Hyperspectral Image Classification With Multi-Attention Transformer and Adaptive Superpixel Segmentation-Based Active Learning

Chunhui Zhao, Boao Qin, *Student Member, IEEE*, Shou Feng, *Member, IEEE*, Wenxiang Zhu, *Member, IEEE*, Weiwei Sun, *Senior Member, IEEE*, Wei Li, *Senior Member, IEEE*, and Xiuping Jia, *Fellow, IEEE*

***Abstract*— Deep learning (DL) based methods represented by convolutional neural networks (CNNs) are widely used in hyperspectral image classification (HSIC). Some of these methods have strong ability to extract local information, but the extraction of long-range features is slightly inefficient, while others are just the opposite. For example, limited by the receptive fields, CNN is difficult to capture the contextual spectral-spatial features from a long-range spectral-spatial relationship. Besides, the success of DL-based methods is greatly attributed to numerous labeled samples, whose acquisition are time-consuming and cost-consuming. To resolve these problems, a hyperspectral classification framework based on multi-attention Transformer (MAT) and adaptive superpixel segmentation-based active learning (MAT-ASSAL) is proposed, which successfully achieves excellent classification performance, especially under the condition of small-size samples. Firstly, a multi-attention Transformer network is built for HSIC. Specifically, the self-attention module of Transformer is applied to model long-range contextual dependency between spectral-spatial embedding. Moreover, in order to capture local features, an outlook-attention module which can efficiently encode fine-level features and contexts into tokens is utilized to improve the correlation between the center spectral-spatial embedding and its surroundings. Secondly, aiming to train a excellent MAT model through limited labeled samples, a novel active learning (AL) based on superpixel segmentation is proposed to select important samples for MAT. Finally, to better integrate local spatial similarity into active learning, an adaptive superpixel (SP) segmentation algorithm, which can save SPs in uninformative regions and preserve edge details in complex regions, is employed to generate better local spatial constraints for AL. Quantitative and qualitative results indicate that the MAT-ASSAL outperforms seven state-of-the-art methods on three HSI datasets.**

***Index Terms*— Hyperspectral image classification, multi-attention transformer, active learning, adoptive superpixel segmentation.**

Manuscript received 13 October 2022; revised 6 March 2023 and 28 April 2023; accepted 14 June 2023. Date of publication 27 June 2023; date of current version 5 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62002083, Grant 61971153, and Grant 62071136; in part by the Open Fund of State Key Laboratory of Remote Sensing Science under Grant OFSLRSS202210; and in part by the Heilongjiang Provincial Natural Science Foundation of China under Grant LH2021F012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Clinton Fookes. *(Corresponding author: Shou Feng.)*

Chunhui Zhao, Boao Qin, and Wenxiang Zhu are with the College of Information and Communication Engineering and the Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin 150001, China (e-mail: zhaochunhui@hrbeu.edu.cn; qinboao@hrbeu.edu.cn; zhuwenxiang@hrbeu.edu.cn).

Shou Feng is with the College of Information and Communication Engineering and the Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin 150001, China, and also with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: fengshou@hrbeu.edu.cn).

Weiwei Sun is with the College of Architectural Engineering, Civil Engineering and Environment, Ningbo University, Ningbo 315211, China (e-mail: nbsww@outlook.com).

Wei Li is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: liwei089@ieee.org).

Xiuping Jia is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: x.jia@adfa.edu.au).

Digital Object Identifier 10.1109/TIP.2023.3287738

I. Introduction

HYPERSPECTRAL images (HSIs) with abundant spatial-spectral features play an important role in image target detection [1], anomaly detection [2] and image classification. In recent years, HSI classification (HSIC) [3], [4], a hot research field which assigns each pixel with predefined labels by using some manually labeled samples [5], is widely used in agriculture [6], ecological monitoring [7], national defense [8], urban construction [9], and et al. Initially, the traditional machine learning (ML) methods that only take the spectral information into consideration such as support vector machines (SVM) [10] and sparse representation [11] were applied in HSIC, but these ML-based methods have a limitation on learning practical features and cannot achieve satisfactory performance. With the rise of deep learning (DL), a large number of DL-based methods are used to tackle hyperspectral image classification tasks [12], [13]. Some features in hyperspectral images may be highly unstructured, DL-based methods aim to construct multiple continuous latent spaces to find a linear separable feature space as much as possible by the feature transformation. Besides, they begin to utilize the spatial information of HSIs and achieved remarkable performance, where they mainly benefit from an assumption that HSIs are usually locally homogeneous. For instance, in order to extract more discriminative spatial-spectral features, a unified network that adopts a one-dimensional (1D) convolution kernel to extract spectral features and a two-dimensional (2D) convolution kernel to extract spatial features was proposed by [14]. Hamida et al. [15] directly utilized a 3D convolutional

neural network (CNN) for the HSIC task. Roy et al. [16] combined the 2D and 3D convolutional kernels to build a hierarchical 3D–2D CNN for HSIC. To better preserve detailed features of spatial edges, Zhao et al. [17] proposed a method based on kernel-guide deformable convolution and double-window joint bilateral filter. Encouraged by the full convolution network [18], some recent works begin to explore the patch-free global learning framework. For instance, Zheng et al. [19] first proposed the new global stochastic stratified sampling strategy and patch-free global learning framework. Then, Zhu et al. [20] proposed a spectral-spatial-dependent global learning framework for insufficient and imbalanced HSIC. Furthermore, to extract deeper and more convincing features, the contextual deep CNN with shortcut connection, such as ResNet [21] and DenseNet [22], was utilized in HSIC.

However, CNN-based methods, especially with the small-size convolution kernels, only focus on capturing the fine-level features in a limited reception field. For the long-range contextual (high-level) features, CNN needs a deep network to capture them, but limited hyperspectral data is prone to lead such deep network overfitting. In recent years, Transformer [23] has become the de-facto standard for natural language processing (NLP) tasks owing to the strong global modeling capability of its self-attention mechanism. Subsequently, some attention-based modules were applied in HSIC, such as 3D attention networks [24], hierarchical homogeneity-attention network [25], attention-based dense CNN [26], attention-based second-order pooling network [27], attention-based adaptive spectral–spatial kernel ResNet [28] and spatial-spectral self-attention network [29]. These methods either applied the self-attention in conjunction with convolutional networks, or used self-attention to replace certain components of convolutional networks while keeping their overall structure in place. However, the classification performance of these methods still strongly depend on their CNN-based backbones. Vision Transformer (ViT) [30], which splits an image into multiple patches and treats these patches as tokens (words) of Transformer, successfully applied a pure Transformer to computer vision for the first time. In order to apply the Transformer-based backbone in HSIC, Li et al. [31] first proposed a HSIC framework based on the encoder of Transformer, named HSI-BERT, which employed the dynamic input regions to jointly train Transformer. Subsequently, a spectral–spatial Transformer network (SSTN) is proposed in [32], which consists of spatial attention and spectral association modules. Motivated by ViT, a SpectralFormer [33] which splits the spectral bands of HSIs into multiple groups (tokens) is proposed for HSIC.

The excellent performance of CNN-based and Transformer-based networks is attributed to adequate training samples, that is, the accuracy of the model is determined by the quantity and quality of training samples [34]. Nevertheless, obtaining a sufficient number of training samples for an HSIC task is expensive and time-consuming [35]. Thus, some semi-supervised methods [36] are used to alleviate the HSIC problem of low classification accuracy under the condition of small sample size. For example, to train the semi-supervised model, the method proposed in [36] adopted a strategy that pre-training models by enlarged pseudo-labels and then fine-tuning by true labels. In order to obtain a excellent model and save the cost of manual labeling, active learning (AL) strategy, which selects important training samples for the classifier according to the prediction of model, is used in HSIC [37], [38], [39], [40]. Unlike the methods that directly use the original training datasets, AL based methods augment training sample sets by selecting unlabeled samples with high informativeness under certain criteria. Sun et al. proposed an AL method combined with the Gaussian process classifier for HSIC [41]. Later, Cao et al. [42] combined AL and CNN to reduce the number of labeled samples required for training. Liu et al. [43] employed the AL to train a deep contextual residual network for HSIC. Taking the spatial information of HSIs into account, some AL approaches based on local spatial assumption, which assumes that neighboring pixels share the same class labels, are proposed to refine the sample collection.

These methods integrate spatial information into AL strategies and achieve good performance, but their local similarity spatial assumption usually comes in the form of a rectangular or square window. To acquire multiple homogeneous regions, superpixel segmentation can be used to generate spatial-adaptive regions with similar spectral features [44]. Superpixel-based methods are mainly based on the assumption that the pixels within a superpixel should share the same class labels, which can provide a stronger and more informative inductive bias to constrain the HSIC task. These superpixel-based methods are primarily used for semi-supervised learning. A line of works aims to utilize the superpixel segmentation mainly for graph-based learning [45], [46] or graph convolutional networks (GCNs) [47], [48], [49]. They typically treat a superpixel as a node to construct a global graph structure and perform superpixel-wise feature extraction or label propagation. For instance, Liu et al. [50] constructed a hyperspectral GCN model to extract the superpixel-wise spatial feature and then used CNN's local information to generate complementary spectral-spatial features at the pixel and superpixel levels simultaneously. Sellars et al. [51] proposed a semi-supervised classification framework based on superpixel graph learning (SGL), which implements semi-supervised classification through constructing a graph using superpixels and propagating labels. Another line of works focuses on using the spatial constraints of superpixels to generate pseudo-labels to augment training samples [52], [53] or combining superpixel segmentation into representation learning to constrain pixels in homogeneous regions [54], [55]. For example, Zheng et al. [56] used a superpixel segmentation method to guide the augmentation of training samples, enabling classification under small-sample conditions. Yang et al. [57] achieved superpixel-guided discriminative low-rank representation by using classification-guided superpixel segmentation to group homogeneous pixels into clusters and then feeding them into their novel discriminative low-rank representation.

To integrate local spatial similarity of irregular homogeneous regions into active learning, Liu et al. [58] proposed

a HSIC method based on superpixel (SP) segmentation to select unlabeled samples, and Lu and Wei [59] subsequently proposed an AL method based on multiscale SP segmentation. However, these SP-based methods usually have a prohibitive cost to select a reasonable number of superpixels by handling heuristically as the topology of HSIs. That is, how to set the segmentation scale for these methods also depends on human experience [60], [61]. Besides, since the number of superpixels is fixed, these methods are difficult to save SPs in uninformative regions and preserve edge details in complex regions simultaneously. To sum up, the HSIC task is mainly limited by two aspects. The first is that the local spectral-spatial relationship and the high-level global features with discriminative semantic information are hard to balance. Among the existing learners, the high-level features generally require a deeper convolution network to learn, which is not suitable for the such small-size hyperspectral data. Although a lightweight ViT can directly construct the global spectral-spatial representation, the encoding of local fine-level spatial-spectral features is inefficient. The other is that the size of the hyperspectral training samples is hard to match the VC-dimension of the model. The models may exhibit undesirable behaviors such as over-fitting and memorization due to a small amount of training samples. Therefore, even if a well-design model is built, it is often difficult for the model to achieve the expected performance when the number of training samples is insufficient.

In order to tackle with aforementioned problems, an HSIC framework based on multi-attention Transformer and adaptive superpixel segmentation active learning (MAT-ASSAL) is proposed in this paper. First, the multi-attention Transformer (MAT) is built in the proposed HSIC framewrok. Specifically, to model long-range contextual spectral-spatial dependencies of an HSI, the self-attention module of Transformer is utilized for HSIC. Meanwhile, to capture the locally contextual spectral-spatial features, an outlook-attention [62] module is adopted to encode fine-level features and contexts into tokens. Secondly, to train an excellent MAT model with limited samples, a novel AL framework based on superpixel segmentation is proposed to select important samples for the training of MAT. Finally, to generate better spatially superpixel-based constraint, an adaptive superpixel segmentation method is integrated into the AL strategy to improve the confidence of pseudo labels within SPs and augment the number of training samples. The main contributions of this paper can be summarized as follows:

1) To extract features from a sequential perspective, a multi-attention Transformer is built for hyperspectral image classification in this paper. The self-attention of Transformer is adopted to model the long-range contextual dependencies between spectral-spatial embeddings. In addition, in order to overcome the low efficacy of self-attention in encoding local contextual features, an outlook-attention module is utilized to aggregate neighboring spectral-spatial features.
2) To train an excellent MAT model with limited samples, an AL framework based on SP segmentation is proposed. By virtue of the AL strategy, the important training samples for MAT classifier can be selected via an informativeness criterion. The proposed AL framework based on superpixel segmentation can reduce the number of labeled samples required for training, so as to effectively improve the prediction accuracy of the model when the number of labeled samples is limited.
3) In order to integrate local spatial similarity into active learning, an adaptive superpixel segmentation (ASS) method which can adaptively generate local and global details of HSIs is applied in AL. Specifically, ASS can save SPs in uninformative regions and preserve edge details in complex regions, which can provide assistance to the active learning to generate pseudo labels with high confidence within large irregular regions (superpixels). Furthermore, ASS can automatically select the number of superpixels, thus effectively avoiding the adverse influence of setting SP scales by human experience on segmentation results.

The remainder of this paper is organized as follows. In Section II, the model of multi-attention Transformer (MAT) and the framework of adaptive superpixel segmentation algorithm active learning (ASSAL) will be presented in detail. Sections III presents and analyzes the experimental results in detail. Finally, conclusions are reported in Section IV.

II. Proposed Method

This section introduces the proposed MAT-ASSAL in detail and the flowchart of MAT-ASSAL is presented in Fig.1. Firstly, the dimension-reduced HSI obtained by PCA is fed into the adaptive superpixel segmentation algorithm (ASS). Then, the segmented superpixel map and initial training samples are transmitted to active learning process. After the MAT predicts the classification results of each iteration, the most informative samples are selected by a sample query rule and the pseudo labels with high confidence are derived from the corresponding superpixels. These assigned samples with high informativeness and confidence are added to the training set for the next iteration. After the iterative process, the proposed MAT-ASSAL outputs the final classification map.

A. Adaptive Superpixel Segmentation

For an HSI $\mathbf{H} \in \mathbb{R}^{H\times W\times B}$, its reduced image after PCA is denoted as $\hat{\mathbf{H}} \in \mathbb{R}^{H\times W\times \hat{B}}$, where $N = H \times W$ is the number of pixels, B is the number of spectra and $\hat{B}$ is the number of principal components. Let $\boldsymbol{x}_i = (\boldsymbol{c}_i, \boldsymbol{b}_i)$ denote the measurement of the pixel i, where $\boldsymbol{c}_i \in \mathbb{R}^N$ is the location and $\boldsymbol{b}_i \in \mathbb{R}^{\hat{B}}$ is the principal component. The main task of segmentation is to partition each pixel $\boldsymbol{x}_i$ into one of K disjoint clusters $\mathbf{S} = \{S_1, S_2, \ldots, S_K\}$. Assuming l_i is the label of $\boldsymbol{x}_i$, which is the measurement-to-cluster assignment, the cluster (*i.e.*, associated superpixel) S_k contains a set of pixels labeled S_k (*i.e.*, $\{\boldsymbol{x}_i : l_i = S_k\}$),and thus l_i is also a pixel-to-superpixel assignment. Inspired by Bayesian adaptive superpixel segmentation (BASS) [63], the adaptive superpixel segmentation adopts a two-step process for generating $\boldsymbol{l}$ given Ω, where $p((\Omega_j)_{j=1}^K)$ is a Dirichlet distribution.

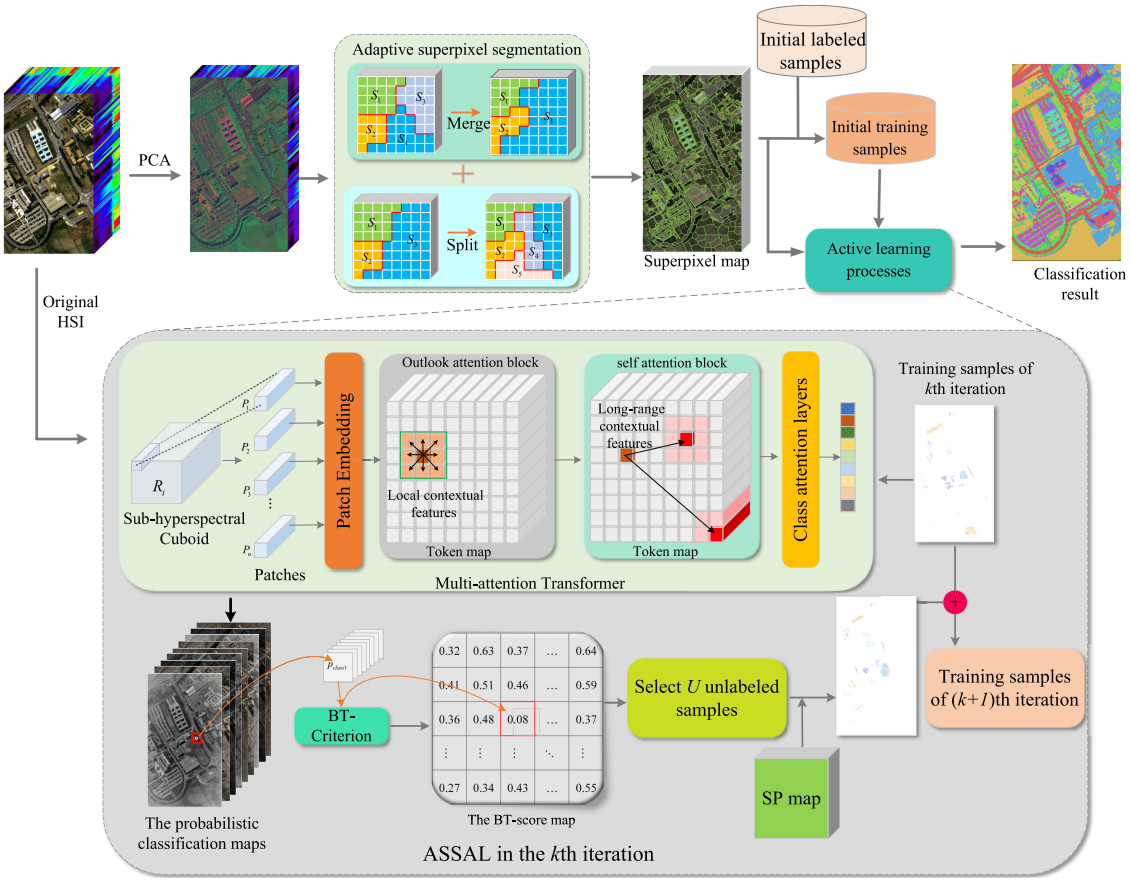


Fig. 1. The flowchart of the MAT-ASSAL.

1) Generation of Superpixel Labels: Initially, N temporary labels $\widetilde{\boldsymbol{l}} = (\widetilde{l}_1, \widetilde{l}_2, \ldots, \widetilde{l}_N)$ are the drawn *i.i.d.* from Ω. Then the actual label $\boldsymbol{l} = (l_1, l_2, \ldots, l_N)$ is obtained by the action of a random permutation matrix Π (an $N \times N$ matrix and has one entry of 1 in each row and each column and zeros elsewhere), which can be written as $\boldsymbol{l} = \Pi\widetilde{\boldsymbol{l}}$. Given $\widetilde{\boldsymbol{l}}$, the Π can be drown by

$$p(\boldsymbol{\Pi} \mid \widetilde{\boldsymbol{l}}) \propto \mathbb{F}_{\text{valid}}(\boldsymbol{l}) \exp\left(-\beta \sum_{i \sim i'} \mathbb{F}_{l_i \neq l_{i'}}(\boldsymbol{l})\right). \tag{1}$$

where the $\mathbb{F}_{\text{valid}}$ is the indicator function of event of valid segmentation, and the probabilities of any invalid segmentation is assigned as zero. The i' is the neighboring pixels of i in a predefined graph G. In addition, the $\exp\left(-\beta \sum_{i \sim i'} \mathbb{F}_{l_i \neq l_{i'}}(\boldsymbol{l})\right)$ with the form of Potts model is to enhance the spatial coherence of $\boldsymbol{l}$ and the β is a balance parameter. To generate K superpixel region, the probability density function (pdf) of a K-component Gaussian Mixture Model (GMM) can be drawn by

$$p\left(\boldsymbol{x}; \left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \Omega_j\right)_{j=1}^{K}\right) = \sum_{j=1}^{K} \Omega_j \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{2}$$

where $\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$ is a Gaussian pdf, with mean $\mu_j \in \mathbb{R}^n$ and an $n \times n$ covariance matrix $\boldsymbol{\mu}_j$. Therefore, if $\boldsymbol{l}$ is given, the spectral and location of an HSI are assumed as conditionally independence and the spectral covariance is isotropic. The Gaussian pdf is denoted by

$$\mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \Sigma_j\right) = \mathcal{N}\left(\boldsymbol{c}_i; \boldsymbol{\mu}_{j,\boldsymbol{c}}, \boldsymbol{\Sigma}_{j,c}\right) \mathcal{N}\left(\boldsymbol{b}_i; \boldsymbol{\mu}_{j,\boldsymbol{b}}, \sigma_{j,\mathbf{b}}^2 \boldsymbol{I}_{3\times 3}\right) \tag{3}$$

The mean and covariance matrix is defined as

$$\boldsymbol{\mu}_j = \begin{bmatrix} \boldsymbol{\mu}_{j,c} \\ \boldsymbol{\mu}_{j,b} \end{bmatrix}, \quad \boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_{j,c} & \mathbf{0}_{2\times 3} \\ \mathbf{0}_{3\times 2} & \sigma_{j,b}^2 \boldsymbol{I}_{3\times 3} \end{bmatrix} \tag{4}$$

where the $\boldsymbol{u}_{j,c}, \boldsymbol{\Sigma}_{j,\boldsymbol{c}}, \boldsymbol{\mu}_{j,\boldsymbol{b}}, \sigma_{j,\boldsymbol{b}}^2$ are latent, random, and possibly-dependent on j. Besides, $\left(\boldsymbol{\mu}_{j,\boldsymbol{c}}, \boldsymbol{\Sigma}_{j,c}\right)$ is a Normal-Inverse Wishart (NIW) prior and $\left(\boldsymbol{\mu}_{j,\boldsymbol{a}}, \sigma_{j,\boldsymbol{b}}^2\right)$ is a multivariate Normal-Inverse-Gamma (NIG) prior. To summarized as follows:

$$\begin{aligned} p\left(\boldsymbol{\mu}_{j,c}, \boldsymbol{\Sigma}_{j,c}\right) &= \text{NIW}\left(\boldsymbol{\mu}_{j,c}, \boldsymbol{\Sigma}_{j,c}; \boldsymbol{m}_{j,c}, \kappa_{j,c}, \boldsymbol{\Lambda}_{j,c}, \nu_{j,c}\right) \\ p\left(\boldsymbol{\mu}_{j,b}, \sigma_{j,b}^2\right) &= \text{NIG}\left(\boldsymbol{\mu}_{jb}, \sigma_{j,b}^2; \boldsymbol{m}_{j,b}, \kappa_{j,b}, \hat{a}_{j,b}, \hat{b}_{j,b}\right); \\ p\left(\left(\boldsymbol{\Omega}_j\right)_{j=1}^{K}\right) &= \text{Dir}\left(\left(\boldsymbol{\Omega}_j\right)_{j=1}^{K}; \alpha\right). \end{aligned} \tag{5}$$

where $p\left(\left(\boldsymbol{\Omega}_j\right)_{j=1}^{K}\right)$ is a Dirichlet distribution.

2) Updating Parameters and Labels: In order to achieve valid segmentation, BASS merges the uninformative superpixels and splits superpixels with rich spectral details. Initially, Iterated Conditional Modes (ICM) is exploited to speed up the convergence of sampling, and the acceptance of the

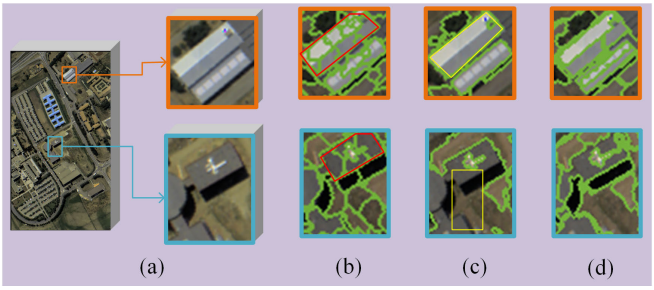


Fig. 2. The local results of different superpixel segmentation scales and ASS. (a) is the original HSI; (b) is the results of the small segmentation scale; (c) is the results of the large segmentation scale; (d) is the results of ASS. Notably, the results of (b) and (c) are segmented by a conventional SLIC-based method [51]. The regions of the red box waste a lot of SPs; The regions of the yellow box did not preserve edge details of HSI; The ASS can save SPs in uninformative regions and preserve edge details in complex regions.

merges/splits is determined by the Hastings' ratio exceeding 1. Assuming that the initial number of superpixels is K_0, different from K_0 in paper [63] which is a fixed value, the setting of K_0 in this paper is related to the spatial prior of hyperspectral images. However, the setting of K_0 is trivial for BASS due to the later split/merge, but K_0 still needs to be in a reasonable range because of the quite different sizes of different HSIs. Therefore, the K_0 can be defined as

$$K_0 = \frac{H \times W}{h} \tag{6}$$

where the h is the size of hyperspectral cube, which means that a $h \times h$ hyperspectral patch share the same land-cover. The update of labels respects the spatial constraint while satisfying topological constraints, and a superpixel which is connected and has no holes is called *simply-connected*. Thus, the valid segmentation means that all superpixels are *simply-connected*. The conditional models of parameter updates can be drown by (the '$*$' represents the updated parameters)

$$\left(\frac{\alpha_j^* - 1}{\sum_{j'=1}^{K} \alpha_{j'}^* - K}\right)_{j=1}^{K} = \underset{(\Omega_j)_{j=1}^{K}}{\arg\max}\ p\left((\Omega_j)_{j=1}^{K} \mid z, (\boldsymbol{x}_i)_{i=1}^{N}\right) \tag{7}$$

The parameters in formula (5) also need to be updated, and the details see [63].

After the parameter update is completed, the most important thing is to update the label of the superpixel. In order to avoid that the connectivity is destroyed by only updating the label of each single superpixel, simply-connected points with one pixel apart are updated in parallel. A single label l_i is updated by the conditional mode of $p\left(l_i \mid (\boldsymbol{\theta}_j, \pi_j)_{j=1}^{K}, (\boldsymbol{x}_i)_{i=1}^{N}, l_{/i}\right)$:

$$l_i = \underset{j:\exists i' \in \eta_4(i) \text{ s.t. } l_{i'}=j}{\arg\max}\ \Omega_j \mathcal{N}\left(\boldsymbol{c}_i; \boldsymbol{\mu}_j^{\boldsymbol{d}}, \boldsymbol{\Sigma}_j^{\boldsymbol{d}}\right) \mathcal{N}\left(\boldsymbol{a}_i; \boldsymbol{\mu}_j^{a}, \sigma_j^2 \boldsymbol{I}_{3\times 3}\right) \times \exp\left(-\beta \sum_{i'':i\sim i''} \mathbb{F}_{j\neq l_{i''}}(\boldsymbol{l})\right) \tag{8}$$

The splits and merges are determined according to the corresponding Hastings ratio. Finally, the final segmented map can be generated until each superpixel is *simple-connected*.

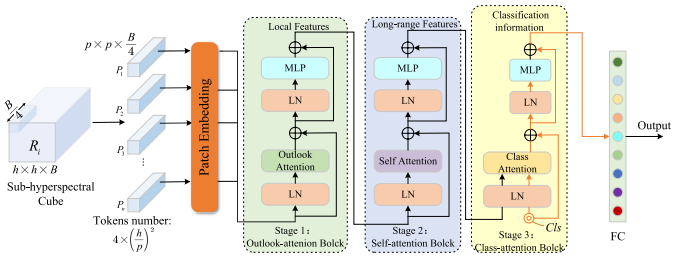


Fig. 3. An overview illustration of the MAT for the HSI classification task. MAT consists of three stages, Outlook-attention block, self-attention block and Class-attention block.

B. Multi-Attention Transformer for HSIC

To model the long-range contextual dependencies from the larger receptive field while extracting local contextual features simultaneously, a Transformer based on multi-attention is applied in HSIC. By taking the local information and the long-range dependency into account simultaneously, the features learned by MAT are prone to be linearly separable in the latent space. The self-attention module is used to capture long-range interactions through constructing a global self-attention mechanism to model the correlations between local patches (tokens) and other tokens. Transformer has strong capability of capturing the long-range dependency to obtain a global representation of the single sample's feature, and long-range relationships are prone to contain the overall semantic information which can help the model find a better feature transformation path (i.e., extract more discriminative spectral-spatial features), then the final output features are prone to be linearly separable in the feature space. Besides, the outlook attention [62] which can efficiently encode finer-level features and contexts into tokens is utilized to improve the correlation between the center token and surroundings. In addition, the class attention [64] is used to integrate useful classification information for the later classification tasks. The detailed architecture of multi-attention Transformer (MAT) is shown in Fig.3, which is a three-stage network including an outlook-attention block, an self-attention block (Transformer block) and a class-attention block.

For an HSI $\mathbf{H} \in \mathbb{R}^{H\times W\times B}$, it is cast into a set of $h \times h$ subhyperspectral cubes $\mathbf{R} \in \mathbb{R}^{h\times h\times B}$. The main purpose of the MAT is to seek the most distinctiveness features on a $h \times h$ receptive field and to assign the labels to each pixel in the cube. The representations of HSIs are spatially sparse while spectrally correlated, hence we separate the subhyperspectral cube into tokens along the spatial and spectral dimensions to jointly capture spectral-spatial representation. $\mathbf{R}$ is first separated into non-overlapping spectral-spatial patches by a patch splitting module. Each spectral-spatial patch $R_i \in \mathbb{R}^{p\times p\times b}$ is treated as a "token" $\mathrm{p}_n \in \mathbb{R}^C$, where the b is the size of spectral dimension and the C is the embedding dimension. Following the ViT [30], the learnable and absolute positional encoding is added into the embedding vectors to provide the contextual relationship. The b is an empirical parameter which is set as $b = (B/4)$ in this paper, hence the number of tokens is $4 \times (\frac{h}{p})^2$. Defined $h_t = \frac{h}{p}$, then the token map can be expressed as $\mathbf{P} \in \mathbb{R}^{2h_t\times 2h_t\times C}$. It should be noted that the first two stages did not add the extra class token due to

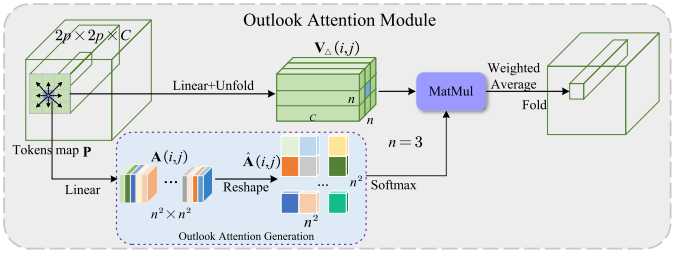


Fig. 4. An illustration of the outlook attention module.

the subsequent class-attention block. Then, the consecutive outlook-attention blocks $\mathbf{P}_l$ can be computed as

$$\begin{aligned}\hat{\mathbf{P}}_{l-1} &= \text{Out\_A}(\text{LN}(\mathbf{P}_{l-1})) + \mathbf{P}_{l-1},\\ \mathbf{P}_l &= \text{MLP}(\text{LN}(\hat{\mathbf{P}}_{l-1})) + \hat{\mathbf{P}}_{l-1}.\end{aligned} \tag{9}$$

where the LN is LayerNorm and the MLP is Muti-Layer Perception. The details of Outlook Attention (Out_A) is shown in Fig.4. Given a center token $\mathrm{p}_{i,j} \in \mathbb{R}^C$ with spatial location (i, j), the local spectral-spatial relationship is built via generating a window of size $n \times n$, and the n is generally set as 3 to aggregate the information of surroundings. For the attention weights, Out_A directly generates via linear layers of weights $\mathbf{W}_\text{A}$ and a reshaping operation rather than the Query-Key matrix multiplication of self-attention. Specifically, the attention weights can be written as

$$\mathbf{A} = \mathbf{P}\mathbf{W}_A \quad \mathbf{W}_A \in \mathbb{R}^{C\times n^4} \tag{10}$$

Then the attention weight for the value aggregation is generated by reshape the $\mathbf{A}$ to $\hat{\mathbf{A}} \in \mathbb{R}^{2h_t\times 2h_t\times n^2\times n^2}$. The value is computed via a linear layer $\mathbf{W}_\text{V} \in \mathbb{R}^{C\times C}$, and $\mathbf{V} \in \mathbb{R}^{2h_t\times 2h_t\times C}$. Let $\hat{\mathbf{A}}(i, j) \in \mathbb{R}^{n^2\times n^2}$ represent the weight at location (i, j) and $\mathbf{V}_\Delta \in \mathbb{R}^{C\times n^2}$ represent all the values which consists of the center value and neighbor values within the local window. Thus, the value projection procedure can be represented as

$$\mathbf{Y}_\Delta(i, j) = \text{MatMul}(\text{SoftMax}(\hat{\mathbf{A}}(i, j)), \mathbf{V}_\Delta(i, j)) \tag{11}$$

The correlation of center and its neighborhoods $\mathbf{Y}(i, j)$ can be summed up as

$$\mathbf{Y}(i, j) = \frac{\sum_{0\le i,j\le n} \mathbf{Y}_\Delta(i, j)}{n^2}. \tag{12}$$

To allow the model to focus on abundant information, an ensemble multi-heads attention is implemented to jointly attend to information from different representation subspaces at different position. The adjustment of multi-heads outlook attention is that the weight are set to multiple matrices, hence the $\mathbf{A}$ and $\mathbf{V}$ are adjusted to multiple $\mathbf{A}_m \in \mathbb{R}^{2h_t\times 2h_t\times n^4}$ and $\mathbf{V}_m \in \mathbb{R}^{2h_t\times 2h_t\times C/M}$, where the M is the number of heads.

After extracting the local features by Out_A, the Transformer block based on multi-heads Self_A is used to model the global information of the sub-hyperspectral cube. Before that, the token map $\mathbf{P}^l$ output from Out_A need to be flattened as a concatenation of the raw tokens $x_0^{patches} \in \mathbb{R}^{4h_t^2\times C}$.The Transformer block based on self-attention (Self_A) and the classification block based on class-attention (Cl_A) can be drawn as

$$\begin{aligned}\hat{x}_{l-1}^{\text{patches}} &= \text{Self\_A}(\text{LN}\left(x_{l-1}^{\text{patches}}\right)) + x_{l-1}^{\text{patches}},\\ x_l^{\text{patches}} &= \text{MLP}((\text{LN}(\hat{x}_{l-1}^{\text{patches}})) + \hat{x}_{l-1}^{\text{patches}},\\ \hat{x}_{l-1}^{\text{class}} &= \text{Cl\_A}(\text{LN}(\theta)) + x_{l-1}^{\text{class}},\\ x_l^{\text{class}} &= \text{MLP}((\text{LN}(\hat{x}_{l-1}^{\text{class}}) + \hat{x}_{l-1}^{\text{class}}.\end{aligned} \tag{13}$$

where the x^{class} is the class token and the $\theta = \left[x_{\text{class}}, x_{\text{patches}}\right]$. The self-attention produces a weighted average of values via Query-Key matrix multiplication, and its Query, Key and Value are generated only from the x_{patches}. The layer of Cl_A is identical to Self_A but Cl_A adds a class embedding. Therefore, the Out_A and self_A aim to learn the contextual relationship between patch embeddings, and the Cl_A is leveraged to summarize the classification information from spectral-spatial tokens into class token. The goal of Cl_A [64] is to avoid conflicts that may arise from inconsistent learning objectives between patch embeddings and the class token, so the Cl_A freezes patch embeddings and only updates the class token. In turn, the Out_A and Self_A only need to update the patch embeddings, which is the reason why the class token was not augmented to the first two stages. Therefore, the class-attention weights can also be obtained by the Query-Key matrix multiplication, but Query is obtained by x_{class} and {Key, Value} are generated by the $\theta = [x_{\text{class}}, x_{\text{patches}}]$. The CL_A layer extracts the useful information from the patch embeddings to the class embedding. There is at least a key corresponding to the query because θ includes x_{class} and x_{patches}. In the end, the class probability of each pixel can be predicted by linear layers. In the end, the class probability of each pixel can be predicted by linear layers.

C. Active Learning Based on ASS

Active learning (AL) is an effective semi-supervised way to select the important and informative unlabeled samples for the classifier. AL assumes that all training samples are not equally important to the same classifier. More specifically, only a few samples define the separating surface and most other ones are redundant to the classifier. Therefore, the AL is formulated according to certain sampling criteria, and then requires experienced *oracles* to annotate them. In the proposed MAT-ASSAL method, the adaptive superpixel segmentation is used to provide local spatial similarity for AL.

Suppose that $G = \{(r_i, y_i)\}_{i=1}^{e}$ and $G^{\text{Initial}} = \{(r_i, y_i)\}_{i=1}^{e_{\text{Initial}}}$ are the set of data samples and selected training samples, where e is number of all samples and e_{Initial} is the number of initial labeled samples. The training samples of the 1th iteration $G_0^{\text{training}} = G^{\text{Initial}}$. First, denoted the $\tilde{C}$ is the number of classes, a set of *posteriori* probability maps of kth iteration $\mathcal{P}_k \in \mathbb{R}^{H\times W\times \tilde{C}}$ can be produced via MAT classifier, which predicts the class probability of each pixel $\rho_{r_i} \in \mathbb{R}^{\tilde{C}}$ by training the samples G^{training}. Then, the *Break-Ties* (BT) criterion which focuses on the boundary regions between two classes is used in ASSAL for seeking the most informative unlabeled samples as desired training samples. Specifically, assuming that the unlabeled samples of kth iteration is

$G_k^{\text{pool}} = G - G_k^{\text{training}}$ and $\mathbf{F}_\Theta$ is a nonlinear function implemented by MAT model with parameter Θ, the BT criterion selects the most uncertain samples of the MAT classifier by

$$G_k^{BT}(u) = \arg \min_{y_i \in \mathcal{D}_{\text{pool}}} \{\max_{\tilde{c} \in \tilde{C}} p(y_i = \tilde{c} \mid r_i, \mathbf{F}_\Theta) - \max_{\tilde{c} \in \tilde{C} \setminus \{\hat{c}\}} p(y_i = \tilde{c} \mid r_i, \mathbf{F}_\Theta)\} \quad (14)$$

where $\hat{c} = \arg\max_{\tilde{c} \in \tilde{C}} p(y_i = \tilde{c} | r_i, \mathbf{F}_\Theta)$. After collecting the most informative samples, the generation of pseudo labels is very important to accelerate convergence and reduce the cost of manual labeling. However, it is difficult to generate the most confident pseudo labels with very limited labels. Taken the spatial constraints of superpixel into consideration, the ASS, which can preserve details of edge pixels while saving the number of superpixels, ensures that the pseudo samples within the same superpixel have high confidence. Accordingly, the pseudo labels $G_u^{\text{Pse}} \in S_u$ corresponding to selected unlabeled samples $G_k^{BT}(u) \in S_u$ can be obtained by assuming that the pixels within the same superpixel S_i share the same class. Next, the class of the center pixel of the S_i is labeled as the class of the entire superpixel. The training samples used for the next iteration can be presented as

$$G_k^{\text{unlabeded}} = \sum_{u=0}^{U} G_u^{\text{Pse}}$$
$$G_{k+1}^{\text{training}} = G_k^{\text{training}} \cup G_k^{\text{unlabeded}} \quad (15)$$

III. Experimental Results and Analysis

In this section, the details of three experimental datasets and experimental settings are first introduced. Then, the results of the proposed method and other state-of-the-art methods on three HSI datasets are compared. To verify the effectiveness of the active learning method based on adaptive superpixel segmentation, AL methods based on different spatial assumptions are analyzed. Finally, the parameters of the MAT model are discussed.

A. Datasets

Three benchmark HSI datasets are selected for experiments, which are University of Pavia, Houston2013 and Yellow River Estuary. These datasets are collected by different sensors over different land covers, and have been selected for experiments by many related papers. The details of these datasets are given as follows.

1) *University of Pavia*: This dataset was acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the urban area surrounding the University of Pavia, Italy. The spatial size of the data is 610×340, and the spatial resolution is as high as 1.3 m/pixel. The image contains 115 spectral channels from 0.43 to 0.86 μm, and 13 spectral bands are removed due to noise. Its ground truth contains 42776 labeled pixels and 9 classes.
2) *Houston* 2013: The Houston 2013 dataset was captured by an airborne spectrographic image sensor, which covers the area of University of Houston and its neighboring urban area. Houston 2013 dataset include 349×1905 pixels with a spatial resolution of 2.5m and consists of 144 spectral channels ranging from 0.38 to 1.05 μm, and the ground truth map utilizes 15029 labeled pixels with 15 different land-cover classes. It adopted the standard training and testing sets given by the 2013 GRSS Data Fusion Contest.
3) *Yellow River Estuary*: The Yellow River Estuary (YRE) dataset which captured by the advanced hyperspectral imager (AHSI) sensor includes 1185×1342 pixels and 6471 labeled pixels with 21 different land-cover classes. It is designed and produced by the Shanghai Institute of Techincal Physics, Chinese Academy of Sciences (CAS), and is mounted on China's Gaofen 5 satellite platform [18]. The YRE consists of 148 visible and near-infrared (VNIR) bands (after removing 2 noisy bands) ranging from 0.4 to 1.0 μm and 132 visible shortwave-infrared (SWIR) bands (after removing 48 noisy bands) ranging from 1.0 to 2.5 μm. The YRE is collected over the yellow river estuary field on November 1, 2018 and the spatial resolution of each pixel is 30m.

Detailed information of different landcover classes and the corresponding false colors of University of Pavia, Houston2013, and YRE) are listed in Table I.

B. Experimental Settings

All the experiments are implemented on the same hardware platform GPU: GTX-3090, CPU: Intel 4210R and memory: 32G. The input size of sub-hyperpsectral cube is 16×16 for the all methods without any data augmentation, and the other method-specific hyper-parameters of all methods obey the original optimal settings. The SGD is adopted as the optimizer, the initial learning rate, the weight decay and the batch size are set to $1e^{-3}$, $1e^{-4}$, 100, respectively. Taking the trade-off between the performance and the computational overhead into comprehensive consideration, the $b = (B/4)$ is chosen as the empirical setting for the subsequent experiments. Meanwhile, three metrics, overall accuracy (OA), class-specific accuracy, and the Kappa coefficient are employed to evaluate the classification performance.

1) Experimental Settings for Comparing With Other State-of-the-Arts: several representative methods belong to different categories with different advantages are selected to compare with the proposed framework MAT-ASSAL, which are SVM [10], 3D-CNN [15], SGL [51], SSAL-SN [58], FAAL [65], HSI-BERT [31], DBDA [66], HybridSN [16] and A^2S^2K-ResNet [28]. In order to show the overall effect of the proposed method, a classical machine learning method SVM and two CNN-based method (i.e., 3D-CNN and HybridSN) are exploited for comparison. To evaluate the classification performance of MAT-ASSAL under the small samples condition, a semi-supervised method SGL is adopted for comparison. Moreover, to show the effectiveness of the ASSAL scheme, two approaches based on active learning: SSAL-SN and FAAL are selected. Notably, SP-based methods include MAT-ASSAL may implicitly used the test data when they utilize the superpixel to augment training sets, but they only utilize the unlabeled data rather than their labeled information. In order

TABLE I

TRAIN-TEST DISTRIBUTION OF SAMPLES AND THE BACKGROUND COLOR OF LAND COVER CLASSER ON UNIVERSITY OF PAVIA, HOUSTON2013, AND YELLOW RIVER ESTUARY

University of Pavia				Houston 2013				Yellow River Estuary			
Color	**Land Cover Type**	**train**	**test**	**Color**	**Land Cover Type**	**train**	**test**	**Color**	**Land Cover Type**	**train**	**test**
	Asphalt	39	6592		Healthy grass	29	1222		Salt marsh	15	378
	Self-Blocking Bricks	21	3661		Commercial	29	1215		Shallow sea	8	203
	Meadows	109	18540		Stressed grass	29	1225		Acquaculture	29	763
	Bitumen	8	1322		Residential	29	1239		Freshwater herbaceous marsh	4	91
	Gravel	12	2087		Synthetic grass	16	681		Mud flat	4	106
	Bare Soil	29	5000		Water	8	317		Seep sea	4	92
	Trees	18	3046		Trees	29	1215		Rice	7	183
	Painted metal sheets	8	1337		Soil	29	1213		Aquatic vegetation	3	80
	Shadows	6	941		Road	29	1223		Reed	8	192
					Parking Lot 1	29	1204		Suaeda salsa	18	451
					Parking Lot 2	11	458		Flood plain	14	347
					Highway	29	1198		Pond	33	897
					Tennis Court	10	418		River	9	231
					Running Track	15	645		Soybean	23	572
					Railway	29	1206		Building	21	532
									Spartina	3	65
									Broomcorn	18	436
									Intertidal saltwater	1	38
									Maize	5	128
									Tamarix	3	69
									Locust	15	362
Total Number		250	42526	Total Number		350	14679	Total Number		255	6216

to verify the advantages of MAT classifier through local and global feature modeling, three attention-based methods are chosen: spatial-only based, HSI-BERT, spectral-spatial based, A^2S^2K-ResNet and DBDA. Note that the DBDA of Mish version is employed to compare in our experiments. In addition, to explore the performance of MAT model in the supervised mode and the boosting of ASSAL strategy, a MAT-only model is employed for comparison. The spatial size of patches of MAT-only and MAT-ASSAL are empirically set as the same 2×2, and the $2\times$out_A, $2\times$self_A and $2\times$Cl_A blocks are selected for MAT-only and MAT-ASSAL. The number of labeled samples is set to 250 (0.58% per class), 350 (2.33% per class), 255 (3.94% per class) for university of Pavia, Houston2013 and YRE, respectively. It is noteworthy that the train and test set listed in Table I are randomly selected for the non-AL methods. To facilitate the description of sampling criteria for AL-based methods, we simplify the process of active sampling into a formula $N_{total} = N_i \times N_c + K \times N$, where the N_{total} is the number of totally selected labeled samples (e.g., 250, 350 and 255 for three datasets), N_i is the number of the initial samples per class, the N_c the number of the classes, the K is the number of iterations and the N is the number of actively augmented samples per iteration. Therefore, for the AL-based methods, the initial training samples N_i are {10,10,5} per class and actively select {32, 20, 30} samples (e.g., N) to augment the training set per iteration for university of Pavia, Houston2013 and YRE datasets, respectively, that is, 250 ($10\times9+32\times5$), 350 ($10\times15+40\times5$), 255 ($5\times21+30\times5$) for the three datasets. Furthermore, the number of iterations K is 5 and 100 epochs for training per iteration, and 300 epochs are set for the non-AL methods.

2) Experimental Settings for Comparing Active Learning: Initially, in order to verify the importance of break-ties criterion of AL for sample selection, the random selection (RS) and the AL methods without any local spatial assumption (AL-Basic) are first selected as comparison. Then, to verify the effectiveness of spatial constraints, AL methods based on different local spatial assumptions are compared: window-based AL (Win-AL) and general superpixel-based AL (SP-AL). In addition, to clearly show the superiority of adaptive superpixel segmentation, the SP-AL is divided into small scale (the number of superpixels) [3000, 5000, 5000], large scale [500, 700, 700] for PaviaU, Houston2013 and YRE, respectively. The learners of all AL methods are MAT, and the sample sizes consistent with previous settings, that is, 250 ($10\times9+32\times5$), 350 ($10\times15+20\times5$), 255 ($5\times21+30\times5$) for university of Pavia, Houston2013 and YRE datasets, respectively.

3) Parameter Analysis Settings: Firstly, to explore the effect of PCA on classification results on three datasets, an ablation experiment about the PCA is designed. On university of pavia and houston2013 datasets, we set up three comparative experiments, including without PCA (original hyperspectral data), 99.8% principal components, and 95% principal components. For the YRE dataset, we retained 98% and 95% principal components, and set up one of the most extreme cases, which selects a pseudo-color image with three bands (20,38,59) for comparison. The PCA method of [51] is utilized in our paper, which is based on the variance criteria to select the principal components. Notably, the sample sizes of ablation experiment about the PCA are consistent with the Table II-IV, that is, 250 ($10\times9+32\times5$), 350 ($10\times15+20\times5$), 255 ($5\times21+30\times5$) for the three datasets. Besides, to compare the computational overhead of different principal components, the processing time (PCA and segmentation) and memory requirements also are employed to evaluate their computational overhead.

In order to analysis the influence of spatial size, the size of subhyperspectral cube h and tokens size p on classification results are discussed. The candidate interval of h and p are set to [12:4:32] and [2:2:8], respectively. Then, an ablation experiment is designed to verify the contribution of different modules. In order to only analyze the performance of the MAT model in the supervised mode, we randomly selected fixed number of labeled samples for training. Specifically, for the parameter analysis of MAT-only model, 5%, 10%, and 15% labeled samples per class were randomly selected form the PaviaU, Houston2013 and YRE to explore the parameter sensitivity of MAT model in supervised mode. It is noted that the ASSAL is removed in the ablation experiment and parameter analysis of MAT, which aims to explore the MAT-only

TABLE II

CLASSIFICATION PERFORMANCE OF THE EIGHT APPROACHES ON THE UNIVERSITY OF PAVIA

Class Name	SVM	3D-CNN	HybridSN	HSI-BERT	DBDA(Mish)	A^2S^2K-ResNet	SGL	SSAL	FAAL	MAT	MAT-ASSAL
Asphalt	82.42±0.41	84.25±4.63	69.22±2.16	62.36±1.68	95.70±1.07	89.04±2.58	92.47±3.32	83.25±1.39	89.76±4.69	92.52±0.41	**99.25±0.19**
Meadows	84.59±1.02	88.23±3.79	91.83±1.60	96.92±2.73	97.38±1.90	95.22±4.57	96.82±1.26	90.56±1.93	93.99±2.77	99.26±0.27	**99.40±0.47**
Gravel	0.00±0.00	49.73±7.30	47.69±1.92	45.50±0.84	85.78±5.59	94.27±2.65	86.27±5.28	76.09±5.12	86.14±2.67	72.53±5.68	**99.07±0.15**
Trees	47.66±24.88	91.17±3.44	65.07±6.02	90.09±1.86	95.49±0.96	91.49±2.05	**99.58±0.35**	89.18±2.80	89.36±8.75	92.78±1.36	95.69±2.02
Painted metal sheets	97.75±0.92	97.10±2.22	95.49±2.01	98.58±1.04	99.65±0.26	97.75±1.89	**99.75±0.32**	99.43±0.15	99.41±0.52	96.71±2.35	99.44±0.50
Bare Soil	27.07±2.49	51.02±6.35	81.87±7.49	82.61±4.34	93.06±0.32	**99.79±0.13**	97.93±1.03	86.38±2.73	90.33±7.33	98.36±0.63	99.04±1.25
Bitumen	0.00±0.00	54.42±7.98	33.22±1.93	83.22±1.95	92.22±6.09	**99.92±0.41**	95.96±5.08	90.34±4.99	98.77±0.85	80.96±5.86	99.11±1.02
Self-Blocking Bricks	73.77±0.75	68.39±5.44	54.39±4.32	82.50±3.86	92.95±2.02	83.95±1.36	82.87± 4.39	72.02±2.83	82.38±3.02	86.51±2.75	**98.57±0.76**
Shadows	**99.83±0.03**	90.97±8.02	10.64±7.48	90.55±2.07	98.75±0.53	96.21±2.03	98.79±1.65	73.38±3.90	97.11±0.70	89.81±4.14	97.84±1.19
OA(%)	74.69±1.18	79.75±4.01	78.11±1.62	85.11±2.49	95.29±1.83	94.62±1.96	94.75±0.49	86.06±0.82	91.28±2.08	94.50±0.37	**98.95±0.58**
AA(%)	57.01±3.39	75.03±6.79	61.05±3.82	81.37±2.17	94.55±3.19	95.29±3.42	94.46±0.86	84.51±2.88	91.92±3.47	91.95±0.41	**98.60±0.84**
Kappa(×100)	64.16±2.09	72.91±4.79	70.09±2.31	80.36±2.46	93.72±2.53	92.97±2.48	93.00±0.66	84.12±0.98	89.65±2.63	92.70±0.49	**98.60±0.78**

TABLE III

CLASSIFICATION PERFORMANCE OF THE EIGHT APPROACHES ON THE HOUSTON2013

Class Name	SVM	3D-CNN	HybridSN	HSI-BERT	DBDA(Mish)	A^2S^2K-ResNet	SGL	SSAL	FAAL	MAT	MAT-ASSAL
Healthy grass	86.64±4.11	87.70±3.98	87.75±2.07	95.39±2.73	90.73±2.57	88.84±5.75	93.38±4.67	89.35±1.33	87.41±1.68	86.87±2.02	**94.43±1.60**
Stressed grass	85.62±1.72	87.52±4.45	89.93±1.89	90.87±1.08	93.04±3.73	76.42±4.61	**98.59±1.56**	91.66±0.31	84.95±3.62	88.16±2.90	94.19±1.68
Synthetic grass	0.00±0.00	84.22±7.52	97.89±1.42	98.20±2.42	99.39±0.70	99.62±3.89	**99.86±0.27**	55.19±5.73	47.24±1.87	94.38±3.19	99.43±0.72
Trees	91.28±1.37	89.07±5.17	83.12±3.86	91.43±3.07	94.71±1.62	80.10±7.35	97.46±2.78	93.62±1.23	90.54±4.57	92.65±1.07	**98.91±0.49**
Soil	88.12±3.38	89.42±1.62	97.26±1.61	98.76±1.51	96.73±2.13	96.19±1.43	96.31±2.90	99.07±0.79	94.61±3.76	96.12±2.81	**99.18±0.26**
Water	88.09±3.63	68.72±14.10	89.20±4.91	61.14±5.78	88.42±2.25	90.73±5.21	99.04±0.38	75.49±17.25	89.85±6.71	85.84±3.24	**99.30±0.62**
Residential	50.83±4.69	75.06±14.10	71.69±3.77	77.80±0.83	87.07±4.15	78.58±5.65	92.73±3.36	88.17±1.74	87.28±4.44	83.42±6.58	**98.56±0.33**
Commercial	43.66±8.54	75.06±10.12	79.37±2.99	82.65±5.46	83.10±6.51	65.92±5.93	95.57±4.92	73.80±7.08	70.20±7.07	77.92±11.93	**97.47±0.94**
Road	57.85±2.04	65.95±8.84	68.98±4.50	82.78±4.08	75.23±13.99	69.44±4.13	89.54±5.82	86.11± 2.89	83.10±7.48	82.69±4.83	**97.74±0.35**
Highway	17.28±12.05	60.05± 17.56	84.48±3.98	73.83±1.25	65.64±13.48	96.25±3.76	80.49±7.62	76.80±1.02	67.64±4.10	79.68±3.26	**98.84±0.35**
Railway	45.95±1.11	61.89±13.42	85.56±4.35	82.66±0.17	74.18±14.82	89.22±5.24	90.49±36.75	85.48±3.12	81.35±3.67	87.91±4.98	**98.66±0.63**
Parking Lot 1	23.90±6.86	61.83±13.84	85.01±2.93	64.79±2.49	70.66±14.13	80.25±5.63	9035±5.03	91.11±0.96	79.04±10.63	78.00±7.80	**97.99±0.77**
Parking Lot 2	5.01±6.47	42.47±14.57	58.30±11.77	66.40±0.41	77.51±10.49	88.14±5.65	86.65±3.19	91.248±4.85	83.29±09.17	79.79±9.24	**98.80±0.48**
Tennis Court	78.16±6.40	83.33±10.55	98.54±1.65	83.46±1.35	98.54±1.33	98.94±0.94	97.88±2.16	98.01±1.51	96.99±1.97	97.71±0.79	**99.64±0.05**
Running Track	97.19±1.00	92.39±8.92	96.74±1.60	98.08±1.87	98.90±0.70	98.72±0.04	**98.89±1.40**	89.54±2.37	83.11±5.82	92.60±3.02	87.97±1.04
OA(%)	60.96±1.83	76.01±8.02	84.73±1.11	84.11±1.65	85.23±4.39	84.44±0.85	92.76±1.49	85.24±1.33	84.56±3.19	86.39±3.10	**97.26±1.29**
AA(%)	67.31±4.23	74.98±11.25	84.92±3.55	83.29±1.40	86.26±6.17	86.49±0.51	93.87±1.07	85.64±3.68	81.77±4.10	86.92±4.51	**98.98±0.62**
Kappa(×100)	67.63±1.99	74.03±8.92	83.48±1.21	82.82±1.32	84.02±4.76	83.21±0.93	92.17±1.61	84.10±1.43	78.94±3.42	85.29±3.36	**96.92±0.91**

TABLE IV

CLASSIFICATION PERFORMANCE OF THE EIGHT APPROACHES ON THE YRE

Class Name	SVM	3D-CNN	HybridSN	HSI-BERT	DBDA(Mish)	A^2S^2K-ResNet	SGL	SSAL	FAAL	MAT	MAT-ASSAL
Salt marsh	0.00±0.00	87.82±6.42	81.09± 8.33	87.4±3.99	95.71±1.78	97.03±2.55	90.24±2.59	90.11±4.93	90.48±2.30	91.12±1.01	**99.09±0.57**
Acquaculture	77.38±2.06	98.09±2.46	94.08±2.16	**100.00±0.00**	99.74±0.05	97.35±2.48	97.36±4.81	98.62±3.61	99.71±0.38	99.61±0.35	98.72±1.81
Mud flat	51.96±0.19	77.92±6.93	74.78±9.61	82.69±1.08	**96.27±3.40**	82.47±10.69	85.92±9.67	85.94±2.72	73.41±22.23	75.93±12.68	92.16±1.43
Rice	58.64±20.06	82.42±5.71	78.14±6.63	87.78±0.77	90.08±3.53	85.36±5.38	79.89±5.21	85.91±3.12	81.48±6.22	83.29±3.70	**95.01±1.12**
Aquatic vegetation	20.07±5.42	83.28±13.59	30.53±14.14	56.41±2.05	78.59±24.67	94.29±5.82	83.94±7.56	80.83±10.46	72.16±2.59	70.52±3.50	**97.80±1.79**
Seep sea	34.60±17.71	76.87±28.63	71.69±17.35	61.54±3.62	**98.21±2.65**	96.87±4.40	75.36±10.17	80.91±8.12	71.82±10.35	82.81±7.87	95.66±4.65
Freshwater herbaceous marsh	0.00±0.00	78.59±16.47	83.34±9.42	85.56±4.61	96.18±3.67	**99.89±0.34**	92.53±5.72	82.81±2.48	71.95±10.50	86.42±5.10	97.93±1.65
Shallow sea	90.70±0.75	89.26±29.76	91.04±2.73	99.50±0.25	99.17±1.15	94.93±4.71	92.86±3.91	98.50±1.12	97.23±3.72	96.73±1.16	**100.00±0.00**
Reed	83.42±11.30	97.88±3.13	94.82±4.29	95.15±4.96	99.03±2.75	97.55±2.36	99.81±0.32	97.56±2.47	97.62±2.51	96.69±1.54	**99.83±0.24**
Pond	79.41±0.96	92.88±5.73	84.52±3.91	**99.32±0.13**	97.74±1.48	93.89±2.95	89.99±4.97	90.75±4.12	95.84±2.07	96.26±1.49	97.84±1.61
Building	88.49±2.47	97.14±1.32	89.66±4.37	99.03±0.52	99.43±0.44	95.97±3.33	81.70±5.65	98.78±1.51	98.12±1.24	98.82±0.97	**99.79±0.30**
Suaeda salsa	98.32±0.49	98.21±3.98	97.46±1.71	100.00±0.00	100.00±0.00	99.46±0.78	99.34±1.00	99.50±0.50	99.81±0.35	99.47±0.47	**100.00±0.00**
Flood plain	87.28±17.26	96.29±5.39	96.41±2.44	100.00±0.00	99.45±0.89	95.39±2.88	89.54±2.89	97.95±1.96	96.10±1.97	98.09±1.99	**100.00±0.00**
River	68.67±7.96	80.93±10.16	64.30±6.05	76.31±1.96	87.23±4.05	78.85±12.34	70.61±5.40	83.35±3.61	82.79±2.63	81.30±3.91	**91.47±1.93**
Soybean	99.96±0.07	97.99±1.57	90.37±2.53	99.64±0.11	99.42±0.39	95.93±3.61	97.59±1.37	99.52±0.30	99.64±0.41	99.07±0.25	**100.00±0.00**
Broomcorn	89.31±2.57	94.51±4.19	91.07±4.57	97.45±4.76	99.14±1.65	98.78±1.14	88.33±4.54	98.82±1.12	98.60±1.17	98.05±0.62	**99.89±0.16**
Maize	79.03±2.41	70.94±22.04	57.99±9.23	82.54±3.96	87.69±4.31	80.44±11.76	79.92±5.93	83.62±5.83	78.90±6.11	80.38±1.73	**89.37±1.64**
Locust	0.00±0.00	70.62±31.49	74.31±4.85	89.94±3.50	93.28±3.05	90.12±7.40	83.86±12.38	81.05±1.35	85.45±5.18	86.08±4.97	**96.46±1.90**
Spartina	0.00±0.00	54.66±31.97	23.88±18.65	64.06±3.25	83.91±4.53	76.63±12.79	**95.54±2.29**	63.54±1.23	46.93±23.41	55.64±19.99	91.39±4.24
Tamarix	0.00±0.00	53.14±18.74	39.53±12.82	69.12±1.76	79.48±10.47	80.23±19.24	87.83±4.13	79.74±7.96	6.23±21.28	85.73±3.65	**99.01±0.70**
Intertidal saltwater	0.00±0.00	50.91±28.54	36.60±22.40	70.27±5.27	75.37±23.48	92.79±10.57	1.00±0.00	90.49±6.35	23.62±7.14	92.35 ±5.62	**99.52±0.68**
OA(%)	76.45±1.22	90.83±3.76	85.86±2.19	94.20±2.95	96.66±6.17	93.83±0.70	90.08±1.06	88.53±3.03	93.07±0.80	94.10±1.21	**98.15±1.01**
AA(%)	51.87±5.24	82.39±13.25	73.60±2.08	86.17±1.99	93.01±4.69	91.63±1.28	88.68±8.07	87.12±2.51	81.80±2.81	89.30±1.04	**97.07±0.84**
Kappa(× 100)	74.11±1.36	90.04±4.09	84.53±2.41	93.70±2.09	96.38±0.67	93.30±0.77	89.25±1.49	88.98±3.81	92.47±0.87	93.59±1.02	**97.99±0.94**

performance in supervised mode. Finally, to evaluate the performance of different methods with different sample sizes, the total sample sizes of {0.5%, 1%, 3%, 5%, 10%} are set for comparison on three datasets. Same as the previous settings, the {0.5%, 1%, 3%, 5%, 10%} samples per class are randomly selected for non-AL based methods. As for the AL-based methods, the N_{total} is set as the integer corresponding to the total number of sample sizes and the settings still obey the $N_{total} = N_i \times N_c + K \times N$. To unify the settings, the N_i is set to 5 (5 samples per class) and the number of iteration K is set to 5. When the $\{N_{total}, N_i, N_c, K\}$ are given, the corresponding number of actively augmented samples per iteration K will be determined. Notably, 0.5% and 1% sample sizes of YRE only include 32 and 74 total samples, so we set the $N_i = 1$ and $K = 2$ for AL-based methods in these two cases. To avoid biased estimation, all experiments were conducted with ten independent tests, and the average values were reported for all the evaluation metrics.

C. Comparison With State-of-the-Art Methods

The quantitative results of different methods are reported in Table II-IV. Compared with other methods, it can be seen that the proposed method achieves the best performance in terms of the OA, AA and the Kappa coefficient. Among

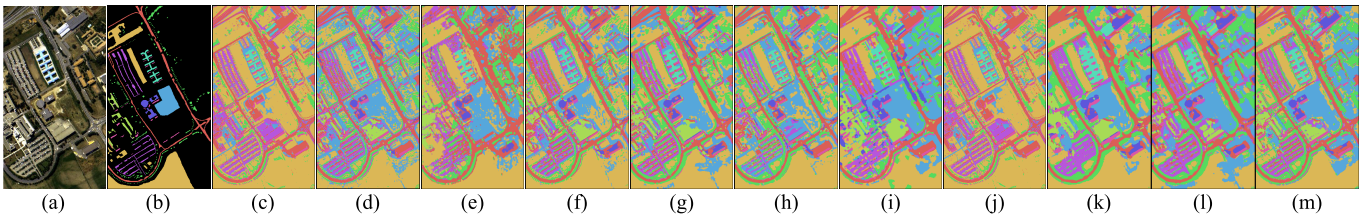


Fig. 5. Classification maps obtained by different classification methods for University of Pavia. (a) Three-band false color composite. (b) Ground truth. (c) SVM. (d) 3D-CNN. (e) HybridSN. (f) HSI-BERT. (g) DBDA. (h) A^2S^2K-ResNet. (i) SGL. (j) SSAL. (k) FAAL. (l) MAT. (m) MAT-ASSAL.

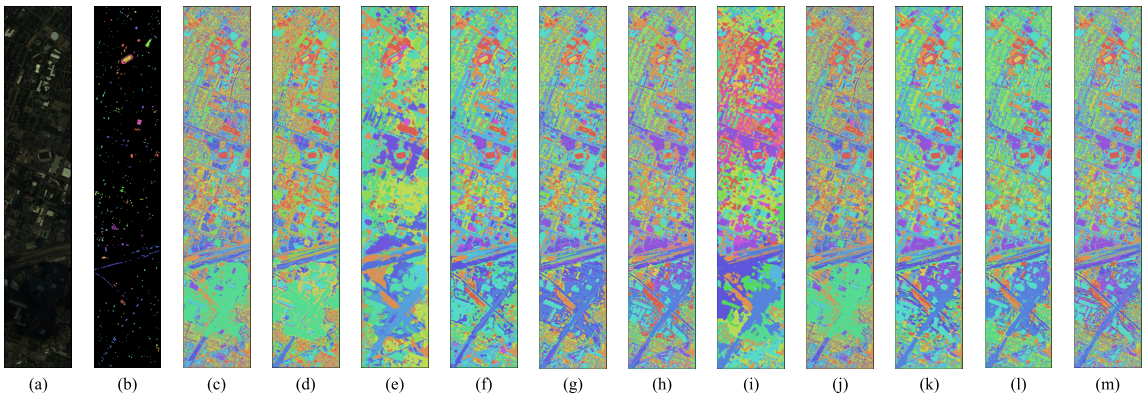


Fig. 6. Classification maps obtained by different classification methods for Houston2013. (a) Three-band false color composite. (b) Ground truth. (c) SVM. (d) 3D-CNN. (e) HybridSN. (f) HSI-BERT. (g) DBDA. (h) A^2S^2K-ResNet. (i) SGL. (j) SSAL. (k) FAAL. (l) MAT. (m) MAT-ASSAL.

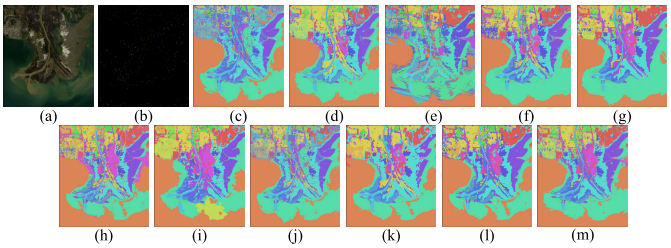


Fig. 7. Classification maps obtained by different classification methods for YRE. (a) Three-band false color composite. (b) Ground truth. (c) SVM. (d) 3D-CNN. (e) HybridSN. (f) HSI-BERT. (g) DBDA. (h) A^2S^2K-ResNet. (i) SGL. (j) SSAL. (k) FAAL. (l) MAT. (m) MAT-ASSAL.

supervised-based methods, SVM, 3D-CNN and HybridSN, are much worse than those of the other methods, indicating that some supervised-based methods may be not adequate for small-sample training. The performance of HSI-BERT is acceptable, but is relatively poor compared with other well-design deep learners due to the spatial-only modeling. DBDA and A^2S^2K-ResNet still achieve the competitive performance even with the limited labeled sample especially the DBDA, which is befit from their excellent designs of backbones. But the standard deviation of DBDA yields large, indicated that the DBDA may be unstable. Notably, in the supervised mode, the MAT-only model did not achieve the best performance, but the performance of MAT is very close to the best model. Especially on the challenging Houston2013 dataset, MAT obtains the best accuracy compared with other supervised deep models, which can be inferred that this design based on the local and long-range contextual relationship modeling is helpful to learn the discriminative spectral-spatial representations.

Compared with the supervised-based methods, the semi-supervised SGL can achieve high accuracy, but its' classification performance is limited when the sample is relatively sufficient (i.e., YRE). Compared with two methods based on active learning SSAL and FAAL, MAT-ASSAL has also achieved significant improvement in classification accuracy particularly for the dataset which contains the complex distribution of land-cover. It is noteworthy that the MAT-ASSAL significantly boosts the performance of the MAT, which can be inferred that the significant improvements benefit from the proposed active learning based on adaptively superpixel segmentation. That is, ASSAL is more capable of selecting the informative samples and enlarging them with high-confidence pseudo labels. It is noteworthy that three experimental datasets are with more complex spatial distribution rather than the simpler Salinas and Indian Pines, and the OA of the proposed MAT-ASSAL exceeds the counterpart of the second highest method 3.66%, 4.53% and 1.49% on the University of Pavia, Houston2013 and YRE, respectively. Furthermore, the standard deviation of DBDA is small, indicating that the performance of MAT-ASSAL is stable, which may benefit from that the model-adaptive informative samples selected by model-free ASSAL are helpful to stability training of models. The quantitative results indicate that the MAT-ASSAL is superior to other methods in HSI classification, especially in the case of a small number of training sets and datasets with complex spatial distribution.

Visual classification maps of the proposed MAT-ASSAL and compared methods on the three datasets are shown in Fig. 8-11. The traditional methods based on statistical machine learning usually learned the pixel-level pattern, but the deep learning based methods tend to model the relationship

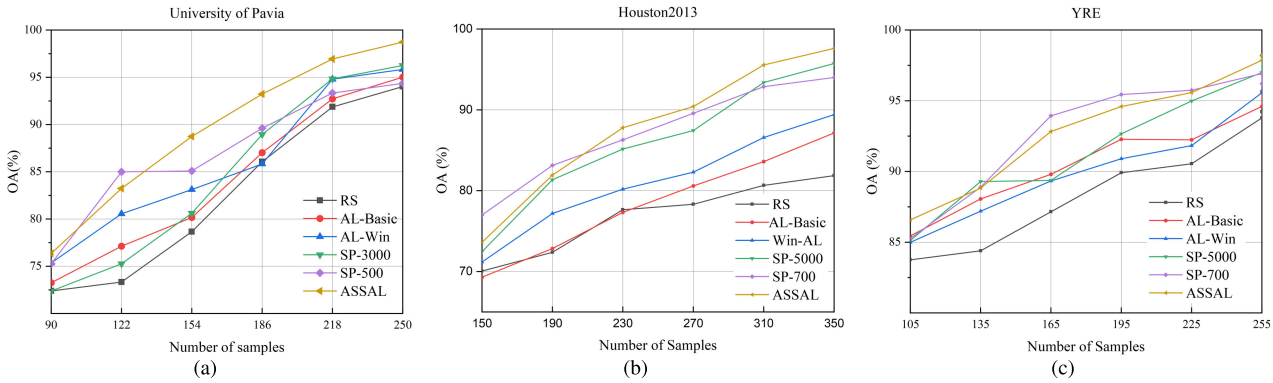


Fig. 8. OA of six AL methods versus different numbers of labeled samples. (a) University of Pavia. (b) Houston2013. (c) YRE.

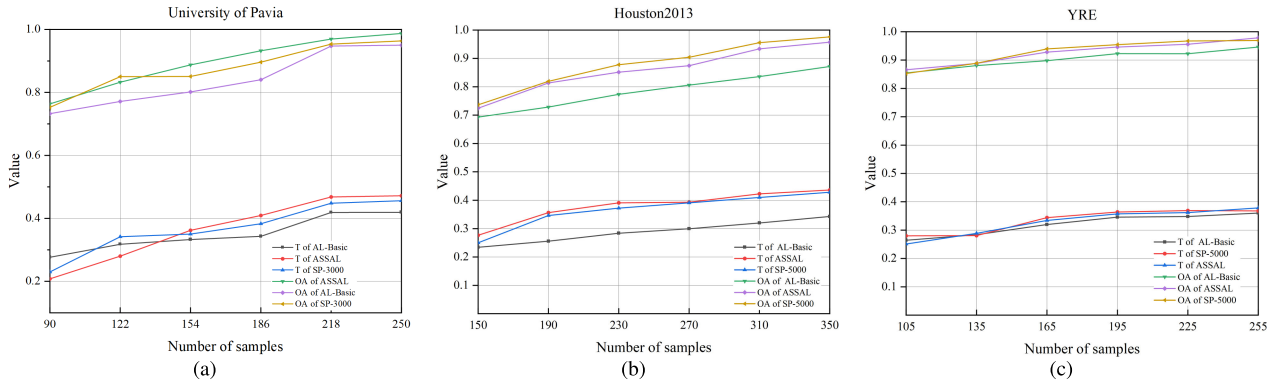


Fig. 9. T and OA of three AL methods versus different numbers of labeled samples. (a) University of Pavia. (b) Houston2013. (c) YRE.

between group-level (patch-level) spatial-spectral features. Consequently, although the classification accuracy of SVM is low, SVM retains the clear edge of an HSI due to the reason that SVM can process each pixel of hyperspectral image directly. A similar situation also appears in the results of the SSAL, because it adopted the multinomial logistic regression (MLR) as the classifier. However, the classification maps of these classifiers based on pixel-level pattern statistics are prone to exist salt and pepper noise (the dissimilarity in local spatial), which are not consistent with real HSIs. The SGL, HybridSN and HSI-BERT cannot retain the clear edge although they have acceptable classification accuracy. In general, DBDA and A^2S^2K-ResNet can achieve good performance, but in some complex regions, such as the right region of houston, their results are not satisfactory. Compared with the above method, the proposed MAT-ASSAL not only achieves good performance on quantitative indicators, but also has the least noise and the clearest object boundary. Strikingly, MAT is employed as our backbone, but the qualitative performance of MAT-ASSAL are not limited by the MAT (i.e., clearer classification maps than MAT). The classification results of the proposed MAT-ASSAL are close to the ground truth maps on three datasets and retain more local details. Thus, there is a reasonable prospect that the model-agnostic ASSAL can effectively boosting the performance of model and the application of outlook-attention module may be helpful to retain the local details.

Notably, the supervised model was trained on randomly sampled subsets of overall HSI datasets, and the AL-based method employed active sampling based on some query

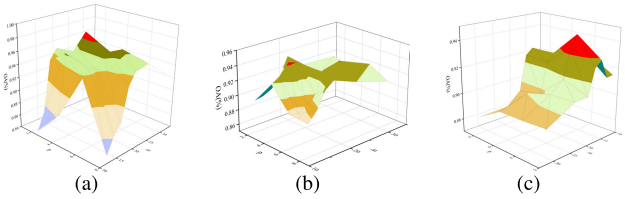


Fig. 10. Sensitivity analysis of parameter h and p for the MAT on three HSI datasets. (a) University of Pavia. (b) Houston2013. (c) YRE. Red regions represent the optimal parameter settings, and the deep brown regions contain the acceptably sub-optimal parameter settings, which suggests that the MAT-ASSAL is not sensitive to the h and p.

criteria. However, since there may be potential overlap between the training and testing datasets, the model will implicitly utilize or observe the testing distribution during training [67], [68], which may result in overfitting to the dataset and learning a shortcut solution. In fully supervised mode, models may exhibit significant performance drops when the testing data is strictly separated from the training data (i.e., disjoint datasets). This is mainly due to the unobservability of testing data during training and the existence of slight distribution shifts between them, which poses a greater challenge for supervised HSIC models. However, MAT-ASSAL mainly focuses on the classification task under relatively worse cases (e.g., with limited labeling conditions), and we aim at improving the model's global fitting ability by actively sampling a small number of samples. When the labeling conditions are limited, the results of the above experiments suggest that the performance of the semi-supervised SGL and AL-based methods, particularly our MAT-ASSAL, is significantly superior to other methods.

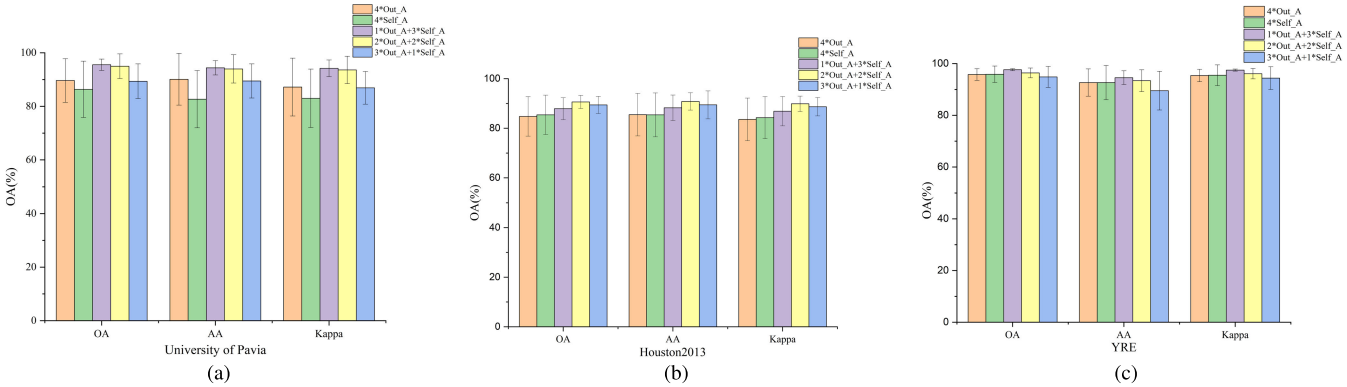


Fig. 11. Influence of different modules on three datasets. (a) University of Pavia. (b) Houston2013. (c) YRE.

D. Analysis of the ASSAL

The classification accuracies obtained by different active learning strategies are reported in Figure 8. It can be observed from the three graphs that the selection strategy based on the BT criterion exceeds the random selection RS, indicating that active learning can select better training samples for the classifier thanks to its sample query rule. The accuracy of AL methods with spatial constraint surpasses those of AL methods without local spatial assumption. It can be inferred that AL integrated the local spatial similarity can achieve better performance. Thanks to the irregular spatial similarity, the AL methods based on superpixel segmentation present higher accuracy and faster convergence compared with the window-based AL. However, the two non-adaptive SP-based AL methods with different scales have their own defects. For the SP-based AL method of large scale, the rate of convergence is faster than small scale, e.g., on Pavia dataset, SP-700 obtains roughly 85% OA when the number of samples is 122, but the SP-3000 can achieve 85% OA only when the number of samples exceed 186. On the contrary, the final accuracy of the SP-based AL method with suitable scale is usually higher than the SP-based AL method with large scale, when the training samples are sufficient. Two conclusions can be drawn from the results of non-adaptive superpixel based methods, the first is that the suboptimal scales may lead the poor results, and the another is that fixed scales even with the optimal settings are hard to achieve the best performance compared with ASS. Overall, the results on three datasets show that the ASSAL methods has the highest classification accuracy. These results suggest that the proposed ASSAL method can be well applied to select the important samples for the model and provide the more reliable spatial constraints for AL, so as to effectively solve the problem of low classification accuracy under the condition of small samples.

To verify the guidance of active learning for informative sample selection, we chose AL-Basic, small-scale SP-based AL, and ASSAL to plot the mean-value curves of BT score (e.g., $T = mean(BT_{map})$) and OA after each iteration. Fig. 5 visualizes the results on three datasets. As can be seen from the trend of curves, T is positively related to the corresponding classification result to some degree. Moreover, a long upward trend exists similarly on the OAs and T, and the T of highest OA is also the maximum. Hence, it is reliable to adopt AL based on the BT criterion to guide the selection of unlabeled samples.

TABLE V

INFLUENCE OF PCA ON CLASSIFICATION RESULTS AND THE CORRESPONDING COMPUTATIONAL OVERHEAD

Dataset	PCs (bands)	OA(%)	Times (CPU / GPU)	memory need
University of Pavia	Original HSI (103)	**98.89±0.66**	**171.42s / 13.12s**	**14.80G**
	99.8% PCs (11)	98.65±0.77	65.13s / 10.68s	3.25G
	95% PCs (3)	98.59±0.84	54.93s / 10.38s	2.87G
Houston2013	Original HSI (144)	97.06±1.87	**863.11s / 97.36s**	**56.69G**
	99.8% PCs (7)	97.23±1.03	178.65s / 30.82s	6.24G
	95% PCs (3)	**97.26±1.18**	162.28s / 28.97s	4.78G
YRE	99.8% PCs (13)	**98.15±0.94**	**512.58s / 95.37s**	**12.10G**
	95% PCs (3)	98.09±1.22	428.57s / 85.96s	8.47G
	Pseudo-color (20,38,59)	97.98±1.18	399.48s / 51.41s	8.47G

E. Parameter Analysis

To better explore the effect and function of the MAT classifier, the size of subhyperspectral cubes h and the size of tokens p are detailed analyzed in this section. Furthermore, an ablation experiment is designed to evaluate the contribution of the outlook attention (Out_A) module and the self attention (Self_A) module for the classification results.

1) Effect of PCA: The experimental results are shown in TABLE V, which suggest that different principal components have little effect on the final classification results due to 0.3% difference is negligible in HSIC. Even employing the segmentation map of pseudo-color image, the final classification accuracy is still promising. These results are expected, since PCA only affects superpixel segmentation results (spatial constraints on active learning strategy). The features extraction of the MAT model is not affected by PCA, because that its input is original hyperspectral data. However, it can be seen from the computational overhead in the Table V that segmentation process adopted the original HSI is unfriendly to implemented platform. Especially for the Houston2013 and YRE datasets, their memory requirements without PCA may be unaffordable. Notably, the memory need of original HSI of Houston2013 dataset is out of memory of our platform (56.69G≥32G), so we split it into two parts for processing and then merge them. Therefore, we generally tend to retain 99.8% principal components, which will bring negligible information loss and relatively stable classification results. More importantly, it will bring small and affordable computational overhead.

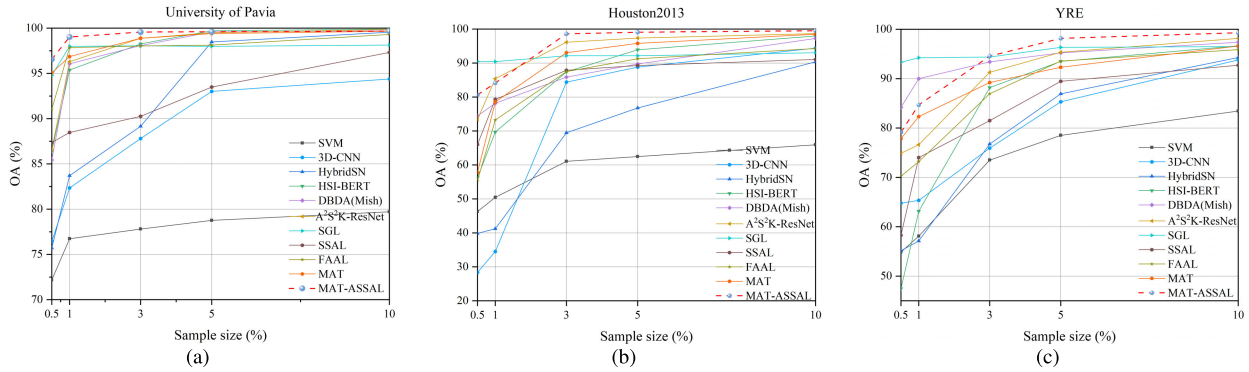


Fig. 12. The results of ten ten comparison methods and our MAT-ASSAL under different sample sizes on the three datasets. (a) University of Pavia. (b) Houston2013. (c) YRE.

TABLE VI
DETAILED SETTINGS OF DIFFERENT MODULES

Module	Outlooker block	Transformer block	Outlook-attention head	Self-attention head
Out_A	×4	-	×4	-
Self_A	-	×4	-	×4
3*Out_A+1*Self_A	×3	×1	×4	×4
2*Out_A+2*Self_A	×2	×2	×4	×4
1*Out_A+3*Self_A	×1	×3	×4	×4

2) Effect of h and p: For the MAT learner, the size of subhyperspectral cubes h and the size of tokens p are jointly analyzed. The results of sensitivity analysis on three datasets are reported in Fig.12. For the university of Pavia dataset, the MAT has the highest accuracy 98.873% when the $h = 18$ and the $p = 4$. In addition, with the increase of h, the OA of MAT shows a significant upward trend, indicating that the larger-size subhyperspectral cubes can provide better spatial information for self-attention module. For the size of tokens p, the MAT yields better performance when the p is smaller. But when the p continues to increase, OA begins to show a downward trend. Thus, it can be inferred that the entire token may lose similarity of local features when p is too large. For the Houston2013 dataset which contains complex distribution of land-cover, the MAT yields the highest OA 94.608% when the $p = 4$ and $h = 16$. It can be inferred that the smaller reception field can help MAT to extract finer-lever contextual features when the spatial distribution of the hyperspectral data is relatively complex. Finally, the MAT achieves the highest OA 94.453% when the $p = 6$ and $h = 18$ on YRE dataset, because that YRE contains not only relatively simple land covers (e.g., seep sea and shallow sea) but also complex buildings. Overall, for the datasets with complex spatial distribution, the smaller p will achieve better performance. In contrast, for the dataset with relatively simple ground object distribution, a larger p may lead to better results.

3) Contribution of the Outlook-Attention Module and Self-Attention Module: To better display the effect of two key components of the MAT classifier for HSIC, an ablation experiment is designed to quantify the Out_A and Self_A modules. The experimental results on three datasets are presented in Fig.4. From the results, it can be seen that the networks combining Out_A and Self_A achieves the highest accuracy and the minimum standard deviation on three datasets. If we only utilized Out_A modules to extract local features or only adopted Self_A modules to extract global features, the final accuracy and standard deviation are significantly worse than 'Self_A+Out_A'. Therefore, it can be inferred that single Out_A module or Self_A module cannot gain best classification performance even if they have the same number of attention heads and MLP layers as 'Self_A+Out_A'. That is, both of the local features from local window and the long-range contextual features from larger reception field are important for hyperspectral image classification. It is noteworthy that setting different number of blocks also affects the final classification results for the 'Self_A+Out_A' networks. More specifically, for PaviaU and YRE datasets which has simple spatial distribution, 1 Out_A module with 3 Self_A modules outperforms the other combinations. For the spatial complex dataset Houston2013, 2 Out_A modules with 2 Self_A modules yield the best performance. It can be demonstrated that datasets with complex spatial distributions require more Out_A layers to extract local features. Quantitative results on the PaviaU dataset present that the OA of '1∗Out_A+3∗Self_A' surpasses other two models, '3∗Out_A+1∗Self_A' and '2∗Out_A+2∗Self_A', 6.163% and 1.554% respectively. Besides, '1∗Out_A+3∗Self_A' outperforms '4∗Out_A and 4∗Self_A' 9.191%, 6.371% on the PaviaU dataset, which can demonstrat that combination Out_A and Self_A are of great help to improve classification accuracy. It can be noted that the combination of '2∗Out_A+2∗Self_A' achieves the highest accuracy on the Houston2013 dataset, but '2∗Out_A+2∗Self_A' only surpass '1∗Out_A+3∗Self_A' 0.289%. There is a reasonable prospect that the combination of '1∗Out_A+3∗Self_A' usually can achieve excellent performance. To sum up, adopting Out_A and Self_A simultaneously in the proposed method for HSIC can achieve excellent performance, and the '1∗Out_A+3∗Self_A' is more capable of capturing beneficial features for classification than other combinations.

4) Experiments With Different Sample Sizes: The results of different methods with sample sizes on three datasets are illustrated in Fig.12. It can be seen that the MAT-ASSAL converges faster to high OA and requires fewer samples. Especially for the University of Pavia and Houston2013 datasets, the MAT-ASSAL requires only 3% of the sample size to achieve ≥99.00% classification accuracy. However, the performance of MAT-ASSAL is not satisfactory when the sample sizes are

0.5% and 1% on YRE dataset, mainly because the ASSAL will not contribute to the overall framework if the sample sizes are too small (e.g., one or two-shot). When the sizes of training samples are sufficient, the deep learning based methods such as DBDA, A^2S^2K-ResNet and so on, achieving the promising results. Notably, the semi-supervised method SGL performed excellently when the sample sizes are very small, as the number of samples increases, however, the performance of SGL did not boost as expected. Overall, compared with supervised and AL-based methods, MAT-ASSAL can achieve excellent performance with fewer samples. Furthermore, unlike some semi-supervised methods with performance saturation, MAT-ASSAL is scalability.

IV. Conclusion

In this paper, a framework combined multi-attention Transformer and adaptive superpixel segmentation (ASS)-based active learning (MAT-ASSAL) is proposed for hyperspectral image classification. Motivated by the advantage of sequential reception fields of the Transformer backbone, the multi-attention Transformer (MAT) is built. Specifically, the self-attention of Transformer is adopted to model long-range contextual dependencies between spectral-spatial embeddings. To capture local features of HSIs, a outlook-attention module is used to encode finer-level features and contexts into tokens. Moreover, to train an excellent model with limited labeled samples, active learning is adopted to select the most important training samples for the MAT leaner. Furthermore, inspired by the Bayesian adaptive superpixel segmentation that can merge uninformative SPs in simple regions and split under-segmentation SPs in complex region, an active learning (AL) method based on adaptive superpixel segmentation (ASSAL) is proposed to locally integrate irregular spatial similarity into active learning. The ASSAL can save SPs in uninformative regions and preserve edge details in complex regions, helping the AL select the most informative samples and generate pseudo-labeled samples with high confidence.

Extensive experiments and analysis suggest that the proposed MAT-ASSAL outperforms the state-of-the-art CNN-based, CNN-AL-based and self-attention-based methods. By adopting ASSAL to train MAT, the MAT can obtain excellent and robust prediction performance even if the number of labeled samples is limited. In the future, we will endeavor to extend the MAT-ASSAL to other remote sensing tasks. Furthermore, although MAT-ASSAL achieves remarkable performance under limited annotation conditions, we will attempt to improve the generalization performance of the MAT-ASSAL on datasets which training data slightly differ from the testing data.

References

[1] S. Yang, Z. Shi, and W. Tang, "Robust hyperspectral image target detection using an inequality constraint," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3389–3404, Jun. 2015.

[2] W. Li and Q. Du, "Collaborative representation for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1463–1474, Mar. 2015.

[3] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[4] J. C. Harsanyi and C.-I. Chang, "Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 4, pp. 779–785, Jul. 1994.

[5] N. He et al., "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 755–769, Feb. 2019.

[6] D. Haboudane, "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture," *Remote Sens. Environ.*, vol. 90, no. 3, pp. 337–352, Apr. 2004.

[7] D. R. A. D. Almeida et al., "Monitoring restored tropical forest diversity and structure through UAV-borne hyperspectral and LiDAR fusion," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112582.

[8] X. Kang, S. Li, L. Fang, M. Li, and J. A. Benediktsson, "Extended random walker-based classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 144–153, Jan. 2015.

[9] Z. Shao, H. Fu, D. Li, O. Altan, and T. Cheng, "Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation," *Remote Sens. Environ.*, vol. 232, Oct. 2019, Art. no. 111338.

[10] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[11] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.

[12] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H^2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112012.

[13] B. Rasti et al., "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020.

[14] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.

[15] A. Ben Hamida, A. Benoit, P. Lambert, and C. Ben Amar, "3D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

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

[17] C. Zhao, W. Zhu, and S. Feng, "Hyperspectral image classification based on kernel-guided deformable convolution and double-window joint bilateral filter," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5506505.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[19] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, "FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5612–5626, Aug. 2020.

[20] Q. Zhu et al., "A spectral–spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 11709–11723, Nov. 2022.

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

[22] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.

[23] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon et al., Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–10.

[24] Q. Shi, X. Tang, T. Yang, R. Liu, and L. Zhang, "Hyperspectral image denoising using a 3D attention denoising network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10348–10363, Dec. 2021.

[25] S. Li, X. Luo, Q. Wang, L. Li, and J. Yin, "H2AN: Hierarchical homogeneity-attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5509816.

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

[27] Z. Xue, M. Zhang, Y. Liu, and P. Du, "Attention-based second-order pooling network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9600–9615, Nov. 2021.

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

[29] J. Li, Q. Du, Y. Li, and W. Li, "Hyperspectral image classification with imbalanced data based on orthogonal complement subspace projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3838–3851, Jul. 2018.

[30] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[31] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

[32] Z. Zhong, Y. Li, L. Ma, J. Li, and W. 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.

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

[34] Y. Xu, B. Du, and L. Zhang, "Self-attention context network: Addressing the threat of adversarial attacks for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 8671–8685, 2021.

[35] B. Liu, A. Yu, X. Yu, R. Wang, K. Gao, and W. Guo, "Deep multiview learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7758–7772, Sep. 2021.

[36] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.

[37] Z. Zhang, E. Pasolli, and M. M. Crawford, "An adaptive multiview active learning approach for spectral–spatial classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2557–2570, Apr. 2020.

[38] K. Y. Ma and C.-I. Chang, "Iterative training sampling coupled with active learning for semisupervised spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8672–8692, Oct. 2021.

[39] M. Xu, Q. Zhao, and S. Jia, "Multiview spatial–spectral active learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512415.

[40] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new Bayesian approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6440–6461, Nov. 2018.

[41] S. Sun, P. Zhong, H. Xiao, and R. Wang, "Active learning with Gaussian process classifier for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1746–1760, Apr. 2015.

[42] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4604–4616, Jul. 2020.

[43] S. Liu, H. Luo, Y. Tu, Z. He, and J. Li, "Wide contextual residual network with active learning for remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 7145–7148.

[44] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6663–6674, Dec. 2015.

[45] M. S. Kotzagiannidis and C.-B. Schönlieb, "Semi-supervised superpixel-based multi-feature graph learning for hyperspectral image data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703612.

[46] S. Jia, X. Deng, M. Xu, J. Zhou, and X. Jia, "Superpixel-level weighted label propagation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 5077–5091, Jul. 2020.

[47] W. Zhu, C. Zhao, S. Feng, and B. Qin, "Multiscale short and long range graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535815.

[48] H. Zhang, J. Zou, and L. Zhang, "EMS-GCN: An end-to-end mix-hop superpixel-based graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5526116.

[49] Z. Gong, L. Tong, J. Zhou, B. Qian, L. Duan, and C. Xiao, "Superpixel spectral–spatial feature fusion graph convolution network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536216.

[50] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.

[51] P. Sellars, A. I. Aviles-Rivero, and C. Schönlieb, "Superpixel contracted graph-based learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4180–4193, Jun. 2020.

[52] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, 2021.

[53] H. Xu, H. Zhang, and L. Zhang, "A superpixel guided sample selection neural network for handling noisy labels in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9486–9503, Nov. 2021.

[54] S. Jia, X. Deng, J. Zhu, M. Xu, J. Zhou, and X. Jia, "Collaborative representation-based multiscale superpixel fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7770–7784, Oct. 2019.

[55] H. Su, Y. Gao, and Q. Du, "Superpixel-based relaxed collaborative representation with band weighting for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5525416.

[56] C. Zheng, N. Wang, and J. Cui, "Hyperspectral image classification with small training sample size using superpixel-guided training sample enlargement," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7307–7316, Oct. 2019.

[57] S. Yang, J. Hou, Y. Jia, S. Mei, and Q. Du, "Superpixel-guided discriminative low-rank representation of hyperspectral images for classification," *IEEE Trans. Image Process.*, vol. 30, pp. 8823–8835, 2021.

[58] C. Liu, J. Li, and L. He, "Superpixel-based semisupervised active learning for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 357–370, Jan. 2019.

[59] Q. Lu and L. Wei, "Multiscale superpixel-based active learning for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5503405.

[60] R. A. Borsoi, T. Imbiriba, J. C. M. Bermudez, and C. Richard, "A blind multiscale spatial regularization framework for kernel-based spectral unmixing," *IEEE Trans. Image Process.*, vol. 29, pp. 4965–4979, 2020.

[61] Y. Zhou, L. Ju, and S. Wang, "Multiscale superpixels and supervoxels based on hierarchical edge-weighted centroidal Voronoi tessellation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3834–3845, Nov. 2015.

[62] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "VOLO: Vision outlooker for visual recognition," 2021, *arXiv:2106.13112*.

[63] R. Uziel, M. Ronen, and O. Freifeld, "Bayesian adaptive superpixel segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8469–8478.

[64] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 32–42.

[65] G. Wang and P. Ren, "Hyperspectral image classification with feature-oriented adversarial active learning," *Remote Sens.*, vol. 12, no. 23, p. 3879, Nov. 2020.

[66] 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, p. 582, Feb. 2020.

[67] N. Audebert, B. L. Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[68] J. Liang, J. Zhou, Y. Qian, L. Wen, X. Bai, and Y. Gao, "On the sampling strategy for evaluation of spectral–spatial methods in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 862–880, Feb. 2017.

Chunhui Zhao received the B.S. and M.S. degrees from Harbin Engineering University, in 1986 and 1989, respectively, and the Ph.D. degree from the Department of Automatic Measure and Control, Harbin Institute of Technology, in 1998. He was a Postdoctoral Research Fellow with the College of Underwater Acoustical Engineering, Harbin Engineering University. Currently, he is with the College of Information and Communication Engineering, Harbin Engineering University, as a Professor and a Doctoral Supervisor. His research interests include digital signal and image processing, mathematical morphology, and hyperspectral remote sensing image processing. He is a Senior Member of the Chinese Electronics Academy.

Boao Qin (Student Member, IEEE) is currently pursuing the Ph.D. degree with Harbin Engineering University, China.

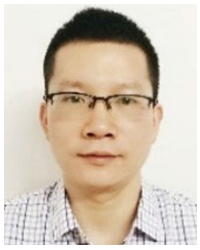
His main research interests include hyperspectral image processing.

Shou Feng (Member, IEEE) received the Ph.D. degree from the Harbin Institute of Technology, China, in 2019.

He is an Associate Professor with Harbin Engineering University, China. His main research interests include remote sensing image processing, data mining, and machine learning.

Wenxiang Zhu (Member, IEEE) is currently pursuing the Ph.D. degree with Harbin Engineering University, Harbin, China.

His research interests include remote sensing image processing and hyperspectral image processing.



Weiwei Sun (Senior Member, IEEE) received the B.S. degree in surveying and mapping and the Ph.D. degree in cartography and geographic information engineering from Tongji University, Shanghai, China, in 2007 and 2013, respectively. From 2011 to 2012, he studied at the Department of Applied Mathematics, University of Maryland, College Park, as a Visiting Scholar with Prof. John Benedetto, to study on the dimensionality reduction of hyperspectral image. From 2014 to 2016, he studied at the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, as a Postdoctoral Researcher to study intelligent processing in hyperspectral imagery. He is currently a Full Professor with Ningbo University, Zhejiang, China. He has published more than 80 journal articles. His current research interests include hyperspectral image processing with machine learning.

Wei Li (Senior Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012. Subsequently, he spent one year as a Postdoctoral Researcher with The University of California at Davis, Davis, CA, USA. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. His research interests include hyperspectral image analysis, pattern recognition, and data compression.

Xiuping Jia (Fellow, IEEE) received the B.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 1982, and the Ph.D. degree in electrical engineering from the University of New South Wales, Canberra, ACT, Australia, in 1996. Since 1988, she has been with the School of Information Technology and Electrical Engineering, University of New South Wales, where she is a Senior Lecturer. She is also a Guest Professor with Harbin Engineering University, Harbin, China; and an Adjunct Researcher with the National Engineering Research Center for Information Technology in Agriculture, Beijing. She is the coauthor of the remote sensing textbook titled *Remote Sensing Digital Image Analysis* (Springer-Verlag, third edition, 1999, and fourth edition, 2006). Her research interests include remote sensing, image processing, and spatial data analysis.