. Inspired by NAS, we propose an MSSP, as shown in Fig. 4. Unlike NAS, which aims to determine the optimal network architecture, MSSP aims to identify an optimal spectrum combination.

MSSP consists of two modules: learnable spectral selection and spectral combination. The former filters out noisy bands to

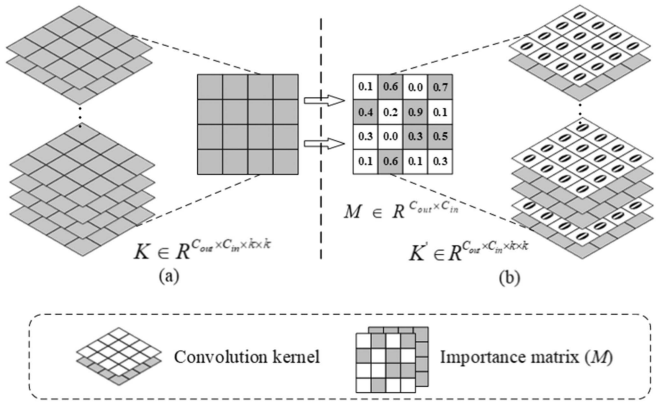


Fig. 5. Learning process of convolutional kernels in the MSSP. (a) Initial convolutional kernels in Stage I. (b) Learned convolutional kernels in Stage II.

obtain class-contributing bands for classification, whereas the latter randomly combines selected bands to search for optimal spectral combinations. We start by defining the input bands as C_{in} and the output bands as C_{out}. The size of the convolution kernel is denoted as $C_{\text{out}} \times C_{\text{in}} \times k \times k$, where k represents the kernel size. Fig. 5 illustrates the learning process of the convolutional kernels in MSSP. In the first stage, we use the entire spectral band as the input for classification, where the information from each band flows into the network and all the convolutional kernels (indicated in gray) are updated with gradient propagation during this stage, as shown in Fig. 5(a). In the second stage, after a certain number of iterations, each band contributes to the classification result with an initial estimate. We simplify the convolutional kernel as a matrix M of size $C_{\text{in}} \times C_{\text{out}}$, where each position in this matrix is computed as the L_1 norm of the corresponding filter $F_{i,j} \in R^{1\times1\times k\times k}$, where i and j represent the input and output channels, respectively. The importance of each band is quantified using the values in matrix M, such as $\sum^{C_{\text{out}}C_{\text{in}}} |F_{i,j}|$. If the L_1 norm of the current position is smaller than that of the other positions, as shown on the left-hand side of Fig. 5(b), we zero all the weights of the corresponding convolutional kernel to mask the ith input band. The right-hand side of Fig. 5(b) illustrates this process, where the white color indicates the convolutional kernels corresponding to the noisy bands that have been zeroed out. To enhance the generalization capability of the MSSP, we introduced a purification factor P (where $P = 1, 2, 3, \ldots$) to adjust the sparsity of the convolutional kernel. The purification factor determines the step size for searching class-specific bands, where $1/P$ of the bands is purified at each step. In the third stage, we fixed the positions of the convolutional kernels corresponding to the optimal bands and conducted further fine-tuning using the sparsified input to achieve a stable and high-performance classification. The process of spectral selection can be represented as

$$F_{\text{subset}} = \delta\left(gn\left(\sum\nolimits_{g=1}^{G}\left(\sum\nolimits_{p=1}^{P} w_p^g F + b_g\right)\right)\right) \quad (1)$$

where δ represents the sigmoid function; gn stands for group norm; G and P denote the number of groups and purification

Algorithm 1: Spectral Process of the MSSP.

Input: $F = \mathbb{R}^{C_{in} \times H \times W}$

Learnable Spectrum Selection :
if Stage I **then**
  $F_{subset} = GroupConv(F; w)$
else if Stage II **then**
  Purify noisy bands to obtain the optimal bands subset.
  Get kernel matrix $M \in \mathbb{R}^{C_{out} \times C_{in}}$
  for $p \in P$ **do**
    $M_n^{C_{out} * C_{in}} = 0, n = (C_{in} * p)/P$
    $w^p_{i_k j_k} = 0$, where $i \in (1, C_{out}), j \in (1, C_{in}), k = 1, 2, \ldots C_{in} * p/P$
    $F_{subset} = GroupConv(F; w^p)$
  end for
else if Stage III **then**
  $F_{subset} = GroupConv(F; w^P)$
end if
Spectral Combination :
Get combination $F_{combin} = Shuffle(F_{subset})$
Get final $F_U = GroupConv(F_{combin})$

Output: $U \in \mathbb{R}^{C_{out} \times H \times W}$

factors, respectively; w is the convolutional kernel weights; b_g represents the convolutional residual; and F is the input feature.

Selecting the optimal subset of class-specific bands is the first stage of spectral purification. Subsequently, we determine the best spectral combination for each class, which involves a search process. The sparse input bands underwent a shuffle operation to randomize the original order. Subsequently, grouped convolution was applied to the shuffled bands for random grouping and feature extraction. By introducing random spectral combinations, we can limit excessive focus on highly significant bands during the selection stage. This helps prevent selection results from being biased toward the mere superposition of significant bands, thereby enhancing the overall performance of MSSP. The process of spectral combination can be represented as follows:

$$U = \delta\left(gn\left(\sum_{g=1}^{G} w^g F_{\text{subset}} + b_g\right)\right) \tag{2}$$

where U represents the feature obtained after spectral purification. We designed the module to maintain consistent output dimensions across three stages for seamless integration with subsequent networks. To clearly illustrate the process described above, we present it in Algorithm 1.

Finally, to mitigate the negative impact of excessive sparsity on classification accuracy, we incorporate a regularization term into the loss function to balance the sparsity of MSSP. Typically, regularization is achieved through L_1

$$L_{sp} = \sum_{g=1}^{G} \sum_{p=1}^{P} \sum^{C_{\text{out}} \times C_{\text{in}}} \left|F_p^g\right|. \tag{3}$$

B. GLMG Module

Convolution and self-attention are two distinct paradigms widely employed for feature extraction. The former is adept at perceiving local regions and effectively capturing local spatial features, such as edges and textures. The latter, which uses a weighted averaging operation based on contextual features, can simultaneously consider various positions in the image, facilitating a better capture of global information and the establishment of long-term dependencies. In our research, we observed a strong underlying relationship between self-attention and convolution. We define the input and output features as $F \in \mathbb{R}^{C_{\text{in}} * H * W}$ and $G \in \mathbb{R}^{C_{\text{out}} * H * W}$, respectively, where C_{in} and C_{out} represent the input and output channels of the convolution, respectively and H and W represent the height and width of the feature map, respectively. Given a kernel size of k. Finally, the traditional convolution can be divided as follows:

$$\overline{g}_{i,j} = \sum_{i,j}^{H,W} \sum_{m,n}^{K} w_{m,n} f_{i,j} \tag{4}$$

$$g_{i,j} = \left(\sum_{i,j}^{H,W} \sum_{m,n}^{K} w_{m,n} f_{i,j}, m - [k/2], n - [k/2]\right) \tag{5}$$

where $f_{i,j} \in \mathbb{R}^{C_{in}}$ represents the position in the input feature map, $g_{i,j} \in \mathbb{R}^{C_{\text{out}}}$ represents the corresponding position in the output feature map, and $w_{m,n} \in \mathbb{R}^{C_{\text{in}} * C_{\text{out}}}$ represents the weight at the current position in the kernel. We then delve into the self-attention portion, where the attention weights are obtained by dynamically calculating the similarity function between related pixel pairs. This flexibility allowed the network to adaptively focus on different regions. Similarly, given $F \in \mathbb{R}^{C_{\text{in}} * H * W}$ and $G \in \mathbb{R}^{C_{\text{out}} * H * W}$, the output calculation of self-attention is computed as follows:

$$g_{i,j}^n = \sum_{n=1}^{N} \sum_{i,j}^{H,W} A(q_{i,j}^n, k_{i,j}^n) v_{i,j}^n \tag{6}$$

$$q_{i,j}^n = \sum_{n}^{N} w_q^n f_{i,j}, k_{i,j}^n = \sum_{n}^{N} w_k^n f_{i,j}, v_{i,j}^n = w_v^n f_{i,j} \tag{7}$$

where N represents the number of attention heads, $q_{i,j}^n, k_{i,j}^n$, and $v_{i,j}^n$, w_q^n, w_k^n, and w_v^n denote the query, key, and value, respectively, and $A(q_{i,j}^n, k_{i,j}^n)$ are the corresponding projection matrices, and represent the attention weights.

The above analysis shows that convolution and self-attention are similar in process: first, the input is mapped to a higher dimensional space by feature learning, and second, the learned features are aggregated. In addition, computational complexity was primarily concentrated in the preliminary mapping phase, whereas the subsequent aggregation phase was lightweight and required almost no additional parameters. Based on this observation, we designed the GLMG shown in Fig. 6. The feature maps in the network were first projected to three times their original dimensions using a 1×1 projection and then divided into N groups, as shown in Fig. 6(a). Subsequently, the mapped information flow enters the designed branch 1, which represents the offset and aggregation operations in the convolution. However, during implementation, we observed that tensors tended to disrupt data locality when moving in different directions, making it challenging to achieve vectorization and severely affecting the propagation efficiency of the network. Therefore, we optimized this process by introducing frozen kernel shift convolution, in which all positions in the convolution kernel

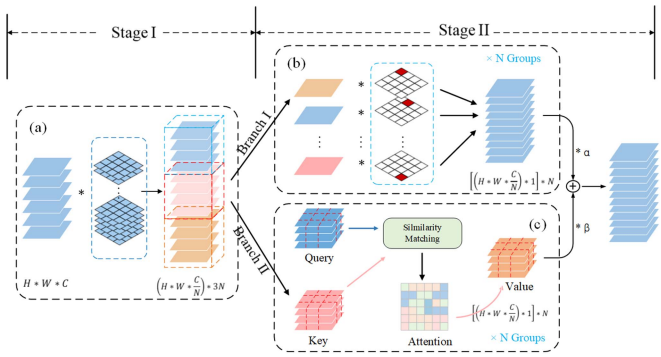


Fig. 6. Structure of GLMG. (a) Feature mapping stage, (b) convolution branch, and (c) edge mask MWSA.

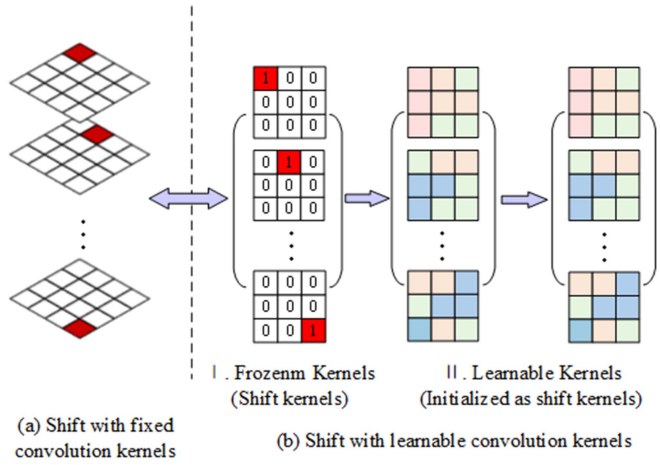


Fig. 7. Frozen kernel shift convolution.

were replaced with 0, except for the direction of movement. This achieved tensor-shifting effects with carefully designed convolution weights specific to offset directions, as illustrated in Fig. 6(b). To further combine features from different directions, we concatenate all input features and convolutional kernels. During training, we initialized the shift kernel to learn the weights of the convolutional kernel and enhance the flexibility of the model. In addition, we incorporated the concept of grouping into modules to further reduce computational complexity, as shown in Fig. 7. We then input the remapping features from Fig. 6(a) into Branch II. Branch II employs MWSA to compute and aggregate the similarity between the remapping features, as depicted in Fig. 6(c). We observed that the increased spatial resolution of the H^2 imagery exacerbates spatial heterogeneity, where adjacent regions in the image are more likely to exhibit similar features, whereas distant areas may contain more noise. Hence, we designed a windowed feature map to perform self-attention operations. However, this lead to a lack of connection between windows. We employ an edge weights method to supplement features, extracting boundary features between windows through a simple mask operation to establish connections between them. Specifically, we first locate the partitioning boundary in the mapped convolution, which is a straightforward operation by averaging the feature map and grouping the excess

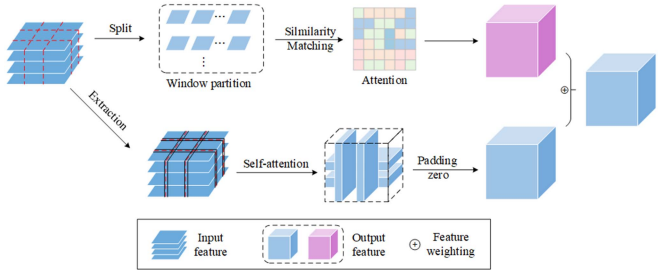


Fig. 8. Edge mask MWSA.

separately. Then, we expand pixels upward and downward from the boundary line, performing self-attention calculations on the areas these pixels belong to. The processed value is zero-padded to match the shape of the value in MWSA, followed by feature weighting, as shown in Fig. 8. In this scenario, the self-attention process is described as follows:

$$g_{i,j}^{n} = \sum_{n=1}^{N} \sum_{a,b \in P_k(i,j)} A(q_{i,j}^{n}, k_{a,b}^{n}) v_{a,b}^{n} \tag{8}$$

where $P_k(i,j)$ denotes a window of size k around the current pixel (i,j). Similarly, we integrated the MWSA with a grouping approach to reduce the complexity of the computation while matching the output dimensions of Branch I. The information flows from both branches were fused and finalized for the output by assigning each branch a learnable weight. In summary, the convolution and self-attention branches share a feature mapping operation, and the intermediate feature maps are reorganized and reused. Branch I compensates for the lack of feature-learning ability and local feature information in self-attention, whereas Branch II supplements the difficulty of establishing long-term dependencies in features owing to the limited receptive field of convolution. These dual branches facilitate the exchange of local and global information in an image through the reorganization and reuse of intermediate feature maps, thereby achieving interactive guidance.

C. DSSP Module

The cleverness of skip connections lies in their ability to transfer lost spatial information from the encoder to the decoder. This assists the decoder in combining semantic information to restore the features to the same spatial resolution as the original image [60]. Encoders primarily capture low-level spatial information, whereas decoders extract high-level semantic information. There is often a significant semantic gap between the mapped features. This disparity is particularly noticeable in the first skip connection (between the first and last decoding layers). Direct fusion of these incompatible features can introduce disturbances during the learning process and affect the outcome. We designed a new DSSP to mitigate abrupt transitions between information flows. DSSP consists of down and up progressive blocks, as illustrated in Fig. 9. We begin by discussing the information flow within the entire module. First, the features after the four pooling layers flowed into the downward progressive block. Information exchange occurs between features from adjacent encoding layers. Given the features of the four pooling layers of

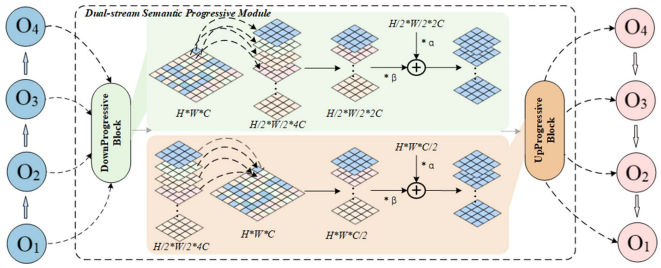


Fig. 9. Structure of DSSP. where $O_i, i \in (1, 2, 3, \ldots)$, is the input feature.

O_1, O_2, O_3, and O_4, we sequentially fused (O_1, O_2), (O_2, O_3), and (O_3, O_4). In this way, O_2 interacted with O_1 before being fused with O_3.

The fused features then flow into the upward progressive block, which involves a backward layerwise interaction on feature. This enables the bidirectional exchange of information between the shallow and deep layers. Consequently, shallow features gradually approach the semantic representation of deep features, thereby reducing semantic gaps. Finally, the exchanged features are concatenated with the features of the decoder. In the downward progressive block, merging between O_1 and O_2 involves a scale transformation. Common pooling operations and strided convolutions may cause information loss. Therefore, we devised a simple yet effective method to mitigate this loss. We uniformly sample the original feature map O_1 in both horizontal and vertical directions, dividing it into multiple subfeature maps of the same size as O_2. After stacking, the channels become four times the original and are then projected to a feature tensor with the same dimensions as O_1 through a 1×1 convolution. Finally, a learnable weight is assigned for weighted summation. Specifically, within any 2×2 window, the sampling of O_1 is stacked and then weighted and summed with O_2, and the calculation formula is as follows:

$$\overrightarrow{O} = [O_{i,j}, O_{i,j+1}, O_{i+,j}, O_{i+1,j+1}], i, j \in R^{h \times w} \quad (9)$$

$$(O_1, O_2) = \alpha * \text{Conv}_1(\overrightarrow{O_1}) + \beta * O_2 \quad (10)$$

where Conv_1 represents the 1×1 projection, and α and β are the learnable weights. The upsampling process follows a similar procedure, and the details have been omitted.

IV. Experiment

In this section, we provide an overview of the experimental data, evaluation metrics, and experimental settings. Finally, a comprehensive comparison of the various methods used in the experiment is conducted.

A. Datasets and Experimental Settings

1) Datasets: WHU-Hi-LongKou dataset [61] was acquired on July 2018, in Longkou Town, Hubei Province, China, using a DJI Matrice 600 Pro UAV platform equipped with a focal-length Headwall Nano-Hyperspec imaging sensor. The image resolution was 550×400 pixels, and spatial resolution was 0.463 m. The Nano-Hyperspec imaging sensor comprises 270 bands covering a wavelength range of 400–1000 nm. Background

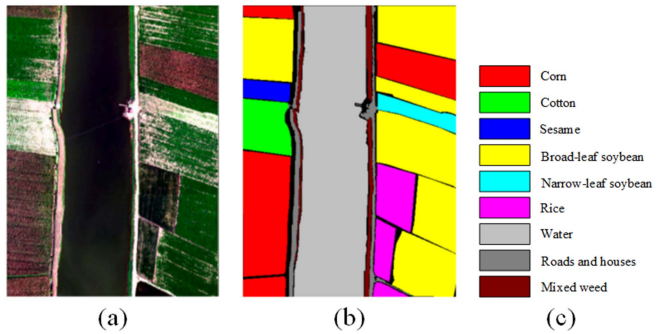


Fig. 10. WHU-Hi-LongKou dataset. (a) H^2 imagery. (b) Ground-truth map. (c) Legend.

TABLE I
Pixels Used for Training and Testing From the LongKou Dataset

No.	Class	Train	Test
C1	Corn	10	34 501
C2	Cotton	10	8364
C3	Sesame	10	3021
C4	Broad-leaf-soybean	10	63 202
C5	Narrow-leaf-soybean	10	4141
C6	Rice	10	11 844
C7	Water	10	67 046
C8	Roads and houses	10	7114
C9	Mixed weed	10	5219

pixels within the image scope were removed, leaving 204 542 labeled pixels, which were classified into nine categories. For convenience, we use C1–C9 to represent these categories in the following figures and tables. A visualization of the dataset is shown in Fig. 10. The training setup used for the LongKou dataset is given in Table I. A total of 10 pixels were randomly selected from each class for training, accounting for 0.04% of the total number of pixels. The remaining pixels were used to test the model performance.

WHU-Hi-HanChuan dataset [61] was collected in June 2016, in Hanchuan, Hubei Province, China. The image resolution was 1217×303 pixels, and spatial resolution was 0.109 m. It consists of 274 bands ranging from 400 to 1000 nm. Within the image scope, land objects were classified into 16 categories. The WHU-Hi-HanChuan dataset contained 257 530 labeled pixels. We use C1–C16 to represent these categories in the following figures and tables A visualization of the dataset is shown in Fig. 11. For the HanChuan dataset, 50 pixels were randomly selected from each class for training, accounting for 0.3% of the total number of pixels. The remaining pixels were used to test model performance, as given in Table II.

WHU-Hi-HongHu dataset [61] was collected in Honghu, Hubei Province, China. The UAV platform travelled at a height of 100 m, resulting in a spatial resolution of 0.043 m and an image resolution of 940×475 pixels. The dataset consists of 270 bands ranging from 400 to 1000 nm. The land parcels within the image scope were noticeably fragmented, and the crop types included different varieties of the same crop. This is a typical area for studying the spatial–spectral heterogeneity of H^2 imagery. The image contained 386 693 labeled pixels classified into 22

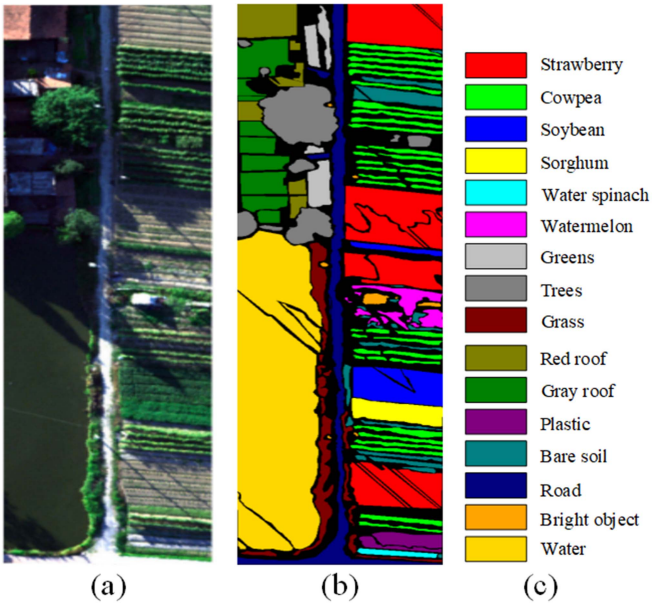


Fig. 11. WHU-Hi-HanChuan dataset. (a) H^2 imagery. (b) Ground-truth map. (c) Legend.

TABLE II
PIXELS USED FOR TRAINING AND TESTING FROM THE HANCHUAN DATASET

No.	Class	Train	Test
C1	Strawberry	50	44 685
C2	Cowpea	50	22 703
C3	Soybean	50	10 237
C4	Sorghum	50	5303
C5	Water spinach	50	1150
C6	Watermelon	50	4483
C7	Greens	50	5853
C8	Trees	50	17 928
C9	Grass	50	9419
C10	Red roof	50	10 466
C11	Gray roof	50	16 861
C12	Plastic	50	3629
C13	Bare soil	50	9066
C14	Road	50	18 510
C15	Bright object	50	1086
C16	Water	50	75 351

categories, as shown in Fig. 12. We use C1–C22 to represent these categories in the following figures and tables. For the HongHu dataset, 50 pixels were randomly selected from the labeled samples for model training, accounting for 0.2% of the total number of pixels, and the remaining pixels were used to test model performance, as given in Table III.

Houston 2013 dataset [62] was collected near the University of Houston in the United States, captured for the 2013 IEEE GRSS Data Fusion Competition by the National Center for Airborne Laser Mapping. The dataset dimensions are 349 × 1905 pixels, with a spatial resolution of 2.5 m and coverage across 144 bands spanning from 380 to 1050 nm. The image includes 15 029 labeled pixels classified into 15 categories. Fig. 13 illustrates the image, ground-truth maps, and legend. For the Houston 2013 dataset, 20 pixels were randomly

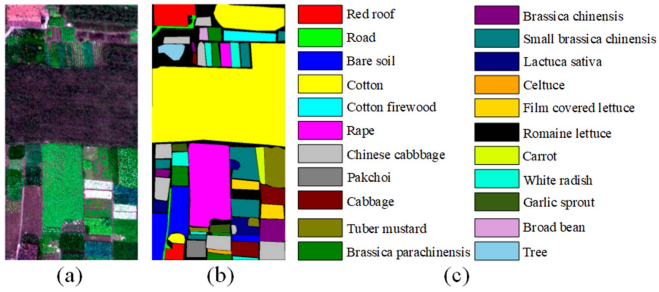


Fig. 12. WHU-Hi-HongHu dataset. (a) H^2 imagery. (b) Ground-truth map. (c) Legend.

TABLE III
PIXELS USED FOR TRAINING AND TESTING FROM THE HONGHU DATASET

No.	Class	Train	Test
C1	Red roof	50	13 991
C2	Road	50	3462
C3	Bare soil	50	21 771
C4	Cotton	50	163 235
C5	Cotton firewood	50	6168
C6	Rape	50	44 507
C7	Chinese cabbage	50	24 053
C8	Pakchoi	50	4004
C9	Cabbage	50	10 769
C10	Tuber mustard	50	12 344
C11	Brassica parachinensis	50	10 965
C12	Brassica chinensis	50	8904
C13	Small Brassica chinensis	50	22 457
C14	Lactuca sativa	50	7306
C15	Celtuce	50	952
C16	Film covered lettuce	50	7212
C17	Romaine lettuce	50	2960
C18	Carrot	50	3167
C19	White radish	50	8662
C20	Garlic sprout	50	3436
C21	Broad bean	50	1278
C22	Tree	50	3990

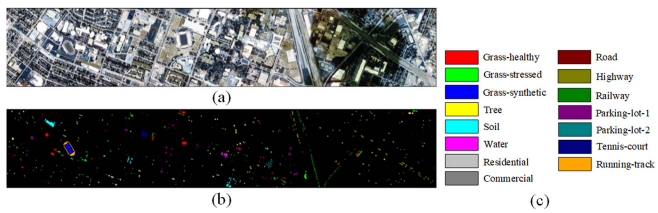


Fig. 13. Houston 2013 dataset. (a) H^2 imagery. (b) Ground-truth map. (c) Legend.

selected from the labeled samples for model training, accounting for 2% of the total number of pixels, and the remaining pixels were used to test model performance, as given in Table IV.

Large-scale Houston 2018 dataset was collected for the 2018 IEEE GRSS Data Fusion Contest (Saux et al. [63]), with spatial dimensions of 601 × 3058 pixels. It covers a spectral range from 380 to 1050 nm across 50 bands. The dataset includes 504 856 labeled pixels categorized into 20 classes. Fig. 14 illustrates the distribution of the dataset. For the Houston 2013 dataset, 50 pixels were randomly selected from the labeled samples

TABLE IV
PIXELS USED FOR TRAINING AND TESTING FROM THE HOUSTON 2013 DATASET

No.	Class	Train	Test
C1	Grass-healthy	20	1231
C2	Grass-stressed	20	1234
C3	Grass-synthetic	20	677
C4	Tree	20	1224
C5	Soil	20	1222
C6	Water	20	305
C7	Residential	20	1248
C8	Commercial	20	1224
C9	Road	20	1232
C10	Highway	20	1207
C11	Railway	20	1215
C12	Parking-lot-1	20	1213
C13	Parking-lot-2	20	449
C14	Tennis-court	20	4008
C15	Running-track	20	640

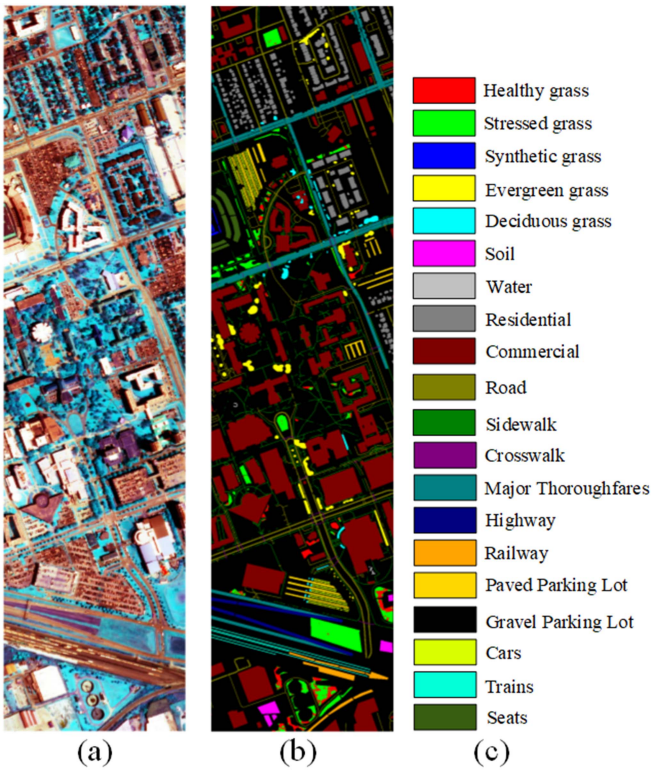


Fig. 14. Large-scale Houston 2018 dataset. (a) H^2 imagery. (b) Ground-truth map. (c) Legend.

for model training, accounting for 0.2% of the total number of pixels, and the remaining pixels were used to test model performance, as given in Table V.

2) Experimental Settings: To validate the superiority of the proposed method, it was compared with several state-of-the-art models, including SVMs with a radial basis function kernel [34], FDSSN [40], SSRN [41], DBMA [43], DBDA [44], A^2S^2KNet [46], SSFTT [51], BS2T [52], FreeNet [55], SSDGL [56], and CLSJE [57]. All patch-free methods were trained for 1000 iterations on the three datasets. We selected

TABLE V
PIXELS USED FOR TRAINING AND TESTING FROM THE LARGE-SCALE HOUSTON 2018 DATASET

No.	Class	Train	Test
C1	Healthy grass	50	9749
C2	Stressed grass	50	32 452
C3	Synthetic grass	50	634
C4	Evergreen grass	50	13 538
C5	Deciduous grass	50	4998
C6	Soil	50	4466
C7	Water	50	216
C8	Residential	50	39 712
C9	Commercial	50	223 634
C10	Road	50	45 760
C11	Sidewalk	50	33 952
C12	Crosswalk	50	1466
C13	Major Thoroughfares	50	46 308
C14	Highway	50	9799
C15	Railway	50	6887
C16	Paved Parking Lot	50	11 425
C17	Gravel Parking Lot	50	99
C18	Cars	50	6528
C19	Trains	50	5315
C20	Seats	50	6774

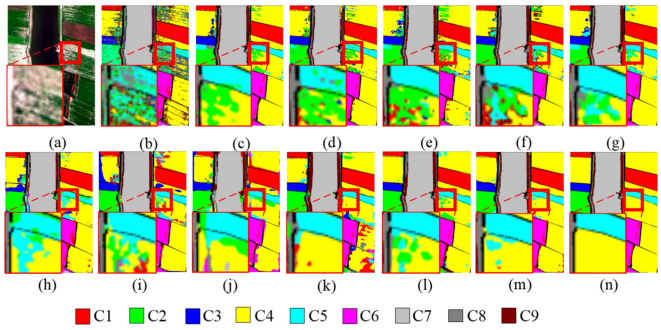


Fig. 15. Classification results for the LongKou dataset. We have enlarged the red boxed areas for a clearer comparison of the effects of different methods. (a) H^2 imagery, (b) SVM, (c) FDSSC, (d) SSRN, (e) A^2S^2KNet, (f) DBMA, (g) DBDA, (h) FreeNet, (i) SSDGL, (j) CLSJE, (k) SSFTT, (l) BS2T, (m) S^2PNet, and (n) Gt.

SGD with a momentum of 0.9 and a weight decay rate of 0.001 as the optimizer. The base learning rate was set to 0.001 and multiplied by, with power = 0.9. For the patch-based methods, each model was trained for 100 epochs with a batch size of ten samples. The window size for A^2S^2KNet, RSSAN, and DBDA was 9×9 and for all others it was 7×7. Finally, the overall accuracy (OA), average accuracy (AA), kappa coefficient (Kappa), and accuracy of each class were adapted to evaluate the performance of each method. All experiments described in this section were conducted using an RTX3090.

B. Experiment Results

1) Experiment on WHU-Hi-LongKou Dataset: Classification results from various methods are depicted in Fig. 15. The spatial–spectral heterogeneity of H^2 imagery pose challenges for classification. For instance, SVM relies on spectral information for classification, misclassifying similar spectral crops, such

TABLE VI
CLASSIFICATION ACCURACY FOR THE WHU-HI-LONGKOU DATASET

Class	SVM	FDSSC	SSRN	A^2S^2KNet	DBMA	DBDA	FreeNet	SSDGL	CLSJE	SSFTT	BS2T	S^2PNet
C1	95.79	99.73	98.07	92.77	97.48	99.62	98.38	89.83	91.96	91.59	**99.76**	98.56
C2	51.35	65.18	64.93	52.91	75.18	81.12	85.45	96.19	90.12	96.76	72.93	**97.02**
C3	21.38	79.85	73.21	56.69	91.88	71.31	96.23	96.26	92.95	98.73	98.81	**99.07**
C4	97.35	99.26	98.23	99.62	99.3	99.47	85.63	78.73	82.23	78.94	**99.60**	95.55
C5	30.47	67.02	82.35	57.14	66.1	48.45	96.98	96.57	92.97	97.75	69.92	**99.54**
C6	84.04	99.29	88.15	96.53	95.97	96.55	95.64	93.26	91.71	90.44	**99.60**	97.1
C7	99.66	99.94	99.82	99.89	99.36	99.79	98.82	99.84	96.43	94.52	**99.98**	99.52
C8	78.17	81.62	83.74	89.4	85.61	87.96	83.1	90.24	85.14	66.07	79.76	**92.37**
C9	33.45	73.89	78.51	82.45	50.56	74.94	80.36	84.96	73.5	91.42	52.33	**91.75**
OA	82.42	95.04	94.25	91.6	93.69	94.75	92.84	90.26	89.66	88.87	94.74	**97.26**
AA	65.74	85.09	85.22	80.82	84.6	84.36	91.18	91.76	88.56	90.36	85.85	**96.72**
Kappa	77.79	93.56	92.53	89.2	91.84	93.19	90.75	87.48	86.64	85.79	93.18	**96.41**

With optimal results in bold.

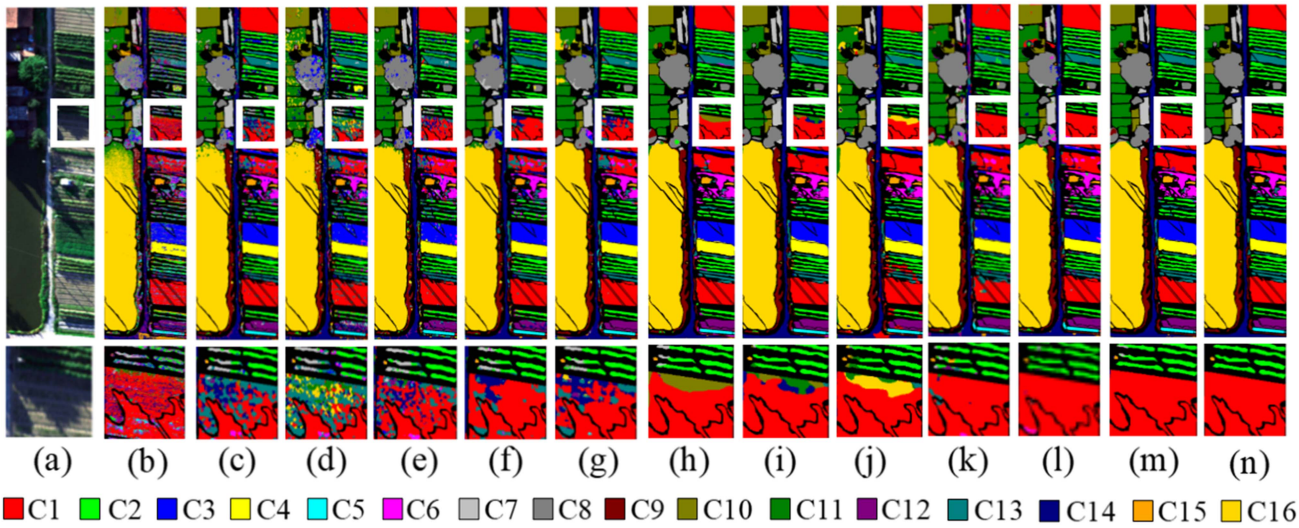


Fig. 16. Classification results for the HanChuan dataset. The first row shows the classification results, and the second is an enlargement of the white box in the first row. (a) H^2 imagery, (b) SVM, (c) FDSSC, (d) SSRN, (e) A^2S^2KNet, (f) DBMA, (g) DBDA, (h) FreeNet, (i) SSDGL, (j) CLSJE, (k) SSFTT, (l) BS2T, (m) S^2PNet, and (n) Gt.

as cotton and broadleaf soybean. SSRN and A^2S^2KNet also exhibited poor performance with numerous pixels misclassified. DBMA and DBDA, which extract spatial–spectral information using two branches, achieved good results. However, there is confusion when classifying the sparsely distributed areas within broadleaf soybean regions, limited by patch size constraints. Patch-free methods reduces isolated misclassifications, but struggled to maintain effective boundary classification. FreeNet and SSDGL were affected by spectral heterogeneity, leading to noise in the classification results in areas containing sesame and broadleaf soybeans. CLSJE exhibits undesirable effects at the boundary between rice and broad-leaf soybean. In addition, in BS2T, some broadleaf soybean pixels were incorrectly classified as cotton. In contrast, S^2PNet demonstrated superior performance with the best visual results and minimal isolated misclassifications. Quantitative results are given in Table VI, with the best accuracy highlighted in bold. S^2PNet accurately identified the pixels indicating sesame, cotton, and narrow-leaf soybeans, which are prone to spectral confusion. OA, AA, and kappa exceeded those of the other methods. In summary, S^2PNet better addresses spectral and spatial heterogeneity than the other tested models in the LongKou dataset.

2) Experiment on WHU-Hi-HanChuan Dataset: The classification performance of the different methods for the HanChuan dataset is shown in Fig. 16(b)–(m). Similar to Experiment 1, the SVM performed the worst with a considerable amount of noise in the classification map. There were many misclassified regions with FDSSC and SSRN. For example, most strawberry pixels were erroneously categorized as bare soil, whereas the bare soil was misclassified as cowpea. For DBMA, cowpea was identified as soybean, and watermelon was misclassified as greens and water spinach. These misclassifications in small areas result from complex planting structures and high spatial heterogeneity, which create challenges when determining the optimal patch size. FreeNet, SSDGL, and CLSJE produced smoother classification maps but still contained some misclassified pixels. For example, the red roof was misclassified as a gray roof, and the gray roof was misclassified as water. This is because pixels in shadowed areas generally have low and similar grayscale values, thereby exacerbating the spectral heterogeneity and increasing the difficulty of classification. SSFTT and BS2T showed promising performance, but still exhibited some flaws between strawberries and watermelons. However, S^2PNet could still effectively identify various land cover types

TABLE VII
CLASSIFICATION ACCURACIES FOR THE WHU-HI-HANCHUAN DATASET

Class	SVM	FDSSC	SSRN	A²S²KNet	DBMA	DBDA	FreeNet	SSGDL	CLSJE	SSFTT	BS2T	S²PNet
C1	91.49	96.13	96.09	95.16	**98.11**	98.1	92.32	94.09	94.18	94.39	96.63	97.99
C2	67.08	97.16	94.7	95.75	**98.41**	97.44	94.94	96.35	89.72	90.50	97.14	96.62
C3	58.05	78.25	60.85	76.95	79.47	91.16	97.51	94.39	95.15	97.92	96.03	**97.93**
C4	82.36	95.58	83.73	89.96	94.31	97.47	98.64	97.57	**99.15**	99.43	98.34	98.28
C5	13.5	83.25	77.9	65.75	80.23	65.23	98.09	**100**	**100**	99.48	72.16	**100**
C6	17.86	64.66	44.65	59.04	65.08	74.35	98.37	96.56	90.85	93.51	60.36	**99.26**
C7	52.73	70.54	59.11	71.25	74.79	81.03	96.22	96.65	96.09	95.34	84.84	**99.56**
C8	69.49	95.04	95.55	90.92	97.06	97.19	90.34	96.13	94.11	87.55	**97.49**	94.77
C9	54.11	83.51	79.11	85.73	89.5	88.07	91.79	95.26	80.86	87.99	86.97	**97.7**
C10	88.63	95.26	99.44	97.11	**99.55**	96.95	98.91	97.91	93.61	99.13	98.94	99.44
C11	69.02	88.03	78.78	87.63	93.1	93.57	98.14	98.23	87.98	95.42	88.74	**99.72**
C12	26.2	62.19	87.69	60	93.15	83.68	98.93	99.7	92.26	99.92	80.96	**100**
C13	41.99	44.58	33.77	52.77	54.17	59.73	86.72	87.77	78.47	80.30	86.12	**91.35**
C14	79.24	83.95	80.91	85.59	84.26	85.99	97.64	99.03	93.08	90.19	95.81	**99.05**
C15	28.64	89.52	64.92	88.84	83	83.99	90.15	**98.9**	97.05	96.59	82.77	94.01
C16	99.78	99.77	93.53	99.82	99.6	96.64	98.71	99.38	97.11	98.70	**100**	99.19
OA	73.99	88.75	81.64	89.2	91.55	92.18	95.75	96.9	93.23	94.42	94.65	**98.08**
AA	58.76	82.96	76.92	81.39	86.48	86.91	95.46	96.75	92.48	94.15	88.95	**97.8**
Kappa	70.03	86.94	78.61	87.43	90.16	90.86	95.04	96.38	92.04	93.49	93.76	**97.76**

With optimal results in bold.

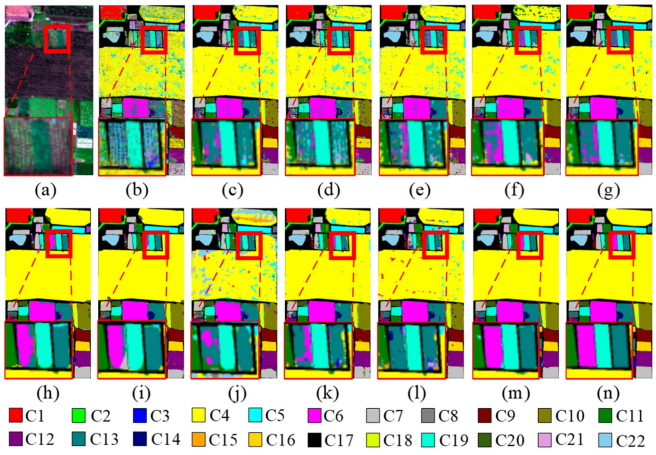


Fig. 17. Classification results for the HongHu dataset. We have enlarged the red boxed areas for a clearer comparison of the effects of different methods. (a) H^2 imagery, (b) SVM, (c) FDSSC, (d) SSRN, (e) A^2S^2KNet, (f) DBMA, (g) DBDA, (h) FreeNet, (i) SSDGL, (j) CLSJE, (k) SSFTT, (l) BS2T, (m) S^2PNet, and (n) Gt.

in shadow areas despite the enhanced spectral variability. The classification accuracies of all the methods are presented in Table VII. S^2PNet significantly improved the classification accuracy of categories, such as gray roofs, red roofs, grass, and the complex planting structure of bare soil in shadow-covered areas. The accuracy of each class demonstrates a certain level of competitiveness.

3) Experiment on WHU-Hi-HongHu Dataset: The qualitative and quantitative results of the different methods on the Hong Hu dataset are shown in Fig. 17 and Table VIII, respectively. Because of the diversity of crops and varieties in the Hong Hu region, there is significant intraclass spectral variability. The SVM classification map contained considerable noise and severe misclassifications. Deep-learning-based methods generally outperform SVM, with the poorly performing SSRN improving the results by 16.06%, 19.79%, and 18.69% compared with the SVM in the OA, AA, and Kappa, respectively. However, the spectral heterogeneity of H^2 imagery remains challenging. For example, cotton pixels are misclassified as cotton firewood, and rape pixels are identified as Brassica parachinensis. In addition, there are many isolated misclassification areas on the map, which indicates that patch size can significantly affect the classification results. FreeNet and SSDGL showed competitive performance with smoother classification graphs, and OA and AA were improved by 25.82% and 39.83% relative to the SVM, respectively. However, misclassification of rape with small Brassica chinensis and white radish with Brassica parachinensis still occurred. In comparison, S^2PNet achieved better results in classes with high spectral complexity, demonstrating the highest classification accuracy and the best visual performance.

4) Experiment on Houston 2013 Dataset: Table IX and Fig. 18 display the classification results for the Houston 2013 dataset. This dataset includes many similar land cover categories with strong spectral heterogeneity and a fragmented, highly discontinuous distribution of test samples. These characteristics result in poor performance for most models that rely on window sizes. For example, SVM, SSDGL, and CLSJE exhibit numerous misclassifications when distinguishing between road and highway. DBMA and DBDA struggle to effectively differentiate between Parking-lot-1 and Parking-lot-2. In contrast, our S^2PNet effectively handles this spectral confusion in secondary classification, flexibly capturing contextual changes around pixels, and achieving optimal and most stable classification results across all metrics.

5) Experiment on Large-Scale Houston 2018 Dataset: Table X and Fig. 19 present the classification results for the large-scale Houston 2018 dataset. As the dataset size increases, all methods exhibit significant performance declines. The three metrics for SVM are only 60.92, 49.47, and 53.94. Among CNN- and attention-based methods, the most stable performer, FreeNet, achieves an OA of only 74.58 and a Kappa of 68.83. Transformer-based methods demonstrate their superiority. The classification accuracy of SSFTT reaches 75.96, which is a 1.38% improvement over FreeNet, with Kappa increasing by

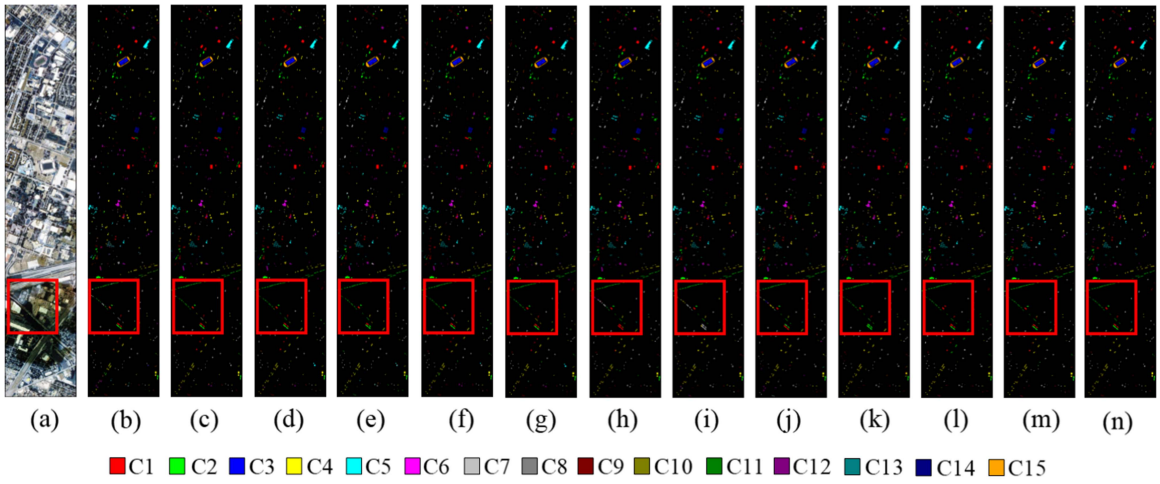


Fig. 18. Classification results for the Houston 2013 dataset. (a) H^2 imagery, (b) SVM, (c) FDSSC, (d) SSRN, (e) A^2S^2KNet, (f) DBMA, (g) DBDA, (h) FreeNet, (i) SSDGL, (j) CLSJE, (k) SSFTT, (l) BS2T, (m) S^2PNet, and (n) Gt.

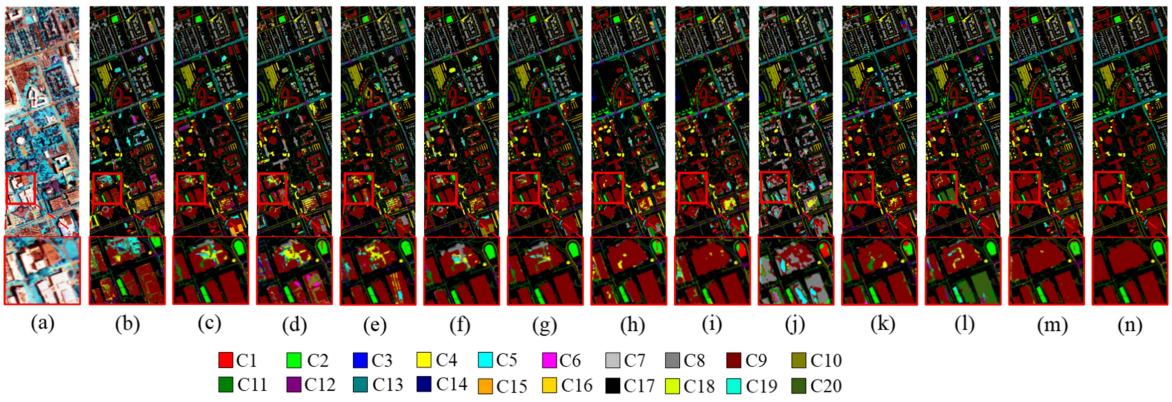


Fig. 19. Classification results for the large-scale Houston 2018 dataset. We have enlarged the red boxed areas for a clearer comparison of the effects of different methods. (a) H^2 imagery, (b) SVM, (c) FDSSC, (d) SSRN, (e) A^2S^2KNet, (f) DBMA, (g) DBDA, (h) FreeNet, (i) SSDGL, (j) CLSJE, (k) SSFTT, (l) BS2T, (m) S^2PNet, and (n) Gt.

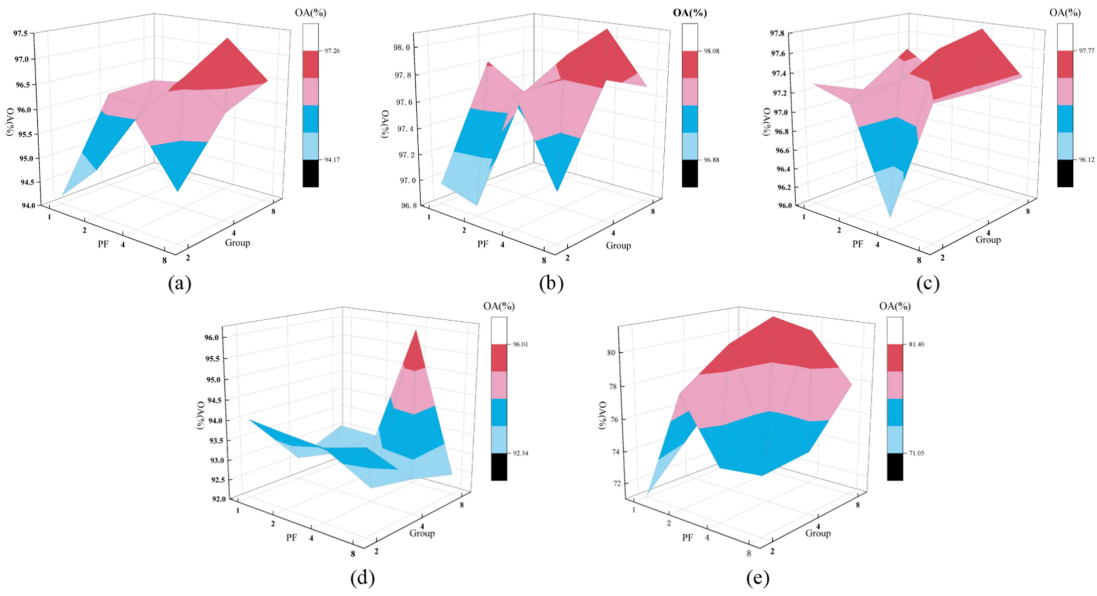


Fig. 20. Comparison of the purification factor and groups numbers. (a)–(e) Impact of different combinations of purification factors and group numbers on classification accuracy across the LongKou, HanChuan, HongHu, Houston 2013, and Houston 2018 datasets. In these plots, the x-axis represents the PF, the y-axis represents the group numbers, and the z-axis represents the OA.

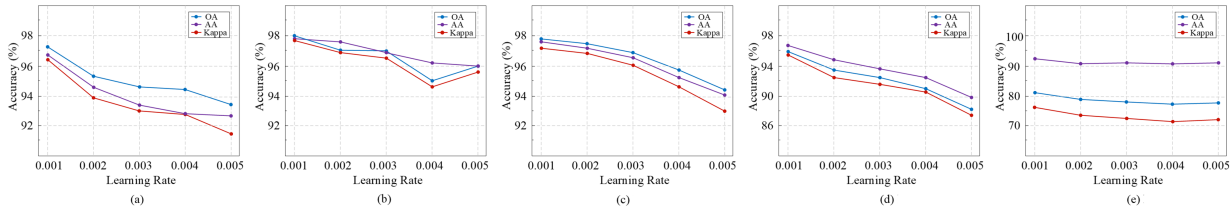


Fig. 21. Impact of learning rate on the model accuracy.

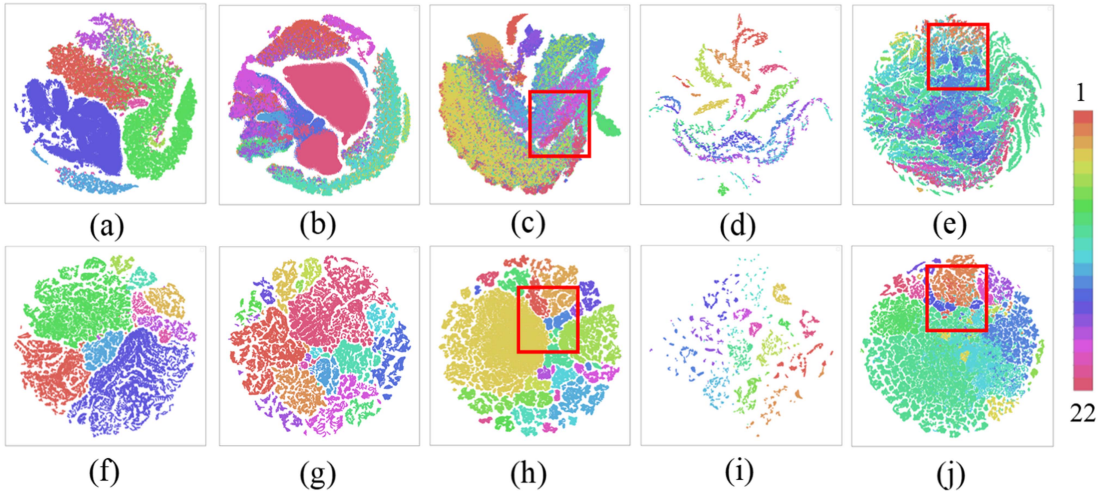


Fig. 22. Visualization of T-SNE features of S^{2}PNet on five datasets, where each sample of different categories is represented by color dots. (a)–(e) Distribution of the original datasets. (f)–(j) Distribution of features extracted by S^{2}PNet.

TABLE VIII
CLASSIFICATION ACCURACIES FOR THE WHU-HI-HONGHU DATASET

Class	SVM	FDSSC	SSRN	A^{2}S^{2}KNet	DBMA	DBDA	FreeNet	SSGDL	CLSJE	SSFTT	BS2T	S^{2}PNet
C1	94.55	98.17	96.92	97.89	98.74	**98.95**	96.38	97.05	96.69	95.77	86.45	97.66
C2	57.81	83.25	78.57	78.02	86.7	87.06	97.31	**98.04**	96.39	95.64	79.44	96.45
C3	92.27	96.7	94.12	96.8	97.13	**98.41**	97.03	96.57	88.03	94.64	96.13	96.22
C4	97.65	99.38	98.97	99.32	99.56	**99.59**	98.77	99.02	78.09	97.43	99.76	99.51
C5	18.46	43.7	29.75	34.27	47.52	58.31	96.16	98.99	98.22	98.12	61.71	**99.79**
C6	90.2	97.25	92.95	96.62	96.96	98.07	95.19	94.05	87.91	95.99	99.30	**98.24**
C7	82.72	91.34	91.26	91.17	93.32	**94.66**	90.41	82.86	71.89	86.74	98.06	92.99
C8	13.98	71.73	41.23	47.89	80.16	73.67	98.85	**99.63**	93.13	99.00	65.26	99.45
C9	96.29	**99.19**	98.12	98.35	98.56	98.78	98.12	98.43	96.03	98.73	99.26	98.44
C10	55.56	94.34	87.89	79.61	96.81	96.83	93.04	96.33	89.65	92.19	98.43	**98.41**
C11	36.19	84.74	69.42	85.3	77.58	91.17	89.03	94.77	74.41	93.52	89.91	**97.92**
C12	43.79	65.43	68.99	70.97	81.15	80.06	95.36	97.47	90.82	95.63	71.00	**97.62**
C13	60.09	80.91	70.74	79.75	83.49	83.39	85.75	83.48	79.59	84.55	83.83	**89.96**
C14	67.82	87.38	87.23	79.7	96.26	95.14	94.44	95.63	91.66	**98.08**	84.40	95.69
C15	7.96	66.55	76.54	81.33	85.23	84.46	**100**	99.9	**100**	99.79	59.77	**100**
C16	86.19	96.82	97	**99.57**	98.3	99.32	97.48	99.38	97.64	95.56	99.70	98.57
C17	58.06	84.36	87.33	82.99	91.3	92.42	99.19	**100**	**100**	98.65	90.33	99.73
C18	27.34	83.69	50.66	62.78	83.76	83.69	97.06	98.29	98.04	98.53	83.43	**99.27**
C19	66.79	93.37	80.68	85.22	91.4	90.38	86.4	90.41	88.26	94.27	94.83	**95.01**
C20	38.08	63.69	63.45	65.71	73.48	78.02	97.88	**98.08**	97	96.22	85.89	97.21
C21	11.06	55.3	43.82	60.67	58.9	68.98	98.67	97.97	**100**	**100**	74.01	**100**
C22	23.3	56.61	56.12	43.71	59.98	71.75	**100**	**100**	96.02	97.34	63.94	**100**
OA	70.07	91.17	86.13	88.12	92.74	94.34	95.89	95.81	83.53	95.29	93.35	**97.77**
AA	55.74	81.54	75.53	78.08	85.29	87.41	95.57	96.2	91.34	95.75	84.77	**97.64**
Kappa	64.16	88.99	82.85	85.32	90.94	92.89	94.82	94.72	80.14	94.07	91.68	**97.18**

With optimal results in bold.

1.23%. We speculate that the increased spatial heterogeneity due to the larger spatial scope favors Transformer-based methods, which excel at handling long-range dependencies, thus achieving higher accuracy over a large area. In comparison, our method still achieves the best results despite the more severe spatial heterogeneity in the large-scale dataset, with OA, AA, and Kappa improving by 1.53%, 3.72%, and 2.15%, respectively, over the second-best method. Our classification result map also more accurately reflects the real land cover situation and retains more edge details.

TABLE IX
CLASSIFICATION ACCURACIES FOR THE HOUSTON 2013 DATASET

Class	SVM	FDSSC	SSRN	A^2S^2KNet	DBMA	DBDA	FreeNet	SSGDL	CLSJE	SSFTT	BS2T	S^2PNet
C1	91.85	93.19	92.46	91.78	92.25	92.03	97.15	90.50	85.22	82.78	86.01	**98.96**
C2	95.47	**98.58**	91.20	93.05	95.31	97.68	88.73	71.31	70.34	93.84	98.42	98.56
C3	100	100	99.84	100.00	100	99.85	100	100	99.70	99.85	99.85	**100**
C4	98.17	**100**	99.29	97.37	98.49	99.73	97.71	81.13	71.41	88.64	97.14	94.86
C5	92.19	99.91	93.74	93.01	96.71	95.49	**100**	98.53	91.00	**100**	99.92	98.07
C6	90.12	**100**	99.66	93.89	97.09	94.65	98.03	98.03	98.03	**100**	89.41	98.46
C7	82.01	94.98	95.66	91.36	95.65	**96.46**	88.54	85.66	49.28	88.54	85.17	91.56
C8	69.83	89.33	89.59	97.30	89.48	88.50	87.33	84.80	71.90	70.92	**99.79**	83.76
C9	71.25	85.02	87.87	94.99	80.03	93.35	84.41	77.27	46.67	82.79	**95.51**	90.73
C10	79.04	75.29	76.78	75.80	81.71	80.22	98.17	93.45	84.76	100	88	**100**
C11	67.26	96.13	92.40	79.28	96.21	96.26	84.85	75.88	76.30	94.90	73.23	**98.30**
C12	69.28	87.40	85.50	88.80	83.22	82.92	97.60	94.39	89.45	90.60	90.24	**97.97**
C13	50.92	92.39	96.93	78.49	79.69	89.96	**100**	88.20	87.53	90.42	92.57	99.08
C14	81.08	94.15	93.50	94.95	100	92.01	100	100	100	100	100	**100**
C15	98.43	98.43	98.58	98.71	98.43	98.59	100	100	94.22	88.13	98.61	**100**
OA	82.55	92.50	91.50	89.63	91.43	92.65	93.66	87.34	77.30	90.26	91.43	**96.01**
AA	82.46	93.65	92.87	91.25	92.28	93.18	94.83	89.28	81.05	91.43	92.92	**96.68**
Kappa	81.13	91.89	90.81	88.79	90.74	92.05	93.15	86.31	75.48	89.48	90.74	**95.69**

With optimal results in bold.

TABLE X
CLASSIFICATION ACCURACIES FOR THE LARGE-SCALE HOUSTON 2018 DATASET

Class	SVM	FDSSC	SSRN	A^2S^2KNet	DBMA	DBDA	FreeNet	SSGDL	CLSJE	SSFTT	BS2T	S^2PNet
C1	74.86	68.86	66.03	55.92	70.76	71.65	88.89	84.92	85.36	87.25	61.36	**92.84**
C2	88.16	93.07	**93.79**	90.44	89.37	92.55	84.52	65.04	64.52	75.01	91.82	88.13
C3	52.57	94.17	81.95	72.51	82.05	81.21	100	100	100	100	85.79	**100**
C4	81.90	82.87	82.57	79.11	83.33	82.73	97.34	98.73	85.95	96.61	71.26	**97.48**
C5	23.69	52.91	47.45	44.65	47.94	48.49	89.62	84.89	84.39	89.54	60.93	**94.88**
C6	27.21	85.08	66.23	72.71	84.42	94.27	100	99.93	100	99.19	74.77	**100**
C7	69.03	95.15	88.16	88.01	56.58	96.43	100	100	100	100	97.74	**100**
C8	59.41	65.78	59.79	59.38	61.42	67.98	93.03	91.99	**96.74**	88.10	68.37	93.30
C9	95.34	96.66	97.78	96.89	96.88	96.87	72.24	75.40	26.55	81.87	**98.27**	80.30
C10	39.52	68.09	54.57	48.91	47.40	59.67	49.00	53.01	42.68	42.74	61.46	**66.60**
C11	39.75	57.65	52.30	39.46	48.28	56.32	50.70	23.86	33.55	46.35	45.33	**59.38**
C12	3.73	6.70	5.43	5.99	5.17	9.03	82.97	78.75	75.55	78.27	7.47	**90.25**
C13	56.43	70.50	63.63	71.78	68.50	76.02	70.49	51.89	56.80	61.08	**83.06**	79.74
C14	52.64	63.80	56.08	48.56	79.31	71.96	98.42	**99.64**	98.66	98.84	56.34	97.36
C15	58.89	55.95	85.65	87.59	75.08	96.23	98.42	98.32	96.91	97.97	94.69	**99.10**
C16	46.85	74.29	62.02	61.40	72.47	86.99	90.79	86.23	82.75	89.82	88.43	**97.83**
C17	7.82	90.57	80.67	25.11	23.24	71.11	100	100	100	100	29.72	**100**
C18	32.03	52.12	41.34	41.27	44.40	71.47	89.64	93.40	94.29	86.47	74.91	**95.16**
C19	29.62	51.93	66.89	69.69	64.85	73.13	96.69	97.52	95.75	92.61	68.29	**99.63**
C20	49.91	57.51	55.24	68.33	71.91	72.59	99.32	**100**	98.75	99.88	33.70	99.91
OA	60.92	76.01	72.07	69.39	72.46	79.84	74.58	71.38	49.70	75.96	75.54	**81.37**
AA	49.47	68.18	65.38	61.39	63.67	73.83	87.60	84.18	80.96	85.58	67.68	**91.32**
Kappa	53.94	70.37	65.89	62.77	66.21	74.62	68.83	64.50	43.49	70.06	69.83	**76.77**

With optimal results in bold.

TABLE XI
CLASSIFICATION ACCURACIES OF S^2PNET WITH DIFFERENT STRUCTURE ON FIVE DATASETS

Method	MSSP	DSSP	GLMG	LongKou			HanChuan			HongHu			Houston 2013			Houston 2018		
				OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
(a)Baseline	-	-	-	92.02	87.50	89.65	92.65	92.37	91.44	93.08	92.98	91.32	91.21	92.49	90.50	71.67	86.38	65.41
(b)S^2PNet w/o DSSP and GLMG	✓	-	-	95.21	95.46	93.76	96.99	96.81	96.49	96.53	96.69	95.63	92.33	93.61	91.71	76.47	89.15	70.95
(c)S^2PNet w/o DSSP and MSSP	-	-	✓	94.65	96	93.07	97.39	97.33	96.95	95.99	96.53	94.95	92.89	94.09	92.31	74.86	88.69	69.14
(d)S^2PNet w/o MSSP and GLMG	-	✓	-	95.03	95.28	93.54	96.34	95.89	95.73	96.39	96.57	95.44	93.51	94.68	92.99	74.77	89.09	69.30
(e) S^2PNet w/o MSSP	-	✓	✓	96.06	96.39	94.85	97.67	97.45	97.27	96.72	97.06	95.86	93.22	94.40	92.67	74.86	88.88	70.32
(f) S^2PNet w/o GLMG	✓	✓	-	95.96	95.66	93.94	96.88	96.93	96.36	96.94	96.92	96.14	94.09	95.19	93.61	75.95	88.88	73.41
(g) S^2PNet w/o DSSP	✓	-	✓	96.12	96.07	94.93	97.43	97.21	97	97.11	96.48	96.63	93.60	94.70	93.08	78.84	89.83	73.64
(h) S^2PNet	✓	✓	✓	**97.26**	**96.72**	**96.41**	**98.08**	**97.80**	**97.76**	**97.77**	**97.64**	**97.18**	**96.01**	**96.68**	**95.69**	**81.37**	**91.32**	**76.77**

V. DISCUSSION

A. Ablation Study for the Structure of S^2PNet

We conducted ablation experiments on five datasets to validate the effectiveness of MSSP, GLMG, and DSSP in H^2 imagery classification. Table XI presents the classification accuracies for different network structures. The GLMG module enhances features by incorporating both global and local information, making them more discriminative. This effect is particularly evident in the HanChuan dataset. The MSSP module performs exceptionally well in the Hong Hu dataset, which has high spectral heterogeneity, significantly improving Kappa and AA

values. This demonstrates its strong ability to mitigate spectral mixing issues. The DSSP module further enhances the model's flexibility and classification performance by addressing the semantic gap. Overall, the inclusion of MSSP, GLMG, and DSSP modules allows S^{2}PNet to significantly outperform the baseline model across all tested datasets. The specific advantages of each module in different datasets demonstrate their adaptability and effectiveness in various scenarios. The synergistic effect of these modules not only improves the classification accuracy but also enhances the robustness and generalization ability of the model, making it more capable of handling H^2 imagery classification tasks in highly heterogeneous spectral environments.

B. Discussion on the Purification Factor and Groups Numbers

We conducted a detailed comparison of the impact of different combinations of purification factors and grouping numbers on classification accuracy across five datasets, as shown in Fig. 20. The results exhibit a "hump-shaped" trend where classification accuracy initially increases and then decreases as the purification factor increases. Specifically, when the purification factor is set to 4, the classification accuracy reaches its peak. This indicates that the model effectively removes noise while retaining sufficient critical information, achieving optimal classification performance. When the purification factor is less than 4, the model retains some noisy bands, which interferes with the classification process and reduces accuracy. On the other hand, when the purification factor exceeds 4, the model begins to obscure too much band information. While this reduces noise, it also leads to the loss of significant information, preventing the model from accurately distinguishing between different classes. This suggests that an overly high purification factor can also negatively impact the classification performance. In addition, as the number of groups increases, the flexibility in band combinations improves, leading to a significant boost in classification accuracy. Specifically, increasing the number of groups reduces the number of bands within each group, allowing the model to select and combine bands with greater granularity. This enables the model to better capture subtle differences between bands, thus enhancing classification accuracy. The grouping mechanism also reduces noise interference during training, further improving the model's robustness.

C. Sensitivity in Relation to the Learning Rate

We discussed the learning rate, and Fig. 21 illustrates the classification accuracy under different learning rates across the three datasets. As the learning rate increased, accuracy tended to decrease. This may be related to the Hughes phenomenon, which is caused by too few training pixels and too many bands. In particular, when dealing with high-dimensional data, excessively high learning rates may cause the model to adapt too quickly to the training data, even exhibiting signs of overfitting in the training set.

D. Feature Visualization

To more intuitively demonstrate the feature representation capability of our method, we utilized T-SNE [64] to map the original high-dimensional data features and the learned spectral spatial features to a 2-D space. As shown in Fig. 22(a)–(e), due to the high spectral heterogeneity of the H^2 imagery, there is a significant overlap between samples from different classes in the datasets. However, our method effectively removes noise bands and extracts spatial information, resulting in more discriminative features. As illustrated in Fig. 22(f)–(j), overlapping classes are better distinguished. Specifically, in the red box of the WHU-Hi-HongHu and Houston 2018 datasets, the original data show severe mixing of all classes. After feature extraction by S^{2}PNet, the distribution of samples within and between classes becomes clearer and more uniform. Similar samples cluster together, reducing intraclass distances, while interclass distances increase. This indicates that S^{2}PNet significantly improves feature representation, leading to more accurate and reliable classification results.

VI. Conclusion

In this study, an interactive learning framework is proposed to address the spatial–spectral heterogeneity challenges posed by H^2 imagery. Specifically, the MSSP was introduced to filter out noisy bands, enabling the search for specific spectral combinations for each class to effectively address the spectral confusion caused by crop subclasses, shadows, and other factors. In addition, the GLMG module is devised to handle spatial heterogeneity by mutually guiding and complementing image and pixel-level features. Finally, the introduction of the DSSP module significantly reduces the semantic gap between features at different stages of the network, facilitating harmonious information flow. Experimental results on five different datasets demonstrate that S^{2}PNet significantly outperforms other state-of-the-art methods in terms of classification performance, particularly in areas with similar crop spectra and complex planting structures, where its classification accuracy is significantly enhanced.

Acknowledgment

The authors would like to thank the Supercomputing Center of Wuhan University for providing computing power to conduct experiments.

References

[1] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[2] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Collaborative learning of lightweight convolutional neural network and deep clustering for hyperspectral image semi-supervised classification with limited training samples," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 164–178, 2020.

[3] B. Zhang, "Advancement of hyperspectral image processing and information extraction," *J. Remote Sens.*, vol. 20, no. 5, pp. 1062–1090, 2016.

[4] Z. Yanfei et al., "Hyperspectral with high-spatial resolution remote sensing from observation, processing to applications," *Acta Geodaetica et Cartographica Sinica*, vol. 52, no. 7, 2023, Art. no. 1212.

[5] K. Y. Ma and C.-I. Chang, "Kernel-based constrained energy minimization for hyperspectral mixed pixel classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5510723.

[6] R. Li, K. Cui, R. H. Chan, and R. J. Plemmons, "Classification of hyperspectral images using SVM with shape-adaptive reconstruction and smoothed total variation," in *Proc. IGARSS 2022 IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 1368–1371.

[7] D. K. Pathak, S. K. Kalita, and D. K. Bhattacharya, "Hyperspectral image classification using support vector machine: A spectral spatial feature based approach," *Evol. Intell.*, vol. 15, pp. 1809–1823, 2022.

[8] J. Xia, P. Du, X. He, and J. Chanussot, "Hyperspectral remote sensing image classification based on rotation forest," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 239–243, Jan. 2013.

[9] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, 2016.

[10] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[11] Y. Qian, M. Ye, and J. Zhou, "Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2276–2291, Apr. 2013.

[12] J. Xu, J. Zhao, and C. Liu, "An effective hyperspectral image classification approach based on discrete wavelet transform and dense CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6011705.

[13] Y. Zhang, X. Wang, X. Jiang, and Y. Zhou, "Robust dual graph self-representation for unsupervised hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5538513.

[14] S. Jia, J. Hu, Y. Xie, L. Shen, X. Jia, and Q. Li, "Gabor cube selection based multitask joint sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3174–3187, Jun. 2016.

[15] L. He, J. Li, A. Plaza, and Y. Li, "Discriminative low-rank Gabor filtering for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1381–1395, Mar. 2017.

[16] Y. Xu, Z. Wu, and Z. Wei, "Spectral–spatial classification of hyperspectral image based on low-rank decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2370–2380, Jun. 2015.

[17] C. Li et al., "CasFormer: Cascaded transformers for fusion-aware computational hyperspectral imaging," *Inf. Fusion*, vol. 108, 2024, Art. no. 102408.

[18] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5527812.

[19] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[20] J. Liu, J. Xiang, Y. Jin, R. Liu, J. Yan, and L. Wang, "Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: A survey," *Remote Sens.*, vol. 13, no. 21, 2021, Art. no. 4387.

[21] J. Zabalza et al., "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, 2016.

[22] 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.

[23] H. Gao, Y. Yang, C. Li, L. Gao, and B. Zhang, "Multiscale residual network with mixed depthwise convolution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3396–3408, Apr. 2021.

[24] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[25] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.

[26] Z. Feng, S. Yang, M. Wang, and L. Jiao, "Learning dual geometric low-rank structure for semisupervised hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 346–358, Jan. 2021.

[27] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 55, pp. 1–21, 2019.

[28] W. Sun and Q. Du, "Hyperspectral band selection: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 118–139, Jun. 2019.

[29] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[30] C.-I. Chang and K.-H. Liu, "Progressive band selection of spectral unmixing for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 2002–2017, Apr. 2014.

[31] W. Sun, L. Zhang, B. Du, W. Li, and Y. M. Lai, "Band selection using improved sparse subspace clustering for hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2784–2797, Jun. 2015.

[32] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1723–1740, Mar. 2019.

[33] M. Zhang, M. Gong, and Y. Chan, "Hyperspectral band selection based on multi-objective optimization with high information and low redundancy," *Appl. Soft Comput.*, vol. 70, pp. 604–621, 2018.

[34] B.-C. Kuo, H.-H. Ho, C.-H. Li, C.-C. Hung, and J.-S. Taur, "A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 317–326, Jan. 2014.

[35] Y. Zhan, D. Hu, H. Xing, and X. Yu, "Hyperspectral band selection based on deep convolutional neural network and distance density," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2365–2369, Dec. 2017.

[36] Y. Cai, X. Liu, and Z. Cai, "BS-Nets: An end-to-end framework for band selection of hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1969–1984, Mar. 2020.

[37] A. Sellami, M. Farah, I. R. Farah, and B. Solaiman, "Hyperspectral imagery classification based on semi-supervised 3-D deep neural network and adaptive band selection," *Expert Syst. Appl.*, vol. 129, pp. 246–259, 2019.

[38] J. Feng et al., "Deep reinforcement learning for semisupervised hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5501719.

[39] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[40] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral–spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1068.

[41] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[42] Q. Yu, W. Wei, Z. Pan, J. He, S. Wang, and D. Hong, "GPF-Net: Graph-polarized fusion network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5519622.

[43] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1307.

[44] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 582.

[45] 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.

[46] 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.

[47] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5501916.

[48] Y. Gao et al., "Fusion classification of HSI and MSI using a spatial-spectral vision transformer for wetland biodiversity estimation," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 850.

[49] A. Yang, M. Li, Y. Ding, D. Hong, Y. Lv, and Y. He, "GTFN: GCN and transformer fusion with spatial-spectral features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 6600115.

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

[51] 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, 2022, Art. no. 5522214.

[52] R. Song, Y. Feng, W. Cheng, Z. Mu, and X. Wang, "BS2T: Bottleneck spatial–spectral transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532117.
[53] H. Xu, Z. Zeng, W. Yao, and J. Lu, "CS2DT: Cross spatial–spectral dense transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5510105.
[54] W. Qi, C. Huang, Y. Wang, X. Zhang, W. Sun, and L. Zhang, "Global–local 3-D convolutional transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510820.
[55] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, "FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5612–5626, Aug. 2020.
[56] Q. Zhu et al., "A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 11709–11723, Nov. 2022.
[57] D. Yu, Q. Li, X. Wang, C. Xu, and Y. Zhou, "A cross-level spectral–spatial joint encode learning framework for imbalanced hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5411717.
[58] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.
[59] D. Wang et al., "HyperSigma: Hyperspectral intelligence comprehension foundation model," 2024, *arXiv:2406.11519*.
[60] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[61] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "Whu-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, 2020, Art. no. 112012.
[62] C. Debes et al., "Hyperspectral and lidar data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.
[63] B. Le Saux, N. Yokoya, R. Hänsch, and S. Prasad, "2018 IEEE GRSS data fusion contest: Multimodal land use classification [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 52–54, Mar. 2018.
[64] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Shuai Zhang received the M.S. degree in surveying and mapping from the School of Geomatics, Liaoning Technical University, Fuxin, China, in 2022. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include hyperspectral classification, computer vision, and remote sensing image processing.

Yonghua Jiang (Member, IEEE) received the B.S. and Ph.D. degrees in remote sensing science and technique from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2010 and 2015, respectively.

Since 2015, he has been with the School of Remote Sensing and Information Engineering, Wuhan University, where he became a Professor in 2023. His research focuses on the geometry processing of spaceborne optical imagery.

Chengjun Wang (Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2022.

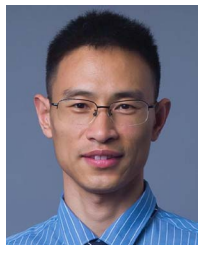
His research interests include hyperspectral image quality improvement, scene classification of high resolution satellite images, deep learning, and computation vision in remote sensing applications.

Meilin Tan (Member, IEEE) received the B.S. degree in remote sensing science and technology and the M.S. degrees in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2010 and 2018.

He is currently an Engineer with the Surveying and Mapping Geographic Information Center, Inner Mongolia Autonomous Region, Hohhot, China.

Bin Du (Member, IEEE) received the B.S. degree in industrial automation from Inner Mongolia University, Hohhot, China, in 1999.

He is currently an Engineer with Inner Mongolia Autonomous Region Surveying and Mapping Geographic Information Center, Hohhot. His research interests include remote sensing data processing and information intelligent extraction.



Feng Tian (Member, IEEE) received the Ph.D. degree from the Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark, in 2016.

Since 2000, he has been a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include vegetation and ecological remote sensing, ecosystem climate change response, and carbon and water cycle.