Pyramidal Dilation Attention Convolutional Network With Active and Self-Paced Learning for Hyperspectral Image Classification

Wenhui Hou, Na Chen, Jiangtao Peng, *Senior Member, IEEE*, Weiwei Sun, *Senior Member, IEEE*, and Qian Du, *Fellow, IEEE*

***Abstract*—In recent years, deep neural networks have been widely used for hyperspectral image (HSI) classification and have shown excellent performance using numerous labeled samples. The acquisition of HSI labels is usually based on the field investigation, which is expensive and time consuming. Hence, the available labels are usually limited, which affects the efficiency of deep HSI classification methods. To improve the classification performance while reducing the labeling cost, this article proposes a semisupervised deep learning (DL) method for HSI classification, named pyramidal dilation attention convolutional network with active and self-paced learning (PDAC-ASPL), which integrates active learning (AL), self-paced learning (SPL), and DL into a unified framework. First, a densely connected pyramidal dilation attention convolutional network is trained with a limited number of labeled samples. Then, the most informative samples from the unlabeled set are selected by AL and queried real labels, and the highest confidence samples with corresponding pseudo labels are extracted by SPL. Finally, the samples from AL and SPL are added to the training set to retrain the network. Compared with some DL- and AL-based HSI classification methods, our PDAC-ASPL achieves better performance on four HSI datasets.**

***Index Terms*—Active learning (AL), deep learning (DL), hyperspectral image (HSI) classification, self-paced learning (SPL).**

I. Introduction

HYPERSPECTRAL imaging is an important technique in remote sensing that collects the electromagnetic spectrum from visible to near-infrared wavelength ranges and can provide hundreds of narrow spectral band images of the same region for Earth observation. In a hyperspectral image (HSI), each pixel can be considered as a high-dimensional vector with entries corresponding to the spectral reflectance in a particular wavelength [1]. The HSI has the advantage of distinguishing subtle spectral differences and has been widely used in many fields such as precision agriculture, land use mapping, urban planning, etc.

HSI classification is important for HSI analysis and has received much attention in the past few decades. According to previously available works, HSI classification methods can use spectral feature, spatial feature, and spectral–spatial features [2], [3]. Spectral feature is the fundamental characteristic of HSI, and spectral-based methods only use spectral information in the classification process. In the early days of HSI classification research, researchers focused on methods solely based on spectral features and simply performed classification on pixel vectors, such as principal component analysis (PCA) [4], linear discriminant analysis [5], etc. However, spectral-based methods ignore the rich spatial information of HSIs. The spatial information of a pixel mainly reflects the relationship between the pixel and its spatial neighbors, which can greatly improve the robustness of the model [6]. To simultaneously use spectral and spatial features, spectral–spatial-based approaches are proposed. These methods include filter-based method [7], [8], morphological methods [9], composite kernel methods [10], [11], sparse or low-rank representation methods [12], [13], [14], deep learning (DL) methods [3], [15], etc.

Recently, DL has been gradually applied to HSI classification. DL methods, such as convolutional neural network (CNN) [16] and DenseNet [17], can automatically learn deep spectral–spatial features from HSIs and have achieved excellent classification performance. However, DL methods usually need a large number of labeled samples to train the network [6]. In practice, the collection of labeling samples requires human involvement, and the process is labor intensive and costly [18], [19]. Therefore, one of the problems facing DL is the inability of obtaining sufficient labeled samples. To solve this problem, experts and scholars have proposed a series of deep HSI classification methods for small-sample problems, such as data augmentation (DA) strategy [20], [21], [22], lightweight networks [23], [24], etc.

DA is a popular technique to improve the generalization ability of deep neural networks by generating more training samples. Traditional DA strategies, such as translation, clipping, flip, rotation, and adding noise, are utilized to increase both the amount and diversity of samples. In recent years, some new

Manuscript received 19 November 2022; revised 30 December 2022; accepted 13 January 2023. Date of publication 17 January 2023; date of current version 27 January 2023. This work was supported in part by the Natural Science Foundation of Hubei Province under Grant 2021CFA087, in part by the National Natural Science Foundation of China under Grant 42171351, Grant 42122009, and Grant 41971296, and in part by the National Key Research and Development Program of China under Grant 2020YFA0714200. *(Corresponding authors: Jiangtao Peng; Weiwei Sun.)*

Wenhui Hou, Na Chen, and Jiangtao Peng are with the Hubei Key Laboratory of Applied Mathematics, Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China (e-mail: 320645980@qq.com; chenna0407@aliyun.com; pengjt1982@126.com).

Weiwei Sun is with the Department of Geography and Spatial Information Techniques, Ningbo University, Ningbo 315211, China (e-mail: nbsww@outlook.com).

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: du@ece.msstate.edu).

Digital Object Identifier 10.1109/JSTARS.2023.3237566

DA strategies have been proposed. Li et al. [20] proposed a pixel-block-pair (PBP) DA method, which greatly increases the number of training samples while using a deep CNN to extract PBP features and decision fusion for final label assignment. Haut et al. [21] used a random occlusion data augmentation (RODA) method during CNN training to randomly occlude pixels in different rectangular spatial regions in the HSI to generate training images with various levels of occlusion and reduce the risk of overfitting.

Lightweight networks have the advantages of fewer parameters, less computation, and shorter inference time compared with typical deep neural networks, which reduce the dependence of labeled samples by pruning, distillation, and group convolution without degrading performance. In HSI classification, Zhang et al. [23] proposed an end-to-end 3-D lightweight convolutional neural network (3-D-LWnet) and alleviated the small-sample problem by using cross-sensor and cross-modal strategies. The LiteDepthwiseNet proposed by Cui et al. [24] decomposes standard convolution into depthwise convolution and pointwise convolution based on 3-D depthwise convolution and removes the ReLu layer and the batch normalization layer in the original 3-D depthwise convolution, which not only alleviates the overfitting phenomenon of the model on small-sized datasets, but also achieves high classification performance using minimal parameters.

The aforementioned methods have achieved good performance in small-sample classification problems, but still cannot solve the problem of labeled sample scarcity. Recently, active learning (AL) has become a hot research topic. AL selects useful samples for labeling and theoretically guarantees a significant reduction in label usage. It assumes that each sample is of different importance. In other words, only fewer samples are important to the classifier, while others are redundant. In general, AL is used as an auxiliary strategy to select more valuable samples. Moreover, AL is more versatile and can be combined with various classifiers, such as support vector machine (SVM) [25], [26], CNN [27], ResNet [28], generative adversarial networks [29], etc.

Although all these AL methods reduce the use of labeled samples to some extent, the lack of labels still largely limits the classification performance. In order to obtain more labeled samples without consuming more resources, this article expands the training set by self-paced learning (SPL). SPL is a new learning mechanism proposed in recent years that gradually incorporates easy to complex samples into the training by simulating the human learning process [30], [31]. Contrary to AL, SPL prefers to select samples with high confidence, so their combination can obtain better information about the dataset and, thus, improve the classification accuracy.

In order to alleviate the problem of insufficient labeled samples in the training of DL-based HSI classification methods, in this article, using the thoughts of semisupervised learning [32], we integrate the DL, AL, and SPL into a unified framework and propose a pyramidal dilation attention convolutional network with active and self-paced learning (PDAC-ASPL) model. In the DL part, we propose a pyramidal dilation attention convolutional (PDAC) network, which improves the original pyramidal dilation convolutional (PDC) network [33] by incorporating the squeeze-and-excitation (SE) attention modules [34], [35] into different PDC blocks. It can effectively suppress the features of unimportant channels and, meanwhile, allows the adaptive selection of spatial information by focusing on the features of the central pixel and its neighboring pixels. Except for the SE-based attention, other attention mechanisms, such as self-attention mechanism in the transformer model, can also be considered [36], [37]. By embedding the AL and SPL strategies into the PDAC main network, we can select the most informative samples with the highest uncertainty to assign true labels and query the samples with the highest confidence to assign pseudo labels. The labeled informative samples and pseudo-labeled high-confidence samples are added to the training set to refine the model training. The contributions of the proposed PDAC-ASPL are mainly threefold.

1) We integrate the DL, AL, and SPL into a unified framework and propose a PDAC-ASPL model for small-sample HSI classification. The PDAC network can effectively extract spatial–spectral features for classification. AL and SPL strategies can select the most informative samples and high-confidence samples to enlarge the training set for refining model training.
2) A PDAC network is designed for spatial–spectral feature extraction and classification. To make full use of valid information for all PDC blocks and to better use the output information from the last PDC block, an SE attention mechanism is incorporated before the first PDC block and after the last PDC block, respectively.
3) A new SPL strategy is constructed for the selection of high-confidence samples. It can effectively solve the unbalanced class problem by designing a class budget and alleviate the effect of noisy pseudo labels by using a weighted symmetric cross-entropy (SCE) loss.

The rest of this article is assigned as follows. In Section II, the related works of AL and SPL are described. In Section III, the proposed model is presented in detail. In Section IV, a series of experiments are conducted on four HSI benchmark datasets. Section V provides the ablation experiments and parameter analysis. Finally, Section VI concludes this article.

II. Related Work

A. AL Methods

In recent years, AL has been extensively studied in HSI classification. For example, Rajan et al. [38] applied the AL method to single-image classification and knowledge transfer, validating the reasonableness of a maximum-likelihood classifier and a binary hierarchical classifier. Wang et al. [26] used the supervised clustering technique and the labeling process of classification results to discover the representativeness and differentiation of samples and assigned pseudo labels to unlabeled data based on the clustering and classification results. The samples that were not assigned pseudo labels in each iteration are also considered as candidates for AL, and finally, a semisupervised AL method was designed based on the SVM. Xu et al. [25] constructed leave-one-class-out multiviews and designed a sample

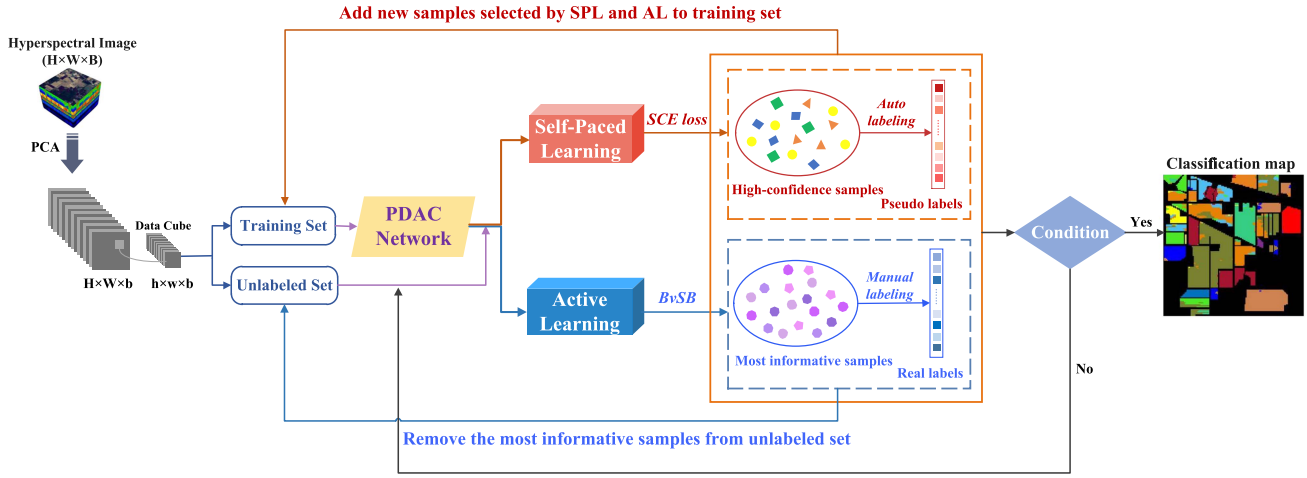


Fig. 1. Framework of PDAC-ASPL. It mainly contains three parts, i.e, PDAC network for deep spatial–spectral feature extraction, SPL for assigning pseudo labels to high-confidence samples (the selected pseudo-labeled samples are added to the training set), and AL for picking up the most informative samples from the unlabeled set (the selected informative samples are added to the training set and meanwhile removed from the unlabeled set).

query strategy from the perspective of classification confidence and training contribution. The most inconsistent high-quality samples are filtered out by making full use of the iterative prediction information and spatial–spectral features of the HSI. Then, the target samples are obtained by AL in each iteration through two-layer screening, and SVM is used to obtain the final classification results. Cao et al. [27] integrated AL and DL into a unified framework, in which a CNN is trained using a limited number of labeled pixels, and the most informative pixels from the candidate pool are selected by AL and then added to the original training set to fine-tune the CNN. Ding et al. [39] proposed a clustering-inspired AL method, which selects highly informative and diverse samples from unlabeled samples in the candidate set by fast search and finding of peaks clustering methods for manual labeling and pretrains the CNN by the k-means clustering-based pseudo-labeling scheme.

Although the aforementioned works have achieved better classification results, they have not fundamentally solved the lack of labeled samples. Therefore, we combine AL with SPL to solve this problem at the root.

B. Self-Paced Learning

SPL has been previously used in the field of HSI classification. Peng et al. [31] combined SPL with sparse representation to construct a self-paced joint sparse representation model for HSI classification. It learns the weights of neighboring pixels using SPL and selects neighboring pixels with nonzero weights (i.e., easy pixels) to be added to the JSR learning process in each iteration. Yang et al. [40] proposed an SPL-based probability subspace projection (SL-PSP) method for HSI classification. After assigning a probability label to each pixel and a risk to each labeled pixel, the two regularizers are developed in SL-PSP for classification from an SPL maximum marginal and probability label graph, respectively.

Recently, there have been some studies combining SPL with AL in other fields, such as in computer vision, Lin et al. [41] first proposed to combine AL with SPL for face recognition. Subsequently, Ren et al. [42] applied AL and SPL with DL to synthetic aperture radar automatic target recognition. In each of these fields, the combination of SPL and AL has yielded more desirable results.

III. Proposed Method

The overall framework of the proposed PDAC-ASPL is shown in Fig. 1, which combines AL and SPL with DL to achieve better results with limited labeled samples. The input hyperspectral dataset D is first preprocessed by the PCA to reduce its dimensionality from B to b. Then, a densely connected PDAC network is trained with a limited number of labeled samples and outputs features for a labeled training set and an unlabeled set. Based on the features, the AL strategy picks up the most informative samples from the unlabeled set and assigns them real labels, and SPL selects high-confidence samples and assigns them pseudo labels. The samples selected by AL and SPL are added to the training set to retrain the network.

A. PDAC Network

For feature extraction, a densely connected PDAC network is designed, as shown in Fig. 2(a). In the PDAC network, all the layers in the dense convolutional network are directly connected to ensure the maximum transmission of information [33], and the dilation convolution is used to integrate the multiscale context information of the HSI.

The main structure of the PDAC network is the PDC block, as shown in Fig. 2(b). It consists of several PDC layers, and dense connections are adopted between different PDC layers. The PDC layers are composed of dilated convolution layers with different dilated factors as [33]

$$N_k = n_1^1 \wedge n_2^2 \wedge n_3^4 \wedge \cdots \wedge n_k^d \tag{1}$$

where N_k represents the kth PDC layer, n_k^d indicates the kth subdilated convolutional layer with dilated factor $d = 2^{k-1}$ in

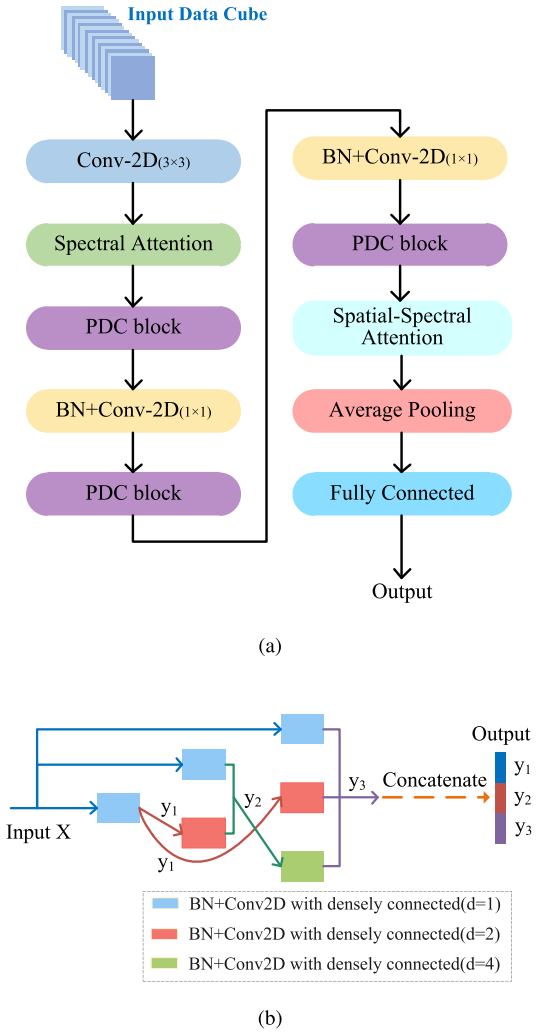


Fig. 2. PDAC network. (a) Framework of the PDAC network. (b) PDC block.

the kth PDC layer, and $\wedge$ represents the stacking of subdilated convolutional layers. Different skip connections correspond to different dilation factors. In general, a shallow skip connection corresponds to a small dilated factor. The width of the network will increase as the number of PDC layers increases. The advantage of the structure is that more and larger ranges of spatial information can be obtained [33]. In this article, our network uses three PDC layers in each PDC block for feature extraction.

To make full use of valid information for all the PDC blocks and to better use the output information from the last PDC block, an SE attention mechanism is incorporated before the first PDC block and after the last PDC block [34], [35], respectively.

1) Spectral Attention Module: Considering that the importance of different channels is different, an SE attention is first imposed on spectral feature channels, as shown in Fig. 3. In detail, for a feature map of $h \times w \times b$ from the first 2-D convolution layer of the network, a global average pooling (pooling size is $h \times w$) is used to generate a feature map of $1 \times 1 \times b$. Then, it passes through two fully connected layers, where the number of neurons in the first fully connected layer is $b/16$ and that in the second fully connected layer is b. Next,

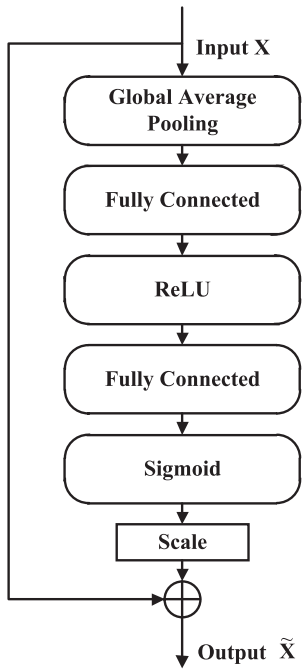


Fig. 3. Spectral attention module.

the feature map of $1 \times 1 \times b$ is obtained through the sigmoid layer. Finally, the original feature $h \times w \times b$ and the generated attention map of $1 \times 1 \times b$ are fully multiplied to obtain the feature map with different channel importance.

2) Spatial–Spectral Attention Module: For the feature map $\mathbf{h}$ of size $h \times w \times b$ output by the third PDC layer, we generate a 1-D spectral attention map M_{se} (size $1 \times 1 \times b$) and a 2-D spatial attention map M_{sa} (size $h \times w \times 1$).

Here, the spectral attention is slightly different from the spectral attention in the last subsection. It concatenates a global maximum pooling with the global average pooling and uses the combination of output results as input to the next layer for the purpose of better complementing the global information. The spectral attention can be expressed as

$$\boldsymbol{h}'_{\boldsymbol{k}} = F_{\text{scale}}(\boldsymbol{h}_{\boldsymbol{k}}, s_k) = s_k \cdot \boldsymbol{h}_{\boldsymbol{k}}, \ k = 1, \ldots, b \tag{2}$$

where s_k is the sum of the average pooling s_k^{avg} and maximum pooling $s_k^{\max}$, and $F_{\text{scale}}(\boldsymbol{h}_{\boldsymbol{k}}, s_k)$ refers to spectralwise multiplication between the feature map $\boldsymbol{h}_{\boldsymbol{k}}$ and the scalar s_k.

For spatial attention, we first perform global average pooling and maximum pooling on input features and combine the results of both pooling and input them into the convolution layer and then use the sigmoid activation function to obtain the output $\boldsymbol{sa}$:

$$\boldsymbol{sa} = \text{sigm}\left([\boldsymbol{sa}^{\text{avg}}, \boldsymbol{sa}^{\max}] * \boldsymbol{W}\right) \tag{3}$$

where $*$ is the convolution operation and $\boldsymbol{W}$ is a learnable built-in parameter.

The final output is $\boldsymbol{h''} = \boldsymbol{h'} \star \boldsymbol{sa}$, where $\boldsymbol{h'} = [\boldsymbol{h'_1}, \boldsymbol{h'_2}, \ldots, \boldsymbol{h'_b}]$, and $\star$ represents the spatialwise multiplication operation between the feature map $\boldsymbol{h}'_{i,j}$ (size $1 \times 1 \times b$) and the scalar $\boldsymbol{sa}_{i,j}$.

B. SPL for Selecting High-Confidence Samples

SPL is a learning mechanism that borrows the idea from human learning from simple to complex [30]. The traditional SPL is limited to select "simple" samples with high confidence derived from the network. It may cause the problem that "simple" samples are from the same class and eventually result in an overfitting phenomenon. Here, we set a class budget for sample selection and design a new loss to alleviate the effect of noisy pseudo labels in SPL.

After predicting the unlabeled samples by the PDAC network, for each sample $\mathbf{x}_i$, a predicted one-hot label $\hat{\mathbf{y}}_i$ can be obtained. SPL assigns a weight v_i to each sample $\mathbf{x}_i$ through the following optimization model:

$$\min_{\mathbf{w},\mathbf{v}} E(\mathbf{w},\mathbf{v};\lambda)=\sum_{i=1}^{n} v_i \ell\left(\hat{\mathbf{y}}_i, f(\mathbf{x}_i,\mathbf{w})\right)-\lambda\sum_{i=1}^{n} v_i$$
$$\text{s.t.}\quad 0\le v_i\le 1,\ i=1,\ldots,n \tag{4}$$

where λ is a parameter, $\mathbf{w}$ is the model parameter, and ℓ is the loss function.

It is easy to obtain the weight v_i as

$$v_i=\begin{cases}1, & \text{if } \ell\left(\hat{\mathbf{y}}_i, f(\mathbf{x}_i,\mathbf{w})\right)<\lambda\\ 0, & \text{otherwise}\end{cases}. \tag{5}$$

The samples whose losses are smaller than λ can be considered as "easy" or high-confidence samples. However, the samples with smaller losses may come from several "easy-to-classify" categories, which may cause the unbalanced classification problem. To avoid selecting too many high-confidence samples from one class, we design a class budget M, i.e., selecting at most M samples in each class. That is, for each class, the selected high-confidence samples should have losses smaller than λ, and the number of selected high-confidence samples is smaller than M.

In the SPL step, we use a new SCE loss function to select high-confidence samples. The general cross-entropy (CE) function is $\ell_{\text{ce}}=-\sum_{k=1}^{K} q(k|\mathbf{x})\log p(k|\mathbf{x})$, where $p(k|\mathbf{x})$ and $q(k|\mathbf{x})$ are the predictive and true probability distributions for sample $\mathbf{x}$, respectively. The CE loss can be intuitively understood as an effort to increase the predicted probability value of the sample corresponding to the label category. Wang et al. [43] indicated that the CE loss is difficult to adjust the effect of noisy labels and proposed an SCE loss as

$$\ell_{\text{sce}}=\ell_{\text{ce}}+\ell_{\text{rce}} \tag{6}$$

where $\ell_{\text{rce}}=-\sum_{k=1}^{K} p(k|\mathbf{x})\log q(k|\mathbf{x})$. In simple terms, ℓ_{rce} is to swap the label and the predicted value. In the SPL step, it is clear that the sample's label $\hat{\mathbf{y}}_i$ is a one-hot pseudo label, so there are noisy labels. To alleviate the effect of noisy labels, we use a new SCE loss function as

$$\ell_{\text{sce}}=\alpha\ell_{\text{ce}}+(1-\alpha)\ell_{\text{rce}} \tag{7}$$

where α is a weight parameter. As ℓ_{rce} plays an auxiliary role in alleviating the effect of noisy labels, the weight α is set as 0.7 in the experiments.

C. AL for Selecting Most Informative Samples

AL is a common machine learning method to query sample information through iterations and then to label some informative samples. It aims to use fewer labels to obtain better learning performance. AL is generally based on the uncertainty criterion and the diversity criterion to select informative samples. Existing AL-based hyperspectral classifiers normally employ off-the-shelf uncertainty-based algorithms [44], such as least confidence [45], entropy sampling [45], best versus Second best (BvSB) [46], Bayesian active learning disagreement [47], etc.

For each sample $\mathbf{x}_i$, the PDAC network produces a vector $\mathbf{z}_i$ of size $K\times 1$, which can be viewed as the class probability matrix of sample $\mathbf{x}_i$. In order to combine the vector $\mathbf{z}_i$ generated by the PDAC network for sample selection by AL, we adopt the BvSB strategy based on the uncertainty criterion for querying.

The BvSB criterion is specifically designed for multiclassification problems; thus, it is well suited for HSI classification. In this criterion, we only need to consider two classes with the highest classification probability. The criterion is defined as

$$\text{BvSB}(\mathbf{z}_i)=P_B(\mathbf{z}_i)-P_{SB}(\mathbf{z}_i) \tag{8}$$

where $P_B(\mathbf{z}_i)$ and $P_{SB}(\mathbf{z}_i)$ denote the highest and second highest class membership probability of unlabeled sample $\mathbf{x}_i$, respectively. For this strategy, a smaller value of $\text{BvSB}(\mathbf{z}_i)$ indicates that the best and second affiliation probabilities are closer. That is, the sample has higher uncertainty; therefore, it will be selected by AL.

D. Whole Training Process

Our proposed PDAC-ASPL method combines DL, AL, and SPL to achieve better results with fewer labeled samples. In the training process, the PCA is first used to reduce the dimensionality of hyperspectral dataset D (size $H\times W\times B$, total K classes) from B to b. Then, data cubes are constructed and divided into a training set T (m labeled samples per class, total Km labeled samples) and an unlabeled set U (n unlabeled samples). In the next, we use Km labeled samples in T for the initial training on the PDAC network and predict n unlabeled samples in U, which eventually outputs a matrix of size $n\times K$ that can be regarded as a probability matrix. For the AL branch, we use the BvSB strategy to select N_{AL} samples with more information, assign them real labels, and add the selected samples with their labels to T, while removing them from U. For the SPL branch, we use the SCE loss function to calculate the loss value of each sample, select the samples with lower loss, and assign them pseudo labels. Meanwhile, we consider the class budget M and finally select N_{SPL} samples ($N_{\text{SPL}}\le KM$) and then add these samples with their pseudo labels to T. At this point, the training set T completes one round of updating and serves as the new training set for the next round. These algorithms execute R rounds and are implemented iteratively together until the termination condition is met.

The entire procedure of PDAC-ASPL is summarized in Algorithm 1.

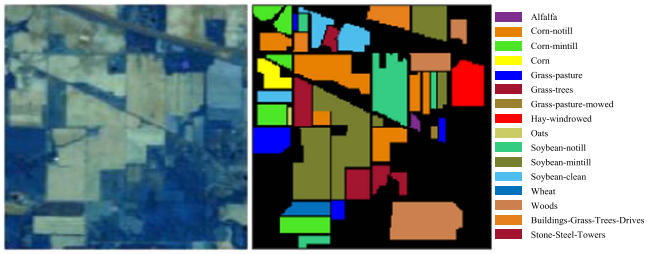


Fig. 4. Pseudo-color composite image and ground-truth map of IP.

Algorithm 1: PDAC-ASPL.

Input: Initial labeled training set T, unlabeled set U, the number of round R, the number of samples selected by AL per round N_{AL}, and the class budget M of SPL.
Output: Predicted label Y.
Begin
1. PCA dimension reduction
Initialization: $r = 1$
While $r < R$ or stopping criterion is not satisfied do:
1) Training PDAC network using samples in T.
2) Calculating the class probability matrix of the samples in U.
3) Selecting N_{AL} samples from U using the BvSB and giving them true labels.
4) Selecting $(N_{SPL} \leq KM)$ samples from U using the SPL and giving them pseudo labels.
5) Adding the selected $N_{\text{AL}} + N_{\text{SPL}}$ samples with N_{AL} true labels and N_{SPL} pseudo labels to T.
6) Removing the N_{AL} samples selected by AL from U.
End

IV. Experiments

A. Datasets

Four hyperspectral datasets are used in the experiments, namely, Indian Pines (IP), University of Pavia (UP), Salinas Valley (SA), and HuangHeKou (HHK).

1) Indian Pines: This dataset was gathered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the IP test site in Northwestern Indiana in 1992. It consists of 145 $\times$ 145 pixels and 220 spectral bands in the wavelength range of 400–2500 nm. Among the pixels, only 10 249 pixels are feature pixels, and the remaining 10 776 pixels are background pixels. The IP scene contains two-third of agricultural land and one-third of forest or other natural perennial vegetation. The ground truth available is designated into 16 classes. The pseudo-color composite image and ground truth map are shown in Fig. 4.

2) University of Pavia: The dataset was acquired by the Reflective Optics System Imaging Spectrometer sensor during a flight campaign over Pavia, Northern Italy, on July 8, 2002. The number of spectral bands is 103 ranging from 430 to 860 nm. The scene has the size of 610 $\times$ 340 with very high spatial resolution of 1.3 m per pixel. There are nine land cover classes. The pseudo-color composite image and ground truth map are shown in Fig. 5.

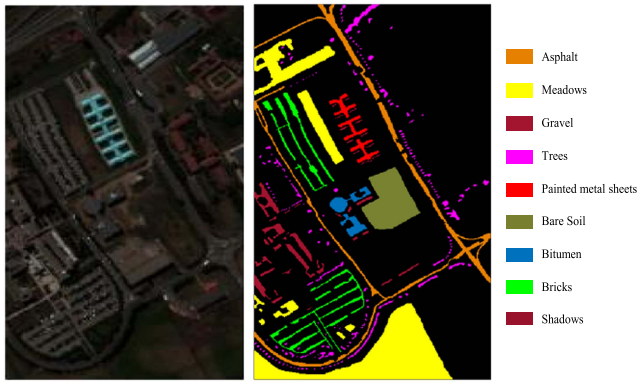


Fig. 5. Pseudo-color composite image and ground-truth map of UP.

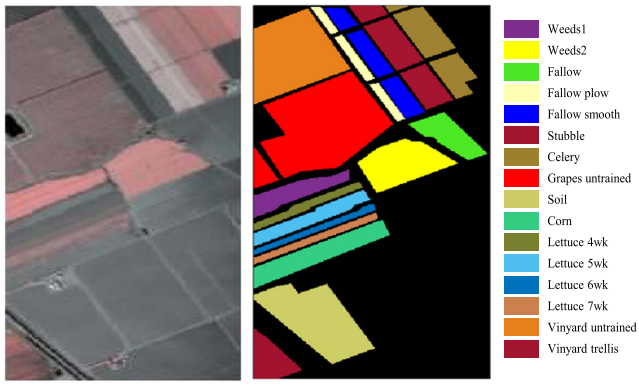


Fig. 6. Pseudo-color composite image and ground-truth map of SA.

3) Salinas (SA): The SA dataset was collected by the AVIRIS sensor in Salinas Valley, CA, USA. The size of the Salinas image is 512 $\times$ 217, and the spatial resolution is 3.7 m. There are 224 spectral bands ranging from 400 to 2500 nm, in which 20 water absorption bands are removed before classification. The dataset contains 16 ground objects and 54 129 labeled pixels. The pseudo-color composite image and ground truth map are shown in Fig. 6.

4) HuangHeKou: The HHK dataset was captured in 2019 by the GF5_AHSI in the area around the Yellow River Estuary ("Huanghekou" in Chinese) in China. The overall image contains 330 spectral bands in the wavelength range of 390–1029 nm (VNIR) and 1005–2513 nm (SWIR). Forty-five bad bands were eliminated and the remaining 285 bands were used for classification. The dataset has the size of 1185 $\times$ 1342 pixels. There are 21 types of materials. The pseudo-color composite image and ground truth map are shown in Fig. 7.

The categories and corresponding number of pixels in four datasets are shown in Tables I–IV.

B. Method Comparison and Parameter Settings

We compare the proposed PDAC-ASPL method with other ten methods for HSI classification on four datasets. For each comparison method, we mostly used the original parameters of the referenced article. The ten compared methods and their parameter settings are listed as follows.

Fig. 7. Pseudo-color composite image and ground-truth map of HHK.

TABLE I
NUMBER OF PIXELS IN THE IP DATASET

No.	Class	Pixels
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93
Total		10249

TABLE II
NUMBER OF PIXELS IN THE UP DATASET

No.	Class	Pixels
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Metal sheets	1345
6	Bare soil	5029
7	Bitumen	1330
8	Bricks	3682
9	Shadows	947
Total		42776

TABLE III
NUMBER OF PIXELS IN THE SA DATASET

No.	Class	Pixels
1	Brocoli green weeds 1	2009
2	Brocoli green weeds 2	3726
3	Fallow	1976
4	Fallow rough plow	1394
5	Fallow smooth	483
6	Stubble	3959
7	Celery	3579
8	Grapes untrained	11271
9	Soil vinyard develop	6203
10	Corn senesced green weeds	3278
11	Lettuce romaine 4wk	1068
12	Lettuce romaine 5wk	1927
13	Lettuce romaine 6wk	916
14	Lettuce romaine 7wk	1071
15	Vinyard untrained	7268
16	Vinyard vertical trellis	1807
Total		54129

1) SVM [48] is one of the representatives of traditional HSI classification algorithms.
2) RODA [21] is a DA method. In this method, the spatial size of data cube is 23×23 and the learning rate is 0.001.
3) 3-D lightweight Siamese network (3DLSN) [49] is a lightweight DL method. In this method, the spatial size of data cube is 7×7 and the learning rate is 0.001.
4) HybridSN [50] is a classical convolutional network combining 3-D CNN and 2-D CNN. In this method, the spatial size of data cube is 25×25 and the learning rate is 0.001.
5) Double-branch multiattention mechanism network (DBMA) [51] is a two-branch spatial–spectral DL classification method based on the attention mechanism. In this method, the spatial size of data cube is 7×7 and the learning rate is 0.0005.
6) Spectral–spatial residual network (SSRN) [52] is a supervised deep residual network. In this method, the spatial size of data cube is 7×7 and the learning rate is 0.0003.
7) Multiview spatial–spectral active learning (MVSS-AL) [25] is a multiview AL-based method that uses SVM as a basic classifier.
8) Feature-oriented adversarial active learning with PCA (FAAL) [29] is an AL method based on adversarial learning. In this method, the spatial size of data cube is 25×25 and the learning rate is 0.001.
9) SpectralFormer (SF) [36] is a transformer-based method. SF is able to learn spectral local sequence information from neighboring bands of the HSI to produce grouped spectral embeddings. In this method, the spatial size of data cube is 9×9 and the learning rate is 0.0005.
10) Spectral–spatial feature tokenization transformer (SSFTT) [37] is a transformer-based method that can capture spectral–spatial features and high-level semantic features. In this method, the spatial size of data cube is 9×9 and the learning rate is 0.001.

TABLE IV
NUMBER OF PIXELS IN THE HHK DATASET

No.	Class	Pixels
1	Salt marsh	393
2	Acquaculture	796
3	Mud flat	110
4	Rice	190
5	Aquatic vegetation	83
6	Seep sea	96
7	Freshwater herbaceous marsh	95
8	Shallow sea	211
9	Reed	200
10	Pond	936
11	Build up	553
12	Suaeda salsa	469
13	Flood plain	361
14	River	240
15	Soybean	595
16	Broomcorn	454
17	Maize	133
18	Locust	377
19	Spartina	68
20	Tamarix	72
21	Intertidal saltwater	39
Total		6471

C. Experimental Settings

In the PDAC-ASPL, the reduced spectral dimension is set to $b = 30$ using the PCA, and the spatial size of data cube is 11×11. The learning rate is 0.001. All the experiments are trained on a computer with a 2.70 GHz and 128-GB RAM CPU and two NVIDIA GeForce RTX 2080Ti GPUs based on Python 3.7 to get the results and computational time.

On the IP dataset, for our method and other two AL-based methods (i.e., MVSS-AL and FAAL), the initial training set contains 160 labeled samples (i.e., ten labeled samples per class), and the top 50 informative samples are selected and assigned true labels in each of the following four rounds of AL. Totally, 360 labeled samples are used for three AL-based methods. For the other eight algorithms (i.e., SVM, RODA, 3DLSN, HybridSN, DBMA, SSRN, SF, and SSFTT), we randomly select 360 labeled samples as the training set and the remaining samples for testing.

On the UP dataset, for our method and other two AL-based methods, the initial training set contains 45 labeled samples (i.e., five labeled samples per class), and the top 50 informative samples are selected and assigned true labels in each of the following four rounds of AL. Totally, 245 labeled samples are used for three AL-based methods. For other eight algorithms, we randomly select 245 labeled samples as the training set and the remaining samples for testing.

On the SA dataset, for three AL-based methods, the initial training set contains 48 labeled samples (i.e., three samples per class), and the top 30 informative samples are selected and assigned true labels in each of the following four rounds of AL. Totally, 168 labeled samples are used for three AL-based methods. For other eight algorithms, we randomly select 168 labeled samples as the training set and the remaining samples for testing.

On the HHK dataset, for three AL-based methods, the initial training set contains 105 labeled samples (i.e., five samples per class), and the top 30 informative samples are selected and assigned true labels in each of the following four rounds of AL. Totally, 225 labeled samples are used for three AL-based methods. For other eight algorithms, we randomly select 225 labeled samples as the training set and the remaining samples for testing.

The overall accuracy (OA), class accuracy (CA), average accuracy (AA), and kappa coefficient (κ) on the testing set are used to evaluate the classification performance of each method. To ensure the stability of the experimental results, we conduct ten random experiments for each method.

D. Classification Results

1) IP Dataset: On the IP dataset, 360 labeled samples are used for each method, and the averaged results are shown in Table V.

From Table V, we can see that our proposed PDAC-ASPL method achieves optimal performance in terms of OA, AA, and κ. Due to the joint use of SPL and AL, the proposed method is likely to select high-confidence samples for all categories, which enlarges the training set to increase the classification performance. Hence, the proposed PDAC-ASPL generates the best averaged accuracy on all categories. In addition, for the "Grass-pasture-mowed" and "Oats" categories with limited labeled samples, PDAC-ASPL and FAAL methods correctly classify all samples, which demonstrates that the AL strategy can select informative samples in these categories to improve the small-sample classification performance.

Fig. 8 visually shows the classification maps of different methods. It can be clearly seen that the classification maps of FAAL and PDAC-ASPL are much more consistent with the ground truth than the map of other methods. Compared with FAAL, our PDAC-ASPL provides much better results in the "Soybean-mintill" category as shown in blackish green color.

2) UP Dataset: On the UP dataset, 245 labeled samples are used for each method. Table VI shows the per class accuracy, OA, AA, and κ of different methods, where the bolded values indicate the best value. It can be seen that our method achieves optimal performance in terms of OA, AA, and κ using only 245 (0.6%) labeled samples. The higher κ coefficient shows that the prediction labels of our PDAC-ASPL are highly consistent with true labels. Compared with two AL-based methods, the proposed PDAC-ASPL shows 19% and 6% performance improvement in OA over MVSS-AL and FAAL, respectively. This demonstrates that both the AL strategy for selecting informative samples and the SPL strategy for selecting high-confidence samples in our PDAC-ASPL are effective. Compared with the classical DL method, such as HybirdSN and SSRN, our PDAC-ASPL also improves the OA by at least 3%. Compared with recently proposed transformer-based methods, the PDAC-ASPL improves the OA by at least 4.5%. For UP data, the categories "Asphalt" and

TABLE V
CLASSIFICATION RESULTS (%) ON THE IP DATASET

Class	SVM	RODA	3DLSN	DBMA	HybridSN	SSRN	MVSS-AL	FAAL	SF	SSFTT	PDAC-ASPL
1	26.72	46.59	17.42	70.80	65.22	**98.25**	83.00	96.08	10.61	68.18	91.64
2	57.87	68.29	59.17	81.77	82.40	80.24	59.59	**84.78**	49.69	82.92	84.45
3	65.26	72.97	48.35	85.32	84.28	77.04	47.09	**89.10**	40.46	83.35	83.03
4	44.78	76.64	26.93	**92.15**	77.92	86.34	51.01	89.26	15.25	69.26	92.03
5	85.24	78.86	68.38	94.82	92.69	**96.11**	85.39	84.86	67.46	83.43	85.03
6	81.56	81.32	95.84	97.25	**99.64**	91.98	88.26	96.74	84.31	95.16	97.18
7	64.71	55.56	11.11	41.75	53.57	33.33	87.04	**100**	9.88	64.20	**100**
8	84.43	72.99	98.84	97.84	96.23	91.57	90.10	**99.61**	88.18	98.34	97.38
9	28.28	44.74	8.77	48.61	65.00	33.33	82.96	**100**	10.52	38.60	**100**
10	62.70	82.25	49.61	79.79	84.16	85.53	59.45	79.73	53.58	**86.53**	83.91
11	67.50	83.05	77.19	85.13	86.03	**94.39**	69.76	86.06	74.57	88.74	92.58
12	54.05	68.44	32.52	77.46	66.55	**86.82**	39.94	74.87	37.16	67.75	77.59
13	86.96	86.62	93.10	**100**	98.37	98.86	95.87	92.70	81.22	96.44	98.50
14	90.74	91.85	90.96	93.92	**97.97**	96.57	88.33	94.27	94.98	97.95	95.44
15	65.43	84.72	43.97	80.60	65.51	**95.95**	41.82	69.46	26.64	79.36	94.97
16	97.50	92.22	77.78	98.91	68.46	98.62	90.31	**100**	31.21	97.74	97.82
OA	70.50	79.31	68.10	86.22	86.22	86.81	68.49	87.00	63.68	87.03	**89.81**
AA	66.48	74.19	56.25	82.89	80.25	84.06	72.49	89.84	48.48	81.12	**91.97**
κ	66.13	76.37	63.14	84.24	84.27	85.07	63.95	85.18	58.08	85.20	**88.35**

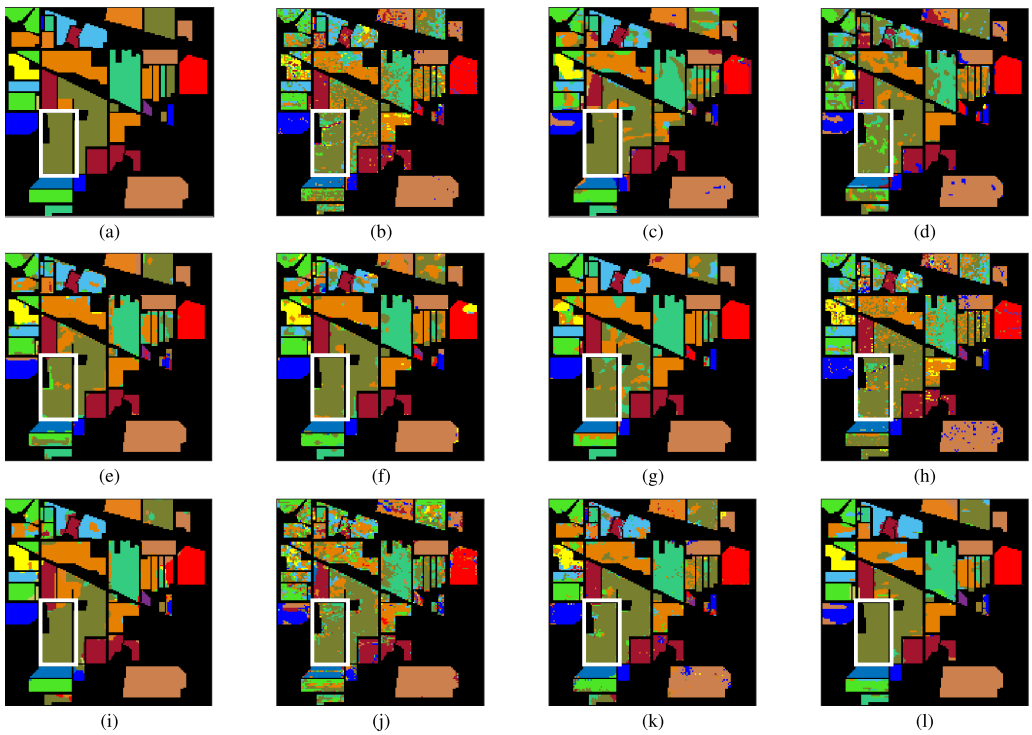


Fig. 8. Classification map on IP. (a) Ground truth. (b) SVM. (c) RODA. (d) 3DLSN. (e) DBMA. (f) HybridSN. (g) SSRN. (h) MVSS-AL. (i) FAAL. (j) SF. (k) SSFTT. (l) PDAC-ASPL.

"Bitumen" are similar materials and are difficult to be classified from each other. MVSS-AL shows poor results on these two categories. FAAL can well classify the "Bitumen" category but performs poor on the "Asphalt" category. In contrast, the proposed PDAC-ASPL provides consistently good results on these two categories.

Fig. 9 visually shows the classification maps of different methods. It can be seen that the map of PDAC-ASPL is more similar to the ground truth map than maps of comparison methods. In particular, for categories "Bare Soil" in blackish green color and "Bricks" in green color, our method shows much better results.

TABLE VI
CLASSIFICATION RESULTS (%) ON THE UP DATASET

Class	SVM	RODA	3DLSN	DBMA	HybridSN	SSRN	MVSS-AL	FAAL	SF	SSFTT	PDAC-ASPL
1	88.46	79.80	91.91	87.60	93.39	95.57	80.43	85.86	69.56	93.38	**98.22**
2	88.89	89.08	92.96	98.30	97.81	98.64	90.76	98.84	93.57	99.22	**99.41**
3	65.84	55.00	78.44	78.90	82.20	**98.65**	64.05	87.56	19.92	76.33	88.81
4	94.60	87.17	83.40	97.72	83.56	**99.82**	86.71	79.45	74.62	87.36	98.05
5	94.80	85.72	98.90	99.65	**100**	99.61	98.98	98.38	64.43	99.88	99.75
6	81.67	56.37	40.53	94.61	87.54	92.35	41.98	96.13	40.13	92.84	**97.95**
7	60.66	61.93	61.40	97.94	59.10	98.22	69.64	**99.27**	16.51	93.67	98.75
8	77.45	86.30	83.87	85.35	56.53	74.43	59.08	81.49	47.04	76.90	**94.07**
9	99.89	82.58	96.11	**99.93**	33.02	99.89	91.18	49.82	53.10	94.12	97.94
OA	85.54	72.32	83.73	93.16	88.01	94.54	78.72	91.99	70.38	93.40	**97.94**
AA	83.58	77.96	80.83	93.33	77.02	95.24	75.87	86.31	53.21	90.41	**96.99**
κ	80.59	84.35	77.99	90.91	83.97	92.77	71.43	89.31	59.52	91.22	**97.26**

TABLE VII
CLASSIFICATION RESULTS (%) ON THE SA DATASET

Class	SVM	RODA	3DLSN	DBMA	HybridSN	SSRN	MVSS-AL	FAAL	SF	SSFTT	PDAC-ASPL
1	97.75	99.08	97.35	93.62	99.05	95.97	98.32	**99.97**	14.50	99.00	99.92
2	98.72	93.99	99.08	98.74	99.70	98.18	95.68	**100**	20.43	98.98	98.93
3	89.44	87.41	81.22	90.58	96.90	90.58	76.30	**100**	19.01	97.78	99.28
4	97.50	97.00	91.61	90.58	73.10	98.28	**98.71**	80.86	18.79	98.37	96.01
5	92.13	89.29	94.93	96.79	90.45	94.87	97.57	**99.83**	15.61	95.73	95.84
6	99.88	90.63	98.82	99.58	99.86	**99.94**	99.22	99.69	35.38	97.94	99.81
7	95.41	92.83	99.66	94.00	**99.86**	98.25	99.40	99.82	28.58	99.35	99.31
8	72.38	74.43	80.98	76.93	86.04	85.54	87.78	93.23	67.50	85.69	**93.63**
9	98.52	90.18	99.72	96.64	99.86	97.33	96.83	**99.94**	83.99	99.58	**99.94**
10	82.65	85.73	84.63	92.42	88.19	96.67	80.71	95.61	30.78	94.43	**97.98**
11	84.21	46.25	84.10	90.74	75.72	95.62	83.03	**99.93**	42.01	99.00	89.25
12	95.21	84.02	98.02	99.39	99.34	87.17	98.45	95.86	40.93	92.56	**99.46**
13	89.01	85.36	98.25	96.71	90.17	85.43	98.46	81.59	34.94	82.33	**99.41**
14	92.08	82.57	92.97	95.66	65.30	**100**	91.09	96.82	18.42	87.09	97.89
15	65.64	69.21	68.92	77.20	75.96	66.42	67.13	82.33	29.45	86.63	**87.84**
16	97.84	35.78	91.25	96.41	88.63	**98.42**	88.17	88.49	13.21	94.58	98.14
OA	86.12	81.51	88.64	88.04	90.05	86.41	85.20	94.50	41.55	93.19	**96.30**
AA	90.52	81.48	91.34	92.87	89.26	93.04	89.26	94.62	32.10	94.31	**97.67**
κ	84.53	79.41	87.33	86.65	88.91	84.96	83.43	93.86	33.93	92.42	**95.87**

3) SA Dataset: On the SA dataset, 168 labeled samples are used for each method. Table VII shows the accuracy per class, OA, AA, and κ of different methods on the SA dataset, where the bolded values indicate the best value. It can be seen that our method shows excellent performance in multiple classes. For categories with similar materials (e.g., categories 11–14 are subclass of "Lettuce"), the proposed PDAC-ASPL provides the best overall results. For categories with similar characteristics (e.g., category 8 "Grape untrained" and category 15 "Vinyard untrained" are spatially adjacent and similar), our PDAC-ASPL also shows the best results on these two categories. The excellent performance of the proposed method in distinguishing pixels with similar materials or characteristics demonstrates that the samples selected by AL and SPL are representative and discriminative.

Fig. 10 visually shows the classification maps, where the proposed PDAC-ASPL provides much better results than comparison methods in the left bottom region of the map (mainly categories 11–14, subclass of "Lettuce"). In addition, in the top left corner of the map (i.e., categories 8 and 15), the results of PDAC-ASPL are more consistent with the ground truth than other methods.

4) HHK Dataset: On the HHK dataset, 225 labeled samples are used for each method. Table VIII shows the accuracy per class, OA, AA, and κ of different methods on the HHK dataset. It can be seen that our method also produces the best overall results. As the spectral resolution of HHK image is very high, some traditional algorithms (e.g., SVM and DBMA) also show good classification performance on this dataset. However, our algorithm has at least 1.42% improvement in OA compared to the more advanced AL algorithms.

5) Effect of Training Rounds: In the experiments, the proposed PDAC-ASPL method is trained in five rounds. To illustrate the accuracy of PDAC-ASPL model at each training round more

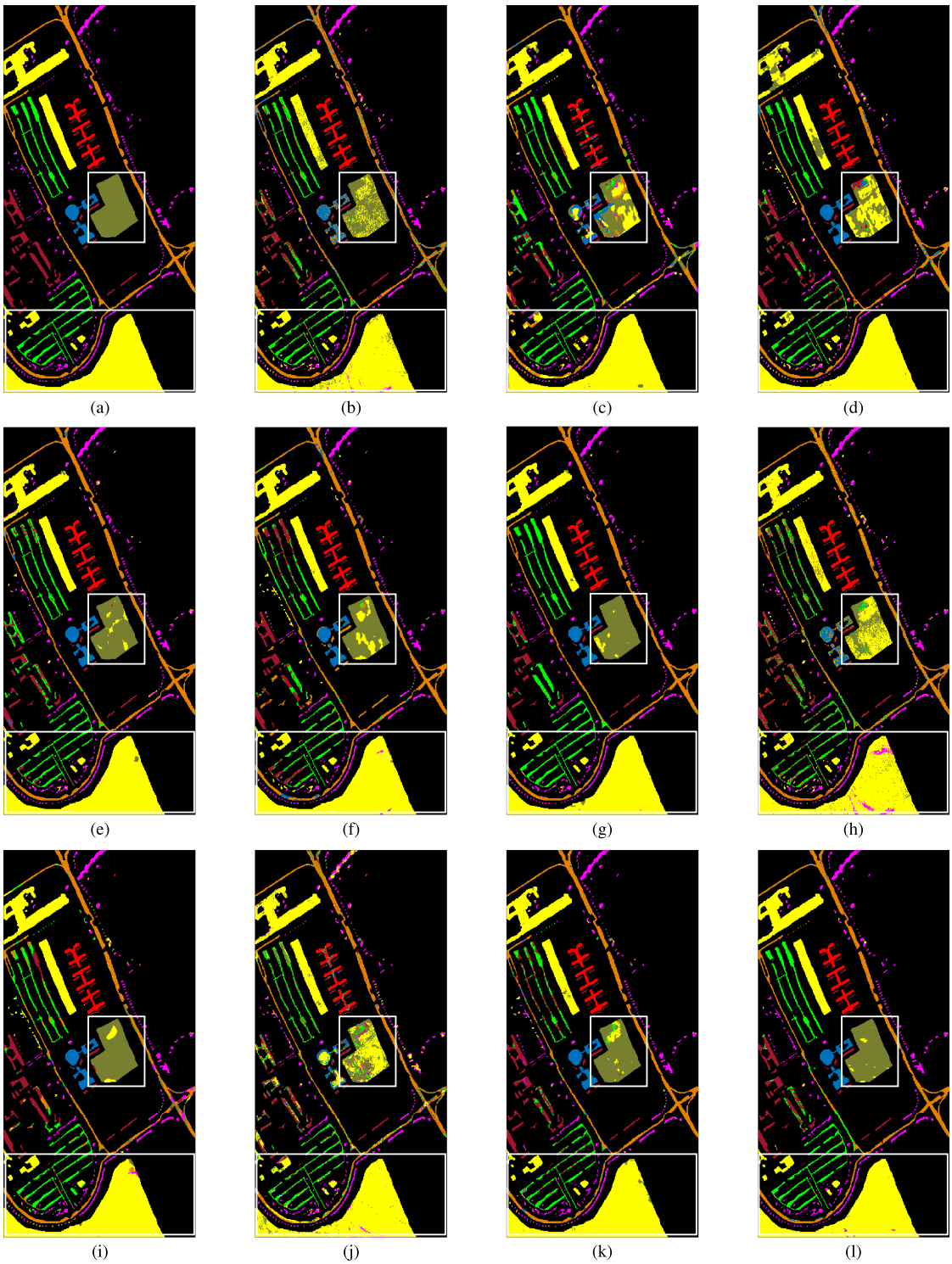


Fig. 9. Classification map on UP. (a) Ground truth. (b) SVM. (c) RODA. (d) 3DLSN. (e) DBMA. (f) HybridSN. (g) SSRN. (h) MVSS-AL. (i) FAAL. (j) SF. (k) SSFTT. (l) PDAC-ASPL.

intuitively, we conduct experiments on four datasets separately, and the results are shown in Fig. 11. It can be seen that the accuracy on four datasets gradually increases as the number of training rounds increases, which indicates that the samples select by AL and SPL in each round are useful. In addition, we can see that the OA increases slowly after four round training because the model is well trained using the selected high-confidence or informative samples.

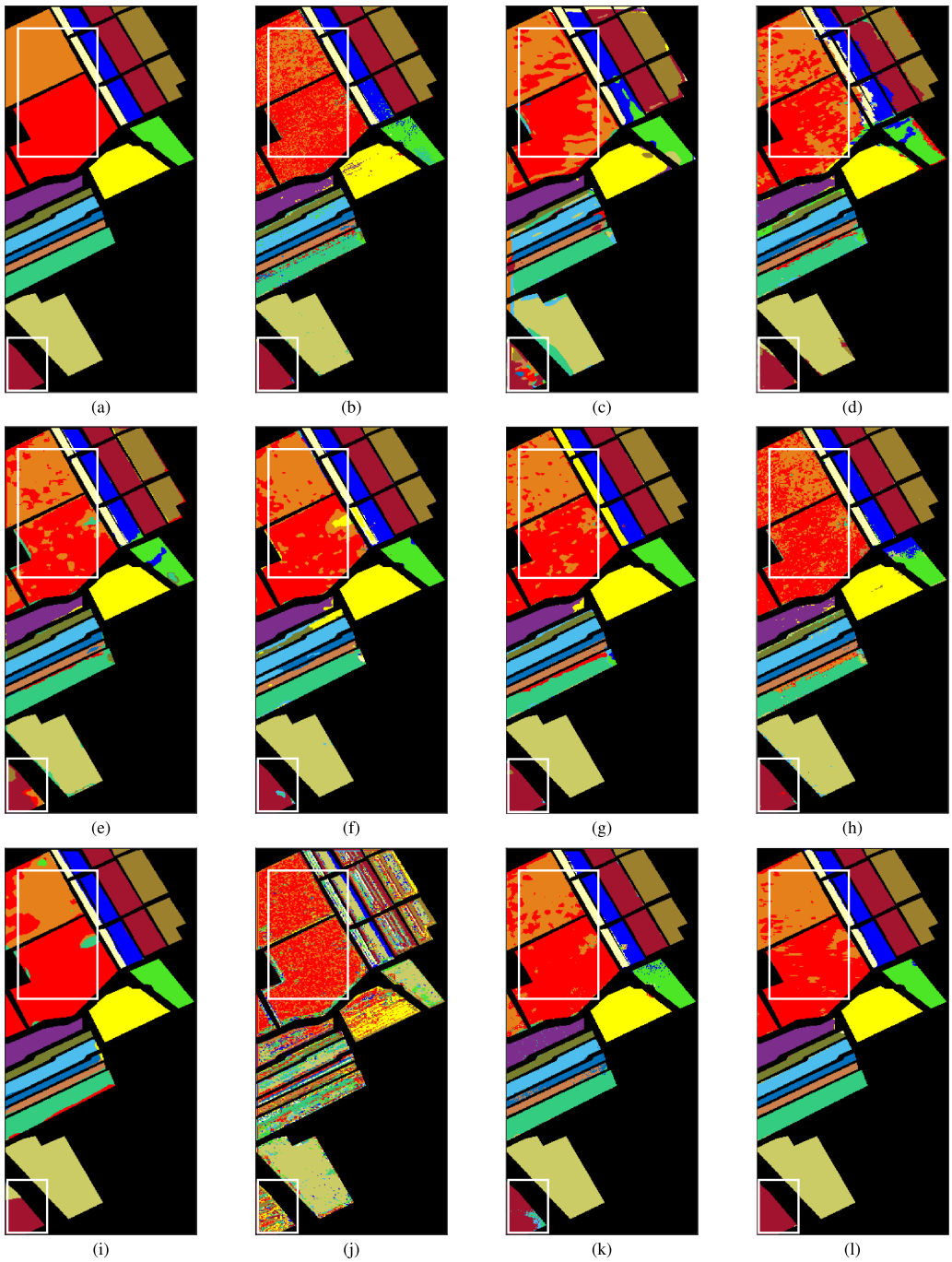


Fig. 10. Classification map on SA. (a) Ground truth. (b) SVM. (c) RODA. (d) 3DLSN. (e) DBMA. (f) HybridSN. (g) SSRN. (h) MVSS-AL. (i) FAAL. (j) SF. (k) SSFTT. (l) PDAC-ASPL.

V. Discussion

A. Computational Time for Each Method

The computational times (in seconds) of the PDAC-ASPL and the other ten comparative methods on the four hyperspectral datasets are shown in Table IX. All machine learning methods were trained on the same device to obtain the computational times. For the eight methods except for the AL-based ones, the time includes model training and testing time. For three AL-based methods, the time refers to the total time of classifier training, testing, and sample selection and iterative training.

According to Table IX, it can be seen that the overall computational time of traditional DL methods is short, and our method runs at least 739.28 s faster than two small-sample HSI classification methods (i.e., RODA and 3DLSN) on four datasets. On the HHK dataset, which has less samples than other datasets, our method is much faster than the existing AL methods MVSS-AL and FAAL. However, since our method needs to calculate the loss between the one-hot label predicted and the

TABLE VIII
CLASSIFICATION RESULTS (%) ON THE HHK DATASET

Class	SVM	RODA	3DLSN	DBMA	HybridSN	SSRN	MVSS-AL	FAAL	SF	SSFTT	PDAC-ASPL
1	91.29	64.64	91.60	**99.55**	91.14	94.25	90.39	63.43	14.16	87.92	94.66
2	**100**	98.50	99.10	**100**	94.07	**100**	**100**	95.64	99.74	99.48	**100**
3	71.47	52.83	92.06	70.96	77.36	59.57	81.77	89.33	8.18	74.12	**94.54**
4	80.21	18.30	88.86	86.88	67.93	85.57	88.55	89.54	5.46	85.38	**90.20**
5	86.13	34.38	31.48	89.19	44.57	93.61	61.60	79.29	3.33	51.86	**95.22**
6	97.09	22.58	87.64	88.78	83.51	99.12	99.44	89.33	11.59	86.04	**99.89**
7	99.52	34.24	91.85	96.71	90.94	**100**	99.43	**100**	14.65	95.93	99.89
8	99.68	55.88	**100**	96.92	98.20	**100**	97.43	93.30	24.30	95.07	96.99
9	**100**	40.16	**100**	98.01	93.26	**100**	99.48	92.47	14.06	97.56	**100**
10	86.12	88.04	**99.34**	96.80	90.94	98.47	95.10	88.83	26.13	96.81	94.40
11	96.21	79.59	99.63	96.71	**99.75**	88.34	98.88	96.53	8.19	99.19	99.40
12	97.43	**100**	**100**	98.81	**100**	97.82	**100**	**100**	7.74	99.41	**100**
13	97.58	49.14	99.72	98.33	97.99	98.38	96.61	95.70	25.57	98.94	**99.80**
14	**87.83**	29.53	79.19	80.68	57.47	86.36	80.13	49.50	15.73	66.01	85.62
15	95.34	90.94	**100**	97.76	96.52	**100**	96.57	93.92	34.90	94.92	97.08
16	96.64	91.78	**100**	97.65	97.72	97.61	99.22	91.89	7.99	96.79	99.53
17	**95.49**	28.12	89.54	66.68	68.99	77.80	91.85	72.02	5.99	82.17	93.92
18	91.41	66.48	84.84	90.08	73.97	74.98	**94.71**	74.43	9.55	82.51	94.33
19	73.17	27.27	81.82	78.37	75.76	91.52	**93.36**	92.45	21.03	86.67	91.37
20	**96.17**	55.80	55.72	70.05	61.90	83.20	86.31	90.35	11.59	77.04	93.98
21	95.93	71.05	47.37	90.11	77.19	66.67	95.59	**100**	7.21	94.57	97.59
OA	93.13	73.02	94.72	94.37	89.58	93.34	95.42	88.43	27.00	92.83	**96.84**
AA	92.13	57.11	86.66	89.95	82.82	90.16	92.69	87.52	17.96	88.02	**96.12**
κ	92.52	70.62	94.25	93.89	88.68	92.78	95.03	87.45	19.13	92.21	**96.57**

TABLE IX
COMPUTATIONAL TIMES (IN SECONDS) OF 11 METHODS ON FOUR DATASETS

Method	SVM	RODA	3DLSN	DBMA	HybirdSN	SSRN	MVSS-AL	FAAL	SF	SSFTT	PDAC-ASPL
IP	17.34	2193.93	1488.68	85.07	107.01	360.01	77.95	980.19	217.47	10.37	749.40
UP	8.72	2342.87	2342.83	33.20	35.09	434.81	316.50	841.90	72.96	9.37	1432.11
SA	10.97	2420.19	3669.61	47.87	34.93	393.44	313.02	757.68	128.79	8.07	1590.77
HHK	6.49	2174.15	4411.21	56.84	24.58	494.29	29812.16	870.35	9944.55	7.82	661.34

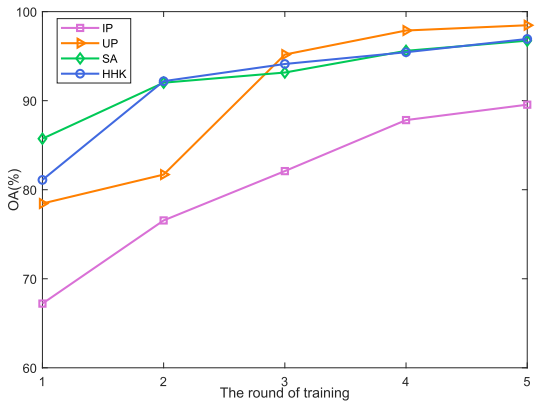


Fig. 11. OA versus the number of training rounds on four datasets.

vector output from the network for each sample and the BvSB values, our method takes longer time for the datasets UP and SA with large data volume.

TABLE X
ABLATION EXPERIMENT ON THE IP DATASET

Module	OA
PDC	81.67
PDAC (PDC+Att)	86.33
PDAC+AL	87.91
PDAC+AL+SPL	89.81

B. Ablation Experiment

The proposed PDAC-ASPL contains three main modules, i.e., PDAC, AL, and SPL modules. The PDAC modifies the original PDC by embedding the SE-based spectral and spatial attention mechanism (Att) into different PDC blocks. To verify the validity of each module in the proposed PDAC-ASPL, we conduct experiment on IP dataset and show results in Table X.

For the PDC-based classification, we randomly select 360 labeled samples as the training set and the rest samples as

TABLE XI
OA (%) UNDER DIFFERENT ROUND OF SPL ON THE IP DATASET

Round	2	3	4	5
OA	88.05	89.81	88.69	88.63

testing set. For PDAC with AL or SPL, we take the initial training set as ten randomly selected samples per class (160 in total) and 50 samples per round for AL. It can be seen that the attention mechanism can improve the OA of the original PDC network by 4.66%. By gradually considering the AL and SPL strategies, the OA is increased by 1.58% and 1.9%, respectively. The results demonstrate that each module can improve the overall performance of the model.

C. Parameter Analysis

1) Training Round of SPL: The whole PDAC-ASPL model is trained in five rounds, in which the first round is the initial training. After the first round of initial training, the samples that meet the AL and SPL criteria with the corresponding labels/pseudo labels are selected and added to the second round of training, and so on, until the end of five rounds. However, one of the possible problems of adding pseudo labels after the first round may be that the model is easily underfitted due to the small number of initial training samples. That is, the prediction ability of the model is poor in the first round and the confidence of samples is low, so the new samples and corresponding pseudo labels added at this time are close to noise, which will degrade subsequent model training.

It is very important to choose an appropriate time to add pseudo labels. We analyze the OA of PDAC-ASPL on the IP dataset when SPL is first used from the second, third, fourth, and fifth rounds, respectively. The result is shown in Table XI. When SPL is added in the second round, the OA is 88.05%. When SPL is added beginning from the third round, the OA is 89.81%, which demonstrates that SPL plays a role in the case of a well-fitted model. When SPL is added beginning from the fourth or fifth round, the AL model has been relatively well trained; the samples and their pseudo labels of SPL are not significantly improved after the addition. Therefore, we choose to add SPL in the third round for four datasets to take into account both training effect of the model and confidence of the samples.

2) Threshold λ in SPL: The threshold parameter λ determines the number of samples selected by SPL. If it is too large, SPL is likely to select more samples with low confidence. If it is too small, SPL will select very few samples or even no samples. Fig. 12 shows the OA of PDAC-ASPL versus parameter λ, where λ changes in the set $\{10^{-6}, 5\times10^{-6}, 10^{-5}, 10^{-4}\}$.

As shown in Fig. 12, when λ is 1×10^{-6}, the OA is low because the threshold is too small, resulting in insufficient samples being selected. At this time, the SPL does not work, which can be seen from Table X. On the contrary, when λ is 1×10^{-4}, the threshold is too large, producing too much inclusion, and thus, the model accuracy decreases. In the experiments, we set the value of λ to 5×10^{-6} for all four datasets.

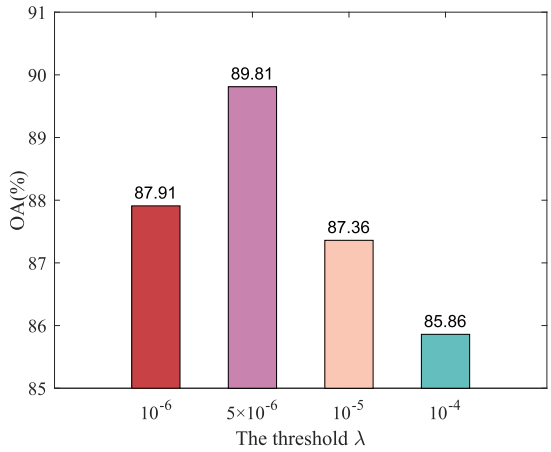


Fig. 12. OA versus threshold parameter λ on IP.

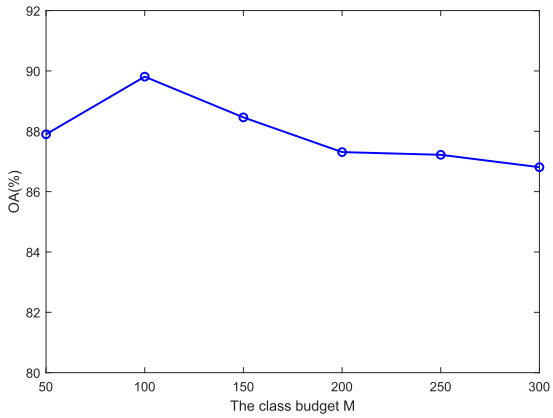


Fig. 13. OA versus the class budget M on IP.

3) Class Budget M in SPL: Considering the problem of unbalanced sample classes, we design a class budget M, i.e., selecting at most M samples in each class. Fig. 13 shows the OA versus the class budget M, where M chooses in the set $\{50, 100, 150, 200, 250, 300\}$. As the selected samples in SPL are not removed from the unlabeled set, part of selected M samples in different training rounds can be the same. As shown in Fig. 13, when the number of samples selected for each class is insufficient, the model does not achieve the desired performance. When too many samples are selected from each category, the OA decreases due to the involving of more low confidence samples. In the experiments, we set the class budget M as 100 for four datasets.

VI. CONCLUSION

In this article, we proposed a PDAC-ASPL model for small-sample HSI classification. The proposed model effectively integrates DL for spatial–spectral feature extraction and classification, AL for the selection of informative samples with true labels, and SPL for the selection of high-confidence samples with pseudo labels. Due to the gradual and effective selection of samples from unlabeled set to training set, the network performance has dramatically improved even with very limited

labeled samples for initial training. Experimental results on four HSI datasets show that the proposed method can obtain better classification results with fewer labeled samples.

In future research, we may consider expanding the application range of this semisupervised method. Specifically, we may further apply it to large-region or cross-scene image classification problems.

REFERENCES

[1] A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.

[2] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[3] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[4] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 447–451, May 2012.

[5] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

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

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

[8] S. Jia et al., "3-D Gabor convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5509216.

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

[10] G. Camps-Valls, L. Gomez-Chova, J. Muñoz Maré, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[11] J. Peng, Y. Zhou, and C. P. Chen, "Region-kernel-based support vector machines for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4810–4824, Sep. 2015.

[12] Y. Chen, N. Nasrabadi, and T. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[13] J. Peng, L. Li, and Y. Y. Tang, "Maximum likelihood estimation based joint sparse representation for the classification of hyperspectral remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1790–1802, Jun. 2019.

[14] J. Peng et al., "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 10–43, Mar. 2022.

[15] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[16] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[18] J. Peng, Y. Huang, W. Sun, N. Chen, Y. Ning, and Q. Du, "Domain adaptation in remote sensing image classification: A survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9842–9859, 2022.

[19] Y. Huang et al., "Two-branch attention adversarial domain adaptation network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540813.

[20] W. Li, C. Chen, M. Zhang, H. Li, and Q. Du, "Data augmentation for hyperspectral image classification with deep CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 593–597, Apr. 2019.

[21] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Hyperspectral image classification using random occlusion data augmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1751–1755, Nov. 2019.

[22] 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. 57, no. 7, pp. 3838–3851, Jul. 2018.

[23] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.

[24] B. Cui, X.-M. Dong, Q. Zhan, J. Peng, and W. Sun, "LiteDepthwiseNet: A lightweight network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502915.

[25] 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.

[26] Z. Wang, B. Du, L. Zhang, L. Zhang, and X. Jia, "A novel semisupervised active-learning algorithm for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3071–3083, Jun. 2017.

[27] 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.

[28] Z. Lei, Y. Zeng, P. Liu, and X. Su, "Active deep learning for hyperspectral image classification with uncertainty learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5502405.

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

[30] M. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.

[31] J. Peng, W. Sun, and Q. Du, "Self-paced joint sparse representation for the classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1183–1194, Feb. 2019.

[32] O. L. F. de Carvalho et al., "Bounding box-free instance segmentation using semi-supervised iterative learning for vehicle detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3403–3420, 2022.

[33] F. Zhao, J. Zhang, Z. Meng, and H. Liu, "Densely connected pyramidal dilated convolutional network for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 17, 2021, Art. no. 3396.

[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[35] L. Wang, J. Peng, and W. Sun, "Spatial–spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 884.

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

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

[38] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1231–1242, Apr. 2008.

[39] C. Ding et al., "Hyperspectral image classification promotion using clustering inspired active learning," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 596.

[40] S. Yang, Z. Feng, M. Wang, and K. Zhang, "Self-paced learning-based probability subspace projection for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 630–635, Feb. 2019.

[41] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 7–19, Jan. 2018.

[42] H. Ren, X. Yu, L. Bruzzone, Y. Zhang, L. Zou, and X. Wang, "A Bayesian approach to active self-paced deep learning for SAR automatic target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 4005705.

[43] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 322–330.

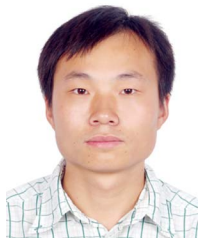
[44] K. Jedoui, R. Krishna, M. Bernstein, and L. Fei-Fei, "Deep Bayesian active learning for multiple correct outputs," 2019, *arXiv:1912.01119*.
[45] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *Proc. Int. Joint Conf. Neural Netw.*, 2014, pp. 112–119.
[46] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2372–2379.
[47] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1183–1192.
[48] 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.
[49] S. Jiang and S. Jia, "A 3D lightweight Siamese network for hyperspectral image classification with limited samples," in *Proc. 10th Int. Conf. Comput. Pattern Recognit.*, 2021, pp. 142–148.
[50] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.
[51] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, 2019, Art. no. 1307.
[52] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

Wenhui Hou received the B.S. degree in mathematics and applied mathematics in 2021 from the Faculty of Mathematics and Statistics, Hubei University, Wuhan, China, where she is currently working toward the M.S. degree in applied mathematics.

Her research interests include machine learning and hyperspectral image processing.

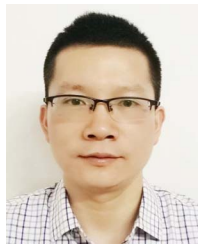
Na Chen received the B.S. degree in information and computing science and the M.S. degree in applied mathematics from Hubei University, Wuhan, China, in 2006 and 2009, respectively, and the Ph.D. degree in computational mathematics from the Huazhong University of Science and Technology, Wuhan, China, in 2012.

She is currently an Associate Professor with the Faculty of Mathematics and Statistics, Hubei University. Her research interests include statistical learning theory and image processing.



Jiangtao Peng (Senior Member, IEEE) received the B.S. degree in information and computing science and the M.S. degree in applied mathematics from Hubei University, Wuhan, China, in 2005 and 2008, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He is currently a Professor with the Faculty of Mathematics and Statistics, Hubei University. His research interests include machine learning 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 was with the Department of Applied Mathematics, University of Maryland, College Park, MD, USA, working as a Visiting Scholar with the famous professor John Benedetto to study on the dimensionality reduction of hyperspectral image. From 2014 to 2016, he was with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, working as a Postdoctoral Researcher to study intelligent processing in Hyperspectral imagery. From 2017 to 2018, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. He is currently a Full Professor with Ningbo University, Ningbo, China. He has authored or coauthored more than 70 journal papers. His current research interests include hyperspectral image processing with manifold learning, anomaly detection, and target recognition of remote sensing imagery using compressive sensing.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland-Baltimore County, Baltimore, MD, USA, in 2000.

She is currently a Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS, USA. Her research interests include hyperspectral remote sensing image analysis and applications, pattern classification, data compression, and neural networks.

Dr. Du is a Fellow of the International Society for Optics and Photonics. She was a recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society (GRSS). She was a Co-Chair of the Data Fusion Technical Committee of the IEEE GRSS from 2009 to 2013. She was the Chair of the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014. She was the General Chair for the fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Shanghai, China, in 2012. She was an Associate Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS), *Journal of Applied Remote Sensing*, and IEEE SIGNAL PROCESSING LETTERS. From 2016 to 2020, she was the Editor-in-Chief for IEEE JSTARS. She is a Member of the IEEE Periodicals Review and Advisory Committee.