

# Unseen Feature Extraction: Spatial Mapping Expansion With Spectral Compression Network for Hyperspectral Image Classification

Chunyan Yu<sup>✉</sup>, Senior Member, IEEE, Yuanchen Zhu, Meiping Song<sup>✉</sup>, Member, IEEE,  
Yulei Wang<sup>✉</sup>, Member, IEEE, and Qiang Zhang<sup>✉</sup>, Member, IEEE

**Abstract**—Hyperspectral image classification (HSIC) models have made remarkable progress in the last decade. Nevertheless, the downsized mapping in the convolutional neural network (CNN) and down-sampled mechanism in the transformer-based approach amplify the loss of hidden knowledge in the subpixel that encompasses crucial yet unseen features within a single pixel. Considering this aspect, the mentioned popular solutions for HSIC contradict the inherent characteristic of hyperspectral data. To address this issue, we rethink the size factor in CNN and propose a novel spatial mapping expansion with spectral compression (SMESC) network for HSIC. Specifically, the SMESC builds a mapping expansion network to mine unseen information in subpixels with enlarged feature maps. A channel modulation residual block (CMRB) is developed to compress spectral redundancy and promote salient channels with modulation information. Moreover, we design a multiple-size training strategy to substitute the traditional multiple feature extraction (FE) branches and improve the model adaptation to the different sizes of the testing samples. The extensive experimental results and analysis of four hyperspectral image (HSI) datasets demonstrate the superiority of the proposed architecture compared to other advanced HSIC methods. Our code will be released at <https://github.com/Chirsycy/SMESC>.

**Index Terms**—Channel modulation residual, hyperspectral image classification (HSIC), multiple size training, spatial expansion.

## I. INTRODUCTION

HYPERSPECTRAL image (HSI) [1], [2], [3] provides a detailed spectral signature for each terrain category with hundreds or even thousands of contiguous spectral bands, which is an irreplaceable technology for ground target recognition due to the distinctive characteristic. HSI classification (HSIC) that aims to assign a unique label for each pixel in the HSI is a fundamental task of remote sensing interpretation research [4], [5], [6]. Nowadays, HSIC is a powerful tool

Manuscript received 15 August 2023; revised 15 May 2024; accepted 15 June 2024. Date of publication 27 June 2024; date of current version 9 July 2024. This work was supported by the National Natural Science Foundation of China under Grant 61971082, Grant 42271355, and Grant 42101350. The work of Chunyan Yu was supported by the Science Foundation of Liaoning Province through the Surface Project under Grant LJKZ0065. (*Corresponding author: Yuanchen Zhu*)

The authors are with the Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian 116026, China (e-mail: yucy@dlmu.edu.cn; zhuyuanchen@dlmu.edu.cn; smping@163.com; wangyulei@dlmu.edu.cn; qzhang95@dlmu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3420137

for material identification and classification in various fields, such as precision agriculture, environmental monitoring, and military surveillance.

Typically, HSIC models are composed of a feature extraction (FE) network and a classifier. In specific, the FE is responsible for capturing discriminative features from the original hyperspectral data, while the classifier assigns labels in terms of the designed classification criterion. In the last decade, popular FE models [7], [8], [9], [10], [11], [12] based on convolutional neural networks (CNNs), recurrent neural networks (RNNs), autoencoders (AEs), and deep belief networks (DBNs) have been proposed to extract informative features from HSIs. Out of these models, the CNN-based approaches [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] have gained significant attraction owing to the capability of simultaneously extracting both spatial and spectral features of HSIs. The architecture in [13] implemented a combined architecture with 2-D CNN and 3-D CNN to extract spatial and spectral features for HSIC. Jiang et al. [11] proposed a method that combined 3-D CNNs and extended morphological profiles for feature representation for HSI. Yu et al. [15] extracted HSI features from a CNN and proposed a locality-sensitive hashing technique to cluster the feature vectors for classification. Especially, due to the intrinsic attribute of spectral-spatial integration, spectral and spatial feature fusion is critical in HSIC. Ma et al. [16] compared the performance of different feature fusion strategies in HSIC and mentioned that parallel fusion was effective in few-shot HSIC. The approach in [17] combines spatial and spectral features on a multilevel to generate a new vector for HSIC. A multifeature fusion approach is proposed in [18], which integrates spectral, spatial, and attention features for HSIC.

Besides, graph convolutional networks (GCNs) have shown great potential for HSIC due to their ability to handle the spectral-spatial information of hyperspectral data as a graph structure. Qin et al. [24] utilized a graph model to represent the spatial-spectral information of the hyperspectral data and performed convolutional operations on the graph to extract spectral-spatial features. Yu et al. [25] proposed a novel method that combines an edge-inferring graph neural network with a dynamic task-guided self-diagnosis mechanism for few-shot HSIC. Wan et al. [26] proposed a dynamic GCN for HSIC, which models the spectral-spatial information as

a graph and learns the graph structure dynamically during training.

Currently, the most advanced research in HSIC is centered around multisize feature expression [27], [28], [29], multimodal feature integration [30], [31], [32], [33], [34], and attention mechanisms [35], [36], [37], [38], [39], [40], [41], [42], [43] for FE, resulting in more robust and informative spatial-spectral space construction. In [27], a multisize feature fusion network with the GCN is established to enhance classification accuracy. The joint feature learning network in [30] explores the feature integration of LiDAR and HSI data. Besides, several studies have proposed novel attention networks for HSIC, such as the spatial-spectral attention network (CDSFT) [35] and the self-pooling attention network (SPFormer) [36]. These models incorporate attention mechanisms to selectively focus on informative spectral and spatial features for classification accuracy improvement. Remarkably, transformer-based models, such as the hyperspectral transformer network (HTN) [37] and multibranch attention transformer networks [38], have also been proposed for HSIC. These models employ a multihead attention mechanism to capture long-range dependencies between spectral and spatial features, leading to enhanced performance in classification tasks. Although the self-attention mechanism effectively extracts long-range dependencies and captures global feature information, the large amount of image tokens leads to severe computational complexity. One approach to address this issue is to apply self-attention to images with lower resolutions, which decreases the complexity caused by self-attention with fewer tokens and captures the cross-size similarity patterns.

Overall, the existing CNN-based FE model utilizes convolution operation to obtain the downsize feature, and the transformer-based FE model focuses on the long-range information expression, which usually employs down-sampling to decrease the complexity of the self-attention computation. Obviously, the reduced mapping mechanism in FE does not obey the unique characteristic of HSI data, that is, spatial resolution and subpixel. Based on this foundation, in this article, we present the spatial mapping expansion with spectral compression (SMESC) network for HSIC, which offers a novel approach to explicitly extract the concealed features. Specifically, the SMESC builds a fully transposed convolution network to extract the spatial features that are hidden in the subpixels and decrease the spectral redundancy. To compensate for the limitations that may arise from relying solely on spatial information, we present a channel modulation residual mechanism that focuses on spectral information with different weights. Furthermore, to enhance the adaptability of the HSIC model, we design a serial training strategy that leverages multisize samples to address the issue of inconsistent sizes between training and testing samples.

The main contributions of this article are summarized as follows.

- 1) To break the barrier of decreased mapping in FE against subpixel characteristics, the spatial expansion and spectral compression network that employs transposed convolution for feature assembling is proposed for the first attempt to explicitly activate the spatial

information involved in the subpixel of HSIs. In essence, SMESC provides a novel feature expression schema innovatively that abides by the intrinsic characteristic of HSI and contributes to the spatial expansion architecture for HSIC conceptually and structurally.

- 2) As a valuable addition, we have developed a new channel modulation residual block (CMRB) that enhances spectral information refinement by building a batch-driven modulation way for channel compression. The creative pattern emphasizes the importance of the harmonious balance between spatial and spectral features, which effectively supplements the deficiency of spectral expression in SMESC while enlarging spatial features.
- 3) Instead of the parallel fusion with the multisize FE branches in the training stage, a serial training strategy is devised to prompt the model adaptability to different sizes of testing samples. We integrate multiscale information during the training process to substitute building multiscale architecture for spatial information extraction, which limits the model parameter increase and brings the insensitivity of the different sizes of the testing samples.

The remainder of this article is organized as follows. The motivation and the details of the proposed SMESC architecture are described in Sections II and III, respectively. Section IV provides the experimental results. Analysis and conclusions are drawn in Section V.

## II. MOTIVATION

Natural images are often characterized by high resolution, whereas HSIs have unique features, such as low spatial resolution and ubiquitous subpixels. Consequently, the accuracy of natural image classification (NIC) heavily relies on the effectiveness of high-resolution FE, whereas the accuracy of HSIC is significantly influenced by both subpixel and spectral variability.

Although the CNN-based and transformer-based models have demonstrated impressive results in natural image analysis, these models may not be appropriate for analyzing complex HSI due to inherent subpixel and low spatial resolution characteristics. The reason is that CNN-based models depend on convolution operations to extract features, which results in the loss of critical spatial information due to the gradual decrease in spatial size as the network depth increases. Likewise, the transformer-based approach for HSIC adopts downsampling to decrease the size to guarantee the complexity of the self-attention computation of global context information, which may cause the loss of important subpixel information. Therefore, it is necessary to develop a new FE mechanism to substitute for the standard CNN approach that is adequate for subpixels with feature size expansion instead of the decreased-size embedding maps. Additionally, we also observe that discrepancies in the sizes of training and test samples have a detrimental effect on the performance of the HSIC model, which highlights the significance of the size factor not only in the training phase but also in the test phase. To sum up, the spatial size factor plays a vital role in HSIC, it is crucial to

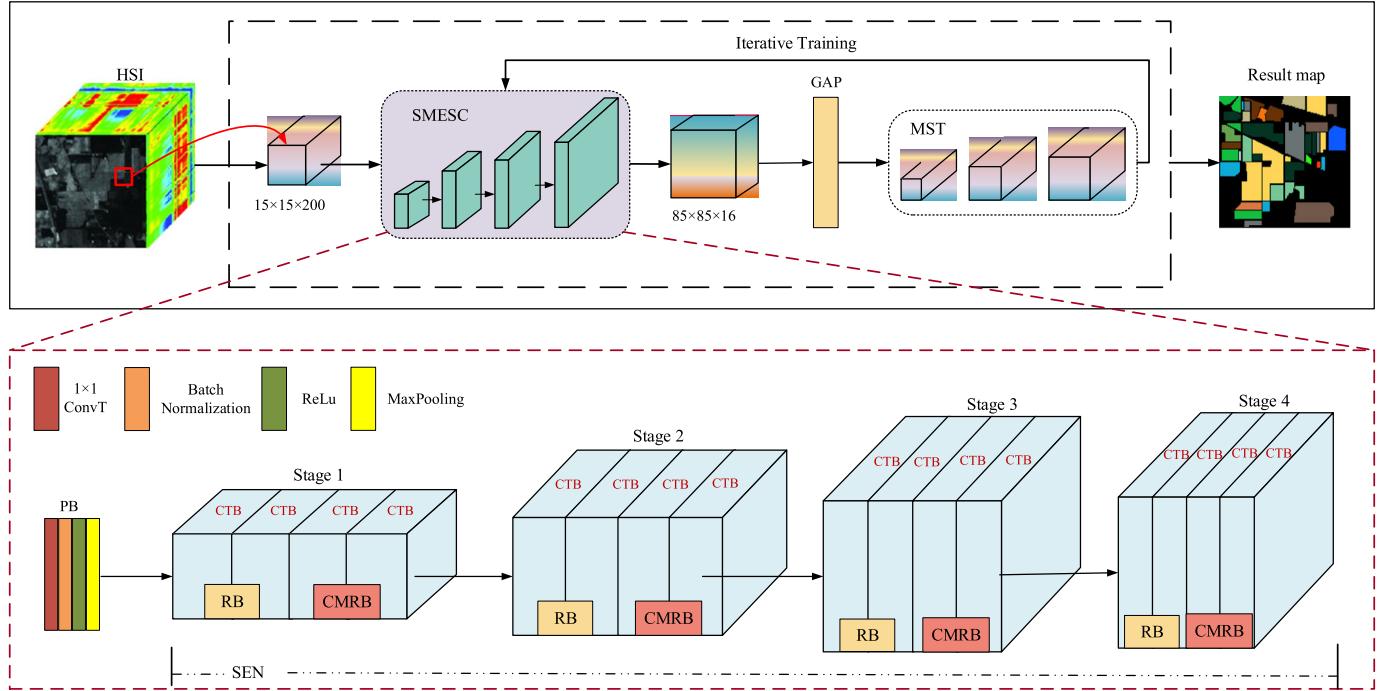


Fig. 1. Flowchart of the proposed SMESC. The framework is mainly composed of the PB, a SEN with spectral compression, and a GAP layer. On the whole, the SEN contains the size preservation for stable learning ability and the size expansion module for the refinement of the hidden information. Importantly, transposed convolutional block (CTB) is the fundamental ingredient of each stage in SEN, RB is responsible for feature skip, and CMRB aims to decrease spectral redundancy. Finally, the model employs GAP instead of the fully connected layer for the final feature embedding. Notably, during the training phase, the MST strategy is designed to implement the model optimization iteratively.

rethink the size factor and design the HSIC model customized with the inherent data characteristics, rather than solely relying on the NIC models. Besides, the preservation of spatial size, the adoption of skip connections, the careful balance between spatial and spectral features, and the training strategy are also vital considerations in HSIC network design.

### III. PROPOSED APPROACH

Fig. 1 illustrates the overall flowchart of the proposed SMESC, which consists of three main blocks: the pre-processing block (PB), the backbone, and the classifier. Initially, the input samples are processed by the PB block to extract preliminary features. Afterward, the obtained features are fed into the backbone of the SMESC network for feature expansion and refinement, which is accomplished by amplifying spatial information and decreasing spectral redundancies. The subsequent global average pooling (GAP) layer replaces the original fully connected layers to generate the category embedding. Finally, the cross-entropy loss is employed as the final optimization function in the multiple-size training phase.

Technically, the proposed SMESC implements hidden feature mining with size preservation and expansion network. Specifically, the size preservation module (SPM) maintains the stable learning ability, and the embedding enlargement is implemented in the expansion network. The CMRB is presented as a channel attention module to decrease spectral redundancy. Moreover, the multisize training strategy is put forward to enhance the size insensitivity and adaptivity of the HSIC model. Further details of each component are described in Sections III-A–III-F.

#### A. Transposed Convolution

Instead of adopting traditional convolution operations (Conv), the proposed method employs transposed convolution (ConvT) that is designed to restore low-dimensional maps to high-dimensional space for feature expansion.

To illustrate the difference between Conv and ConvT, we provide the basic calculations for the two operations. With the stride of 1 and padding of 0, (1) and (2) define the Conv and ConvT calculations, respectively,

$$C(X, K_{n,q,s,d}) = \left| \sum_{i=1}^p x_{ii} * k_{ii} \right|_{(p-q+1, p-q+1)} \quad (1)$$

where

$$X = \begin{vmatrix} x_{11} & \cdots & x_{1p} \\ & x_{ii} & \\ x_{p1} & \cdots & x_{pp} \end{vmatrix}_{(p,p)}$$

is the fed map with the size of  $p \times p$ , and

$$K = \begin{vmatrix} k_{11} & \cdots & k_{1q} \\ & k_{ii} & \\ k_{q1} & \cdots & k_{qq} \end{vmatrix}_{(q,q)}$$

is the convolutional kernel with the size of  $q \times q$ ,  $n$  denotes the kernel number,  $s$  means the striding step, and  $d$  is the

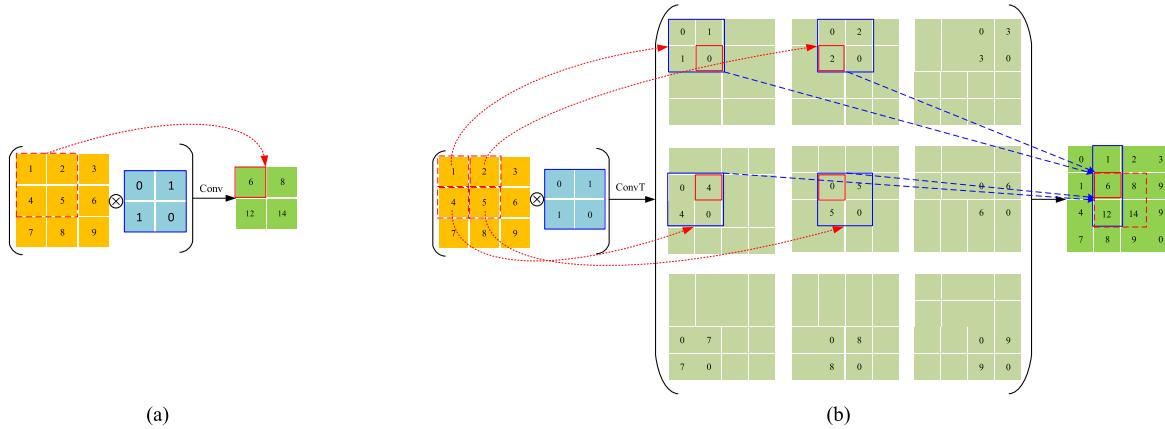


Fig. 2. Illustration of (a) Conv operation and (b) ConvT operation.

padding number

$$\text{CT}(X, K_{n,q,s,d}) = \prod_{i=1}^{p*2-1} \begin{vmatrix} 0, & 0 & \dots & 0, & 0 \\ 0, & x_{ii} * k_{11} & \dots & x_{ii} * k_{1q}, & 0 \\ 0, & x_{ii} * k_{q1} & \dots & x_{ii} * k_{qq}, & 0 \\ 0, & 0 & \dots & 0, & 0 \end{vmatrix}_{(p+q-1, p+q-1)} \quad (2)$$

where  $\prod$  means matrix addition operation.

Vividly, the illustrations of the Conv and ConvT operations are demonstrated in Fig. 2. As can be observed, the size of the fed input feature  $X$  with  $p = 3$ , and  $K$  with ( $n = 1, q = 3, s = 1, d = 0$ ) applying the Conv operation results in a generated map that is reduced to  $2 \times 2$  as shown in Fig. 2(a). In contrast, assuming the same input of  $X$  and  $K$ , the ConvT operation expands the obtained map to the size of  $4 \times 4$ . Besides, the red region of the right side in Fig. 2(b) is the same feature as the right side in Fig. 2(a), which means that the ConvT operation acquires more features than the Conv operation.

Based on the mentioned truth that the ConvT operation is capable of FE with an enlarged map, we believe that the ConvT operation has the potential ability to mine the hidden information of the subpixel of HSI compared to the Conv operation. In the proposed SMESC model, we construct the FE for HSIC with size preservation and size expansion module, which are built upon the ConvT operation.

### B. PB Module

In the initial part of SMESC, PB is composed of a 2-D ConvT layer, batch normalization (BN) layer, ReLu, and a max-pooling layer. Assuming that  $x$  is the input sample,  $x_g$  is the output of the PB module. The input feature is polished by the following equation:

$$x_f = \text{CT}(x, K_{n,q,s,d}) \quad (3)$$

$$x_g = \max(\text{MP}(\text{BN}(x_f)), 0) \quad (4)$$

where  $\text{BN}(\cdot)$  denotes the BN operation and  $\text{MP}(\cdot)$  represents the max-pooling operation with a kernel size of  $3 \times 3$ , stride of 2, and padding of 1. The CT operation in (3) is implemented with a kernel setting of  $K$  ( $n = 100, q = 1, s = 1, d = 0$ ).

### C. Backbone of the SMESC

1) *Size Preservation Module*: On the whole, the backbone network with the input of  $x_g$  is divided into four cascaded stages, each with the SPM, which is the crucial component of the proposed SMESC.

Within each SPM, we adopt the identical mapping mechanism to avoid changing the size of the feature map and reduce the difficulty of model learning. Conveniently, we record the block with a ConvT layer, a BN layer, and a ReLu function as a ConvT block (CTB), which is denoted as the  $\text{CTB}(\cdot)$  operation integrally. As shown in Fig. 3(a), the SPM block is composed of four CTB, where the size of the map and the number of channels is fixed. Specifically, the feature in the  $i$ th CTB is extracted with the following formula:

$$x_i = \text{CTB}_i(x_{i-1}) \quad (5)$$

where  $x_i$  is the output of the  $i$ th CTB,  $i \in \{1, 2, 3, 4\}$ , especially,  $x_0$  denotes the input of the CTB1. In this equation, the CT operation in CTB is implemented with  $q = 3, d = 1$ , while  $s$  and  $n$  are varied in different SPM, the specific settings of the hyperparameters are listed below.

Besides, we enhance feature skipping by incorporating a residual block (RB) between the first and third CTB within the SPM module. The extracted feature is obtained using the following equation:

$$x'_i = x_i \oplus \text{RB}(x_{i-2}) \quad (6)$$

$$\text{RB}(x) = \text{BN}(\text{CT}(x, K_{n,q,s,d})) \quad (7)$$

where  $x_i$  represents the output of the  $i$ th CTB when  $i$  equals 3. In (7), the CT operation in CTB is implemented with a setting of ( $q = 3, s = 1$ , and  $d = 1$ ), and  $\oplus$  denotes the feature addition.

2) *Size Expansion Network (SEN)*: The SPM blocks are stacked to form a cascaded structure to build the SEN as shown in Fig. 3(b), and the feature maps generated by SPM are expanded between the stages to gradually refine the details of the hidden feature.

Specifically, in the  $j$ th stage, the feature expansion is accomplished by the  $\text{CT}(\cdot)$  operation by modifying the parameters of  $n$  and  $s$ . In this way, the SEM expands the receptive field of the subpixels and increases the size of the feature

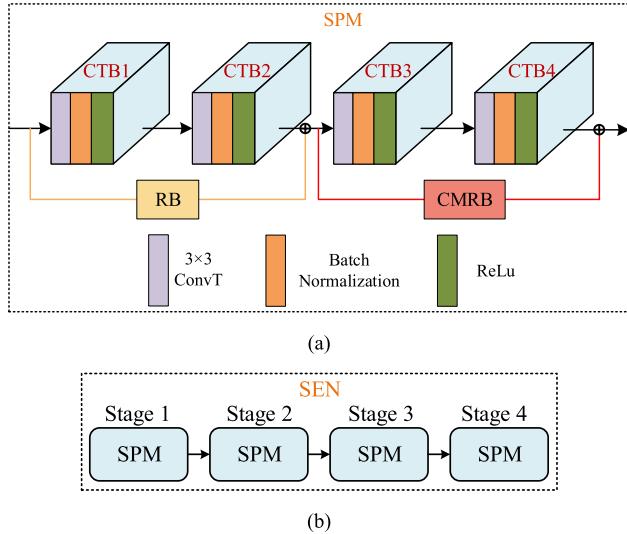


Fig. 3. Framework of SPM and SEN. The architecture of (a) SPM and (b) SEN.

TABLE I  
PARAMETER SETTING AND OUTPUT SIZE IN THE SEN

Block		Size of output ( $c, h, w$ )	$K(n, q, s, d)$
Stage 1	CTB 1	(100, 15, 15)	(100, 3, 1, 1)
	CTB 2	(100, 15, 15)	(100, 3, 1, 1)
	CTB 3	(100, 15, 15)	(100, 3, 1, 1)
	CTB 4	(100, 15, 15)	(100, 3, 1, 1)
Stage 2	CTB 1	(50, 43, 43)	(50, 3, 3, 1)
	CTB 2	(50, 43, 43)	(50, 3, 3, 1)
	CTB 3	(50, 43, 43)	(50, 3, 3, 1)
	CTB 4	(50, 43, 43)	(50, 3, 3, 1)
Stage 3	CTB 1	(25, 85, 85)	(25, 3, 2, 1)
	CTB 2	(25, 85, 85)	(25, 3, 2, 1)
	CTB 3	(25, 85, 85)	(25, 3, 2, 1)
	CTB 4	(25, 85, 85)	(25, 3, 2, 1)
Stage 4	CTB 1	(16, 85, 85)	(16, 3, 1, 1)
	CTB 2	(16, 85, 85)	(16, 3, 1, 1)
	CTB 3	(16, 85, 85)	(16, 3, 1, 1)
	CTB 4	(16, 85, 85)	(16, 3, 1, 1)

maps, enabling the model to learn more abstract and rich information from HSIs. Considering the HSI dataset with a sample size of  $15 \times 15$  and 16 distinct classes, the hyperparameter setting and sizes of the specific layers of SEN are listed in Table I.

Additionally, while expanding the spatial mapping to extract the latent information between SPMs, we incorporate the CMRB to capture channel information and optimize channel reduction. The specific CMRB is described in Section III-D, and the feature fusion with CMRB is defined as follows:

$$x_o = \max(\text{BN}(\text{CT}(x_4)) \oplus \text{CM}(x_2), 0) \quad (8)$$

where  $\text{CM}(\cdot)$  denotes the processing of the CMRB, and  $x_o$  is the output of the SPM.

Afterward, the extracted features are fed into the GAP layer to obtain an embedded representation of the corresponding class. Notably, in our model, the GAP layer replaces the original fully connected layers to generate the embedding with the channel number of the categories that need to be classified. Thus, the last GAP layer serves as the classifier for

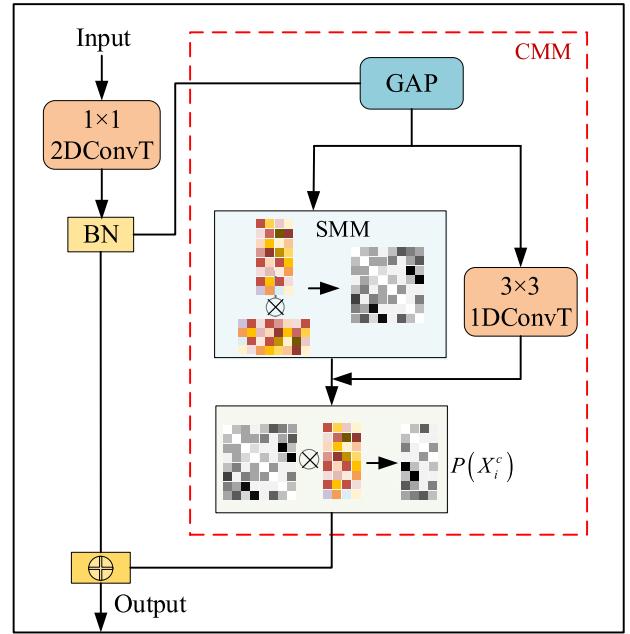


Fig. 4. Illustration of CMRB.

the proposed model. In the training phase, the cross-entropy loss  $J$  is employed for model optimization.

#### D. Spectral Compression by Channel Modulation

To mine the effective channels of the proposed model, we present CMRB to reduce spectral redundancy and refine the feature with channel saliency. Remarkably, the CMRB identifies the jointly important channels across different data of one batch to enhance the model learning ability by the specific channels. As shown in Fig. 4, the key part of the CMRB is the channel modulation module (CMM), which mainly consists of the GAP layer, and the saliency modulation module (SMM).

First, assume the feature  $X_i$  is the output of the last CTB block, which represents the input feature matrix of a mini-batch with the dimension of  $(b, c, h, w)$ ,  $b$  represents the batch number,  $c$  is the number of channels,  $h$  means the height, and  $w$  is the width of the matrix, respectively. Afterward, GAP is employed to leverage the global-level information to generate the matrix  $X_i^c$  with the dimension of  $(b, c)$ , which retains only the channel dimension for the following calculation.

Next, the SMM block is responsible for calculating channel saliency by measuring the similarities between different data within a batch dynamically, which is implemented by the following equation:

$$\text{SMM}(X_i^c) = \frac{\text{Softmax}\left(X_i^c \otimes (X_i^c)^T\right)}{\text{DIM}(X_i^c)} \quad (9)$$

where  $\otimes$  is the matrix multiplication,  $\text{DIM}(\cdot)$  is the second-order norm calculation, and  $\text{softmax}(\cdot)$  is the activation function, which is intended to prevent numerical overflow in the computation.

Subsequently, the channel saliency matrix denoted  $P(X_i^c)$  is calculated with the following formulas:

$$X_i^* = \text{CT}^\Theta(\text{EPD}(X_i^c)) \quad (10)$$

$$P(X_i^c) = \frac{\text{SMM}(X_i^c) \otimes X_i^*}{\text{DIM}(X_i^*)} \quad (11)$$

where  $\text{CT}^\Theta(\cdot)$  represents the 1-D CT operation,  $\text{EPD}(\cdot)$  denotes dimension expansion transformation, which is necessary for the dimension transformation of  $X_i^c$  from  $(b, c)$  to  $(b, 1, c)$ .

The values in  $P(X_i^c)$  reflect the modulated channel feature constrained by all the samples in one batch, which focuses on the characteristics of the channel and the shared information across different samples. Lastly, the feature fused with the channel modulation is defined as the following equation:

$$\text{CM}(X_i) = X_i + P(X_i^c) \odot X_i \quad (12)$$

where  $\text{CM}(\cdot)$  denotes the process of CMRB, and  $\odot$  means the dot product operation.

Unlike the implementation with Softmax to activate the weight of each channel, CMRB is specified with wide adaptability by the involvement of all the data in one batch.

#### E. Multisize Training Strategy

In this article, we present a multiple-size training (MST) strategy for the HSIC model, which is the first attempt to transform the multisize feature processing from the model construction to the training phase. The MST is a simple yet powerful approach that significantly enhances classification performance. Besides, SST is beneficial to the adaptability of handling varying sizes of test samples, which generates stable performance even in situations where the size of the test sample is different from the training samples.

Assuming that the training set is denoted as  $\{S_1, S_2, \dots, S_N\}$ , the validation set is denoted as  $\{T_1, T_2, \dots, T_N\}$ , where  $S_k = \{(x_i^{l_t}, y_i)\}$ ,  $T_k = \{(x_i^{l_t}, y_i)\}$ ,  $i \in \{1, \dots, n\}$ ,  $x_i^{l_t}$  is the  $i$ th sample with the patch size of  $l_t \times l_t$ ,  $y_i$  is the label, and  $n$  is the sample number.  $t \in \{1, \dots, m\}$ ,  $m$  is the number of the size sizes. As shown in Fig. 5, the training order should be determined first, then fed the corresponding samples  $x_i^{l_t}$  into SMESC, and the Adam is adopted to optimize the model subsequently. After training with the  $l_t \times l_t$ , we evaluate the model on a validation set, and the training followed up with  $l_{t+1} \times l_{t+1}$  is finetuned with the best performance on the  $l_t \times l_t$ . Notably, the training order in terms of  $l_t$  is flexible, that is, the sort of  $\{l_t\}$  can be in ascending or descending order, even out of order. In the experimental section, we conduct a series of experiments to evaluate the performance of different orders in MST. Besides, it is worth emphasizing that MST is not exclusively restricted to the proposed HSIC network and exhibits effectiveness when applied to other networks as well.

**Advantage:** MST adopts the sequential training order, which means the samples with different sizes are fed into the model successively. The MST policy also benefits from an iteration property that allows the model to learn and integrate information about multisize features, ultimately leading to a

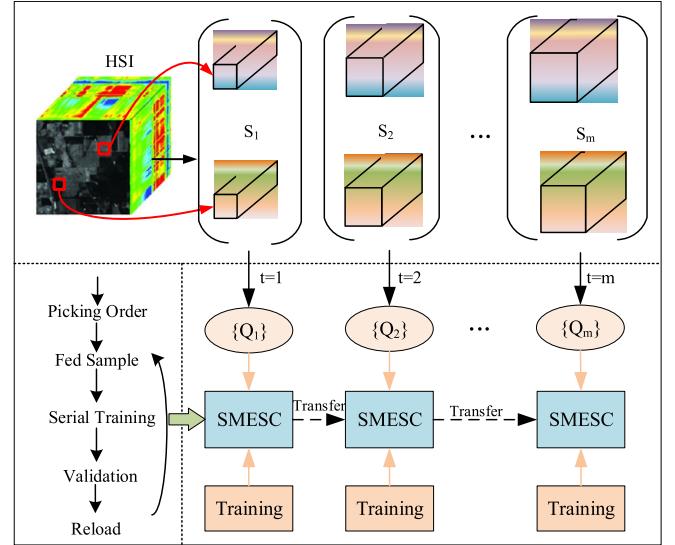


Fig. 5. Illustration of MST procedure.

more robust feature space. This iterative approach enables the model to build a more complete and informative representation of the specific category, especially in the few-shot situation. Particularly, the simple but powerful training strategy plays a crucial role in ensuring the size insensitivity of the model, which enables the model to handle testing samples with different sizes and achieve robust performance.

#### F. Algorithm of the SMESC

Overall, the algorithm of the proposed SMESC with the parameters of  $\theta$  trained with MST is listed below.

---

##### Algorithm 1 Training SMESC With MST

---

```

Input:  $\{S_1, S_2, \dots, S_N\}$ , iteration number  $N$ ,  $\theta$ 
Output:  $\theta$ 
Initialize  $\theta$  with random Gaussian values
For  $i = 1$  to  $N$  do
     $\{Q_i\} \leftarrow$  Get training set  $S_i$  from  $\{S_1, S_2, \dots, S_N\}$ 
     $\{T_i\} \leftarrow$  Get validation set  $T_i$  from  $\{T_1, T_2, \dots, T_N\}$ 
    For  $e = 1$  to Epoch do
        For  $b = 1$  to Batch do
            Randomly generate a mini-batch sample
            Feed samples into SMESC and get the prediction
             $\mathcal{J} \leftarrow$  Calculate the prediction loss
             $\theta = \nabla_\theta(\mathcal{J})$  Update  $\theta$  via Adam optimizer
        End
        Evaluate on  $T_i$  and Save the best model
    End
    Reload the model
End

```

---

## IV. EXPERIMENT AND ANALYSIS

### A. Data Description

We evaluate the performance of our proposed approach on four widely used HSI datasets.

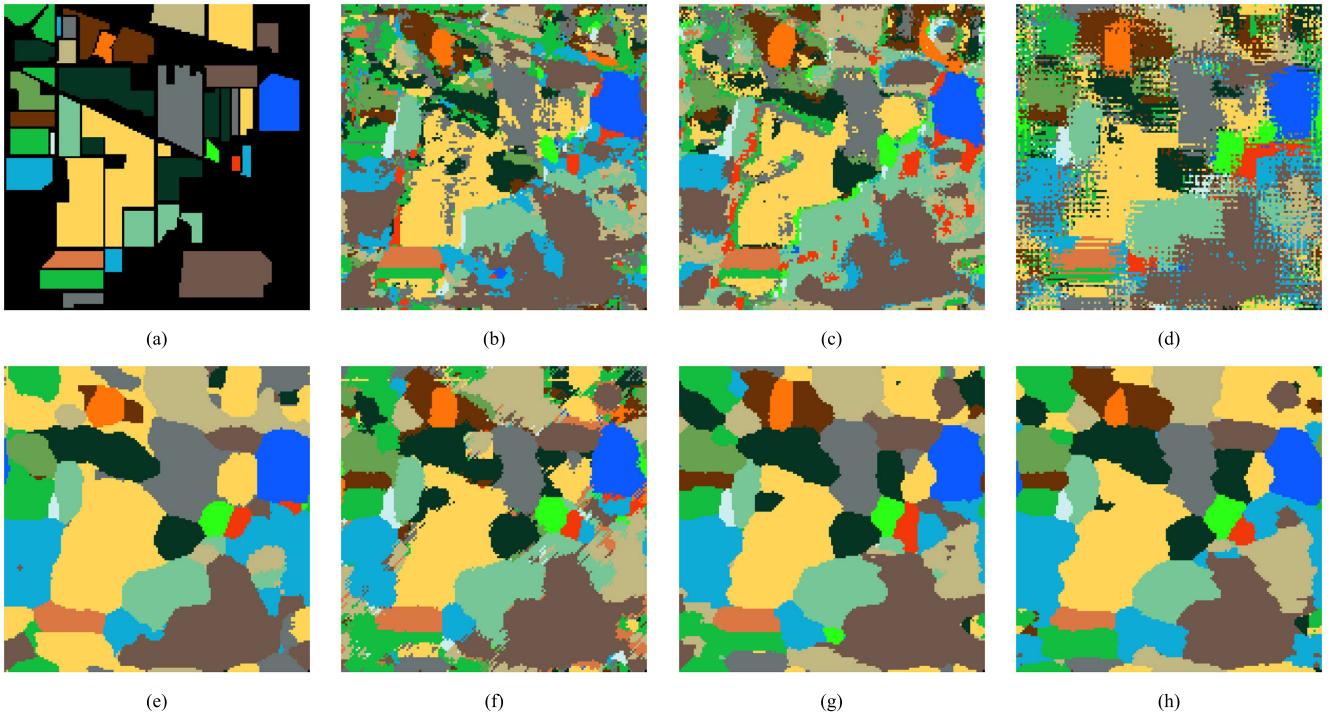


Fig. 6. Classification maps for the IP dataset. (a) Ground-truth map. (b) RSSAN. (c) pResNet. (d) SSMTR. (e) SSTN. (f) SSFTT. (g) SMESC-S. (h) SMESC.

- 1) *Indian Pines (IP)*: The IP dataset was acquired by the AVIRIS sensor over the IP test site in northwestern Indiana in 1992. The imagery captures 200 spectral bands across  $145 \times 145$  pixels, with a spatial resolution of 20 m and 16 distinct classes after removing water absorption bands.
- 2) *Houston*: The Houston dataset was collected with the ITRES CASI-1500 sensor. This scene covers the University of Houston campus with a resolution of  $349 \times 1905$ , 144 spectral bands, and 15 classes of ground objects.
- 3) *KSC*: The KSC dataset was captured using the AVIRIS sensor at the Kennedy Space Center in Florida on March 23, 1996. The imagery contains 176 spectral bands after removing water noise, with a spatial resolution of 18 m and 13 categories.
- 4) *Botswana*: The Botswana dataset was captured by the HYPERION sensor on board the EO-1 satellite of the Okavango Delta between 2001 and 2004, which covers 145 spectral bands to distinguish 14 types of ground objects. The resulting scene contains  $256 \times 1476$  pixels.

### B. Experimental Configuration

To verify the effectiveness of the proposed SMESC, the methods including RSSAN [44], pResNst [45], SSMTR [46], SSTN [47], and SSFTT [48] are chosen as the comparison methods, which are replicated by the codes supplied by authors with the best parameter settings. Besides, we record the approach of SMESC without MST as SMESC-S in the comparison.

- 1) *CNN-Based Methods*: RSSAN and P-ResNst.
- 2) *Transformer-Based Methods*: SSMTR, SSTN, and SSFTT.

In the experimental part, we randomly select 20 samples for each class as the training set, 5% of the samples as the validation set, and utilize all samples as the testing set for all the datasets. Especially, we use five samples as the validation set for the Botswana data, and the sample size of all four datasets is  $15 \times 15$ . For SMESC-S and SMESC, we adopt the Adam optimizer and set the batch size to 16, the learning rate to 0.0001, and the epoch to 300. The SMESC model is trained with the MST strategy with the order of {3, 5, 7, 9, 11, 13, 15} as training sequences, and the rest of the settings are the same as SMESC-S.

In the comparison and analysis, overall accuracy (OA), average accuracy (AA), and Kappa coefficient are the main criteria adopted to evaluate classification performance in the following experiments. To ensure a fair comparison, all the experiments are conducted by repeating five times, and the best performances are recorded in the following tables.

### C. Comparison and Analysis

In this section, we present a comparison of our method with other approaches, the classification results are reported in Tables III–VI and Figs. 6–9. Based on the results, we have made the following analyses.

1) *Compared to the CNN-Based Models*: As observed from the tables, RSSAN results in a relatively limited performance improvement, which incorporates a spectral and spatial fusion module, and does not effectively capture spatial information. The performance on the IP dataset is limited to 70.65% due to the low spatial resolution and severe sample imbalance. While the pResNet approach leverages spatial contextual information fusion to enhance spatial information. Since pResNet fails to maintain a sufficient spatial feature size, leading to limited spatial information capture and poor performance,

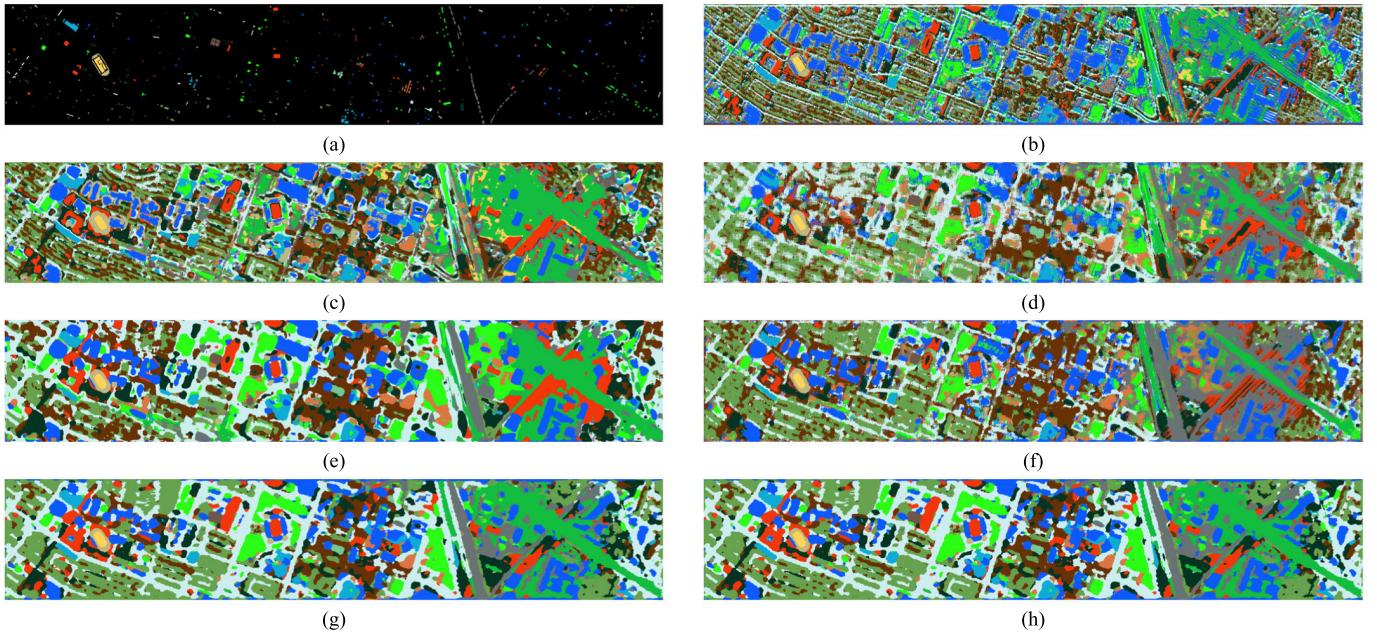


Fig. 7. Classification maps for the Houston dataset. (a) Ground-truth map. (b) RSSAN. (c) pResNet. (d) SSMTR. (e) SSTN. (f) SSFTT. (g) SMESC-S. (h) SMESC.

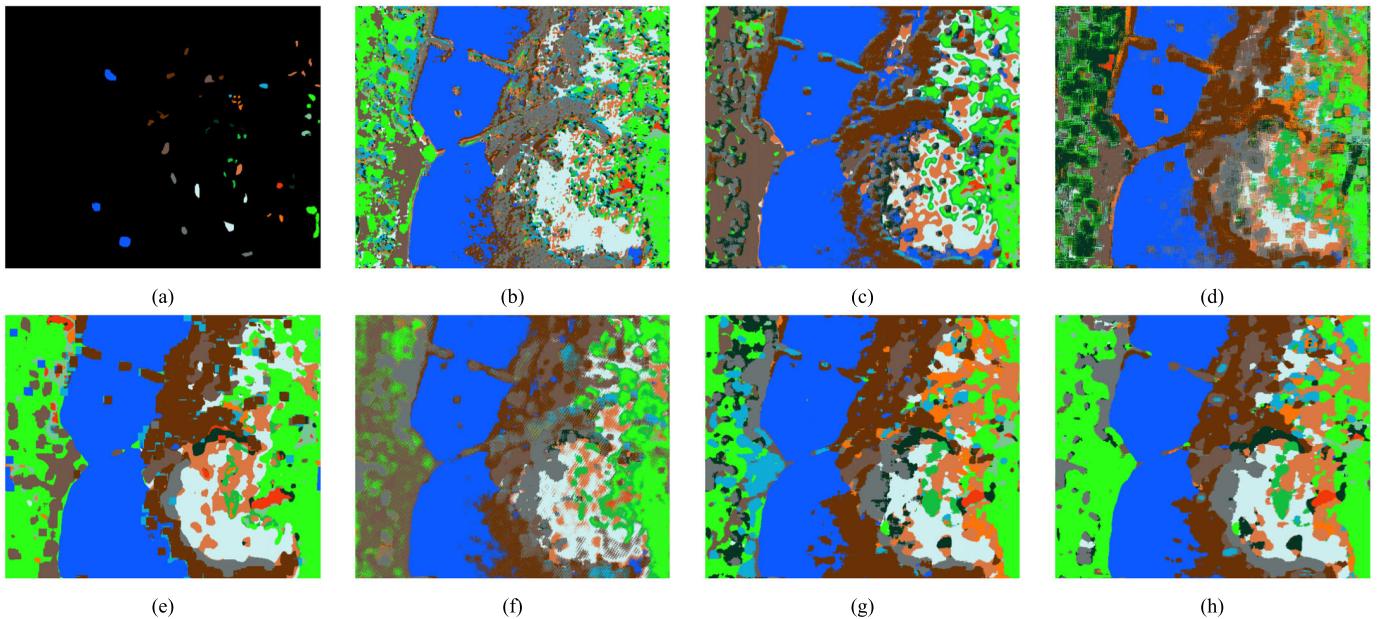


Fig. 8. Classification maps for the KSC dataset. (a) Ground-truth map. (b) RSSAN. (c) pResNet. (d) SSMTR. (e) SSTN. (f) SSFTT. (g) SMESC-S. (h) SMESC.

TABLE II  
ABLATION STUDY ON THE FOUR DATASETS

SEN	✓	✓	✓	✓
CMRB		✓		✓
MST			✓	✓
Indian Pines	92.53±0.45	94.41±0.65	94.65±0.70	97.31±0.46
Houston	93.60±0.29	94.30±1.06	96.15±0.48	97.51±0.31
Botswana	94.34±0.79	95.20±1.16	96.80±0.80	98.17±0.75
KSC	99.38±0.25	99.63±0.17	99.75±0.19	99.93±0.04

particularly on the Indian dataset. In contrast, our proposed method maintains a sufficient spatial size, enabling the model to capture more comprehensive spatial information and achieve superior classification results.

2) *Compared to the Transformer-Based Models:* Tables IV–VII illustrate that the transformer-based model including the SSMTR, SSTN, and SSFTT frameworks yields less competitive performance than our method. In specific, among the three models, SSFTT achieves the highest OA, which are 89.06%, 91.31%, and 93.75% for the IP data, Houston, and Botswana, respectively. For the KSC data, the SSTN model is the most effective and achieves an OA of 91.98%. While SSMTR has attempted to capture advanced semantic information by utilizing reconstructed image elements, the extracted features with limited samples may not be sufficient for the category representation, ultimately leading to reduced classification accuracy. In contrast, our approach has demonstrated

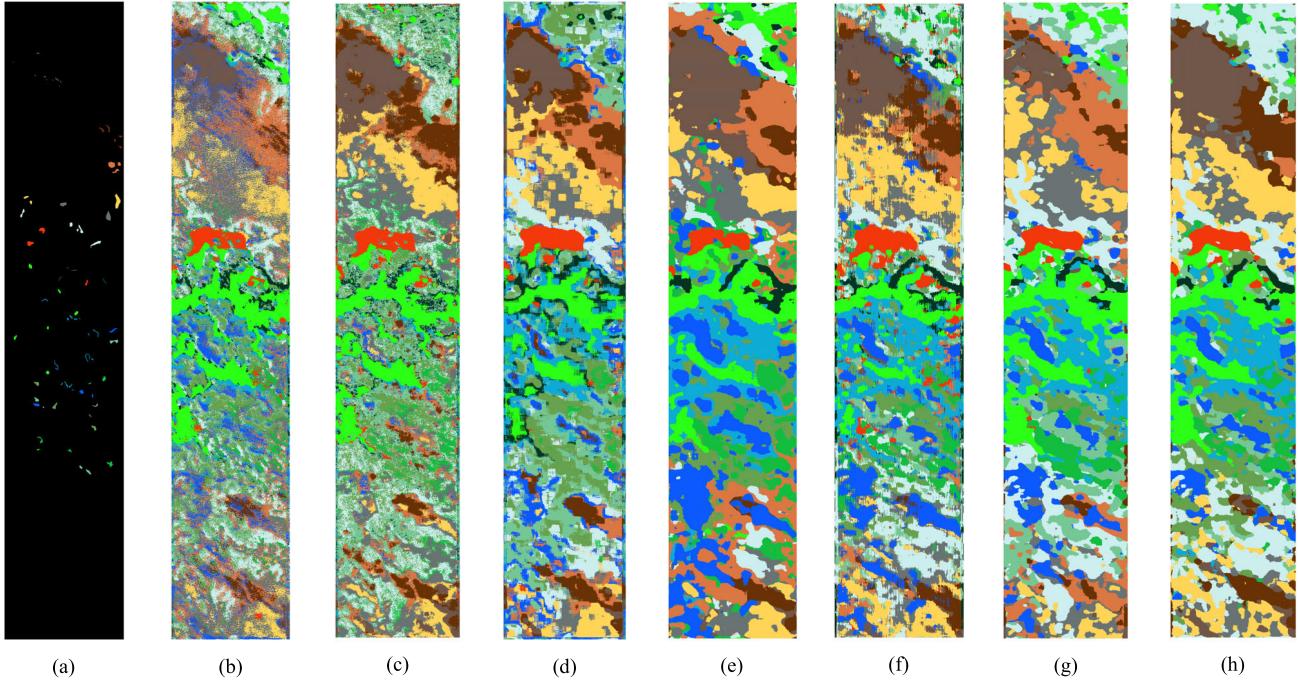


Fig. 9. Classification maps for the Botswana dataset. (a) Ground-truth map. (b) RSSAN. (c) pResNet. (d) SSMTR. (e) SSTN. (f) SSFTT. (g) SMESC-S. (h) SMESC.

TABLE III  
CLASSIFICATION RESULTS (%) WITH COMPARED APPROACHES ON THE IP DATASET

	Color	Class	RSSAN	pResNet	SSMTR	SSTN	SSFTT	SMESC-S	SMESC
1		Alfalfa	76.98	<b>100</b>	90.81	<b>100</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
2		Corn-notill	81.34	65.2	84.69	65.55	83.33	93.70	<b>97.69</b>
3		Corn-mintill	87.8	79.88	92.4	80.72	95.54	<b>97.47</b>	96.63
4		Corn	77.25	96.2	83.44	97.89	94.09	<b>100.0</b>	<b>100.0</b>
5		Grass-pasture	99.52	91.93	93.24	88.2	95.03	<b>98.76</b>	97.10
6		Grass-tress	83.38	98.36	85.23	96.58	94.24	<b>98.90</b>	96.85
7		Grass-pasture-mowed	52.52	<b>100</b>	47.4	100	100.0	<b>100.0</b>	71.43
8		Hay-windowed	53.94	<b>100</b>	57.48	98.12	95.39	<b>100.0</b>	<b>100.0</b>
9		Oats	55.91	<b>100</b>	67.65	<b>100</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b> ,
10		Soybean-notill	76.37	69.96	64.14	73.87	69.96	89.61	<b>96.50</b>
11		Soybean-mintill	57.89	77.68	49.55	79.51	86.52	96.99	<b>98.62</b>
12		Soybean-clean	54.99	85.5	69.75	80.27	84.99	94.10	<b>97.81</b>
13		Wheat	36.25	99.51	42.64	99.51	98.05	<b>100.0</b>	<b>100.0</b>
14		Woods	96.96	81.66	95.56	94.94	97.94	99.60	<b>99.76</b>
15		Buildings	91.36	100	97.58	86.27	<b>99.74</b>	99.22	98.19
16		Stone	76.98	98.02	90.81	<b>100</b>	<b>100.0</b>	97.67	<b>100.0</b>
OA	/	/	70.65	81.62	73.13	83.02	89.06	96.64	<b>98.07</b>
AA	/	/	72.16	90.3	74.77	90.09	93.42	<b>97.95</b>	96.84
Kappa	/	/	68.27	79.18	70.97	80.73	87.56	96.18	<b>97.80</b>

the best performance across all four datasets. After analyzing the results, we have identified several factors that may have contributed to this phenomenon. First, we did not utilize PCA dimensionality reduction or normalization operations in all the HSIC models, which could have increased the classification difficulty. Additionally, while SSMTR and SSTN use larger sample sizes of  $27 \times 27$  and  $9 \times 9$ , respectively, we opted for a smaller sample size of  $15 \times 15$ , which may have led to performance degradation. Furthermore, the relatively few samples may cause the limited performance of transformer-based models, which typically require a larger number of samples for optimal results. Notably, the long-tailed distribution problem in the Indian dataset represents another key factor that impedes the effectiveness of the three frameworks.

3) *Comparison of Different Datasets:* Our proposed approaches including SMESC-S and SMESC, demonstrate superior performance across all datasets. In particular, SMESC proves to be more effective than SMESC-S due to the utilization of the MST strategy. Among the compared methods, the SSFTT method outperforms others for the IP and Houston datasets. Conversely, for the Botswana and KSC datasets, the pResNet model achieves higher accuracy compared to other models, with accuracies of 98.37% and 94.32%, respectively. In addition, since different land cover classes in the KSC dataset typically exhibit significant differences in spectral features and fewer testing samples, our methods yield better performance on KSC than other datasets.

TABLE IV  
CLASSIFICATION RESULTS (%) WITH COMPARED APPROACHES ON THE HOUSTON DATASET

	Color	Class	RSSAN	pResNet	SSMTR	SSTN	SSFTT	SMESC-S	SMESC
1		Healthy	96.48	95.84	72.66	98.56	87.05	98.64	<b>99.76</b>
2		Stressed	81.26	97.13	91.87	78.79	98.48	97.45	<b>99.12</b>
3		Synthetic	95.41	96.56	97.85	92.40	97.13	99.71	<b>100.0</b>
4		Trees	87.22	93.89	84.73	87.14	93.89	<b>99.36</b>	94.69
5		Soil	99.11	99.76	97.67	97.91	94.85	<b>100.0</b>	<b>100.0</b>
6		Water	85.23	92.92	96.31	94.77	84.62	<b>100.0</b>	<b>100.0</b>
7		Residential	89.59	85.65	80.13	85.49	86.99	93.85	<b>95.90</b>
8		Commercial	55.06	82.56	77.25	79.50	86.90	93.57	<b>96.46</b>
9		Road	68.45	82.91	75.48	67.73	81.95	94.81	<b>95.85</b>
10		Highway	69.76	77.75	90.87	<b>100.0</b>	98.61	<b>100.0</b>	<b>100.0</b>
11		Railway	61.46	84.05	85.34	71.26	93.68	99.27	<b>99.84</b>
12		Parking Lot	79.48	87.51	77.94	81.27	84.10	90.59	<b>98.05</b>
13		Parking Lot	69.94	94.46	90.41	92.96	85.29	<b>96.16</b>	95.74
14		Tennis	89.25	<b>100.0</b>	89.49	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
15		Running	92.73	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
OA	/	/	80.37	90.17	85.45	86.68	91.31	97.17	<b>98.18</b>
AA	/	/	81.36	91.4	87.20	88.52	91.57	97.56	<b>98.36</b>
Kappa	/	/	78.78	89.38	84.30	85.61	90.61	96.94	<b>98.03</b>

TABLE V  
CLASSIFICATION RESULTS (%) WITH COMPARED APPROACHES ON THE BOTSWANA DATASET

	Color	Class	RSSAN	pResNet	SSMTR	SSTN	SSFTT	SMESC-S	SMESC
1		Water	98.52	<b>100.0</b>	82.96	97.78	99.63	<b>100.0</b>	99.63
2		Hippo grass	91.09	<b>100.0</b>	96.04	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
3		Floodplain	75.70	<b>100.0</b>	93.63	<b>100.0</b>	93.63	99.60	<b>100.0</b>
4		Floodplain	98.14	<b>100.0</b>	97.67	<b>100.0</b>	99.53	<b>100.0</b>	<b>100.0</b>
5		Reeds	53.16	92.19	63.94	65.06	75.84	94.05	<b>96.28</b>
6		Riparian	86.99	92.19	95.91	43.12	91.82	99.63	95.91
7		Firescar	<b>100.0</b>	<b>100.0</b>	96.91	<b>100.0</b>	<b>100.0</b>	99.61	<b>100.0</b>
8		Island	92.61	<b>100.0</b>	58.13	58.13	96.55	88.67	<b>100.0</b>
9		Acacia	90.13	98.41	<b>100.0</b>	<b>100.0</b>	95.22	<b>100.0</b>	<b>100.0</b>
10		Acacia	<b>100.0</b>	<b>100.0</b>	97.98	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	98.79
11		Acacia	90.16	99.02	78.36	76.39	76.39	99.67	<b>100.0</b>
12		Short	91.16	<b>100.0</b>	99.45	<b>100.0</b>	97.79	97.79	<b>100.0</b>
13		Mixed	90.30	98.88	93.66	86.19	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
14		Exposed	87.37	<b>100.0</b>	88.42	83.16	<b>100.0</b>	<b>100.0</b>	96.84
OA	/	/	88.64	98.37	88.55	85.75	93.75	98.55	<b>99.14</b>
AA	/	/	88.95	98.62	88.79	86.42	94.74	98.50	<b>99.10</b>
Kappa	/	/	87.70	98.23	87.60	84.56	93.23	98.43	<b>99.07</b>

TABLE VI  
CLASSIFICATION RESULTS (%) WITH COMPARED APPROACHES ON THE KSC DATASET

	Color	Class	RSSAN	pResNet	SSMTR	SSTN	SSFTT	SMESC-S	SMESC
1		Scrub	99.21	97.90	90.93	99.21	97.77	98.29	<b>100.0</b>
2		Willow swamp	72.84	93.83	85.60	92.59	20.58	<b>94.29</b>	94.24
3		CP hammock	67.97	79.30	90.23	<b>100.0</b>	82.42	<b>100.0</b>	<b>100.0</b>
4		Slash pine	52.38	91.27	75.40	0.0	43.65	<b>97.62</b>	94.44
5		Oak/Broadleaf	90.06	96.89	78.26	89.44	59.63	83.85	<b>100.0</b>
6		Hardwood	81.22	85.59	79.48	95.2	25.33	<b>99.13</b>	<b>99.13</b>
7		Swamp	99.05	98.10	97.14	97.14	94.29	<b>100.0</b>	<b>100.0</b>
8		Graminoid	79.12	90.26	54.52	95.59	54.99	<b>100.0</b>	<b>100.0</b>
9		Spartina marsh	90.77	98.46	92.69	91.15	85.19	<b>100.0</b>	<b>100.0</b>
10		Cattail marsh	44.55	91.58	92.33	94.80	94.06	<b>100.0</b>	<b>100.0</b>
11		Salt marsh	99.76	99.52	97.85	<b>100.0</b>	99.76	<b>100.0</b>	<b>100.0</b>
12		Mud flats	63.02	88.07	59.64	95.03	92.84	99.40	<b>100.0</b>
13		Water	98.27	99.57	99.14	100.0	99.89	<b>100.0</b>	<b>100.0</b>
OA	/	/	82.75	94.32	85.40	91.98	81.31	98.77	<b>99.42</b>
AA	/	/	79.86	93.10	84.09	88.47	73.11	97.89	<b>99.06</b>
Kappa	/	/	80.80	93.67	83.77	91.07	79.02	98.63	<b>99.36</b>

#### D. Ablation Studies

To demonstrate the contribution of the presented models of the SMESC method, we conducted ablation studies on the four datasets, and the average OAs and standard deviations of

the five times are shown in Table II. In the experiments, the baseline of SEN represents the presented expansion network without the CMRB module. As can be observed, the approach with SEN generates the lowest OAs for the four datasets. The

TABLE VII  
COMPARISON OF THE COMPUTATION COST (MB)

	RSSAN	pResNet	SSMTR	SSTN	SSFTT	SMESC-S	SMESC
P	0.19	1.12	1.49	0.03	0.83	0.42	0.42
F	0.32	0.92	1.17	0.11	2.35	7.35	7.36

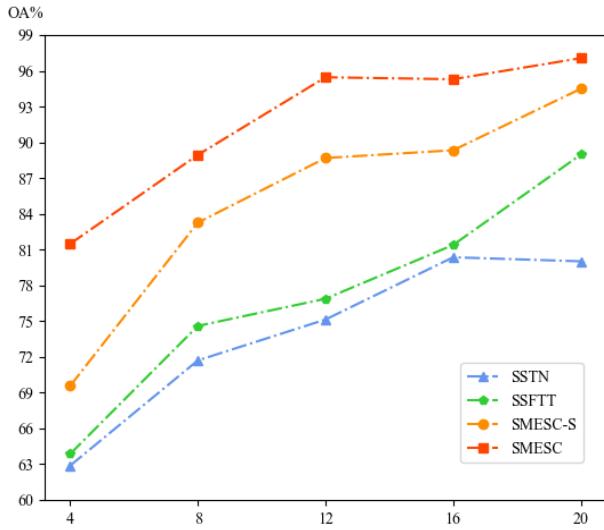


Fig. 10. OA with different numbers of training samples of each category on the compared methods.

SEN + CMRB approach acquires higher OA than SEN, and the SEN + CMRB + MST obtains better performance than SEN + MST, which illustrates the effectiveness of the CMRB for all the datasets. The results also show the CMRB module has a positive effect on the dataset with more bands, that is, the IP dataset, the approaches with CMRB have superior performance than the implementation without it.

The approaches with the MST strategy achieve 1%–3% improvement compared to SEN and SEN + CMRB, respectively. On the IP dataset, SMESC achieves nearly a 5% improvement over SEN. Furthermore, with the CMRB and the MST strategy, the model yields the best performance. In specific, the accuracies of the IP, Houston, and Botswana improve by 4.85%, 3.09%, and 2.83%.

#### E. Analysis of Parameters and Robustness Evaluation

1) *Impact of Different Numbers of Samples:* To evaluate the effect of different numbers of samples on the proposed SMESC model, we conduct the experiments on the IP dataset with the training number of each category varied in the set of {4, 8, 12, 16, 20}. In this section, the sample size is fixed to  $15 \times 15$ , and the SSTN and SSFTT methods that performed better than other compared methods are utilized for comparison. We repeat each experiment five times and record the average OA in Fig. 10. As can be seen, the values of OA of all methods gradually improve as the number of samples increases. Notably, our proposed method consistently outperforms the other methods by achieving higher classification accuracy compared to the other two methods across different sample sizes. Besides, our SMESC-S and SMESC methods generate satisfactory results even in the situation with

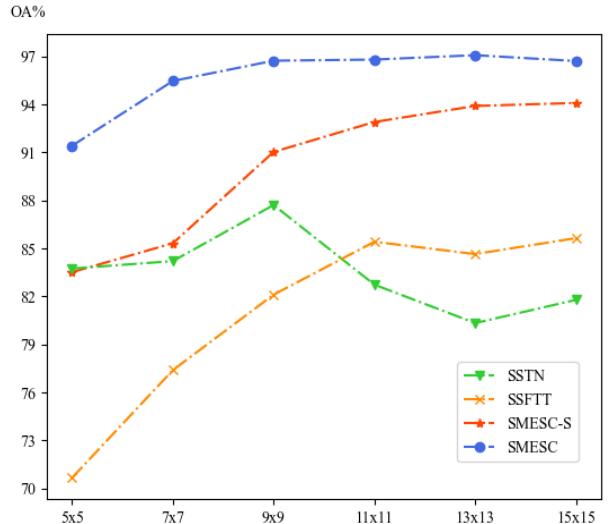


Fig. 11. OA with different sizes of training samples on the compared methods.

only four samples. In particular, the SMESC yields an OA of 81%, which exceeds the SSTN model by nearly 15%. These results indicate that the proposed method effectively utilizes sample information and multisize information even with fewer labeled samples.

2) *Impact of Different Sizes of Samples on the Model:* To assess the impact of sample size on our method, we conducted tests on the IP dataset with the patch size of  $k \times k$ , which  $k \in \{5, 7, 9, 11, 13, 15\}$ . Specifically, we randomly selected samples with different sizes for training and testing and repeated each experiment five times for each sample size. The SSFTT and SSTN methods are employed for comparison, and the final experiment results with the average OA are presented in Fig. 11.

As observed, smaller sample sizes containing less spatial information result in lower classification accuracy observed for all four methods. As the sample size increases, the OA of the SSTN method initially improves but begins to decline later, with the best performance achieved with a sample size of  $9 \times 9$ . Although the classification performance of SSFTT improves with increasing sample size, its accuracy still falls short of our method, particularly when compared to the SMESC method, which is lower by nearly 12%. These results further validate the robustness of our proposed method, which maintains the spatial size and effectively captures spatial information to achieve superior classification performance.

Furthermore, the SMESC method outperforms the other methods across all sample sizes. In particular, it leads by nearly 7% with a sample size of  $5 \times 5$ , indicating that the multisize training strategy enables the model to fully absorb the spatial information embedded in samples of different sizes, resulting in better classification results.

3) *Generalizability of CMRB:* In addition, to verify the generalizability of the CMRB proposed in this article, we apply it to the widely used ResNet network and conduct five experiments on the IP dataset, with the results and average OA values shown in Fig. 12. Specifically, we employ CMRB in the residual stage of ResNet18 by replacing the 1-D ConvT in

TABLE VIII  
CLASSIFICATION RESULTS (%) WITH COMPARED APPROACHES WITH AND WITHOUT MST ON THE IP DATASET

size	RSSAN-S	RSSA-N	pResnet-S	pResNet	SSMTR-S	SSMTR	SSTN-S	SSTN	SSFTT-S	SSFTT	SMESC-S	SMESC
3×3	2.91	69.23	5.31	70.1	2.37	80.31	5.74	89.17	5.03	70.21	80.42	84.77
5×5	3.78	74.53	4.27	74.93	3.08	82.2	11.85	92.33	13.17	75.95	93.61	95.13
7×7	10.21	77.21	12.03	79.61	18.09	87.71	16.09	93.22	10.56	85.03	95.64	97.57
9×9	17.45	79.03	30.57	82.64	16.43	89.31	44.04	93.07	19.65	90.63	<b>96.07</b>	98.14
11×11	27.13	82.61	41.50	84.61	27.46	88.56	35.53	91.88	27.06	93.03	95.78	<b>98.37</b>
13×13	35.36	84.32	37.03	85.03	35.30	87.46	39.00	90.01	38.61	93.52	94.81	98.04
15×15	<b>64.51</b>	<b>85.88</b>	<b>82.51</b>	<b>84.93</b>	<b>70.01</b>	<b>85.45</b>	<b>82.04</b>	<b>88.19</b>	<b>89.76</b>	<b>93.24</b>	93.26	97.52
17×17	30.27	84.09	43.05	83.22	33.29	84.32	36.32	87.52	31.78	92.89	91.19	96.34
19×19	20.04	83.61	30.02	82.17	27.30	82.17	21.03	86.28	27.33	92.37	89.27	94.64

TABLE IX  
CLASSIFICATION RESULTS (%) WITH COMPARED APPROACHES WITH AND WITHOUT MST ON THE HOUSTON DATASET

size	RSSAN-S	RSSA-N	pResnet-S	pResNet	SSMTR-S	SSMTR	SSTN-S	SSTN	SSFTT-S	SSFTT	SMESC-S	SMESC
3×3	1.30	71.34	3.97	67.71	4.40	74.77	4.24	88.58	2.26	64.97	83.88	92.74
5×5	6.89	80.19	5.32	76.59	7.12	84.66	11.15	91.02	3.66	74.1	96.43	99.39
7×7	13.86	87.09	21.61	85.78	17.94	91.46	24.48	92.47	14.32	85.69	98.66	<b>99.68</b>
9×9	23.34	91.38	18.60	91.22	33.92	95.06	38.87	93.15	25.14	93.07	99.53	99.64
11×11	35.01	92.29	28.40	93.33	39.67	95.3	46.08	93.51	30.34	94.54	99.63	99.53
13×13	37.02	90.78	43.52	93.47	55.92	94.38	56.39	92.3	50.49	93.1	99.55	99.31
15×15	<b>79.42</b>	<b>88.45</b>	<b>89.56</b>	<b>92.47</b>	<b>84.42</b>	<b>92.7</b>	<b>86.02</b>	<b>90.54</b>	<b>91.27</b>	<b>88.59</b>	<b>99.31</b>	98.42
17×17	30.21	87.21	37.23	92.02	36.77	91.56	37.55	89.37	42.79	87.63	98.70	97.05
19×19	25.32	86.59	35.04	91.53	36.10	91.03	35.23	89.09	36.56	86.73	97.34	95.73

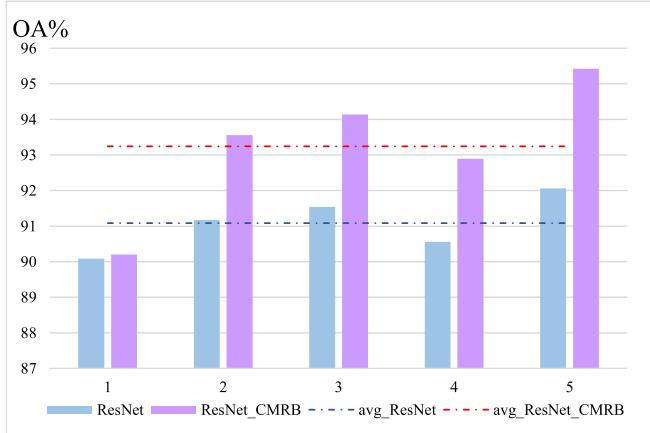


Fig. 12. OA comparison in terms of CMRB.

the CMRB with a 1-D Conv operation and maintaining the rest of the parameters unchanged. For convenience, we refer to this model as the ResNet-CMRB model. The average OA of the ResNet-CMRB is nearly 2% higher than that of the original ResNet, which proves the effectiveness and generalization ability of the CMRB module. Except in the SMESC model, the presented CMRB effectively captures salient channel information in other networks and provides valuable feedback to the model, which ensures that the specific network learns and utilizes relevant spectral information for accurate classification.

4) Computational Cost: In this section, we present the computational cost of the compared methods in Table IX, including the number of floating-point operations (FLOPs) denoted as F and network parameters (Param) denoted as P. As shown in Table VII, the SMESC-S and SMESC mod-

els in this article strike a balance between the number of parameters, computation, and performance. Compared to the original ResNet model, the SEN model has less than 10% of the number of parameters with better performance. Apparently, the CMRB module proposed in this article has minimal impact on the number of parameters.

Compared to other methods, the SSTN model has lower Param and FLOPs, while the performance is reduced severely. The RSSAN and pResNet models have a reasonable number of Param and FLOPs with unsatisfactory accuracies. Similarly, the SSMTR and SSFTT models have increased the number of parameters, but the performance is not further improved.

Overall, since the models focus on the spatial size information expansion, the SMESC-S and SMESC have better classification performance, and balance the relationship between performance, number of parameters, and computation.

#### F. Analysis of the Multisize Training Strategy

In this section, we conduct experiments to evaluate the impact of training order on the performance of MST with different patch sizes ( $k \times k$ ). We set the patch size  $k$  chosen in the set of  $\{5, 7, 9, 11, 13, 15\}$ , and the patch size for the testing samples is  $15 \times 15$ . The average OA of five experiments for each training sequence is recorded in Fig. 13.

As observed, the  $a$  and  $b$  sequences perform slightly better than the other sequences by achieving higher classification accuracies. We speculate that the model iteratively learns feature information and different-sized samples contain unique information that complements other sizes. In this way, the SMESC model learns multisize knowledge and obtains more comprehensive high-level semantic information to improve the classification performance on different samples. In addition, in sequences  $c$  and  $d$ , we trained the model with training

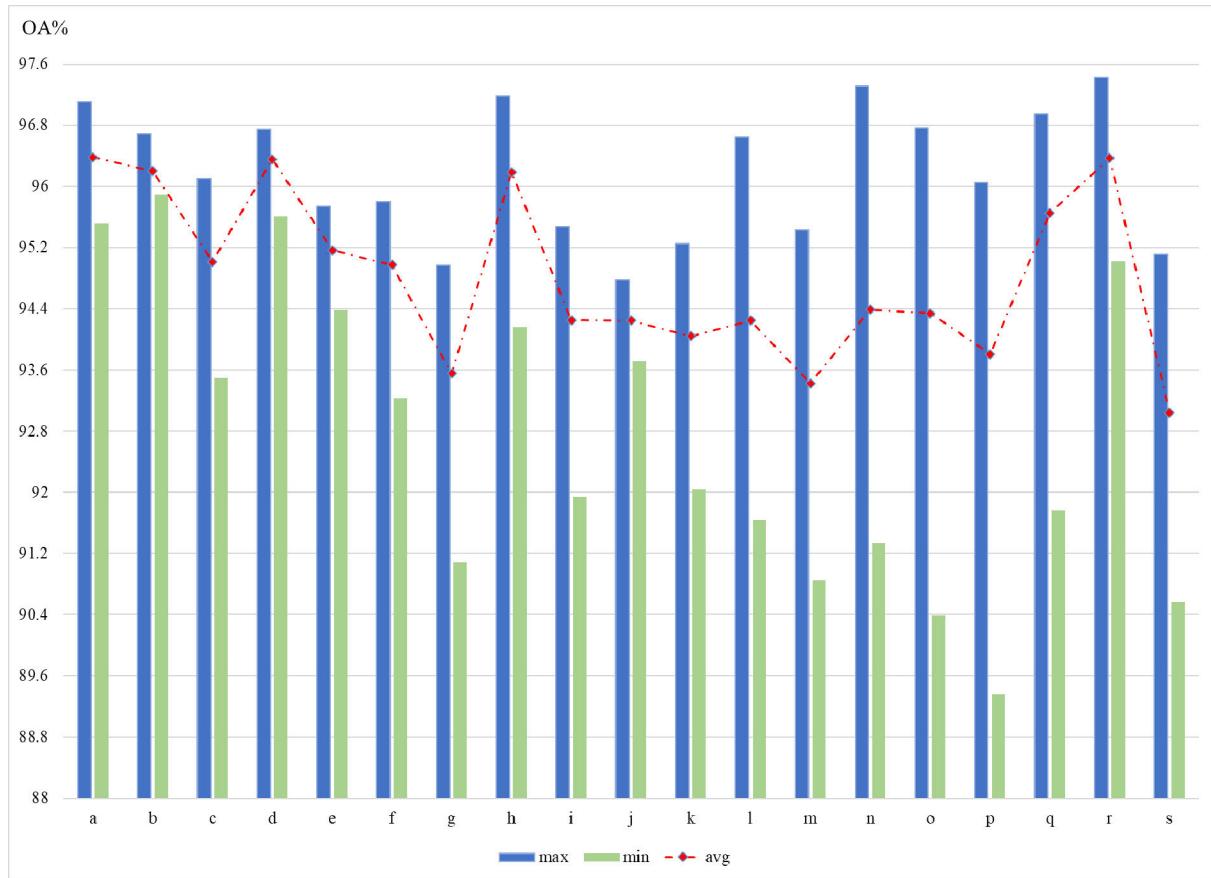


Fig. 13. OA comparison with different orders of the MST with SMESC for the IP data.

samples of size  $13 \times 13$ ,  $15 \times 15$ , and  $11 \times 11$ ,  $15 \times 15$ , respectively, and the final results were different. We assume that the reason is that different-sized samples contain varying background information, and the model is susceptible to complex background features, which can affect the final classification results. This phenomenon is also observed in the experiments with sequences *e*, *f*, and *g*.

For the sequences *i*, *j*, *k*, *l*, *n*, and *o*, we conduct the experiments with the training samples of  $15 \times 15$  and test the model on  $15 \times 15$  samples, as can be seen, the results are relatively inferior. It should be noted that the accuracies did not decrease by a large margin since our method maintains enough spatial information. Besides, sequences *h* and *r* are unordered and partially sampled, and the results also demonstrate good classification accuracy. Since the unordered sequence increases the learning difficulties, the performance is slightly lower than the accomplishment with the situation of *a* and *b*. Although it is challenging to identify an optimal solution among various combinations of sequences, we attempted to search for an optimal solution in this experiment. In terms of the results, the increasing and decreasing order suites a step-by-step learning approach for different samples, which reduce the fluctuation for the model training.

Besides, we have conducted experiments with the compared models to validate the adaptability of the approaches with MST on the IP and Houston datasets. The experimental results are demonstrated in Tables VIII and IX. Specifically, we train the model with samples of size  $15 \times 15$  and evaluate the model on different patch sizes ( $k \times k$ ,  $k = 5, 7, 9, 11, 13, 15$ ). In the tables, the model with the suffix “-S” indicates the same model without the MST policy. As observed, each row displays the highest accuracy of five executions of the test samples with the corresponding sample size indicated in the first column. The highest accuracies are obtained when the testing size matches the training size, which indicates that non-MST approaches are sensitive to the sample size, whereas the model with MST effectively addressed this issue and achieved favorable results on all the scales. The comparison results highlight the advantages of the MST in enhancing the adaptability of the HSIC model and absorbing multiscale information of HSI data. Besides, as observed in the columns with a green background and their adjacent columns in the two tables, it is evident that the model incorporating the multiscale information with the MST strategy consistently outperforms the non-MST model, which also illustrates the effectiveness of our proposed training strategy.

## V. CONCLUSION

In this article, we rethink and explore the importance of the spatial size factor. Based on the foundation that the original convolution operation is contrary to the inherent characteristic of HSI data, we employ the transposed convolution operation to extract the hidden information of the subpixel and a SMESC framework is presented for HSIC. Separately, the proposed SMESC has the following components in terms of spatial size factor.

- 1) The SEN is designed to extract hidden information from subpixels with enlarged feature maps, where the size preservation block is built to ensure that the HSIC model maintains a consistent mapping to facilitate stable learning.
- 2) A CMRB is responsible for reducing spectral redundancy while ensuring the focus on the spatial information of the model.

A straightforward yet highly effective multiple-size training strategy is designed to replace the conventional multiple-size FE branches. The proposed architecture is illustrated to outperform other advanced methods through extensive experimentation and analysis of four HSI datasets.

In future research, we intend to explore the integration of the proposed expansion features with the self-attention mechanism in the graph neural network. By incorporating expansion spatial-spectral features and leveraging the self-attention mechanism in graph-structured data, we aim to improve the ability of the model to capture intricate spatial and spectral relationships and further enhance the discriminative representation of subpixels for HSIC.

## REFERENCES

- [1] Y. Li, W. Xie, and H. Li, "Hyperspectral image reconstruction by deep convolutional neural network for classification," *Pattern Recognit.*, vol. 63, pp. 371–383, Mar. 2017.
- [2] Q. Zhang, Y. Zheng, Q. Yuan, M. Song, H. Yu, and Y. Xiao, "Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven," *IEEE Trans. Neural Netw. Learn. Syst.*, no. 6, Jun. 2023, Art. no. 3278866.
- [3] Y. Wang, X. Chen, E. Zhao, and M. Song, "Self-supervised spectral-level contrastive learning for hyperspectral target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510515.
- [4] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," *Remote Sens.*, vol. 10, no. 11, p. 1700, Oct. 2018.
- [5] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023.
- [6] Q. Zhang, Q. Yuan, M. Song, H. Yu, and L. Zhang, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, 2022.
- [7] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
- [8] K. Jiang et al., "Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining," *IEEE Trans. Image Process.*, vol. 30, pp. 7404–7418, 2021.
- [9] L. Zhang, J. Li, and J. M. Bioucas-Dias, "Hyperspectral image classification using deep convolutional autoencoder features," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2565–2576, Oct. 2020.
- [10] 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.
- [11] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.
- [12] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610819.
- [13] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial-spectral fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2485–2501, 2020.
- [14] S. Ghaderizadeh, D. Abbasi-Moghadam, A. Sharifi, N. Zhao, and A. Tariq, "Hyperspectral image classification using a hybrid 3D-2D convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7570–7588, Jul. 2021.
- [15] C. Yu et al., "Hyperspectral image classification method based on CNN architecture embedding with hashing semantic feature," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1866–1881, May 2019.
- [16] Z. Ma, Z. Jiang, and H. Zhang, "Hyperspectral image classification using feature fusion hypergraph convolution neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517314.
- [17] 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.
- [18] Y. Zheng, S. Liu, and L. Bruzzone, "An attention-enhanced feature fusion network (AeF<sup>2</sup>N) for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023, Art. no. 5511005.
- [19] J. Li, X. Zhao, Y. Li, Q. Du, B. Xi, and J. Hu, "Classification of hyperspectral imagery using a new fully convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 292–296, Feb. 2018.
- [20] C. Yu et al., "Distillation-constrained prototype representation network for hyperspectral image incremental classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 3359629.
- [21] Q. Zhang, Y. Dong, Q. Yuan, M. Song, and H. Yu, "Combined deep priors with low-rank tensor factorization for hyperspectral image restoration," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [22] C. Zhao, B. Qin, S. Feng, W. Zhu, L. Zhang, and J. Ren, "An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5546216.
- [23] C. Yu, B. Gong, M. Song, E. Zhao, and C.-I. Chang, "Multi-view calibrated prototype learning for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5544713.
- [24] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral-spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019.
- [25] C. Yu, J. Huang, M. Song, Y. Wang, and C.-I. Chang, "Edge-inferring graph neural network with dynamic task-guided self-diagnosis for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535613.
- [26] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.
- [27] L. Wang, L. Zhang, and B. Du, "Hyperspectral image classification based on multi-scale feature fusion network and graph convolutional network," *Remote Sens.*, vol. 13, no. 24, p. 5227, 5227.
- [28] H. Zhang, H. Yu, Z. Xu, K. Zheng, and L. Gao, "A novel classification framework for hyperspectral image classification based on multi-scale dense network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2238–2241.
- [29] X. Wang and Y. Fan, "Multiscale densely connected attention network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1617–1628, 2022.
- [30] S. K. Roy, A. Deria, D. Hong, M. Ahmad, A. Plaza, and J. Chanussot, "Hyperspectral and LiDAR data classification using joint CNNs and morphological feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530416.
- [31] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global-local transformer network for HSI and LiDAR data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541213.

- [32] Y. Fan et al., "MSLAENet: Multiscale learning and attention enhancement network for fusion classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 10041–10054, 2022.
- [33] J. Li, Y. Ma, R. Song, B. Xi, D. Hong, and Q. Du, "A triplet semisupervised deep network for fusion classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540513.
- [34] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 552–556, Mar. 2015.
- [35] Z. Qiu, J. Xu, J. Peng, and W. Sun, "Cross-channel dynamic spatial-spectral fusion transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5528112.
- [36] Z. Li, Z. Xue, Q. Xu, L. Zhang, T. Zhu, and M. Zhang, "SPFormer: Self-pooling transformer for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502019.
- [37] H. Ge et al., "Pyramidal multiscale convolutional network with polarized self-attention for pixel-wise hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5504018.
- [38] Z. Zhao, H. Wang, and X. Yu, "Spectral-spatial graph attention network for semisupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5503905.
- [39] J. Bai et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535317.
- [40] X. Tang et al., "Hyperspectral image classification based on 3-D octave convolution with spatial-spectral attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2430–2447, Mar. 2021.
- [41] R. Shang et al., "Hyperspectral image classification based on pyramid coordinate attention and weighted self-distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5544316.
- [42] T. Song, Y. Wang, C. Gao, H. Chen, and J. Li, "MSLAN: A two-branch multidirectional spectral-spatial LSTM attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528814.
- [43] Y. Cui, Z. Yu, J. Han, S. Gao, and L. Wang, "Dual-triple attention network for hyperspectral image classification using limited training samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [44] 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.
- [45] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.
- [46] L. Huang, Y. Chen, and X. He, "Spectral-spatial masked transformer with supervised and contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508718.
- [47] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715.
- [48] 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.



**Chunyan Yu** (Senior Member, IEEE) received the Ph.D. degree in environmental engineering from Dalian Maritime University, Dalian, China, in 2012.

She is currently an Associate Professor with the Information Science and Technology College, Dalian Maritime University. Her research interests include image segmentation, hyperspectral image classification, and pattern recognition.



**Yuanchen Zhu** received the bachelor's degree from Qingdao University of Technology, Qingdao, China, in 2022. He is currently pursuing the master's degree in computer science and technology with Dalian Maritime University, Dalian, China.

His research interests include hyperspectral image processing and deep learning.



**Meiping Song** (Member, IEEE) received the Ph.D. degree from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2006.

From 2013 to 2014, she was a Visiting Associate Research Scholar with the University of Maryland, Baltimore County, Baltimore, MD, USA. She is currently a Professor with the College of Information Science and Technology, Dalian Maritime University, Dalian, China. Her research interests include remote sensing and hyperspectral image processing.



**Yulei Wang** (Member, IEEE) received the B.S. and Ph.D. degrees in signal and information processing from Harbin Engineering University, Harbin, China, in 2009 and 2015, respectively.

In 2011, she was awarded by the China Scholarship Council to study at the Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore County, Baltimore, MD, USA, as a joint Ph.D. Student, for two years. She is currently an Associate Professor with the Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. Her research interests include hyperspectral image processing and vital signs signal processing.



**Qiang Zhang** (Member, IEEE) received the B.E. degree in surveying and mapping engineering and the M.E. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017, 2019, and 2022, respectively.

He is currently an Xinghai Associate Professor with the Center for Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. He has authored more than 20 journal articles in the *IEEE TRANSACTIONS ON IMAGE PROCESSING*, *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, *Earth System Science Data*, and *ISPRS Journal of Photogrammetry and Remote Sensing*. His research interests include remote sensing information processing, computer vision, and machine learning. More details could be found at <https://qzhang95.github.io>.