Expansion Spectral–Spatial Attention Network for Hyperspectral Image Classification

Shuo Wang, Zhengjun Liu, Yiming Chen, Chengchao Hou, Aixia Liu, and Zhenbei Zhang

***Abstract*—Deep learning is increasingly used for the classification of hyperspectral images (HSI), thanks to its ability to completely utilize the rich characteristics of this type of imagery. However, at present, most classification models proposed for processing HSI data are based on standard convolution neural networks, which prefer to learn local information rather than global information, so that it is difficult to achieve ideal accuracy in the case of insufficient training samples in real applications. In this article, we propose a novel expansion spectral–spatial attention network (ESSAN) for HSI data classification in cases of insufficient training samples. First, a dual-branch network based on expansion convolution is employed as the model backbone to extract spectral and spatial information. All feature maps produced during the dual-branch process are superimposed to combine deep and shallow features by the ResNet concept. With the design philosophy of the superposition of expansion convolutional layers, the network can increase the receptive field to gather more global contextual information. Second, the model also includes a coordinate attention block, which directs the network to weight features according to their significance and suppresses those that are irrelevant. Finally, the method was tested on the four datasets from Matiwan Village, Pavia Center, Pavia University, and Shenzhen University, utilizing 1%, 1%, 5%, and 0.2% training samples, respectively. The results showed the overall accuracies, in order, 97.96%, 99.12%, 98.73%, and 99.36%. The preliminary results demonstrate the higher efficacy and accuracy of the proposed ESSAN in HSI data classification than the other state-of-the-art.**

***Index Terms*—Convolutional neural network (CNN), deep learning, expansion spectral–spatial attention network (ESSAN), hyperspectral image (HSI) classification.**

I. Introduction

The recent rapid development of the hyperspectral sensing technology has improved the data availability and quality of hyperspectral images (HSIs). With these advances, HSI applications have excelled in a variety of fields, including crop monitoring, the estimation of crop leaf area index inversion, the prediction of soil organic carbon, and the prediction of soil organic carbon [1], [2], [3], [4].

The purpose of remote sensing image classification is to categorize each type of feature presented in an image. There have been numerous studies on HSI classification due to its high spatial and spectral resolutions and wealth of information features. Initially, the traditional machine learning algorithms have been widely applied to HSI classification, such as using support vector machines to classify the reduced-dimensional data [5], Zhang et al. [6] proposed a spatial–spectral joint classification method based on the random forest for classification. The redundancy between a large number of HSI bands and adjacent bands leads to an increase in noise and uncertainty, which may limit classification accuracy in the case of limited training samples [7], [8]. Therefore, feature extraction and dimensionality reduction techniques have been developed. The traditional methods are used in the early days, including principal component analysis (PCA) for linear dimensionality reduction [9], [10], linear discriminant analysis [11], nonlinear dimensionality reduction kernel PCA [12], isometric feature mapping [13], and extended morphological profiles [14], [15]. However, some advanced band selection and dimensionality reduction techniques have also been proposed. Zhang et al. [16] proposed a new spectral–spatial and SuperPCA method to reduce dimensionality and extract effective low-dimensional features of HSI. He et al. [17] proposed a dual global–local attention network band selection method for high-dimensional hyperspectral data reduction.

As the volume of tasks and data continues to grow, if the training features are manually selected inappropriately, there may be misclassification, resulting in the accuracy not meeting the expected results. Therefore, a new method of machine learning, deep learning, has emerged. Convolutional neural networks (CNNs) have been somewhat successful in the categorization of ground feature objects from HSI imagery. 1-D CNN is straightforward and requires little hardware configuration [18]. Wei et al. [19] parsed raw hyperspectral data using 1D-CNN to extract and classify hierarchical spectral features. However, 1D-CNN only uses the 1-D vector pixel information, whereas 2-D CNNs [20] can fully utilize the rich spectral values or the spatial information in HSI. In comparison with 1-D and 2-D, 3D-CNN [21], [22], [23] combines spectral and spatial information to improve classification results, but the complicated network topology increases hardware configuration requirements.

Manuscript received 24 April 2023; revised 30 May 2023; accepted 12 June 2023. Date of publication 22 June 2023; date of current version 19 July 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0504504 and in part by the Funded Project of Fundamental Scientific Research Business Expenses of the Chinese Academy of Surveying and Mapping under Grant AR2203. *(Corresponding author: Zhengjun Liu.)*

Shuo Wang, Zhengjun Liu, Yiming Chen, and Chengchao Hou are with the Chinese Academy of Surveying and Mapping, Beijing 100830, China (e-mail: shuowang7738@163.com; zjliu@casm.ac.cn; chenym@casm.ac.cn; houchengchao@163.com).

Aixia Liu is with the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing 100048, China (e-mail: liuaixia@lasac.cn).

Zhenbei Zhang is with the State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment (TPESRE), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China (e-mail: zhangzb@itpcas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3288521

Pi et al. [24] suggested a shallow GDIF-3D-CNN classification model using 3-D convolution to classify pure and mixed pixel sets by tweaking the parameters. Lee and Kwon [25] suggested extracting features by combining spatial–spectral contextual information with a Context-Deep CNN (CDCNN). Theoretically, a deeper network can gather more information characteristics and produce better results; nevertheless, the deeper the network gets, the greater the chance that gradient disappearance and gradient explosion will occur, which will worsen the outcomes. To address the mentioned issues, He et al. [26] proposed the ResNet residual structure. The spectral–spatial residual network (SSRN) was proposed by Zhong et al. [27] using the ResNet residual block as the primary structure. Inspired by ResNet, Wang et al. [28] created the fast dense spectral–spatial convolution (FDSSC) in which the network feeds all of the feature maps output in the previous module into the next module via dense connections to achieve the accurate classification; however, the huge amount of parameters increases the running training time. Combining convolutional layers of different dimensions into the same model can better capture the spatial and spectral information of multidimensional data, thereby improving the accuracy and generalization ability of the model. The model HybridSN proposed by Roy et al. [29] consists of a 3-D convolutional block that extracts spectral information, followed by a 2-D convolutional block that extracts spatial information. Compared with using only 3D-CNN, the use of HybridSN can reduce the complexity of the model. Tinega et al. [30] suggested a deep 3-D/2-D genome graph-based network (HybridGBN-SR) that is acceptable for small sample data and does not exhibit overfitting. Yang et al. [31] proposed a synergistic CNN that combines a hybrid convolutional module with a data interaction module.

All of the methods described above are implemented on a single branch and cannot extract information from several channels and spaces at the same time. Therefore, some researchers have proposed multibranch networks to extract the desired features separately. For instance, to categorize photographs of coastal wetlands, Xie et al. [32] created a dual-branch multilayer global spectral–spatial attention network. They employed the extended random walker approach to maximize the classification probability and build the final map. To improve the capability of extracting global information from small HSI sample data, Feng et al. [33] proposed a three-branch mixed spatial–spectral features cascade fusion network, which uses two 3-D residual modules and one 2-D separable residual block to extract features after fusing them to form a cascade fusion model.

Although the traditional convolution can produce accurate classifications, the local operation of the convolution kernel with a fixed shape size cannot obtain a large range of features, and a large amount of parameters significantly increases the computing workload. To overcome this issue, Shi et al. [34] presented the feedback expansion convolution net (FECNet) to introduce holes into the regular convolution kernel to increase the receptive field (RF) and extract more context data. Zhao et al. [35] reduced the computing costs with the hybrid depth separable residual network based on the depth separable convolution.

The vast spectral and spatial features offered by HSI increase information redundancy. The proposed attention technique [36], [37], [38], [39] enables the network to concentrate on more crucial features and enhance model performance. Ma et al. [40] created the dual-branch multiattention (DBMA) by incorporating spectral and spatial attention mechanisms in two branches of the model. Li et al. [41] suggested the dual-branch dual-attention (DBDA), which flexibly employs an adaptive attention mechanism. By mining the characteristics of the HSI spectrum from the viewpoint of a transformer, Hong et al. [42] proposed the SpectralFormer network; however, SpectralFormer does not yield high classification accuracy under small sample HSI. Gong et al. [43] proposed the spectral and spatial attention network model to apply the attention mechanism to HSI-based change detection. In addition, the transformer [44] has also been successfully applied to HSI classification tasks. The transformer uses a self-attention mechanism to learn global features, which can better capture the global relationships and contextual information in the image. Hong et al. [42] proposed a backbone network called SpectralFormer from the perspective of learning spectral sequence information. Based on this backbone network, Sun et al. [45] proposed spectral–spatial feature tokenization transformer to capture spectral–spatial features and high-level semantic features, greatly improving computational efficiency. Liu et al. [46] proposed a hyperspectral image transformer iN transformer method for drawing coastal wetland classification maps on satellite HSIs, which achieved great classification results.

Despite the good results achieved by the existing depth learning algorithms, there are still numerous issues with the classification of HSI features, such as insufficient training samples [47], [48] and a high number of parameters [49], making training slow. This article proposes a novel expansion spectral–spatial attention network (ESSAN) to address these issues and enhance the extraction of HSI global spatial and spectral information with a dual-branch CNN structure with an attention mechanism.

The main work of this article can be summarized as follows.

1) We propose a dual-branch structure based on expansion convolution to extract the features. This method reduces the number of parameters and broadens the RF while preserving the spatial–spectral data produced by each layer.
2) The model incorporates the coordinate attention block (CAB) module, which gives more weight to relevant information and suppresses unfavorable characteristics, thereby improving accuracy and robustness. The experiment demonstrates that CAB can raise the network's overall classification accuracy.
3) ESSAN combines the expanded CNN block and attention block from shallow to deep, which can effectively extract feature information from HSI in the case of insufficient samples. Moreover, ESSAN has fewer parameters. We conducted comprehensive experiments on three public HSI datasets and a self-created SZU dataset, and the results demonstrate that ESSAN outperforms state-of-the-art methods in terms of classification accuracy and training efficiency.

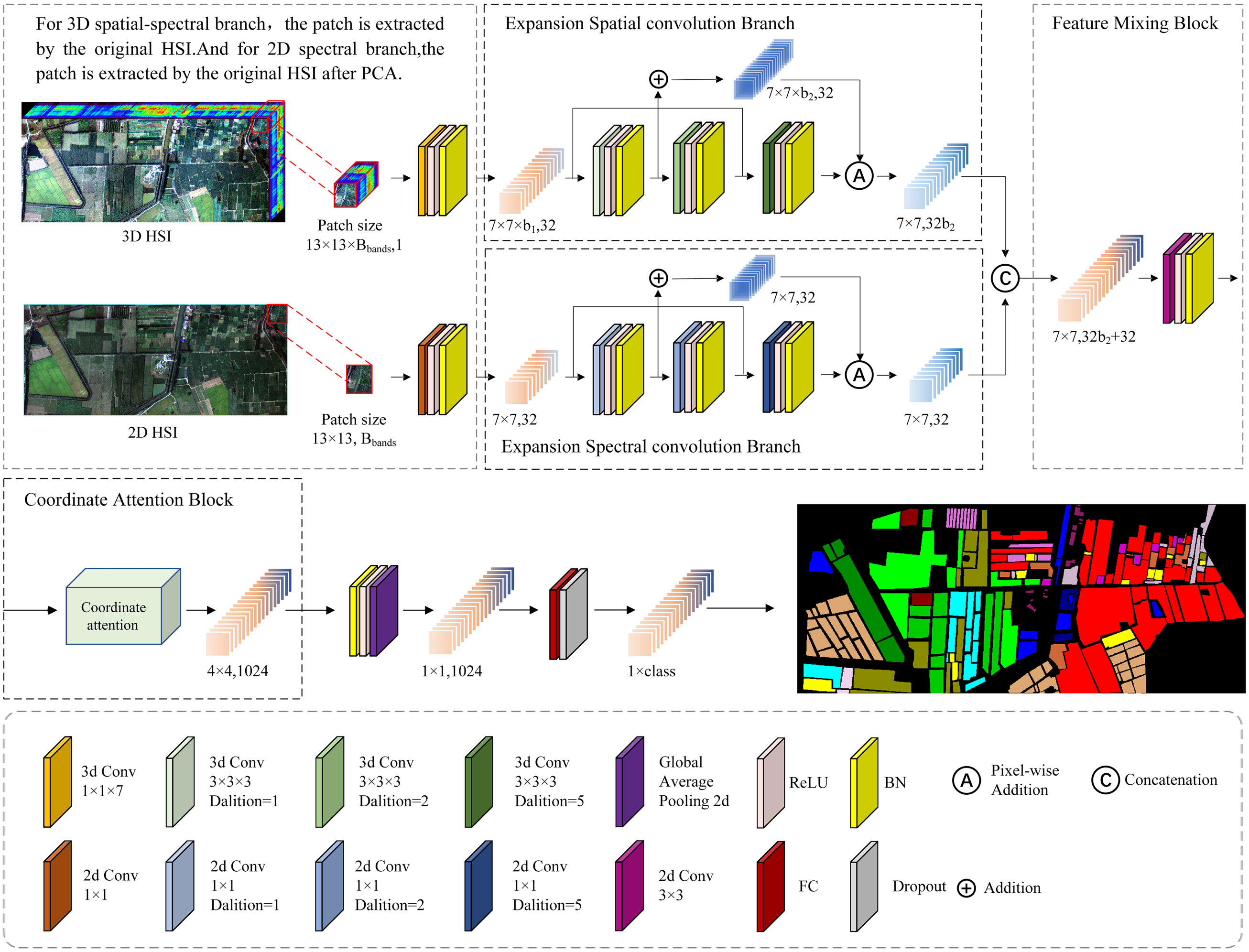


Fig. 1. Overall flowchart of the proposed ESSAN. It mainly consists of three parts: the dual-branch network block, the coordinated attention block, and the expansion convolution basic structure. In addition, the cube of 13×13×band fed into the space branch; sent to the spectral branch is the patch size after PCA dimensionality reduction, that is, the patch size of 13×13.

The rest of this article is organized as follows. Section II provides a detailed description of the proposed ESSAN framework. Section III presents the dataset that was used in this study and contrasts the experimental findings of the suggested method with those of the eight other models. Finally, Section IV concludes this article.

II. Methodology

In this section, we provide a thorough introduction to the ESSAN network framework and all of its elements, including expansion convolution's fundamental structure and design, the dual-branch network module, and the attention mechanism. We also show the advantages of this approach for HSI categorization.

A. ESSAN Framework

The ESSAN framework includes three components (see Fig. 1): the dual-branch network block, the coordinated attention block, and the expansion convolution basic structure.

The area that pixel points in the output feature map on the input image maps is referred to as the RF. When the convolution kernel size is the same, the expansion convolution has a larger RF than the standard convolution. When the RF is the same, the expansion convolution has fewer parameters and a faster calculation speed than the standard convolution. We use two branches to extract the spectral and spatial information of HSI data effectively, and then combine them to derive joint features. First, to extract spatial information, we must create a small cube centered on each pixel of the original image in the three dimensions of height, width, and channel (e.g., a 13×13×band cube, where the band is the number of bands), and then pass these small cubes to the spatial branch. Similarly, for HSI pixels after PCA dimension reduction, we take a patch size centered on this pixel in the height and width dimensions, which is a 13×13 patch, and then pass it to the spectral branch to extract spectral information. Second, using a CAB, we combine spectral and spatial properties to focus on information that is more significant and gives it a higher weight, while ignoring information that is less important and gives it a lower weight. Finally, we use the fully connected

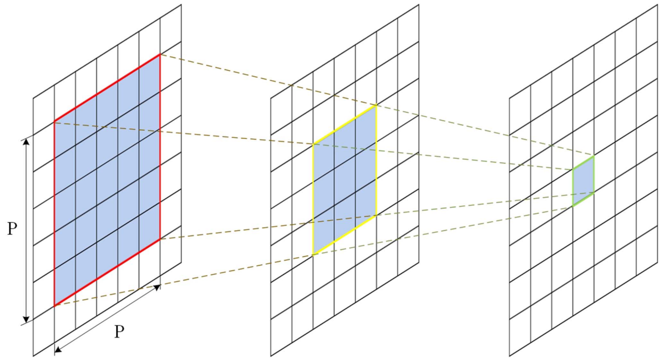


Fig. 2. RF size analysis. Two convolution operations with a convolution kernel of three are performed on the original image. The RF of each pixel in layer 1 is 3 because it was derived from the 3×3 region in the original image; the RF of each pixel in layer 2 is 5 because it was derived from the 5×5 region. (a) Original image. (b) Characteristic layer 1. (c) Characteristic layer 2.

layer to aggregate all features to build classification maps based on the number of land cover categories, preventing the feature locations to impact the classification results.

In this article, we divided the sample data into three categories: training set, test set, and validation set. Samples from the training set are used to train the model and adjust the parameters. The validation set is used to monitor the network performance after each epoch with updated parameters and determine the optimal combination of hyperparameters. The test set is used to evaluate the performance of the model after training is complete and determine the model's generalization ability. Cross-entropy loss is used in the network as the loss function to change the model's parameters. One way to express the multiclassification loss function is given as follows [32]:

$$\text{Loss} = -\frac{1}{C}\sum_{i=0}^{C}\left(X_{\text{target}}\log(X_i) + (1 - X_{\text{target}})\log(1 - X_i)\right) \tag{1}$$

where C represents the total number of categories, and X_i and X_{target} for the predicted labels of each category and the actual labels, respectively.

B. Two-Dimensional and 3-D Expansion Convolution

1) Size of the RF: Fig. 2 shows the analysis process of determining the RF size, calculated as follows:

$$r_{l+1} = r_l + (k_{l+1} - 1) \times \prod_{i=1}^{l} S_i \tag{2}$$

where r_l is the RF size of the lth layer, k_{l+1} is the convolution kernel size of the l+1 layer, and S_i is the stride of the ith layer. Then, the RF needs to be enlarged on a reasonable basis to ensure that the network uses global information rather than simply local information. For instance, if the size of the input image is 13 × 13 and the RF of the pixel in the last layer of the feature map is greater than 13, it indicates that all of the information in the original image were covered by the features that were retrieved during the final classification discrimination of pixels.

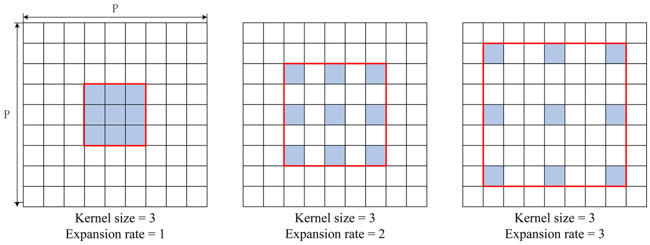


Fig. 3. Working process of the 2-D expansion convolution.

2) Basic Structure of Expansion Convolution: The traditional standard convolution typically employs convolutional layers and pooling layers to improve the RF, but because of limitations, many various convolutions are generated. Among them, to capture multiscale context information, expansion convolution can alter the field of view by modifying the expansion coefficient without altering the size of the feature map. The operation of 2-D expansion convolution is illustrated with an example in Fig. 3. Compared with standard convolution, expansion convolution has an additional hyperparameter called the expansion rate, which describes the number of gaps between the convolution kernel's points. The three images in Fig. 3 each have a convolution kernel size of 3×3, and the expansion rate from left to right are 1, 2, and 3, respectively. The red box represents the size of the equivalent convolution kernel; the blue square represents the position of the convolution kernel; and the white square within the red box represents the holes, which are typically all filled with 0. The RF size is the same as the standard 3×3 convolution kernel size when the expansion rate is 1. When the expansion rate is 2, the RF produced with standard convolution kernels of 5×5 size is equal. When the expansion rate is 3, it is the same size as the RF obtained by a convolution kernel of size 7×7 of standard convolution. The RF will differ when different expansion rates are selected, meaning that multiscale information is collected. To attain the necessary RF size in a practical application, an appropriate expansion rate should be adjusted by the size of the input image.

Equation (2) is the formula for calculating the size of ordinary convolutional RF. By replacing the size of the ordinary convolution kernel in the formula with the equivalent convolution kernel size, the expansion convolutional RF size can be derived. The equivalent convolution kernel size is calculated as follows:

$$K = k + (k - 1)\,(\text{rate} - 1) \tag{3}$$

$$R_{l+1} = R_l + (K_{l+1} - 1) \times \prod_{i=1}^{l} S_i \tag{4}$$

where K is the size of the equivalent convolution kernel, k is the size of the initial convolution kernel, and the rate is the rate of expansion rate. The RF size of layer l+1 is R_{l+1}. The size of the RF expands exponentially as the expansion rate rises. For the same RF, the expansion convolution has fewer parameters than the standard convolution, and the number of parameters falls off exponentially as the expansion rate rises.

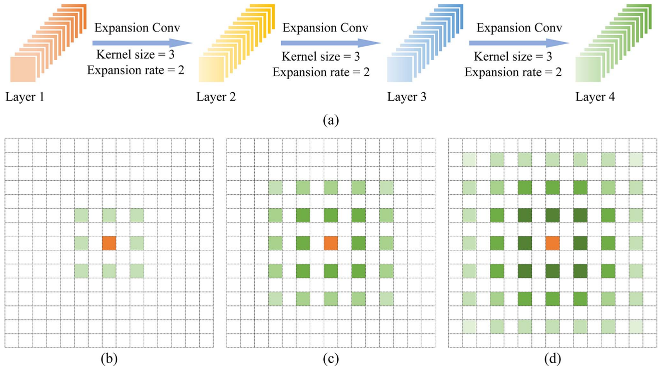


Fig. 4. Gridding effect problems. (a) Original image is Layer1, which is then expanded three times in a row to produce feature maps for Layers 2, 3, and 4. (b)–(d) Three expanded convolutions (both expansion rate = 2) use the pixels in the original image from left to right, with orange representing the position of the center pixel.

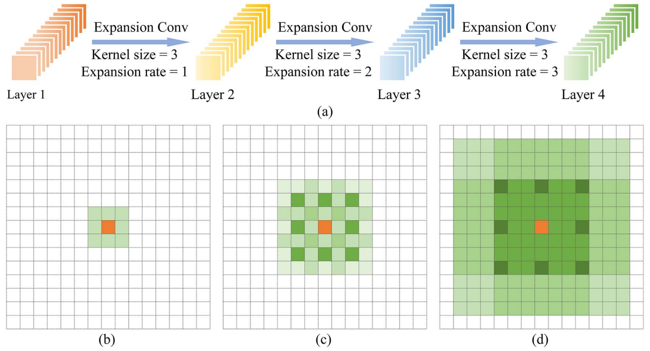


Fig. 5. Continuous convolution with different expansion rates. (a) Original image is Layer1, which is then expanded three times in a row to produce feature maps for Layers 2, 3, and 4. (b)–(d) Three expanded convolutions (The expansion rates are 1, 2, and 3, respectively) use the pixels in the original image from left to right, with orange representing the position of the center pixel.

The 3-D expansion convolution works on the same principles as the 2-D but with a 3-D spatial relationship instead. The 3-D convolution is subject to the same rules of RF and parameter amount as the 2-D convolution.

C. Dual-Branch Network Block

The model employs a dual-branch CNN, as seen in the ESSAN framework flowchart (see Fig. 1). The composition and layout of spatial and spectral branches are thoroughly explained in this section.

1) Expansion Rate of the Dual-Branch Expansion Convolutional Layer: Expansion convolution is frequently used because it can produce a larger RF. However, inappropriate expansion rate settings can result in gridding effect [46] issues when multiple layers of expansion convolution are superimposed. There are three expansion convolutions used sequentially in Figs. 4(a) and 5(a). While the convolution kernel size is 3, the expansion rate choices are different. As demonstrated in Fig. 4(b)–(d), three expansion convolutions with the same expansion rate only employ a portion of the input within their corresponding RF, losing some features and the correlation between information. However, there is no gap between the pixel values when three expansion convolutions [see Fig. 5(b)–(d)] with various expansion rates are stacked since all of the pixel information in its equivalent RF are employed. In all cases, the convolution kernel size and the number of parameters are the same, but the expansion rate varies, with Fig. 5 providing the preferred solution. As a result, it is crucial to design a reasonable expansion coefficient; the distribution of the expansion rate should be zig-zagged.

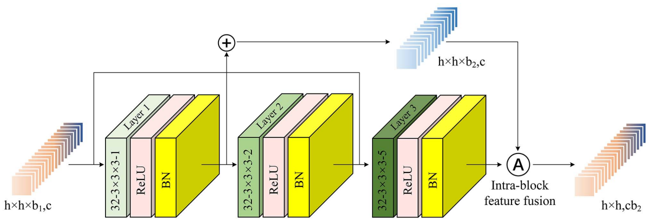


Fig. 6. Spatial branch overall parameter structure.

A straightforward method known as hybrid dilated convolution [50] was suggested, which calls for three convolutional kernel sizes of neighboring convolutional layers, whose expansion rate setting should follow the formula:

$$L_i = \max \left[L_{i+1} - 2d_i, L_{i+1} - 2\left(L_{i+1} - d_i\right), d_i\right]. \quad (5)$$

The goal is to make $L_2 \leq K$, where K is the convolution kernel size and d is the expansion rate, $L_i = d_i$, $i \in \{1,2,3\}$. When the convolution kernel $K = 3$ and the expansion rate of the three convolutional layers $d = [1,2,5]$, $L_2 = 2 < K$, which meets the conditions, so the expansion rates of 3-D and 2-D in both the spatial branch and the spectral branch in this experiments are 1, 2, and 5.

2) Dual-Branch Network Block: It is challenging to train complicated CNNs for HSI classification with small sample sizes, and stacking with many 3-D convolution operations will slow the network. Therefore, we extract features from the spatial and spectral branches using a dual-branch CNN structure (see Fig. 6).

The spatial branch contains three expansion convolutional layers for extracting multiscale features; different expansion rates can obtain information features at various scales, and the third layer expansion rate of 5 can get information at the global level.

The spatial branch has three expansion convolutional layers for extracting multiscale features. The third layer's expansion rate of 5 may extract global information, while different expansion rates can retrieve information features at various scales. The 3-D expansion convolutional layer is denoted by the quantity of output feature maps—the size of the convolution kernel—and the expansion ratio (shown in Fig. 6). For instance, the 3-D convolution represented by 32-3×3×3-1 has 32 feature maps, a convolution kernel size of 3, and an expansion rate of 1. While the spatial branch employs a 3×3 convolution kernel to extract semantic position information, the spectral branch utilizes a 1×1 convolution kernel to filter unnecessary information and concentrate more on the discriminant channel. After each expansion convolutional layer, a batch normalization layer is

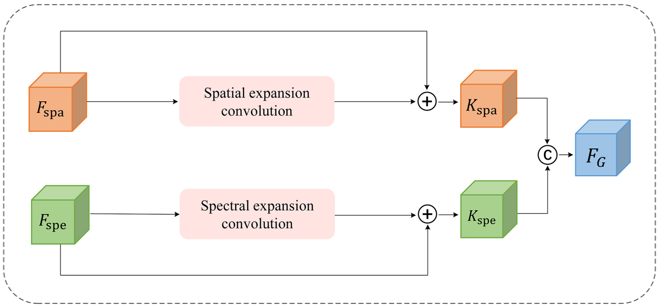


Fig. 7. Dual-branched block.

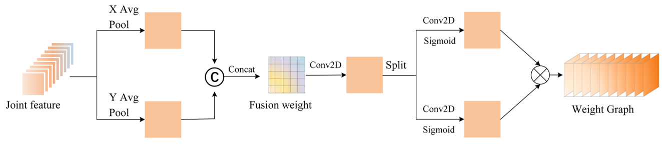


Fig. 8. CAB workflow.

introduced to increase the speed of training and convergence, reduce overfitting during training, and enhance network stability. A rectified linear unit (ReLU) is added between each expansion convolutional layer and the BN layer to increase the nonlinearity of the interaction between the layers. We assemble all the feature maps produced by the expansion convolutional layer in the spirit of ResNet [26], and here, we represent the spatial branch feature map K_{spa} and spectral branch feature map K_{spe} as follows:

$$K_b(x) = F_b(x) + \sum_{i=0}^{l} \sigma(K_b^{l-1}(x) \bowtie W_b^l + r_b^l(x)) \tag{6}$$

where $b\in\{\text{spe, spa}\}$, $1\in\{1,2,3\}$, $F_b(x)$ is represented as the feature of the input space and spectral branch, $K_b^l(x)$ represents the feature map obtained by the lth convolutional layer of the bth branch, and W_b^l represents the convolution kernel size. In the spatial branch, $W_b^l \in \mathrm{R}^{3\times3\times3}$; in the spectral branch, $W_b^l \in R^{1\times1}$. $r_b^l(x)$ is the increased bias; "σ" represents the activation function ReLU.

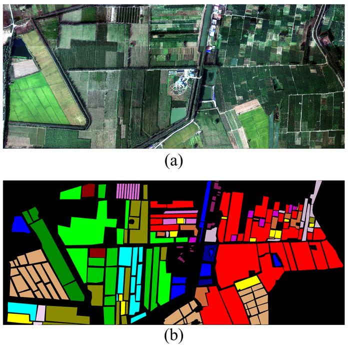
To aggregate spectral information K_{spe} and spatial information K_{spa} efficiently, as seen in Fig. 7 , the two characteristics need to be combined to create the spectral–spatial global joint feature F_G, which is represented as follows:

$$F_G\ (x) = K_{\text{spa}}\ (x) \odot K_{\text{spe}}(x) \tag{7}$$

where $\odot$ represents the concatenation, and the aggregate feature $F_G\in\mathbb{R}^{B\times H\times W}$ includes the extensive spectral and spatial context data. To highlight crucial joint information, reduce unnecessary information, and eliminate noise, the aggregate information is then entered into the attention module to produce a weight map.

D. Coordinate Attention Block

The attention model in the CNN can help give each component of the input *a* different weights, select some crucial information by adjusting the size of the weight, and make each pixel in the model pay more attention to these crucial details, thus improving the training accuracy and effect. A CAB [51] is added to the network to focus the pixels' attention on various categories. The complete flowchart of the CAB framework is shown in Fig. 8.



Label	Color	Class Name	Samples
1	🟧	Paddy	311721
2	🟩	Rice stubble	158099
3	🟦	Water	99070
4	🟨	Meadow	79761
5	🟦	Willow	154132
6	🟪	Vegetable field	22193
7	🟨	Acer multifolia	239169
8	🟩	White wax	217152
9	🟪	House	18005
10	🟫	Maize	58330
11	🟪	Poplar	66687
12	🟥	Koelreuteria paniculata	34450
13	🟩	Sophora japonica	334076
14	🟥	Peach tree	53482
15	🟥	Pear tree	913065
16	🟩	Soybean	2860
17	🟦	Lotus leaf	25107
18	🟪	Elm	17538
		Totle	2804897

(c)

Fig. 9. Original HSI, real land cover, and number of samples in Matiwan Village dataset. (a) Original image. (b) Ground-truth image. (c) Class name and samples.

After entering the spatial–spectral aggregate feature F_G into the CAB, first, it uses the global average pooling to acquire the height feature $M_{\text{ave}}^{\text{Height}}$ and width feature $M_{\text{ave}}^{\text{width}}$. After concatenating the features in the two directions, a 2-D convolutional layer, a BN layer, and an activation function called *h*_swish are coupled to create a remote dependence to combine data in the X and Y directions. Equation (8) illustrates the mixing procedure

$$\begin{cases} M_{\text{ave}}^{\theta} = \ AvgPoolingF_G \\ M_{XY} = \delta[\ \text{Conv}_{2-\text{D}} M_{\text{ave}}^{\text{Height}} \oplus M_{\text{ave}}^{\text{Width}}] \end{cases} \tag{8}$$

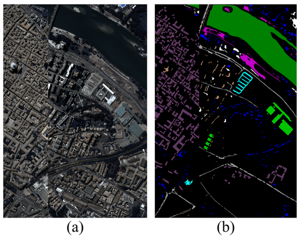
where $\theta\in\{\text{Height, Width}\}$, $\oplus$ represents the feature connection, and "δ" indicates the *h*_swith activation function. The global data are currently present in each dimension of the feature map M_{XY}. A split function is then used to separate the feature map M_{XY}, and the value is then shrunk to between 0 and 1 using the Sigmoid activation function, which can produce two sets of weight maps along the height and width directions. The dot product operation is finally applied to these two sets of weight graphs to obtain the weighted weight maps in the X and Y directions.

III. Experiments and Analysis

A significant number of experiments were conducted on four datasets to evaluate the performance of the ESSAN and the model's ability to recognize insufficient samples.

A. Experimental Datasets

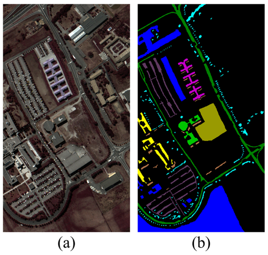
Four hyperspectral datasets were used in this experiment. Three were from widely used public hyperspectral datasets: Matiwan Village [52] in Xiong'an New Area, Pavia Center (PC), Pavia University (PU), and a new land cover categorization database we created, named the Shenzhen University (SZU) HSI dataset. Figs. 9 –12 display the dataset's true color image, the true classification map, the color of each category, and the number of samples.



Label	Color	Class Name	Samples
1	🟩	Water	65971
2	🟦	Trees	7598
3		Asphalt	3090
4	🟦	Self-Bolcking Bricks	2685
5	🟪	Bitumen	6584
6	⬜	Tiles	9248
7	🟩	Shadows	7287
8	🟪	Meadows	42826
9	🟫	Bare soil	2863
		Totle	148152

(c)

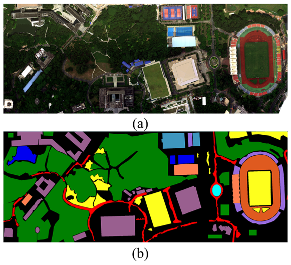
Fig. 10. Original HSI, real land cover, and number of samples in PC dataset. (a) Original image. (b) Ground-truth image. (c) Class name and samples.



Label	Color	Class Name	Samples
1	🟩	Asphalt	6631
2	🟦	Meadows	18649
3	🟨	Gravel	2099
4	🟦	Trees	3064
5	🟪	Painted mental sheets	1345
6	🟨	Bare Soil	5029
7	🟩	Bitumen	1330
8	🟪	Self-Bolcking Bricks	3682
9	🟫	Shadows	947
		Totle	42776

(c)

Fig. 11. Original HSI, real land cover, and number of samples in PU dataset. (a) Original image. (b) Ground-truth image. (c) Class name and samples.



Label	Color	Class Name	Samples
1	🟪	Chair	553947
2	🟦	Court	400909
3	🟦	Water	409759
4	🟨	Grassland	1987197
5	🟥	Asphalt road	1030490
6	🟦	Square	78147
7	🟧	Sunshade	469708
8	🟪	Architecture	2354486
9	🟩	Trees	1739093
10	🟧	Plastic track	1013195
		Totle	10036931

(c)

Fig. 12. Original HSI, real land cover, and number of samples in SZU dataset. (a) Original image. (b) Ground-truth image. (c) Class name and samples.

1) *Matiwan Village (MV):* This dataset was collected using a full-spectrum multimodal imaging spectrometer of the aerial remote sensing system developed under the Speciality Project of High-Resolution Earth Observation System by the Shanghai Institute of Technology Physics, Chinese Academy of Sciences, Xiong'an New Area, Baoding City, Hebei Province, China. The image has a 3750$\times$1580 pixel dimension and a 0.5 m spatial resolution. The number of bands is 256, and the wavelength range is 0.4–1 μm. The image has 18 categories of features, mostly cash crops, as a result of the field campaign of the feature objects. Fig. 9 shows the number of image pixels each category has indicated.
2) *PC:* The ROSIS sensor collected the data for the PC dataset, which covers the center of Pavia in northern Italy. The sensor has 115 bands; however, only 102 of them are present in the PC dataset after excluding 13 noise bands. With a spatial resolution of 1.3 m, the image's spatial size is 1096$\times$715 pixels. Nine land cover feature classes may be found in the photographs.
3) *PU:* The PU dataset was likewise obtained from the ROSIS sensor; 103 bands were kept after 12 noise bands were eliminated. The dimension of the image area is 610 $\times$ 340 pixels. Nine different urban feature categories, each with more than 1000 labeled pixels, are represented on the ground-truth map.
4) *SZU:* SZU is a university in Shenzhen, Guangdong province of China. An unmanned aircraft platform equipped with a Specim FX10 hyperspectral sensor was used to collect data on SZU. This sensor captured 112 bands with a total wavelength range of 0.4–1 μm. The radiometric calibration, geometric correction, and atmospheric correction were applied to the original data during the preprocessing stage. The images have a spatial resolution of 0.1 m and a spatial size of 8757$\times$3373 pixels. The ground-truth data include a total of ten categories.

B. Experimental Setting

1) Sample Settings: For each of the four datasets, an insufficient subset of pixels was chosen as training samples to test the effectiveness of the proposed network model for classification. For MV, PU, PC, and SZU, the training sample proportions were set to 1%, 5%, 1%, and 0.2%, respectively, with SZU having the biggest spatial extent and the fewest samples. Accordingly, the validation and test sample proportions were established at 3%, 10%, 5%, and 0.5%, respectively.

2) Parameter Settings: Pytorch was used to implement all of the networks in this experiment. The input size was set at 13$\times$13 based on prior knowledge; the training period was 100, and Adam was chosen as the optimizer. We tested the five values of 0.001, 0.005, 0.0001, 0.00005, and 0.00005 for the learning rate before settling on 0.0001 as the experiment's learning rate after several iterations of testing. All model results are the average of five experimental results, and the standard deviation of five experimental results is included in the results for each category. All experimental running workstations were configured with Intel(R) Xeon(R) Gold 5218R CPU, NVIDIA GeForce RTX 3080 GPU, machine RAM of 128 GB, and a Windows 10 operating system.

3) Evaluation Factor: The benefits and drawbacks of categorization outcomes were assessed by comparing overall accuracy (OA), average accuracy (AA), and Kappa coefficient. Overall accuracy can be used as a good classification accuracy indicator when the number of samples for each category is balanced. The percentage of samples that the label correctly identified the label relative to samples of actual labels is known as the average accuracy. The degree of correspondence between each category's recognition results and the actual label can be determined using the Kappa coefficient.

C. Comparison Methods

We compared eight commonly used propagation networks, including 3D-CNN, HybridSN, SSRN, CD-CNN, DBMA, DBDA, FDSSC, and FECNet to validate the effectiveness of the proposed method on the dataset.

TABLE I
CLASSIFICATION RESULTS OF 18 CATEGORIES OF THE MV DATASET USING NINE METHODS (%)

Label	Train	Test	Methods								
			3D-CNN	HybridSN	SSRN	CD-CNN	DBMA	DBDA	FDSSC	FECNet	ESSAN
1	3117	9351	99.72±0.05	99.76±0.18	99.53±0.27	99.28±0.43	99.89±0.08	99.27±0.38	99.98±0.03	**99.98±0.02**	99.97±0.03
2	1580	4742	99.77±0.08	99.52±0.28	99.40±0.03	98.72±0.50	99.87±0.12	98.83±0.46	99.97±0.02	99.94±0.03	**99.98±0.02**
3	990	2972	99.64±0.19	99.23±0.50	99.23±0.40	99.41±0.16	99.71±0.15	99.19±0.16	**99.84±0.13**	99.79±0.07	99.80±0.13
4	797	2392	85.41±1.10	80.66±5.73	82.97±3.30	72.24±3.02	85.42±7.23	68.83±0.81	91.92±0.12	91.76±0.53	**92.91±0.15**
5	1541	4623	95.97±0.42	92.51±2.77	96.55±0.96	82.20±6.36	97.29±2.42	89.34±1.21	98.50±0.11	98.39±0.07	**99.49±0.14**
6	221	665	57.44±4.31	43.06±7.32	43.91±10.01	25.86±2.34	46.12±5.62	41.75±1.48	62.76±10.39	54.21±1.85	**66.24±11.2**
7	2391	7174	91.46±0.82	90.40±2.60	92.19±0.48	84.92±4.88	94.48±2.09	87.24±4.44	**97.96±0.71**	97.34±0.21	97.80±0.83
8	2171	6514	92.53±1.30	90.41±3.83	87.52±4.06	78.73±9.39	93.19±1.13	85.19±1.83	96.13±0.61	95.49±0.23	**96.56±0.02**
9	180	540	94.14±0.31	92.10±2.93	94.63±2.18	83.40±11.16	95.62±0.38	91.11±1.70	97.16±0.53	97.13±0.53	**97.97±0.65**
10	583	1749	69.33±3.08	60.55±16.06	62.13±2.81	38.38±6.78	64.06±8.96	45.40±3.25	80.56±5.18	80.02±3.66	**83.13±4.52**
11	666	2000	89.45±1.70	83.83±5.96	82.03±4.90	78.98±2.93	91.83±3.97	78.10±3.55	92.60±2.62	**94.64±1.49**	93.65±2.65
12	344	1033	87.19±1.90	80.67±5.87	78.12±5.08	63.28±7.77	93.48±5.32	62.96±1.73	91.29±4.12	91.49±3.86	**93.56±3.15**
13	3340	10022	93.17±0.25	90.31±2.08	91.91±1.30	83.78±6.45	95.47±1.67	88.62±3.04	97.32±1.41	97.76±1.12	**98.31±0.28**
14	534	1604	87.22±2.17	80.65±11.6	79.28±5.30	69.97±11.66	90.79±9.06	72.03±10.73	**97.59±0.41**	96.61±0.61	97.57±0.50
15	9130	27391	95.14±0.32	93.48±1.08	95.66±0.92	90.49±1.97	96.69±1.32	86.57±1.34	98.74±0.51	98.90±0.06	**98.91±0.54**
16	28	85	56.08±1.47	45.49±17.19	27.06±10.12	16.47±9.98	68.24±6.00	43.14±8.06	83.53±5.88	82.48±3.10	**84.31±4.93**
17	251	753	93.80±0.41	83.18±8.97	90.04±4.32	76.94±15.74	94.25±2.30	87.16±3.57	98.34±0.29	96.85±0.21	**98.41±0.33**
18	174	524	88.30±0.63	77.99±14.44	78.18±7.52	60.81±2.59	86.32±6.31	70.74±4.60	96.90±1.36	95.22±0.71	**98.85±0.19**
OA			93.77±0.20	91.46±2.30	92.55±1.20	86.07±3.59	94.97±1.81	86.87±0.07	96.16±0.53	96.01±0.16	**97.96±0.04**
AA			87.54±0.55	82.43±6.44	82.24±3.11	72.44±5.17	88.48±3.11	77.53±0.46	92.64±1.54	93.32±1.11	**95.17±1.15**
K×100			92.61±0.24	89.87±2.76	91.13±1.44	83.39±4.30	94.02±2.16	84.49±0.11	96.00±0.63	95.00±0.97	**97.56±0.71**
Train Times (s) /epoch			19.24	16.85	23.28	27.88	154.41	47.34	840.75	48.93	32.23

The bold entries represent the highest accuracy values for each class and overall OA, AA, and Kappa coefficients among the 9 methods.

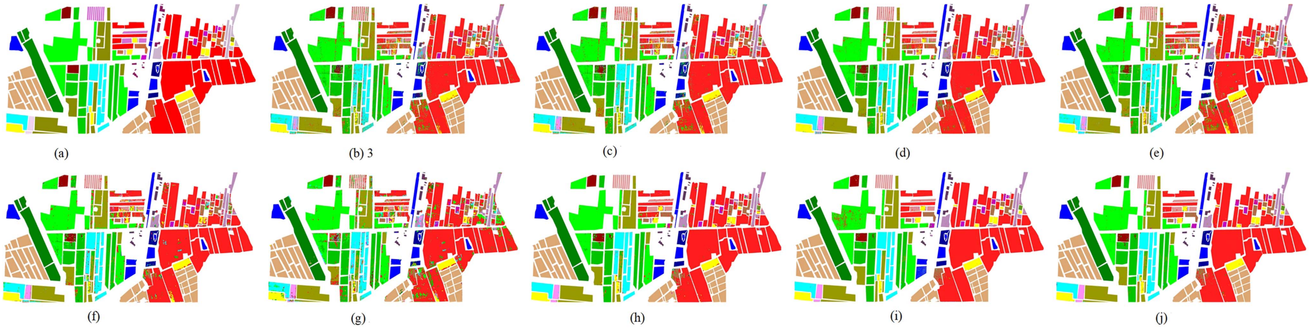

(a) (b) 3 (c) (d) (e)
(f) (g) (h) (i) (j)

Fig. 13. Ground classification map results for MV datasets (1% of training samples). (a) Ground-truth map. (b) Three-dimensional CNN (OA = 93.77%). (c) HybridSN (OA = 91.46%). (d) SSRN (OA = 92.55%). (e) CDCNN (OA = 86.07%). (f) DBMA (OA = 94.97%). (g) DBDA (OA = 86.87%). (h) FDSSC (OA = 96.16%). (i) FECNet(OA = 96.01%). (j) ESSAN (OA = 97.96%).

1) *Three-dimensional CNN [23]:* To extract spatial characteristics, the 3D-CNN framework utilizes 3-D convolutional layers. Three convolutional layers and three maximum pooling layers make up the model. Each convolutional layer also has a BN layer and ReLU added to it.
2) *HybridSN [29]:* To extract joint features of space and spectrum, the HybridSN hybrid network links 3-D convolution and 2-D convolution in series.
3) *SSRN [27]:* It presents the concept of skip connection for residual networks, which can utilize deeper neural networks to enhance classification performance.
4) *CD-CNN [25]:* A deep context network that uses local spatial–spectral properties between nearby vectors of the central pixel to investigate contextual information.
5) *DBMA [40]:* It has two branches to extract spectral and spatial features, and adds an attention mechanism to each of the two branches to make sure that more recognized features can be extracted.
6) *DBDA [41]:* Although it was developed from DBMA, DBDA introduces different attention mechanisms in spectral and spatial branches.
7) *FDSSC [28]:* Uses fast dense space spectrum joint convolution and the tightly coupled structure fully learns each feature to produce an extremely accurate classification.
8) *FECNet [34]:* FECNet increases the RF and extracts more contextual information through expansion convolution and the model includes a feedback mechanism that combines deep and shallow features.

The classification accuracy and feature classification map of ESSAN and eight other models on the MV dataset are shown in Table I and Fig. 13, respectively. The proposed ESSAN in OA (97.96%), AA (95.17%), and Kappa coefficients (97.56%)

TABLE II
CLASSIFICATION RESULTS OF NINE CATEGORIES OF THE PC DATASET USING NINE METHODS (%)

Label	Train	Test	Methods								
			3D-CNN	HybridSN	SSRN	CD-CNN	DBMA	DBDA	FDSSC	FECNet	ESSAN
1	659	3298	99.77±0.04	99.31±0.09	99.33±0.15	99.99±0.01	99.84±0.07	99.50±0.12	99.92±0.04	99.83±0.55	**100.00±0.00**
2	75	379	89.92±1.28	85.64±1.02	70.08±2.01	91.91±1.00	92.79±1.19	70.71±1.14	93.40±0.78	92.19±1.34	**94.16±0.77**
3	30	154	62.47±2.44	61.04±2.36	33.90±3.32	84.20±1.10	71.86±2.45	74.03±3.31	82.68±1.62	93.21±1.66	**93.84±1.96**
4	26	134	52.94±2.64	48.81±1.37	09.40±3.67	67.66±2.46	95.77±1.41	45.52±5.88	95.27±1.27	96.77±0.71	**98.85±1.31**
5	65	329	87.23±1.21	76.96±2.39	28.57±2.28	88.45±2.37	88.35±0.80	78.22±2.98	84.90±0.94	92.56±1.08	**96.19±0.82**
6	92	461	83.51±1.11	88.03±1.00	34.32±2.30	93.64±1.51	98.34±0.51	55.75±0.99	98.77±0.10	**99.20±1.86**	97.65±0.57
7	72	364	84.56±1.14	82.80±2.99	56.59±2.49	85.53±2.48	85.99±0.22	79.12±1.92	86.45±1.28	92.64±1.94	**97.46±0.90**
8	428	2141	98.47±0.10	99.54±0.13	94.32±0.47	99.83±0.06	99.46±0.09	95.78±0.12	99.56±0.09	99.44±2.64	**99.78±0.14**
9	28	143	79.72±1.30	46.29±2.19	51.89±2.81	91.38±1.84	87.65±2.16	63.40±3.35	88.11±0.57	96.35±0.68	**98.82±0.08**
OA			94.56±0.17	93.44±0.19	83.18±0.26	96.83±0.16	97.19±0.14	90.08±0.11	97.41±0.10	98.12±0.81	**99.12±0.14**
AA			82.06±0.46	76.49±0.36	53.16±1.05	89.18±0.41	91.11±0.69	73.56±0.60	92.12±0.39	94.11±1.63	**97.61±0.17**
K×100			92.28±0.24	90.66±0.27	75.79±0.37	95.51±0.23	96.12±0.21	85.98±0.16	96.33±0.13	97.45±0.98	**98.72±0.34**
Train Time (s) /epoch			0.57	0.23	1.35	1.02	2.09	1.65	2.55	1.15	1.13

The bold entries represent the highest accuracy values for each class and overall OA, AA, and Kappa coefficients among the 9 methods.

was higher than those from other approaches (see Table I). While ESSANs training round only takes around 32 s and FDSSCs training round takes about 14 min, both of them produce extremely precise classification results. This is because, during training, the FDSSC model inputs all the feature maps produced by the previous module into the subsequent module, leading to a massive number of parameters and, thus, a slow training process. By utilizing an expansion convolutional residual block, ESSAN and FECNet are able to gather global information while also lowering the number of parameters, speeding up training, and increasing computational efficiency without compromising accuracy. Due to the insufficient number of training samples, HybridSN, SSRN, and CDCNN did not effectively extract "soybean (label16)." DBMA and ESSAN, which introduced the attention mechanism, gave more weight to important information in the case of insufficient samples and achieved higher classification accuracy. As can be observed from Fig. 13, numerous features in the SSRN, DBDA, and other models were incorrectly identified because the MV datasets are all vegetation-based and have similar spectral properties. The suggested ESSAN, however, produces a better feature classification map.

The evaluation metrics and feature classification plots for the PC dataset are displayed in Table II and Fig. 14, respectively. The number of feature categories in PC is nine, which is half as many as in MV; however, the PC dataset performs better in terms of classification than the MV dataset. Table II demonstrates that the proposed method outperforms previous comparison methods in terms of total OA (99.12%), AA (97.61%), and Kappa coefficient (98.72%) for each feature class. FECNet, FDSSC, DBMA, and CD-CNN had the second highest OA after the proposed method, although their AA was 3.5%, 5.49%, 6.5%, and 8.43% lower than MSSANs. Fig. 14 shows that the other compared methods cannot separate "Bitumen (label 5)" well and have all misclassified "Bitumen (label 5)" into "self-locking bricks (label 4)." The 3D-CNN, HybridSN, SSRN, and CD-CNN show a lot of noise on the classification graph with a salt-and-pepper phenomenon. However, the classification graphs of the dual-branch DBMA and DBDA are noticeably superior to those of the other evaluated approaches. The proposed ESSAN accurately captures the spatial and spectral properties of the data using a dual-branch structure, as shown in Table II and Fig. 14, and the obtained classification results are the closest to the ground-truth labels.

The classification evaluation metrics and result plots for the PU dataset are shown in Table III and Fig. 15, respectively. ESSAN still achieved the highest OA, AA, and Kappa values. Compared with other methods, the proposed ESSAN performed significantly better in terms of OA. In this dataset, CD-CNN performed well (OA $= 95.36\%$) and had the highest accuracy in classifying the "Trees (label 4)" class, at 98.28%. Fig. 15 displays the classification results for each method in the PU dataset. Zooming in on the classification plots, we can see that SSRN and HybridSN have more noise, which may be due to the large spectral variation of the same species of features causing severe feature mixing. It is apparent that the classification maps produced by FECNet, FDSSC, DBMA, DBDA, and CD-CNN are superior to the models mentioned above. In comparison, the ESSAN model generates a smoother feature classification map by fully utilizing the incorporation of global information and attention mechanisms.

The results of nine classification methods are displayed in Table IV and Fig. 16. There are ten categories in the SZU dataset and, because of the huge variations between them and the more regular features, all classification methods achieved good OA. However, the ESSAN described in this research produces the greatest OA, AA, and Kappa coefficients in SZU. In the categories of "water (label 3)," DBMA, DBDA, and FDSSC all achieved 100% accuracy, and the final acquired OA for these three models was only 0.46%, 0.33%, and 0.33% lower than the suggested ESSAN. However, the average training round time for the FDSSC is 581 s, which is 38 times longer than the ESSANs 15 s. Therefore, FDSSC training becomes extremely slow when the image space is large and there are many sample pixels, and ESSAN, with the addition of double branching and expansion convolutional residuals block, can ensure that all discriminative features are extracted in complex scenes while also speeding up training. Furthermore, as illustrated in Fig. 16, all approaches appear to mistakenly categorize "trees (label 9)" as "grassland (label 4)," with the suggested method having the fewest errors.

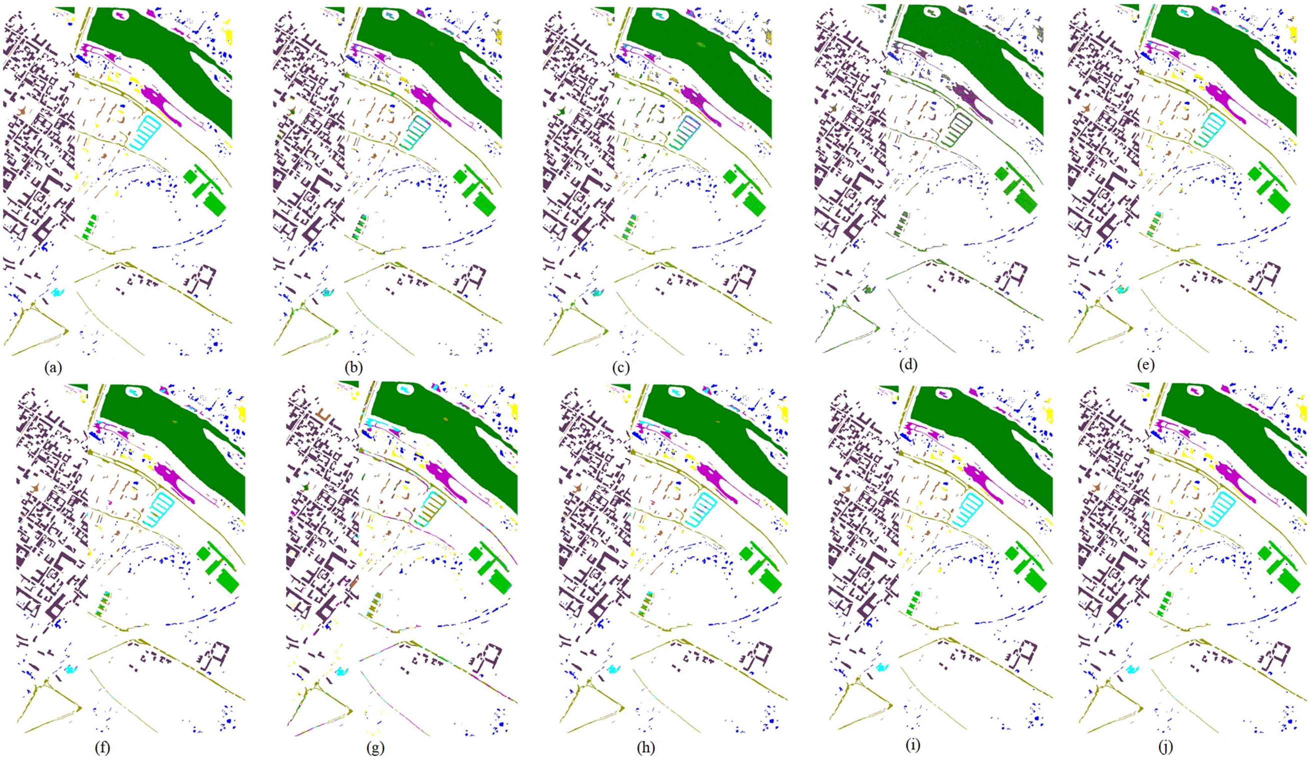


Fig. 14. Ground classification map results for the PC datasets (1% of training samples). (a) Ground-truth map. (b) Three-dimensional CNN (OA = 94.56%). (c) HybridSN (OA = 93.44%). (d) SSRN (OA = 83.18%). (e) CDCNN (OA = 96.83%). (f) DBMA (OA = 97.19%). (g) DBDA (OA = 90.08%). (h) FDSSC (OA = 97.41%). (i) FECNet (OA = 98.12%). (j) ESSAN (OA = 99.12%).

TABLE III
Classification Results of Nine Categories of the PU Dataset Using Nine Methods (%)

Label	Train	Test	Methods								
			3D-CNN	HybridSN	SSRN	CD-CNN	DBMA	DBDA	FDSSC	FECnet	ESSAN
1	331	663	91.44±1.11	93.15±0.79	78.55±0.89	94.26±0.49	85.90±1.88	83.79±1.64	96.88±1.22	96.00±0.45	**98.21±0.71**
2	918	1836	97.28±0.09	95.53±0.85	95.03±0.59	99.06±0.14	95.35±0.91	95.93±0.44	98.66±0.40	99.01±2.44	**99.42±1.63**
3	104	209	70.19±6.71	42.63±2.40	40.71±4.60	79.17±2.97	68.59±3.54	79.81±4.37	83.97±0.91	84.12±1.31	**97.89±0.45**
4	153	306	94.34±1.11	96.08±0.00	89.76±2.41	**98.28±0.82**	87.80±1.63	84.10±1.72	90.41±2.63	98.23±2.92	96.45±1.75
5	67	134	98.51±0.00	96.52±2.81	93.03±0.70	99.00±1.41	99.00±1.41	99.50±0.70	99.00±1.41	98.68±2.09	**99.98±1.34**
6	251	502	89.24±1.63	64.14±1.13	50.33±1.31	94.56±1.96	72.64±1.91	82.20±2.53	98.54±0.50	95.98±2.67	**99.80±1.17**
7	66	133	71.21±3.27	26.77±3.11	14.14±3.98	85.86±4.68	50.00±3.27	54.55±11.0	83.84±2.58	95.31±1.02	**95.68±0.84**
8	184	368	86.78±1.79	80.98±1.93	50.54±1.77	88.95±1.36	87.32±2.23	69.93±1.36	96.38±1.12	91.99±0.71	**97.49±0.55**
9	47	94	97.16±2.01	92.20±2.65	70.92±5.01	**97.16±2.65**	87.23±3.01	75.89±1.00	92.91±2.01	94.16±1.13	93.61±0.82
	OA		92.19±0.18	85.45±0.68	77.15±0.66	95.36±0.62	87.16±0.24	86.89±0.41	96.28±0.17	95.67±0.48	**98.73±0.12**
	AA		88.46±0.65	76.44±1.20	64.78±1.78	92.85±1.34	81.54±0.74	80.63±1.01	93.40±0.24	93.49±0.51	**97.89±0.99**
	K×100		89.66±0.24	80.55±0.90	69.48±0.88	93.86±0.83	82.93±0.30	82.63±0.56	95.08±0.22	95.28±0.86	**98.01±0.36**
Train Times (s) /epoch			0.75	0.32	2.01	1.51	3.35	2.56	4.71	2.12	1.98

The bold entries represent the highest accuracy values for each class and overall OA, AA, and Kappa coefficients among the 9 methods.

In conclusion, the proposed approach ESSAN in this research produced the best OA, AA, and Kappa coefficients on all datasets, as well as the most accurate ground-truth feature classification maps, proving the full potential of ESSAN in HSI data classification.

D. Discussion

1) Ablation Experiments: Some ablation experiments were carried out to confirm the effectiveness of each component of the suggested method. SSC stands for single standard CNN, SEC for single expansion CNN, DSC for double standard CNN, and DEC for double expansion CNN. Table V presents the findings with OA serving as the criterion for accuracy assessment. Concatenating the spatial branch with the spectral branch in the dual-branch network produces the single-branch network used in the experiment. As can be seen, the dual-branch network achieves superior classification accuracy when compared with the single-branch network since it can completely and effectively extract spatial and spectral information from the

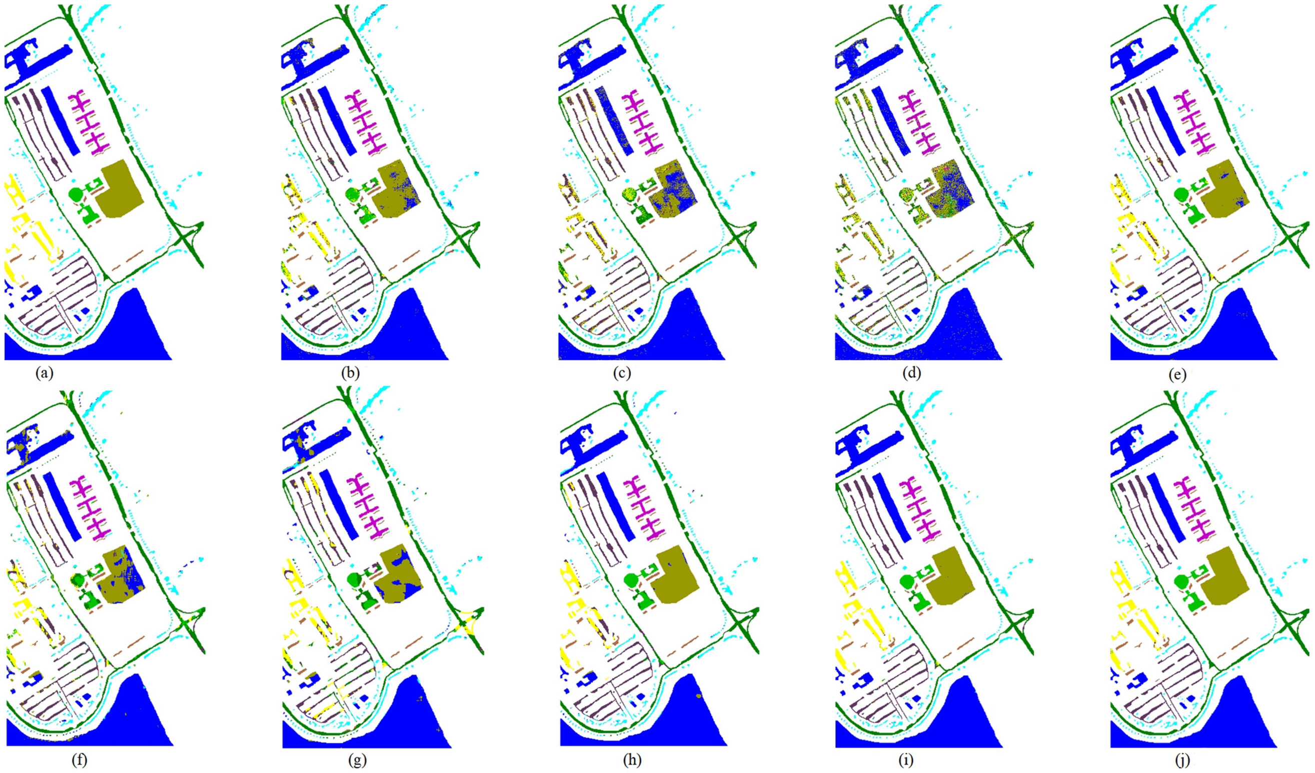


Fig. 15. Ground classification map results for the PU datasets (5% of training samples). (a) Ground-truth map. (b) Three-dimensional CNN (OA = 92.19%). (c) HybridSN (OA = 85.45%). (d) SSRN (OA = 77.15%). (e) CDCNN (OA = 95.36%). (f) DBMA (OA = 87.16%). (g) DBDA (OA = 86.89%). (h) FDSSC (OA = 96.28%). (i) FECNet (OA = 95.67%). (j) ESSAN (OA = 98.73%).

TABLE IV
Classification Results of Ten Categories of SZU Dataset Using Nine Methods (%)

Label	Train	Test	Methods								
			3D-CNN	HybridSN	SSRN	CD-CNN	DBMA	DBDA	FDSSC	FECNet	ESSAN
1	1107	2769	94.08±0.11	93.84±0.20	96.70±0.20	93.84±0.34	97.54±0.29	98.05±0.11	98.41±0.48	98.67±0.14	**98.89±0.38**
2	801	2004	94.66±0.35	95.36±0.15	96.68±0.42	96.41±0.95	97.83±0.62	98.38±0.17	98.30±0.37	98.41±0.20	**98.73±0.17**
3	819	2048	98.24±0.39	99.76±0.05	99.39±0.12	99.78±0.07	**100.00±0.00**	**100.00±0.00**	**100.0±0.09**	99.89±0.02	**100.0±0.00**
4	3974	9935	99.06±0.12	98.69±0.17	99.18±0.01	99.28±0.13	99.27±0.03	99.35±0.06	99.38±0.04	99.36±0.06	**99.52±0.00**
5	2060	5152	93.60±0.05	95.41±0.36	96.87±0.15	96.29±0.19	98.16±0.62	98.55±0.01	97.88±0.12	97.00±0.14	**98.81±0.01**
6	156	390	77.31±3.72	78.97±4.62	88.46±4.36	69.23±1.03	87.18±7.69	95.26±0.13	93.33±2.46	94.31±2.76	**96.15±2.05**
7	939	2348	97.10±0.00	98.00±0.00	99.04±0.15	98.64±0.21	98.98±0.34	**99.40±0.09**	**99.40±1.22**	99.20±0.21	99.21±0.28
8	4708	11772	97.38±0.09	96.36±0.08	97.93±0.21	97.15±0.08	98.74±0.11	98.97±0.03	98.62±0.17	97.99±0.12	**99.35±0.14**
9	3478	8695	99.51±0.08	99.08±0.20	99.60±0.03	99.36±0.24	99.60±0.02	**99.71±0.06**	99.68±0.06	99.70±0.09	99.59±0.07
10	2026	5065	99.54±0.01	99.34±0.01	**99.84±0.02**	99.40±0.05	99.75±0.03	98.92±0.00	99.41±0.05	99.81±0.05	99.77±0.03
	OA		97.49±0.05	97.40±0.10	98.47±0.09	97.84±0.00	98.90±0.06	99.03±0.01	99.03±0.09	98.94±0.06	**99.36±0.03**
	AA		95.05±0.31	95.48±0.53	97.37±0.48	94.94±0.05	97.70±0.79	98.66±0.00	98.50±0.22	98.12±0.23	**98.96±0.22**
	K×100		97.23±0.06	96.93±0.13	98.20±0.10	97.45±0.01	98.71±0.07	98.84±0.01	98.87±0.01	98.69±0.04	**99.22±0.04**
Train Times (s) /epoch			5.45	4.06	7.26	9.29	30.17	15.01	581.74	30.65	15.82

The bold entries represent the highest accuracy values for each class and overall OA, AA, and Kappa coefficients among the 9 methods.

original data. In comparison with the SEC, the OA of DEC is increased in the MV, PC, PU, and SZU datasets by 2.28%, 2.51%, 5.47, and 0.7%, respectively. As can be seen in Fig. 17, the addition of a dual-branch to PU results in the highest gain in OA, although the number of training samples for PU is the smallest. This suggests that the dual-branch block is better suited for improving model accuracy in datasets with limited training samples.

Expansion convolution has enhanced the classification accuracy compared with standard convolution, and DEC and SEC employing it have higher overall accuracy than DSC and SSC. In this experiment, the patch size fed into the network is 13. RF completely covers the information in the patch size when three expansion convolutions running at various expansion rates are utilized continuously. Thus, to increase network accuracy, the expansion convolution can receive global information over

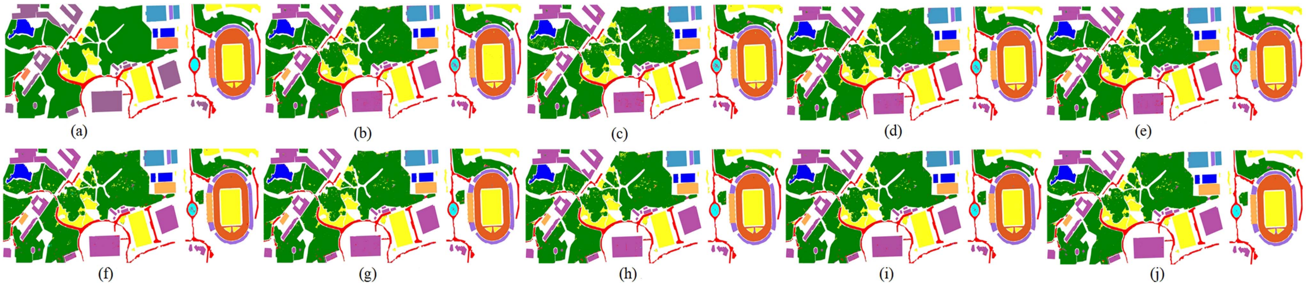


Fig. 16. Ground classification map results for the SZU dataset (0.2% of training samples). (a) Ground-truth map. (b) Three-dimensional CNN (OA = 97.49%). (c) HybridSN (OA = 9.40%). (d) SSRN (OA = 98.47%). (e) CDCNN (OA = 97.84%). (f) DBMA (OA = 98.90%). (g) DBDA (OA = 99.03%). (h) FDSSC (OA = 99.03%). (i) FECNet (OA = 98.94%). (j) Proposed (OA = 99.36%).

TABLE V
ANALYSIS OF ABLATION EXPERIMENTS OF DIFFERENT BLOCKS ON FOUR DATASETS

	MV	PC	PU	SZU
SSC	94.42	95.21	92.12	96.69
SEC	95.17	95.94	92.81	97.41
DSC	97.16	98.02	97.08	98.10
DEC	97.45	98.45	98.28	98.11
DEC + CAB	**97.96**	**99.12**	**98.73**	**99.36**

The bold entries represent the highest accuracy obtained for each column.

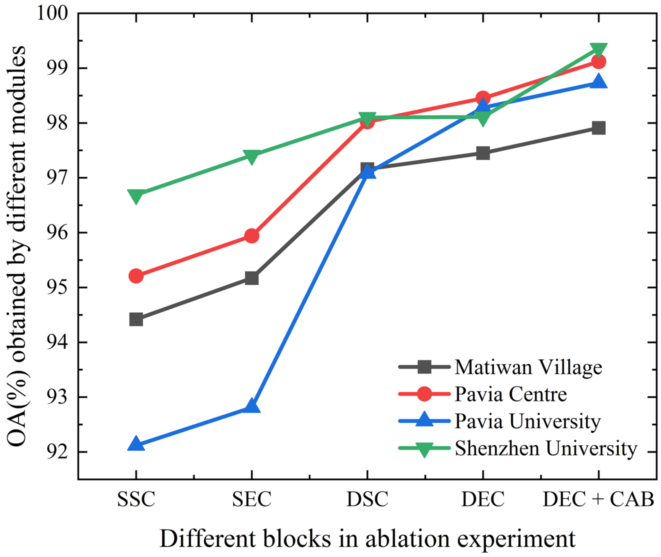


Fig. 17. OA (%) of ablation experiments of different blocks on four datasets.

a larger RF. The highest OA, AA, and Kappa coefficients are obtained in the four datasets when CAB is introduced to the network, demonstrating that adding the attention module to the HSI classification model ESSAN in this article is effective.

2) Effects of Different Attention Models: Two traditional attention models, the convolutional block attention module (CBAM) [36] and the squeeze-and-excitation (SE) [37] attention module, were compared to demonstrate the efficacy of the CAB used in this article. Table VI displays the categorization accuracy results from four datasets using various attention modules. As can be observed, CAB achieved the highest OA on the MV, PC, PU, and SZU datasets, which were 97.96%, 99.12%, 98.73%, and 98.73%, respectively. The CBAM assigns weights in both the channel and the spatial dimensions. The parameters become redundant when many weights are applied to the features, which does not help to increase the model accuracy overall. The interdependence between channels has been established using SE, which increases accuracy and adds a modest bit of computation In the MV dataset, SE earned the greatest AA, and in the PC dataset, the highest OA and Kappa coefficients. Lightweight CABs will not burden network computation because they simply give weight to spatial dimensions. Table VI presents that, in the PU and SZU datasets, the CAB approach had the largest OA and the best classification accuracy.

TABLE VI
ABLATION ANALYSIS OF DIFFERENT ATTENTION MODULES ON FOUR DATASETS (%)

		DEC + CBAM	DEC + SE	DEC + CAB
MV	OA	96.74	97.93	**97.96**
	AA	93.40	**96.11**	95.17
	Kappa	96.13	97.55	**97.56**
PC	OA	98.31	99.11	**99.12**
	AA	95.58	97.18	**97.61**
	Kappa	97.60	**98.81**	98.72
PU	OA	98.34	98.62	**98.73**
	AA	96.88	97.76	**97.89**
	Kappa	97.81	**98.18**	98.01
SZU	OA	98.56	98.73	**99.36**
	AA	97.20	97.34	**98.96**
	Kappa	98.29	98.49	**99.22**

The bold entries represent the highest accuracy obtained for each row.

The output results of the attention model were extracted as semantic features for t-distributed stochastic neighbor embedding (t-SNE) [53] dimensionality reduction, and high-dimensional data were reduced to two dimensions for visualization after applying different attention models to four datasets, as shown in Fig. 18. It can be seen that the addition of CBAM to the model does not lead to correctly distinguish between ground objects. Because of the classification phenomenon of different objects with the same spectrum in the MV dataset, the mixing of different types of features is very serious after adding CBAM.

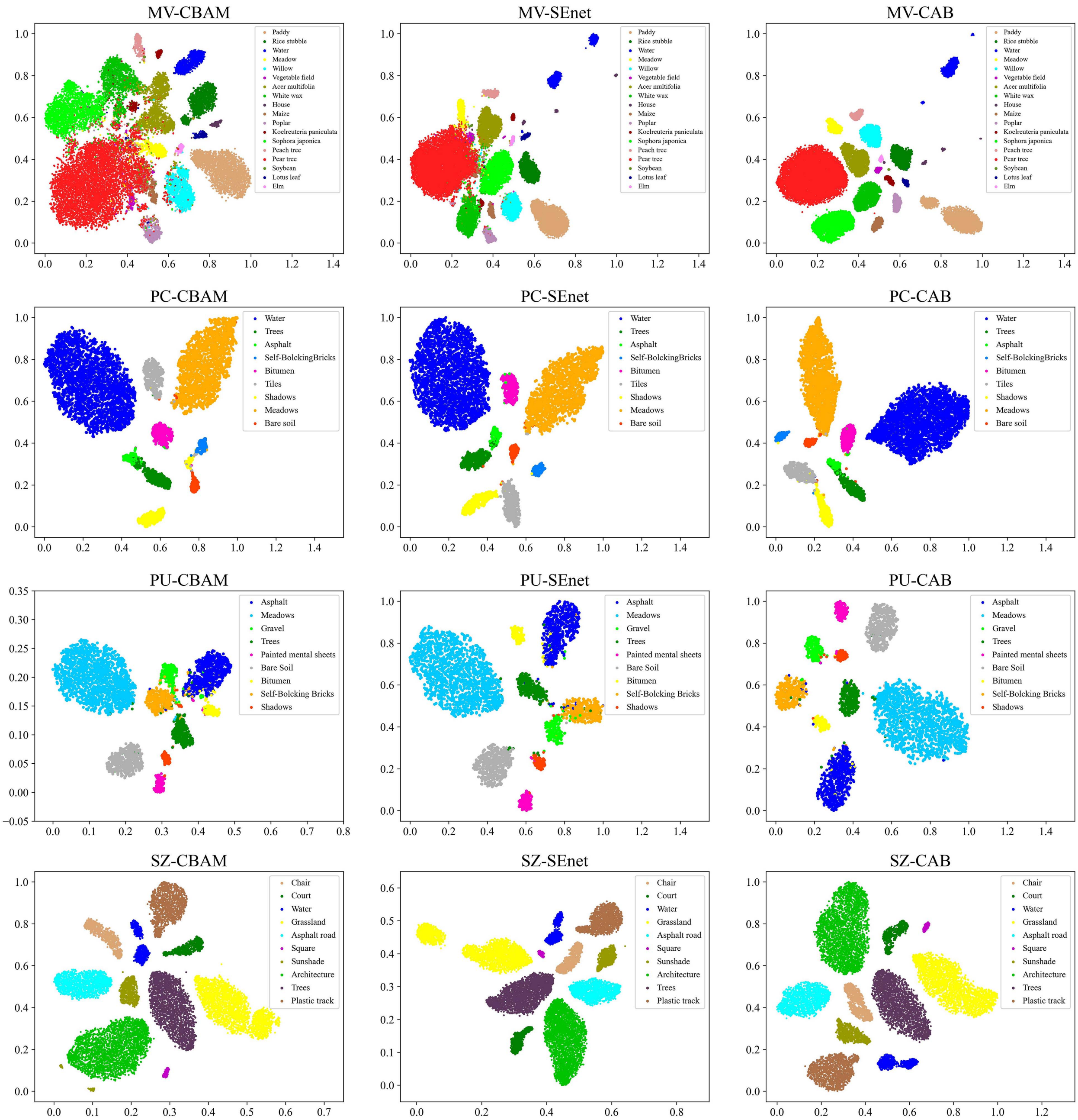


Fig. 18. Two-dimensional t-SNE. Scatter visualization after applying different attention models to four datasets.

SEs visualization is far superior to CBAMs, but there is still some feature mixed in the MV dataset. In the PC dataset, SE and CAB both obtained the same OA and both “Tree” and “Asphalt” features are misclassified. The phenomenon of the same objects with a different spectrum was reduced after CAB was incorporated into the model and the same type of features were grouped into the same area. In particular, the boundaries of each type of feature in the MV dataset are very clear, and the advantages are obvious compared with the other two attention models. This also verifies the effectiveness of adding CABs to ESSAN.

3) Performance of ESSAN on Insufficient Samples: Experiments were carried out with extremely insufficient samples to confirm the efficacy and applicability of the proposed ESSAN model in HSI dataset classification with insufficient training samples. For this experiment, the MV and SZU datasets were chosen, and the training samples for each dataset were reduced by the same amount, taking 0.25%, 0.5%, 0.75%, and 0.05%, 0.1%, and 0.15% of the total samples, respectively. The outcomes of this experiment are shown in Table VII and Fig. 19. As can be observed from Fig. 19, the advantages of ESSAN over other approaches grow as the sample size decreases. When the sample size of the two datasets was reduced to one-quarter of the original, the OA of other comparison methods showed a significant downward trend. In the MV dataset, CD-CNN and DBDA had the worst results, and HybridSN had the worst outcomes in

TABLE VII
OA Results (%) of MV and SZU Datasets in Nine Models With an Insufficient Training Sample Ratio

Models	MV				SZU			
	0.25%	0.50%	0.75%	1.00%	0.05%	0.10%	0.15%	0.20%
3D-CNN	88.00	91.07	93.13	93.52	95.32	96.55	97.26	97.38
HybridSN	84.70	90.12	92.04	88.18	94.70	96.34	97.09	97.36
SSRN	82.07	91.57	93.16	91.62	96.94	97.42	97.88	98.46
CD-CNN	64.07	84.58	86.48	89.07	96.03	97.26	97.51	97.73
DBMA	83.03	93.20	95.46	92.46	97.03	98.25	98.73	98.89
DBDA	68.07	81.12	77.57	86.71	97.48	98.50	98.94	99.13
FDSSC	94.19	96.35	97.03	97.11	97.26	98.09	98.61	99.04
FECNet	90.79	93.22	95.16	97.01	97.51	97.99	98.48	98.94
ESSAN	**94.39**	**96.92**	**97.87**	**97.96**	**98.07**	**98.84**	**99.24**	**99.36**

The bold entries represent the highest accuracy obtained for each row.

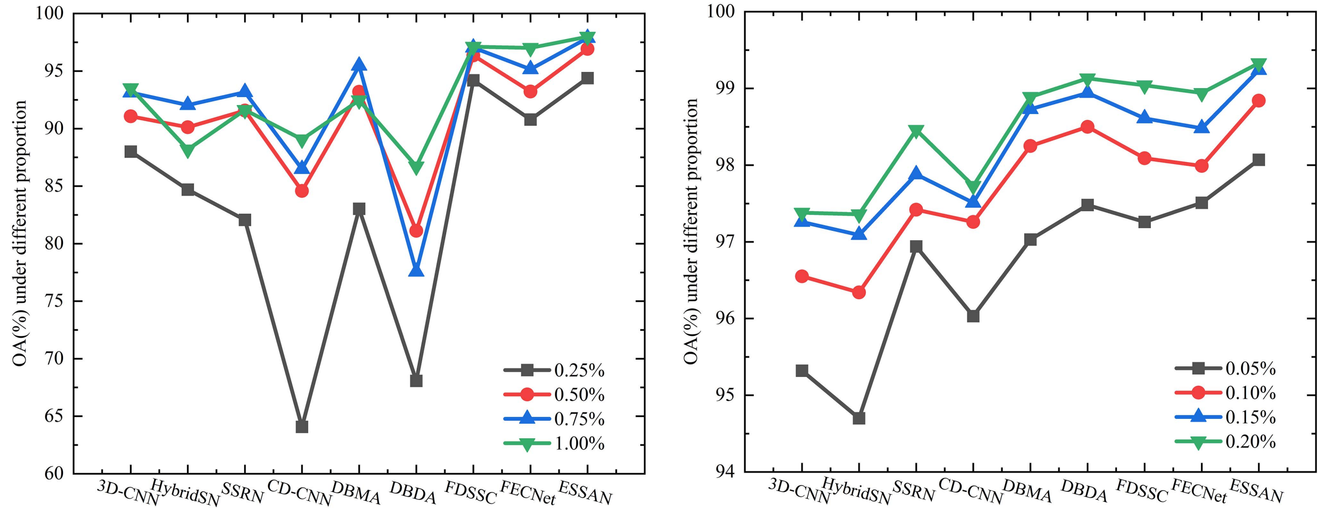


Fig. 19. OA (%) with minimal training samples, MV is on the left and SZU is on the right.

SZU. Overall, only two models—FDSSC and ESSAN—show a slight reduction in OA, and these two models outperform others when dealing with extremely tiny data. However, FDSSCs training time takes significantly longer than ESSANs. Table VII presents that ESSAN has the highest OA in both datasets. As a result, the experimental results in this section further verify the effectiveness of our proposed method ESSAN in insufficient sample situations.

4) Analyzing the Effect of the Dropout Rate: The effect of different learning rates on accuracy was investigated in this experiment using four datasets with dropout rates ranging from 0.1 to 0.9. The OA, AA, and Kappa coefficients for the four datasets at various dropout rates are presented in Table VIII. In the proposed method, PU performs best when the dropout rate is 0.7; when it is 0.4, PC and MV provide the best classification results; SZU chose a dropout rate of 0.5. Table VIII presents that when the dropout rate increases, the classification accuracy of ESSAN approximately follows a rising and then dropping trend. In conclusion, MV, PC, PU, and SZU have dropout rates set at 0.4, 0.4, 0.7, and 0.5, respectively.

5) Impact of the Number of Training Samples: Fig. 20 shows the classification accuracy results for the four datasets in the ESSAN model with various training ratios. Here, the proportions of the training sample are represented by the horizontal coordinates, while the vertical coordinates indicate the overall accuracy. Evidently, as the number of training samples rises, classification performance on all four datasets improves. This further demonstrates the efficiency of the proposed method by showing that it can obtain higher classification performance with adequate training samples.

IV. Conclusion

In this article, we proposed an HSI classification model ESSAN based on the expansion convolution. First, to improve the features' ability to be distinguished from one another, we created a dual-channel structure of joint spatial–spectral features, with both the spatial and spectral branches being blocks of residual structures based on the expansion convolution; the RF was increased by stacking of expanded convolutional layers to gain richer global feature information. In addition, the attention mechanism was added to the network to acquire the weight map to improve the ability of feature extraction, which greatly increased classification accuracy and accelerated the model's

TABLE VIII
ANALYSIS OF DIFFERENT DROPOUT RATES ON FOUR DATASETS (%)

Dropout		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MV	OA	97.13	97.10	97.10	**97.96**	97.07	97.84	97.91	97.67	94.48
	AA	95.41	95.33	95.02	95.17	95.99	95.58	**96.11**	95.22	82.05
	Kappa	96.85	96.55	96.74	**97.56**	96.71	97.43	97.52	97.24	93.43
PC	OA	98.91	99.06	99.08	**99.12**	98.85	98.87	98.94	99.01	96.47
	AA	96.92	97.58	97.52	**97.61**	96.99	96.81	96.97	97.52	87.13
	Kappa	98.45	98.67	98.70	**98.74**	98.38	98.40	98.50	98.73	94.99
PU	OA	98.59	98.56	98.63	98.59	98.63	98.59	**98.73**	98.30	95.43
	AA	96.85	96.86	97.16	97.06	97.24	97.25	**97.89**	96.33	90.76
	Kappa	98.13	98.10	**98.20**	98.13	98.19	98.13	98.01	97.76	93.95
SZU	OA	98.63	98.68	98.98	99.02	**99.36**	99.21	98.60	98.26	97.08
	AA	97.26	97.41	97.31	97.39	**98.96**	97.43	97.60	96.69	93.34
	Kappa	98.39	98.44	98.44	98.49	**99.22**	98.32	98.34	97.94	96.55

The bold entries represent the highest accuracy obtained for each row.

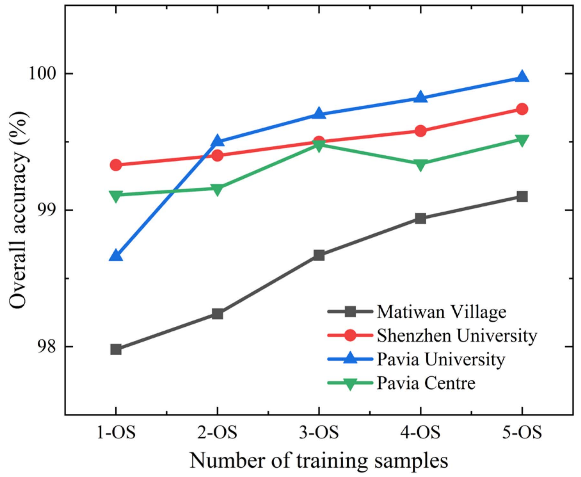


Fig. 20. Classification accuracy of training samples at different scales (%). OS stands for the percentage of the four datasets' original samples, whereas 2-OS stands for twice the original samples.

training efficiency. The comparison of ESSAN and other eight popular deep learning HSI classification algorithms reveals that ESSAN obtains optimum classification results with few training samples and has greater classification efficiency on four different datasets. In MV datasets, in particular, ESSAN provides more overt advantages. This is because all of the species in the MV dataset are tree species, there is little variation existed in the features, and the phenomenon of different objects with the same spectrum exists. As a consequence, the recognition results of the other comparison methods on MV seem to be relatively poor, whereas ESSAN obtains greater OA, AA, and Kappa coefficients. This demonstrates that ESSAN has certain advantages in identifying similar objects in HSI data.

REFERENCES

[1] S. Wang et al., "Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing," *Remote Sens. Environ.*, vol. 271, 2022, Art. no. 112914.

[2] L. Liang et al., "Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method," *Remote Sens. Environ.*, vol. 165, pp. 123–134, 2015.

[3] S. Junttila et al., "Close-range hyperspectral spectroscopy reveals leaf water content dynamics," *Remote Sens. Environ.*, vol. 277, 2022, Art. no. 113071.

[4] E. A. Antipov and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1772–1778, 2012.

[5] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-Based edge-preserving features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7140–7151, Dec. 2017.

[6] H. Zhang, W. Liu, and H. Lv, "Spatial-spectral joint classification of hyperspectral image with locality and edge preserving," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2240–2250, 2020.

[7] C. Jiao, C. Chen, R. G. McGarvey, S. Bohlman, L. Jiao, and A. Zare, "Multiple instance hybrid estimator for hyperspectral target characterization and sub-pixel target detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 235–250, 2018.

[8] W. Xie, J. Lei, J. Yang, Y. Li, Q. Du, and Z. Li, "Deep latent spectral representation learning-based hyperspectral band selection for target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2015–2026, Mar. 2020.

[9] J. Jacques and C. Preda, "Model-based clustering for multivariate functional data," *Comput. Statist. Data Anal.*, vol. 71, pp. 92–106, 2014.

[10] Y. Ait-Sahalia and D. Xiu, "Principal component analysis of high-frequency data," *J. Amer. Stat. Assoc.*, vol. 114, no. 525, pp. 287–303, 2019.

[11] Q. Du, "Modified fisher's linear discriminant analysis for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 503–507, Oct. 2007.

[12] F. Kuang, W. Xu, and S. Zhang, "A novel hybrid KPCA and SVM with GA model for intrusion detection," *Appl. Soft Comput.*, vol. 18, pp. 178–184, 2014.

[13] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[14] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, 2010.

[15] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[16] X. Zhang, X. Jiang, J. Jiang, Y. Zhang, X. Liu, and Z. Cai, "Spectral–spatial and superpixelwise PCA for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502210.

[17] K. He et al., "A dual global-local attention network for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5527613.

[18] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, 2021, Art. no. 107398.

[19] X. Wei, X. Yu, B. Liu, and L. Zhi, "Convolutional neural networks and local binary patterns for hyperspectral image classification," *Eur. J. Remote Sens.*, vol. 52, no. 1, pp. 448–462, 2019.

[20] H. Gao, S. Lin, Y. Yang, C. Li, and M. Yang, "Convolution neural network based on two-dimensional spectrum for hyperspectral image classification," *J. Sensors*, vol. 2018, 2018, Art. no. 8602103.

[21] C. Wang, N. Ma, Y. Ming, Q. Wang, and J. Xia, "Classification of hyperspectral imagery with a 3D convolutional neural network and J-M distance," *Adv. Space Res.*, vol. 64, no. 4, pp. 886–899, 2019.

[22] S. Nezami, E. Khoramshahi, O. Nevalainen, I. Polonen, and E. Honkavaara, "Tree species classification of drone hyperspectral and RGB imagery with deep learning convolutional neural networks," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1070.

[23] S. Mirzaei, H. van Hamme, and S. Khosravani, "Hyperspectral image classification using non-negative tensor factorization and 3D convolutional neural networks," *Signal Process., Image Commun.*, vol. 76, pp. 178–185, 2019.

[24] W. Pi, J. Du, Y. Bi, X. Gao, and X. Zhu, "3D-CNN based UAV hyperspectral imagery for grassland degradation indicator ground object classification research," *Ecol. Inform.*, vol. 62, 2021, Art. no. 101278.

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

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[27] 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.

[28] 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.

[29] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[30] H. C. Tinega, E. Chen, L. Ma, D. O. Nyasaka, and R. M. Mariita, "HybridGBN-SR: A deep 3D/2D genome graph-based network for hyperspectral image classification," *Remote Sens.*, vol. 14, no. 6, 2022, Art. no. 1332.

[31] X. Yang et al., "Synergistic 2D/3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, 2020, Art. no. 2033.

[32] Z. Xie, J. Hu, X. Kang, P. Duan, and S. Li, "Multilayer global spectral-spatial attention network for wetland hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518913.

[33] F. Feng, Y. Zhang, J. Zhang, and B. Liu, "Small sample hyperspectral image classification based on cascade fusion of mixed spatial-spectral features and second-order pooling," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 505.

[34] C. Shi, D. Liao, T. Zhang, and L. Wang, "Hyperspectral image classification based on expansion convolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5528316.

[35] C. Zhao, H. Zhao, G. Wang, and H. Chen, "Hybrid depth-separable residual networks for hyperspectral image classification," *Complexity*, vol. 2020, 2020, Art. no. 4608647.

[36] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[37] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 1, 2020.

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

[39] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[40] 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.

[41] 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.

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

[43] M. Gong et al., "A spectral and spatial attention network for change detection in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5521614.

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

[45] 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.

[46] K. Liu et al., "Mapping coastal wetlands using transformer in transformer deep network on China ZY1-02D hyperspectral satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3891–3903, 2022.

[47] J. Zhu et al., "Survey of few shot learning of deep neural network," *Comput. Eng. Appl.*, vol. 57, no. 7, pp. 22–33, 2021.

[48] S. Jia, S. Jiang, Z. Lin, N. Li, M. Xu, and S. Yu, "A survey: Deep learning for hyperspectral image classification with few labeled samples," *Neurocomputing*, vol. 448, pp. 179–204, 2021.

[49] S. Sengupta et al., "A review of deep learning with special emphasis on architectures, applications and recent trends," *Knowl.-Based Syst.*, vol. 194, 2020, Art. no. 105596.

[50] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460, doi: 10.1109/WACV.2018.00163.

[51] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.

[52] Y. Cen et al., "Aerial hyperspectral remote sensing classification dataset of Xiongan new area (Matiwan village)," *J. Remote Sens.*, vol. 24, no. 11, pp. 1299–1306, 2020.

[53] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Shuo Wang received the B.Sc. degree in geographic information science from the China University of Mining and Technology, Xuzhou, China, in 2021. She is currently working toward the master's degree in photogrammetry and remote sensing with the Chinese Academy of Surveying and Mapping, Beijing, China.

Her research focuses on classification based on hyperspectral image and LiDAR point cloud.

Zhengjun Liu received the Ph.D. degree in cartography and geographical information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2003.

He is currently a Professor with the Chinese Academy of Surveying and Mapping, Beijing, China. His research interests include remote sensing image analysis, mapping application of LiDAR and multisensory fusion, and facility management application with LiDAR and multisensors.

Yiming Chen received the Ph.D. degree in cartography and geographical information system from Beijing Normal University, Beijing, China, in 2018.

He is currently an Assistant Professor Fellow with the Chinese Academy of Surveying and Mapping, Beijing, China. His research interests include air-ground LiDAR data forest resources stereomonitoring survey.

Chengchao Hou received the B.Sc. degree in surveying and mapping engineering from Shijiazhuang Tiedao University, Shaoxing, China, in 2020. He is currently working toward master's degree in photogrammetry and remote sensing with the Chinese Academy of Surveying and Mapping, Beijing, China.

His research focuses on tree species classification based on deep learning from hyperspectral images.

Aixia Liu received the Ph.D. degree in cartography and geographical information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2004.

She is currently a Research Professor with Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, Beijing. Her research interests include satellite remote sensing application, natural resources survey, and remote sensing monitoring.

Zhenbei Zhang received the B.Sc. degree in geographic information science from the China University of Mining and Technology, Xuzhou, China, in 2021. He is currently working toward the master's degree in structural geology with the Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China.