



# Two-Stream Networks for Contrastive Learning in Hyperspectral Image Classification

Shuxiang Xia , Xiaohua Zhang , *Member, IEEE*, Hongyun Meng, Jiaxin Fan, and Licheng Jiao , *Fellow, IEEE*

**Abstract**—In the domain of hyperspectral image (HSI) classification, the majority of deep learning methods have necessitated a substantial number of manually annotated samples to achieve outstanding results. However, the process of annotating HSI is conducted at the pixel-level, rendering it not only time-consuming but also financially burdensome. In light of this circumstance, contrastive learning methods that harness unlabeled samples by assigning pseudolabels through pretext tasks have garnered significant attention. Nevertheless, current contrastive learning methods primarily concentrate on exploring spatial diversity among surface samples in natural images, while neglecting the spectral diversity of point targets in HSI, resulting in insufficiently comprehensive feature exploration. In addition, due to the distinct learning objectives between upstream and downstream tasks, this leads to insufficient generalization when transferring to downstream tasks. To tackle these challenges, we propose a two-stream contrastive learning network for few-shot HSI classification. During the pretraining phase, one stream is deployed to probe spatial diversity among samples, whereas the other stream delves into spectral diversity. Subsequently, for transferring to downstream classification tasks, a multilevel fusion network was introduced. It can integrate shallow network features with higher generalization capabilities and deeper network features that are more task-specific. The fused features exhibit an improved performance when employed for classification tasks. Experimental results on four publicly available datasets illustrate that our approach outperforms state-of-the-art methodologies.

**Index Terms**—Contrastive learning, data augmentation, hyperspectral image (HSI) classification, self-supervised learning.

## I. INTRODUCTION

**H**YPERSPECTRAL images (HSIs) are different from traditional color images composed of RGB channels. They are composed of a broader range of spectral bands, typically ranging from dozens to hundreds of bands, providing abundant

spatial and spectral resolution, making it widely applicable in various fields, such as geological research [1], disease diagnosis and surgical guidance [2], astronomy and space surveillance [3], and food safety inspection [4]. In these applications [5], HSI classification is a common task that aims to provide a classification label for each pixel in the HSIs.

In terms of HSI classification, traditional machine learning approaches focus on extracting easily classifiable features [6] or designing satisfactory classifiers [7] separately, feature extraction methods mainly include extended morphological profile [8], superpixel-based composite kernel method [9]. Simple classifiers mainly involve random forest [10], extreme learning machines [11], and multinomial logistic regression [12]. However, traditional machine learning feature extraction methods typically rely on manual design by domain experts based on relevant domain knowledge. Thus, the features extracted in this way have limited relevance to the specific task, leading to lower classification accuracy when applied to classification tasks.

In recent years, deep learning has rapidly developed, not only does it integrate feature extraction and classifier design within the same framework but it also reduces reliance on expert knowledge. This approach, known as end-to-end learning, has become the basis for advanced hyperspectral classification models [13], [14], [15]. For example, Lee et al. [16] employed a 2-D fully convolutional neural network, eliminating the need for additional preprocessing to extract depth features from HSI. Then, considering the characteristics of the hyperspectral 3-D cubes, Xu et al. [17] proposed a spatial-spectral multiscale 3-D CNN structure to improve the model generalization performance. Since convolution can only extract local information, ignoring global information, Hong et al. [18] designed a transformer structure suitable for HSIs to capture global spectral information. Besides, Zhong et al. [19] proposed the spectral-spatial transformer network. Zhu et al. [20] introduced a short-long graph convolution to extract spatial-spectral features at different scales. Moreover, there are other structures, such as the 3-D convolution network with three-dimensional discrete wavelet transform preprocessing [21], hybrid networks combining 2-D CNN, and recurrent neural networks [22] that achieve good results in the HSI classification. All the mentioned models are based on supervised learning for HSI classification, the model structure is relatively complex, requiring a large number of labeled samples to achieve an excellent classification performance. If the sample size is insufficient, it will lead to model overfitting. However, manual labeling for HSI is time-consuming, expensive, and often contains labeling errors. Therefore, how to achieve good

Manuscript received 20 September 2023; revised 6 November 2023 and 20 November 2023; accepted 6 December 2023. Date of publication 11 December 2023; date of current version 28 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61877066, in part by the Aero-Science Fund under Grant 20175181013, and in part by the Science and Technology Plan Project of Xi'an under Grant 21RGZN0010. (Corresponding author: Xiaohua Zhang.)

Shuxiang Xia, Xiaohua Zhang, and Licheng Jiao are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710126, China (e-mail: 22171214681@stu.xidian.edu.cn; xh\_zhang@mail.xidian.edu.cn; lchjiao@mail.xidian.edu.cn).

Hongyun Meng is with the School of Mathematics and Statistics, Xidian University, Xi'an 710126, China (e-mail: menghy@xidian.edu.cn).

Jiaxin Fan is with the University of New South Wales, Business School, Kensington, NSW 100076, Australia (e-mail: fanapplyforu@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3341338

generalization of the model using a small number of manually labeled samples has become a problem to be solved in HSI classification.

To tackle the issue of scarcity of manually labeled data in HSI classification, researchers have introduced semisupervised [23], [24], meta-learning [25], [26], and transfer-learning [27], [28], [29], [30] frameworks. Semisupervised learning involves training with a dataset composed of limited labeled samples and a large number of unlabeled samples to improve the model performance. Meta-learning enables the network to quickly adapt to new tasks by leveraging experiences from training on multiple different tasks. Transfer-learning allows knowledge learned from a source task to be transferred to a target domain, where the source and target tasks are usually different. All three categories of methods aim to improve model performance in scenarios with limited manually labeled data for the target task.

Semisupervised learning can be categorized into the following approaches: generative models [31], self-training models [32], cotraining models [33], graph-based learning models [34]. Generative models, such as those that utilize generative adversarial networks (GANs), are used to explore the distribution of samples in the target dataset. He et al. [35] improved the original semisupervised GAN by designing an additional classifier to avoid the self-contradictory problem where the discriminator simultaneously performs classification and discrimination; Self-training models train classifiers with labeled samples and assign labels to unlabeled samples. Then, they retrain the classifier with reliable unlabeled samples using a confidence criterion. This process is repeated until all the data have been assigned labels. Pan et al. [36] proposed a 3-D Gabor-based self-training semisupervised framework to mine spatial-spectral features; Cotraining models involve training different classifiers with labeled samples and assigning labels to unlabeled data. Similar to self-training models, cotraining models repeat this process until all the data have been assigned labels. Fang et al. [37] introduced a semisupervised cotraining framework that combines deep convolution with deep clustering; Graph-based learning models seek relationships between labeled samples and unlabeled data and assign labels to the unlabeled data. Yang et al. [38] proposed a graph-based method using superpixel techniques to capture pixel-level and region-level information in hyperspectral data. However, the pseudolabels assigned by the classifier to unlabeled samples may lead to inaccuracies; then, errors will be amplified and accumulate after multiple iterations, resulting in less reliable performance during training and prediction.

Meta-learning allows the model parameters and optimization strategies learned from multiple tasks to be applied to the target task. Li et al. [39] first introduced meta-learning into the hyperspectral domain using a deep residual convolutional network. Subsequently, Pal et al. [40] used Monte Carlo averaging of model parameters learned from different tasks and introduced the prototype network (SPN) to the meta-learning framework for the improved performance. In addition, AL-Alimi et al. [41] combined the new data normalization method QPCA and a novel network structure with mixed multiscale convolution kernels to better process hyperspectral data. Furthermore, meta-learning has been applied with stacked convolutional blocks [42], channel

attention [43], siamese networks [44], and other techniques to improve performance in the hyperspectral domain. However, due to the scarcity of training samples for each task, overfitting often occurs during the model training process.

Transfer learning allows the learning of a general feature extractor from a source domain dataset with abundant manual annotated labels, which can then be applied to the target domain dataset for classification tasks. To achieve class-level feature representation, Wang et al. [45] introduced a class-wise attention metric module in the cross-domain framework to enhance the distinguishability of different class features. To address the issue of varying spectral dimensions in different hyperspectral datasets, Lee et al. [46] designed a universal large model that provides different entry points for different hyperspectral datasets. Zhang et al. [47] introduced supervised contrastive learning loss to help the model extract more generalizable features. Although transfer learning requires only a small number of samples from the target domain, the differences in sample space between the source and target tasks result in significant feature representation disparities. The features learned from the source task may not perform well in the target domain, thereby affecting model generalization performance when transferring it to the target task. Moreover, transfer learning still requires a large amount of manually labeled data from the source task dataset.

Self-supervised learning constructs supervisory information based on the intrinsic features of data to design pretext tasks and then utilizes this supervisory information to train models. In comparison to the few shot methods mentioned earlier, self-supervised learning offers several advantages, as follows.

- 1) Self-supervised learning's supervisory information is designed based on the intrinsic features of data, eliminating the possibility of labeling errors.
- 2) Self-supervision can leverage a large number of unlabeled samples to train models, preventing overfitting.
- 3) Self-supervised learning's unlabeled samples are sampled from downstream task datasets, avoiding the differences in sample space between the upstream and downstream tasks.

Self-supervised learning primarily consists of predictive and contrastive approaches [48]. Predictive approaches mainly involve predicting various aspects of the same image, such as the relative positions of different cubes within the image [49], the rotation angle of the image [50], the content of occluded image blocks [51], and generating a color image from a grayscale one [52], among others. However, the effectiveness of predictive methods in feature extraction heavily depends on the design of the pretext task, and the features are strongly tied to the objectives of the pretext task, resulting in a poor performance when transferred to downstream tasks. In addition, tasks like predicting image block reconstruction often entail pixel-level information, which includes excessive details leading to redundancy in information. On the other hand, contrastive learning methods generate pseudolabels by formulating pretext tasks, such as data augmentation on the same sample [53], [54], [55]. Under this scheme, augmented views of the same sample are assigned to the same class, or different perspectives of the same sample as the same class are compared for contrastive

learning [56], and so forth. Subsequently, contrastive learning, which aims to pull closer between samples of the same class while pushing samples of the different classes farther apart in feature space, compels the model to learn crucial information from the data and discriminative features, leading to better generalization performance. Compared with contrastive learning, metric learning [57], [58], [59] shares the same object, the difference is the latter typically employs the manually annotated class label, but the corresponding labels to the former are dependent on both the design of pretext tasks and downstream classification tasks. Contrastive learning has achieved outstanding results in the field of hyperspectral classification. Zhang et al. [60] employed residual 3-D convolutional networks as feature extractors, Cao et al. [61] combined prototype contrastive learning with an autoencoder, extracting distinct features using different encoders. However, these methods still adopt the contrastive learning framework originally designed for surface samples. To enhance the spatial diversity of samples, techniques, such as flip, cropping, cutout, and other data augmentation methods have been employed. In the context of HSI classification, which involves pixel-level point classification tasks, there are notable differences from surface samples, including distinct geometric shapes and regional characteristics. Consequently, the exploration of spectral feature diversity becomes even more crucial in this context. In addition to this, in contrastive learning, pixels that do not belong to the same class in the pretext task may belong to the same class in the downstream classification task. This discrepancy in task objectives between the upstream pretext task and the downstream classification task can result in suboptimal generalization performance of the features learned during the pretext task when applied to the downstream task.

To fully exploit the spectral–spatial features in hyperspectral data, this article proposes a two-stream contrastive learning framework. We consider spectral vectors and the 3-D cube of HSI as the spectral and spatial modalities of the target pixels for classification. During the pretraining phase, we first apply data augmentation to both modalities, generating spectral-sample pairs and spatial-sample pairs. Next, contrastive learning is employed to train the two modalities using different feature encoders. In the fine-tuning phase, a multilevel fusion network is employed to combine features from different hierarchical networks. This is done to address the issue of overlooking valuable fundamental data features in the upstream pretext task, ultimately benefiting the downstream classification task. The contributions of this article are as follows.

- 1) We propose a data augmentation module specifically for the spectral modality, which enriches spectral diversity and improves the robustness of spectral features.
- 2) We apply a two-stream contrastive learning framework that adequately captures both spatial and spectral information in hyperspectral data.
- 3) We propose a multilevel fusion network to mitigate the impact of disparities between upstream and downstream tasks, thereby enhancing the applicability of fused features for classification tasks.

The rest of this article is organized as follows. Section II shows the details of our model. The experimental description and the result analysis on four public datasets are provided in Section III. Finally, Section IV concludes this article.

## II. METHODOLOGY

The focus of this model lies in the utilization of a two-stream network for extracting spatial–spectral features and a multilevel feature fusion network for integrating beneficial basic data features for the downstream task. The overall architecture of the proposed model is illustrated in Fig. 1. The upper part comprises a pretraining module, consisting of the spectral contrastive learning framework and the spatial contrastive learning framework. The lower part is a fine-tuning module, which incorporates a multilevel fusion network.

### A. Data Augmentation for HSI

Training on a small and insufficiently diverse set of training samples can lead to overfitting, resulting in poor model generalization and subpar performance when transferred to downstream classification tasks. Effective data augmentation involves applying various transformations and augmentations to the data to increase its diversity, thereby enhancing the model’s robustness and generalization capabilities. Existing data augmentation techniques primarily focus on enhancing the spatial diversity of surface samples. However, HSI classification involves pixel-level classification, with spectral features being the intrinsic characteristics. Existing data augmentation methods are not suitable for enhancing hyperspectral point samples. Therefore, as depicted in Fig. 2, we propose four techniques to introduce perturbations within specific spectral bands or randomly select neighboring spectral vector as positive samples. These methods aim to enhance the spectral diversity of samples, thereby attempting to simulate the spectral variations caused by factors like illumination and air humidity in real-world scenarios. The detailed procedural steps are outlined as follows.

Assume we have spectral data  $X_m \in R^{1 \times 1 \times C}$ , where  $C$  represents the number of spectral bands. Fig. 2(a) illustrates the flip operation, where the flipped spectrum  $X_f$  can be obtained as follows:

$$X_f = X_m[:, :, C : 1]. \quad (1)$$

Fig. 2(b) presents the random\_select operation. In HSIs, adjacent pixels often belong to the same class. In addition, HSIs exhibit the characteristic of different spectral for the same material. Therefore, we select a patch of size  $P \times P$  centered around the target pixel, denoted as  $X^p \in R^{P \times P \times C}$ , and randomly choose a spectrum (excluding the center spectrum) within this patch as a positive sample for the center spectrum. The positive sample  $X_s$  is defined as follows:

$$X_s = \text{random\_select}(X^p). \quad (2)$$

Fig. 2(c) demonstrates the addition of Gaussian noise. We generate a Gaussian kernel  $G_b \in R^{1 \times 1 \times d}$  with a mean  $\mu$  and variance  $\theta$ , where  $d$  represents the length of the Gaussian kernel. The

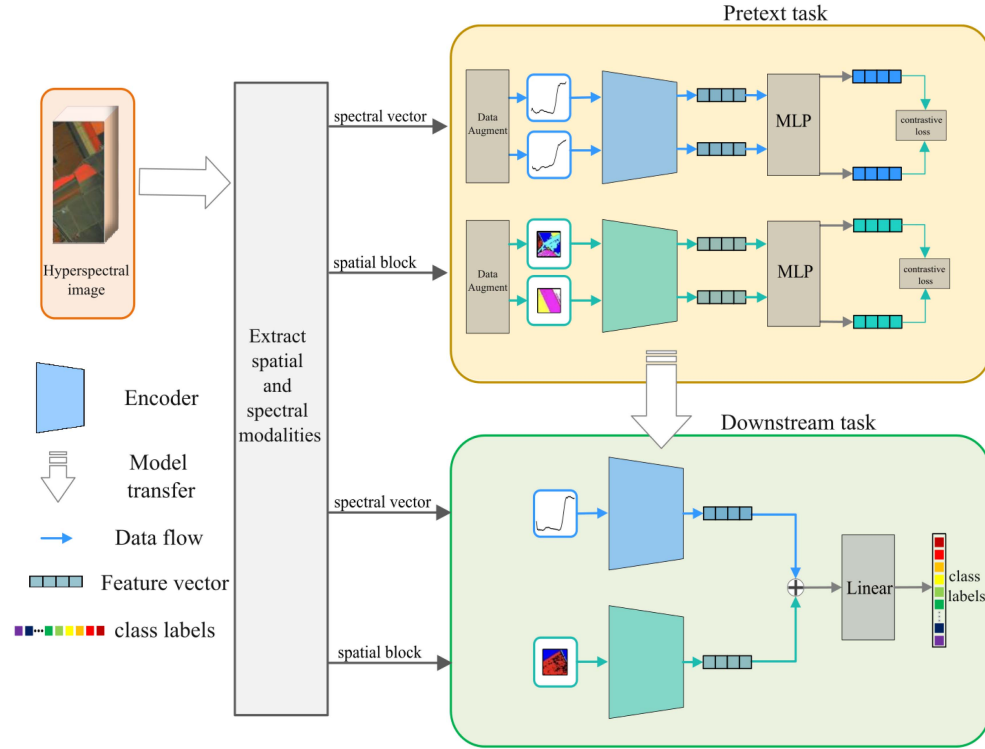


Fig. 1. Whole framework of the model for HSI classification.

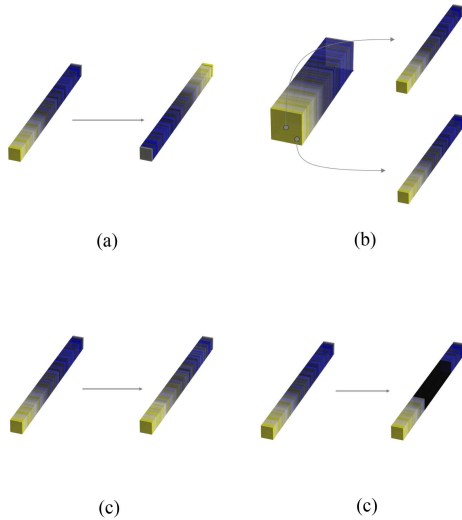


Fig. 2. Four types of data augmentation methods for spectral modality. (a) Flip. (b) Random\_select. (c) Gaussian Blur. (d) Cutout.

resulting positive sample  $X_g$  is computed as follows:

$$X_g = X_m \quad (3)$$

$$X_g = X_g[:, :, i : i + d] + G_b. \quad (4)$$

Here,  $i$  is randomly selected from the interval  $[0, C - d - 1]$ .

Fig. 2(d) shows the data augmentation method of cutout, which sets a band of the spectrum to zero to obtain a positive

sample  $X_c$  as follows:

$$X_c = X_m \quad (5)$$

$$X_c[:, :, i : i + c] = 0. \quad (6)$$

For the spatial features of the images, we employed techniques, such as random cropping, random flipping, color distortion, and grayscale conversion to augment the diversity of spatial characteristics.

### B. Two-Stream Pretraining Networks

Hyperspectral point samples can benefit from incorporating information from neighboring pixels of the same class to reduce the impact of variance on classification. However, it is essential to acknowledge that pixels from different classes in the vicinity can also interfere with classification. The spectral characteristics inherent to point samples provide valuable insights into the chemical composition and constituents of materials. By exploring spectral diversity, we can enhance the accuracy of identifying and classifying various substances or objects. Therefore, we propose a two-stream contrastive learning network to comprehensively exploit the spatial-spectral information in HSI. The detailed procedure is outlined as follows.

Assume there is an HSI denoted as  $X_{\text{img}} \in R^{W \times H \times C}$ , where  $W$  and  $H$  represent the width and height of the image and  $C$  is the number of spectral bands.

For spectral modality processing, a spectral data point  $X_m^i \in R^{1 \times 1 \times C}$  is selected from  $X_{\text{img}}$ , where  $i \in \{1, 2, \dots, N_m\}$ ,  $N_m$  is equal to  $W \times H$ . Data augmentation is applied twice to  $X_m^i$  to obtain positive sample pairs  $(A_m^i, A_m^i)$ . These

TABLE I  
PRETRAINING NETWORK PARAMETER SETTINGS

| Network name            | Layer         | Class                   | Input size               | Kernel size  | Padding      | Activation function | BN  |
|-------------------------|---------------|-------------------------|--------------------------|--------------|--------------|---------------------|-----|
| $f_m(\cdot)$            | 1             |                         | $1 \times 200$           | 8            | -            | ReLU                | Yes |
|                         | 2             | Conv                    | $64 \times 65$           | 3            | -            | ReLU                | Yes |
|                         | 3             |                         | $128 \times 32$          | 3            | -            | ReLU                | Yes |
|                         | 4             |                         | $256 \times 15$          | 3            | -            | ReLU                | Yes |
| MLP1                    | 1             | FC                      | $512 \times 1$           | -            | -            | ReLU                | Yes |
|                         | 2             |                         | $256 \times 1$           | -            | -            | -                   | Yes |
| $f_a(\cdot)$            | 1             | Conv                    | $3 \times 35 \times 35$  | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         | 2             | BasicBlock              | $3 \times 35 \times 35$  | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         |               |                         | $3 \times 35 \times 35$  | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         | 3             | ResidualBlock           | $3 \times 35 \times 35$  | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         |               |                         | $64 \times 18 \times 18$ | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         |               |                         | $64 \times 18 \times 18$ | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         | 4             | ResidualBlock           | $64 \times 18 \times 18$ | $3 \times 3$ | $1 \times 1$ | -                   | Yes |
|                         |               |                         | $64 \times 18 \times 18$ | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         | 5             | ResidualBlock           | $128 \times 9 \times 9$  | $3 \times 3$ | $1 \times 1$ | -                   | Yes |
|                         |               |                         | $128 \times 9 \times 9$  | $3 \times 3$ | $1 \times 1$ | ReLU                | Yes |
|                         | 6             | ResidualBlock           | $128 \times 9 \times 9$  | $3 \times 3$ | $1 \times 1$ | -                   | Yes |
| $128 \times 9 \times 9$ |               |                         | $3 \times 3$             | $1 \times 1$ | ReLU         | Yes                 |     |
| 7                       | ResidualBlock | $256 \times 5 \times 5$ | $3 \times 3$             | $1 \times 1$ | -            | Yes                 |     |
|                         |               | $256 \times 5 \times 5$ | $3 \times 3$             | $1 \times 1$ | ReLU         | Yes                 |     |
| 8                       | ResidualBlock | $256 \times 5 \times 5$ | $3 \times 3$             | $1 \times 1$ | -            | Yes                 |     |
|                         |               | $256 \times 5 \times 5$ | $3 \times 3$             | $1 \times 1$ | ReLU         | Yes                 |     |
| 9                       | ResidualBlock | $512 \times 3 \times 3$ | $3 \times 3$             | $1 \times 1$ | -            | Yes                 |     |
|                         |               | $512 \times 3 \times 3$ | $3 \times 3$             | $1 \times 1$ | ReLU         | Yes                 |     |
| 10                      | ResidualBlock | $512 \times 3 \times 3$ | $3 \times 3$             | $1 \times 1$ | -            | Yes                 |     |
|                         |               | $512 \times 3 \times 3$ | $3 \times 3$             | $1 \times 1$ | ReLU         | Yes                 |     |
| 11                      | Pooling       | $512 \times 1 \times 1$ | $3 \times 3$             | $1 \times 1$ | -            | Yes                 |     |
| MLP2                    | 1             | FC                      | $512 \times 1$           | -            | -            | ReLU                | Yes |
|                         | 2             |                         | $256 \times 1$           | -            | -            | -                   | Yes |

sample pairs are then fed into the feature extractor  $f_m(\cdot)$  with shared weights. The detailed structure of  $f_m(\cdot)$  is presented in Table I, which employs multiple one-dimensional convolutions to model the input spectral sequence and extract local spectral features. The positive sample pairs are processed by  $f_m(\cdot)$  to obtain features  $(f_{m_1^i}, f_{m_2^i})$ . These features are subsequently passed through a multilayer perceptron  $g_m(\cdot)$  to obtain representation vectors  $(Z_{m_1^i}, Z_{m_2^i})$ . The multilayer perceptron  $g_m(\cdot)$  consists of two fully connected layers [48], where ReLU activation functions introduce nonlinear transformations between the layers and batch normalization is applied to normalize the inputs and improve gradient propagation. Through  $g_m(\cdot)$ , the features are projected into a latent space with the objective of making the vectors  $Z_{m_1^i}$  and  $Z_{m_2^i}$  as close as possible while maintaining a significant distance from other vectors  $Z_{m_1^j}$  and  $Z_{m_2^j}$  ( $j \neq i$ ). To impose this constraint, a contrastive loss is utilized

$$\begin{aligned} sim\_all(Z_{m_1^i}) = & \left( \sum_{j=1}^N I_{j \neq i} (\exp(sim(Z_{m_1^i}, Z_{m_1^j})) \right. \\ & + \exp(sim(Z_{m_1^i}, Z_{m_2^j})) \\ & \left. + \exp(sim(Z_{m_1^i}, Z_{m_2^i}))) / \tau \right) \end{aligned} \quad (7)$$

$$l_{Z_{m_1^i}, Z_{m_2^i}} = -\log \frac{\exp(sim(Z_{m_1^i}, Z_{m_2^i})/\tau)}{sim\_all(Z_{m_1^i})}. \quad (8)$$

In this equation,  $\tau$  represents a temperature hyperparameter that controls the smoothness of the logit distribution.  $N$  corresponds to the mini-batch size, wherein  $N$  spectral data points undergo data augmentation to generate  $2N$  samples as inputs for the encoder.  $I_{j \neq i}$  denotes an indicator function that equals 1 when the condition  $j \neq i$  is satisfied, and 0 otherwise. The similarity calculation in this article employs cosine similarity between representative vectors, as shown in the equation

$$sim(Z_{m_1^i}, Z_{m_2^i}) = \frac{Z_{m_1^i} \times Z_{m_2^i}}{|Z_{m_1^i}| \times |Z_{m_2^i}|}. \quad (9)$$

Here,  $|\cdot|$  denotes the vector normalization. The final loss is computed as the sum of losses for all positive sample pairs within a mini-batch, represented by the following equation:

$$L = \frac{1}{2N} \sum_{i=1}^N [l_{(Z_{m_1^i}, Z_{m_2^i})} + l_{(Z_{m_2^i}, Z_{m_1^i})}]. \quad (10)$$

Regarding spectral modality processing, a similar approach is applied to the spatial modality. We applied principal component analysis (PCA) to reduce the dimensionality of  $X_{img}$  to obtain  $X_{img_p}$ , assuming the channel dimension after dimensionality

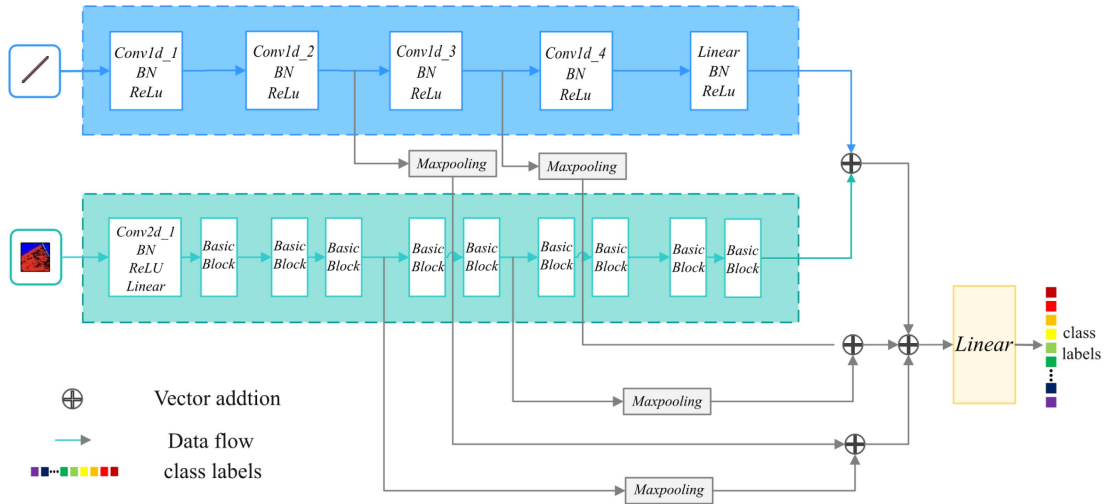


Fig. 3. Illustration of the multilevel fusion module.

reduction is  $p$ ,  $X_{\text{img}_p} \in R^{W \times H \times p}$ . In this article, we set  $p$  to be 3. Then, spatial data patch  $x_a^k \in R^{s \times s \times p}$  with a patch size of  $s \times s$  are extracted from  $X_{\text{img}_p}$ , where  $k \in \{1, 2, \dots, N_a\}$  and  $N_a$  is equal to  $W \times H$ . According to [53], we employed three data augmentation methods: random cropping, random flipping, and color distortion. In addition, we introduced an extra data augmentation method called RandomGrayscale, which involves taking a weighted average of the values in each channel of the image to obtain a grayscale image  $x_{gy}^k$ , as shown in the following equation:

$$x_{gy}^k = \alpha * x_a^k[:, :, 0] + \beta * x_a^k[:, :, 1] + \gamma * x_a^k[:, :, 2]. \quad (11)$$

Then, we apply data augmentation twice for  $x_a^k$  to obtain the sample pair  $(A_a^k, A_{a_2}^k)$ . The spatial encoder  $f_a(\cdot)$  is built based on ResNet18 to obtain  $(f_{a_1}, f_{a_2})$ . The features are then fed into a multilayer perceptron  $g_a(\cdot)$  to project them into the latent space. The loss function remains consistent with the previous equation (10). Both the encoder  $f_m(\cdot)$  and  $f_a(\cdot)$  are trained using spectral and spatial sample pairs for downstream tasks, and their training parameters are saved for future transfer.

### C. Multilevel Fusion Networks

Contrastive learning can extract semantic information beneficial for classification. However, contrastive learning involves learning feature representations of same-class and different-class samples through a pretext task with pseudolabels, whereas downstream tasks involve addressing specific classification tasks based on true manual labels. These two approaches have distinct objectives. According to [55], shallow-level features possess higher generalization capacity, while mid-to-high-level network features are more task-specific. By fusing features from various levels, the consistency between upstream and downstream tasks can be improved. To enhance the model's generalization performance in downstream tasks, we introduce a multilevel network fusion module in Fig. 3. Drawing on the ideas of [62], we propose a multilevel features fusion strategy for the spectral modality as

well. This module integrates features from different hierarchical levels of networks by projecting them into the same dimension through linear and pooling layers before fusion. The detailed procedural steps are outlined as follows.

As illustrated in Fig. 2, model initialization parameters are derived from the saved parameters of upstream contrastive learning. The multilayer perceptron is directly replaced with a classification linear layer.  $M$  spatial samples are selected from  $X_{\text{img}}$ , and the central spectrum of each spatial sample is used as the spectral sample, forming the training set. Shallow features,  $f_{s_m}$  and  $f_{s_a}$ , middle-level features  $f_{m_m}$  and  $f_{m_a}$  are obtained. Finally, final features  $f_m$  and  $f_a$  are obtained from last layer. Due to the one-dimensional of spectral modality and the two-dimensional nature of spatial modality, the dimensions of the two modalities of data do not match. Therefore, a max pooling layer is utilized to compress the spatial modality features, ensuring consistency with the spectral modality dimensions. Fusion features are obtained by summing the respective features, as shown in the following equation:

$$f_S = f_{s_m} + \text{maxpooling}(f_{s_a}) \quad (12)$$

$$f_M = f_{m_m} + \text{maxpooling}(f_{m_a}) \quad (13)$$

$$f_L = f_m + f_a \quad (14)$$

$$f = f_S + f_M + f_L. \quad (15)$$

The final fused features:  $f$ , are then input into the classification layer for classification. The cross-entropy loss function is employed, as shown in the following equation:

$$L_{\text{cross\_entropy}} = - \sum_{i=1}^M y^i \log \frac{\exp(W_i^T \times f)}{\sum_{j=1}^C \exp(W_j^T \times f)}. \quad (16)$$

Here,  $C$  represents the number of classes in the dataset and  $y^i$  denotes the true label of the  $i$ th sample.

Algorithm 1 provides pseudocode for pretraining and fine-tuning.

---

**Algorithm 1: General Procedure of Pretraining and Fine-Tuning.**


---

**Stage 1: Pre-training of two-stream networks****input:** batch size  $N$ , constant  $\tau$ , structure of  $f_m(\cdot)$ ,  $g_m(\cdot)$ , $f_a(\cdot)$ ,  $g_a(\cdot)$ , spectral data  $x_m$ , spatial data  $x_a$ **for** sampled minibatch  $\{x_i^m\}_{i=1}^N$ ,  $\{x_i^a\}_{i=1}^N$  **do****for all**  $i \in \{1, 2, \dots, N\}$  **do**

# Use augmentation to obtain spectral sample pairs.

 $A_{-m_1}^i, A_{-m_2}^i = \text{augment}(X_m^i)$ 

# Obtain their respective features.

 $f_{-m_1}^i, f_{-m_2}^i = f_m(A_{-m_1}^i), f_m(A_{-m_2}^i)$ 

# Project the features into a latent space.

 $z_{-m_1}^i, z_{-m_2}^i = g_m(f_{-m_1}^i), g_m(f_{-m_2}^i)$ 

# Perform the same operations on spatial data.

 $A_{-a_1}^i, A_{-a_2}^i = \text{augment}(X_a^i)$  $f_{-a_1}^i, f_{-a_2}^i = f_a(A_{-a_1}^i)$  $z_{-a_1}^i, z_{-a_2}^i = g_a(f_{-a_1}^i)$ **end for**use (10) update network  $f_m(\cdot)$ ,  $g_m(\cdot)$ ,  $f_a(\cdot)$  and  $g_a(\cdot)$ **end for****return** encoder  $f_m(\cdot)$ ,  $f_a(\cdot)$ **Stage 2: Multi-level Fusion****input:** batch size  $M$ , structure of  $f_m(\cdot)$  and  $f_a(\cdot)$ , spectral data  $x_m$ , spatial data  $x_a$ , classification linear layer  $g(\cdot)$ **Load** pre-trained weights for  $f_m(\cdot)$  and  $f_a(\cdot)$ .**for** sampled minibatch  $\{x_i^m\}_{i=1}^M$ ,  $\{x_i^a\}_{i=1}^M$  **do****for all**  $i \in \{1, 2, \dots, M\}$  **do**

#Obtain features from shallow, middle, and deep layers.

 $f_{s-m}, f_{m-m}, f_m = f_m(x_i^m)$  $f_{s-a}, f_{m-a}, f_a = f_a(x_i^a)$ use (12), (13), (14), (15) to obtain  $f$ #Use  $g(\cdot)$  to obtain one-hot label $\hat{y} = g(f)$ **end for**use (15) to update network  $f_m(\cdot)$ ,  $f_a(\cdot)$  and  $g(\cdot)$ .**end for**

## III. EXPERIMENTS

## A. Dataset Description

This section presents a comprehensive evaluation of our HSI classification model using four publicly available datasets: Indian Pines, Pavia University, WHU-Hi-LongKou, and WHU-Hi-HanChuan. Our model demonstrates its effectiveness across these datasets, showcasing its potential for accurate classification.

1) *Indian Pines*: The Indian Pines dataset comprises images captured by AVIRIS over an Indian pine tree area in Indiana, USA. We extracted a subset of the image with dimensions  $145 \times 145$  for HSI classification. The AVIRIS sensor operates within the wavelength range of  $0.4\text{--}2.5 \mu\text{m}$  and captures data across 220 spectral bands. To account for water reflectance limitations, we excluded 20 bands (104–108, 150–163, and the 220th). Therefore, our study utilized a total of 200 bands, containing 10249 pixels classified into 16 different classes.

TABLE II  
NUMBER OF SAMPLES FOR TRAINING AND TEST OF EACH CLASS ON THE INDIAN PINES DATASET

| Class No. | Categories          | No. of Samples | Training | Test |
|-----------|---------------------|----------------|----------|------|
| C1        | Alfalfa             | 46             | 30       | 16   |
| C2        | Corn-no till        | 1428           | 30       | 1398 |
| C3        | Corn-min till       | 830            | 30       | 800  |
| C4        | Corn                | 237            | 30       | 200  |
| C5        | Grass/pasture       | 483            | 30       | 453  |
| C6        | Grass/trees         | 730            | 30       | 700  |
| C7        | Grass/pasture-mowed | 28             | 15       | 13   |
| C8        | Hay-windrowed       | 478            | 30       | 448  |
| C9        | Oats                | 20             | 15       | 5    |
| C10       | Soybean-no till     | 972            | 30       | 942  |
| C11       | Soybean-min till    | 2455           | 30       | 2425 |
| C12       | Soybean-clean       | 593            | 30       | 563  |
| C13       | Wheat               | 205            | 30       | 175  |
| C14       | Woods               | 1265           | 30       | 1235 |
| C15       | Blg-grass-trees     | 386            | 30       | 356  |
| C16       | Stone-Steel-Towers  | 93             | 30       | 63   |
| Total     |                     | 10249          | 450      | 9799 |

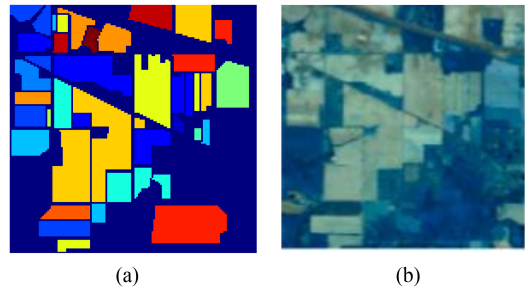


Fig. 4. Visualization of the Indian Pines dataset. (a) Ground-truth map. (b) Pseudocolor image of HSI.

TABLE III  
NUMBER OF SAMPLES FOR TRAINING AND TEST OF EACH CLASS ON THE PAVIA UNIVERSITY DATASET

| Class No. | Categories   | No. of Samples | Training | Test  |
|-----------|--------------|----------------|----------|-------|
| C1        | Asphalt      | 6631           | 30       | 6601  |
| C2        | Meadows      | 18649          | 30       | 18619 |
| C3        | Gravel       | 2099           | 30       | 2069  |
| C4        | Trees        | 3064           | 30       | 3034  |
| C5        | Metal-sheets | 1345           | 30       | 1315  |
| C6        | Bare-soil    | 5029           | 30       | 4999  |
| C7        | Bitumen      | 1330           | 30       | 1300  |
| C8        | Bricks       | 3682           | 30       | 3652  |
| C9        | Shadows      | 947            | 30       | 917   |
| Total     |              | 42776          | 270      | 42506 |

Refer to Table II for the available sample counts for each class and details in Fig. 4.

2) *Pavia University*: The Pavia University dataset was acquired by ROSIS-03, an airborne hyperspectral sensor, in 2003 over the city of Pavia, Italy. The sensor captured continuous imaging across 115 spectral bands within the wavelength range of  $0.43\text{--}0.86 \mu\text{m}$ . Considering noise, 12 bands were excluded, resulting in 103 spectral bands used for analysis. The image size is  $610 \times 340$ , encompassing 207400 pixels, only 42776 pixels containing objects. These pixels belong to nine different classes, including trees, asphalt, bricks, meadows, etc. Detailed information about the sample counts for each class can be found in Table III and Fig. 5.

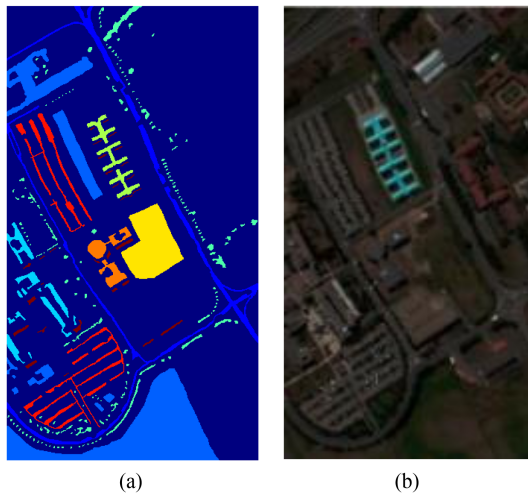


Fig. 5. Visualization of the Pavia University dataset. (a) Ground-truth map. (b) Pseudocolor image of HSI.

TABLE IV  
NUMBER OF SAMPLES FOR TRAINING AND TEST OF EACH CLASS ON THE WHU\_HI\_LONGKOU DATASET

| Class No. | Categories          | No. of Samples | Training | Test    |
|-----------|---------------------|----------------|----------|---------|
| C1        | Corn                | 34511          | 30       | 34481   |
| C2        | Cotton              | 8374           | 30       | 8344    |
| C3        | Sesame              | 3031           | 30       | 3001    |
| C4        | Broad-leaf soybean  | 63212          | 30       | 63182   |
| C5        | Narrow-leaf soybean | 4151           | 30       | 4121    |
| C6        | Rice                | 11854          | 30       | 11824   |
| C7        | Water               | 67056          | 30       | 67026   |
| C8        | Roads and houses    | 7124           | 30       | 7112    |
| C9        | Mixed weed          | 5229           | 30       | 5199    |
|           | Total               | 204542         | 270      | 204,272 |

3) *WHU-Hi-LongKou*: The WHU-Hi-LongKou dataset represents Longkou Town, Hubei Province, China, and was acquired using the Headwall NanoHyperspec sensor. The original dataset comprises 270 spectral bands, with an image size of  $550 \times 400$ . Among these pixels, a total of 204 542 have been identified as suitable for classification purposes. These pixels are distributed across nine distinct categories, which encompass a variety of agricultural crops, including corn, cotton, and others. Please refer to Table IV for a comprehensive overview of the available samples for each class and details in Fig. 6.

4) *WHU-Hi-HanChuan*: The WHU-Hi-HanChuan dataset was acquired in 2016 over HanChuan City, Hubei Province, China, using a Headwall NanoHyperspec sensor equipped on a Leica Aibot X6. This sensor captured 274 bands in the wavelength range of 400–1000 nm, with an image size of  $1217 \times 303$  pixels. Among these pixels, 257 530 pixels were utilized for classification. The dataset comprises 16 categories, including grass, water, road, and more. However, due to data collection occurring during periods of lower sun elevation angles in the afternoon, the images exhibit numerous shadow-covered areas. An overview of this dataset can be found in Fig. 7 and Table V.

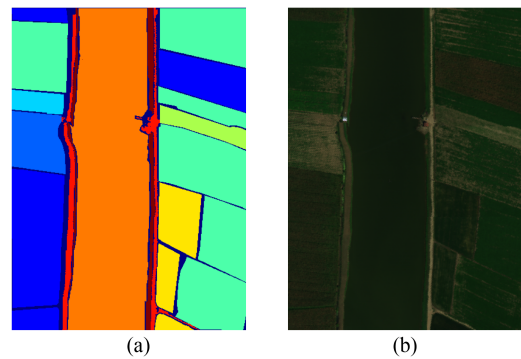


Fig. 6. Visualization of the WHU-Hi-LongKou dataset. (a) Ground-truth map. (b) Pseudocolor image of HSI.

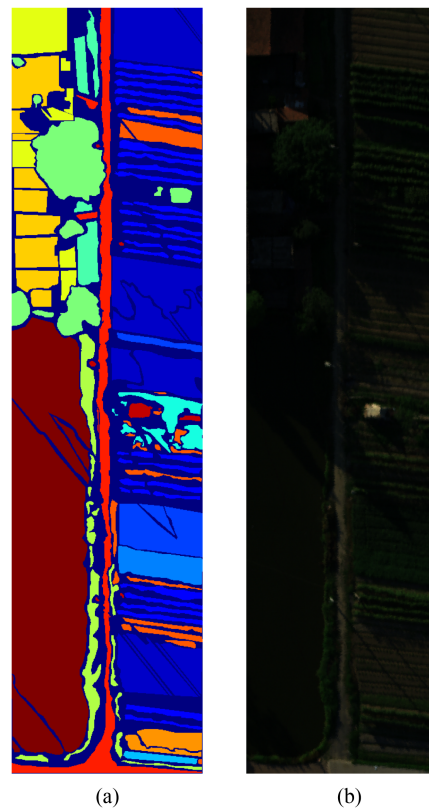


Fig. 7. Visualization of the WHU-Hi-HanChuan dataset. (a) Ground-truth map. (b) Pseudocolor image of HSI.

## B. Experimental Setup

The experimental environment for this model consisted of an Ubuntu 20.04 LTS operating system, two Intel(R) Xeon(R) Silver 4214 CPUs @ 2.20 GHz, 125.6 GB of RAM, and an Nvidia GeForce RTX 3090 GPU for image processing. The deep learning framework PyTorch was employed.

In the pretraining stage, the spectral bands of the spatial samples were reduced to 3 using PCA. The patch size was set to 31 for the Indian Pines, 35 for the WHU-Hi-LongKou and WHU-Hi-HanChuan datasets, and 41 for the Pavia University dataset. The spatial samples were augmented through random cropping,



TABLE V  
NUMBER OF SAMPLES FOR TRAINING AND TEST OF EACH CLASS ON THE  
WHU\_H1\_HANCHUAN DATASET

| Class No. | Categories    | No. of Samples | Training | Test   |
|-----------|---------------|----------------|----------|--------|
| C1        | Strawberry    | 44735          | 30       | 1957   |
| C2        | Cowpea        | 22753          | 30       | 3706   |
| C3        | Soybean       | 10287          | 30       | 1956   |
| C4        | Sorghum       | 5353           | 30       | 1374   |
| C5        | Water spinach | 1200           | 30       | 2658   |
| C6        | Watermelon    | 4533           | 30       | 3939   |
| C7        | Greens        | 5903           | 30       | 3559   |
| C8        | Trees         | 17978          | 30       | 11193  |
| C9        | Grass         | 9469           | 30       | 6177   |
| C10       | Red roof      | 10516          | 30       | 3229   |
| C11       | Gray roof     | 16911          | 30       | 1038   |
| C12       | Plastic       | 3679           | 30       | 1888   |
| C13       | Bare soil     | 9116           | 30       | 889    |
| C14       | Road          | 18560          | 30       | 1041   |
| C15       | Bright object | 1136           | 30       | 7144   |
| C16       | Water         | 75401          | 30       | 1717   |
| Total     |               | 257530         | 480      | 257050 |

TABLE VI  
CLASSIFICATION ACCURACY(%) UNDER DIFFERENT SPECTRAL  
AUGMENTATION POLICY

| Data Augmentation      | OA           | AA           | Kappa        |
|------------------------|--------------|--------------|--------------|
| Blur                   | 94.49        | 95.31        | 92.74        |
| Flip                   | 93.33        | 94.73        | 91.30        |
| Random_select          | <b>97.72</b> | <b>98.16</b> | <b>96.96</b> |
| Cutout                 | 94.06        | 94.78        | 92.18        |
| Random_select + Blur   | 96.88        | 97.87        | 95.89        |
| Random_select + Flip   | 97.38        | 97.67        | 96.54        |
| Random_select + Cutout | 97.49        | 97.41        | 96.68        |

random flipping, color distortion, and conversion to grayscale with a certain probability to construct positive sample pairs. The spectral modality was augmented through random\_select. The patch size was set to 3. The mini-batch size was set to 400, and the number of epochs was set to 200; the learning rate was set to  $1e-3$ .

In the fine-tuning stage, 30 samples were selected from each class for all the datasets. If the sample size is insufficient, take 15 samples from each class. The mini-batch size was set to 64, and the RMSProp optimizer with a decay rate of 0.9 was utilized for training, the number of epochs was set to 100. The performance of different algorithms was evaluated using overall accuracy (OA), average accuracy (AA), and Kappa coefficient.

### C. Ablation Study

Efficient data augmentation facilitates the model in thoroughly exploring the diversity of features, enhancing the dissimilarity between sample pairs, and compelling the encoder to acquire advanced semantic characteristics. Thus, we conducted an evaluation to assess the classification performance of various data augmentation methods within the proposed framework. Table VI presents the classification accuracies achieved using different combinations of data augmentation techniques on the PU dataset, and it is evident that “random\_select” is the most effective data augmentation method. Random\_select involves randomly choosing a spectrum from the vicinity of the target pixel to form a positive sample pair, which, compared to adding Gaussian noise, better captures the influences of noise from

natural sources and light reflection that align with the distribution of the current samples space.

Table VII displays the classification accuracies of various modules in the model. The model consists of the following three parts: spectral encoder, space encoder, and multilevel fusion. The OA obtained from each module is presented in Table VII. In addition, the first column of Table VII indicates whether contrastive learning was used for pretraining, with “√” denoting its usage for that particular module. Due to structural differences and variations in light reflection, different objects may exhibit similar spectral features in the same spectral bands. Therefore, the classification accuracy of the space encoder, which leverages spatial information, is superior to that of the spectral encoder. Specifically, on the PU dataset, the OA are 56.24% and 76.37%, respectively. This indicates that incorporating spatial information into HSI features is more advantageous for image classification. However, when combining the two encoders, further improvement in accuracy has been achieved. For instance, on the IP dataset, the OA of a single space encoder is 73.62%, while the OA accuracy of the combined encoders is 76.54%. This result suggests that fully exploring spatial-spectral features is more conducive to classification. By applying contrastive learning, on the LongKou datasets, the accuracies are improved by 3.27%. This indicates that contrastive learning can unearth latent features of unlabeled samples and enhance the model’s generalization ability. Taking the HanChuan dataset as an example, after incorporating the multilevel fusion module, the accuracy is improved by 0.44%, indicating that features from shallow and middle-level networks contribute to the downstream classification task.

To further substantiate the authenticity of extracting superior features through the exploration of spectral diversity and the utilization of a multilevel fusion network, we present in Fig. 8, visualizations achieved through t-distributed stochastic neighbor embedding of distinct feature sets, which exclusively mine spatial diversity, mining spatial-spectral diversity, employing the multilevel fusion network. Across the IP, PU, LongKou, and HanChuan datasets, we deduce that features capturing space-spectral diversity yield greater discriminability compared to networks focusing on spatial diversity. The distances between samples of different classes have increased, while the distances between samples of the same class have become closer. This underscores the efficacy of harnessing spectral diversity for enhanced feature extraction capabilities. Furthermore, following the incorporation of the multilevel fusion network, a notable reduction in the number of outliers is observed, this reduction signifies the potential of leveraging shallow-to-intermediate network fusion to enhance the model’s generalization performance.

### D. Sensitive Analysis of the Number of Labeled Samples

Due to the significant impact of the sample size on the classification performance of deep networks, this section conducts data sensitivity experiments on nine classification algorithms. Figs. 9–12 displays the OA values of various algorithms on the four datasets under different numbers of training samples. We experimented with sample quantities of 10, 20, 30, and 40 for

TABLE VII  
ABLATION STUDIES ON DIFFERENT DATASETS

| Pretrain | Spectral encoder | Space encoder | Multi-Level fusion | PU           | IP           | LK           | HC           |
|----------|------------------|---------------|--------------------|--------------|--------------|--------------|--------------|
| -        | ✓                | -             | -                  | 56.24        | 40.86        | 85.98        | 65.24        |
| ✓        | ✓                | -             | -                  | 62.56        | 48.06        | 92.76        | 73.68        |
| -        | -                | ✓             | -                  | 76.37        | 73.62        | 90.78        | 81.55        |
| ✓        | -                | ✓             | -                  | 88.11        | 86.15        | 92.83        | 91.14        |
| -        | ✓                | ✓             | -                  | 90.14        | 76.54        | 95.42        | 85.76        |
| ✓        | ✓                | ✓             | -                  | 96.74        | 90.35        | 98.69        | 93.08        |
| ✓        | ✓                | ✓             | ✓                  | <b>97.72</b> | <b>93.23</b> | <b>98.94</b> | <b>93.52</b> |

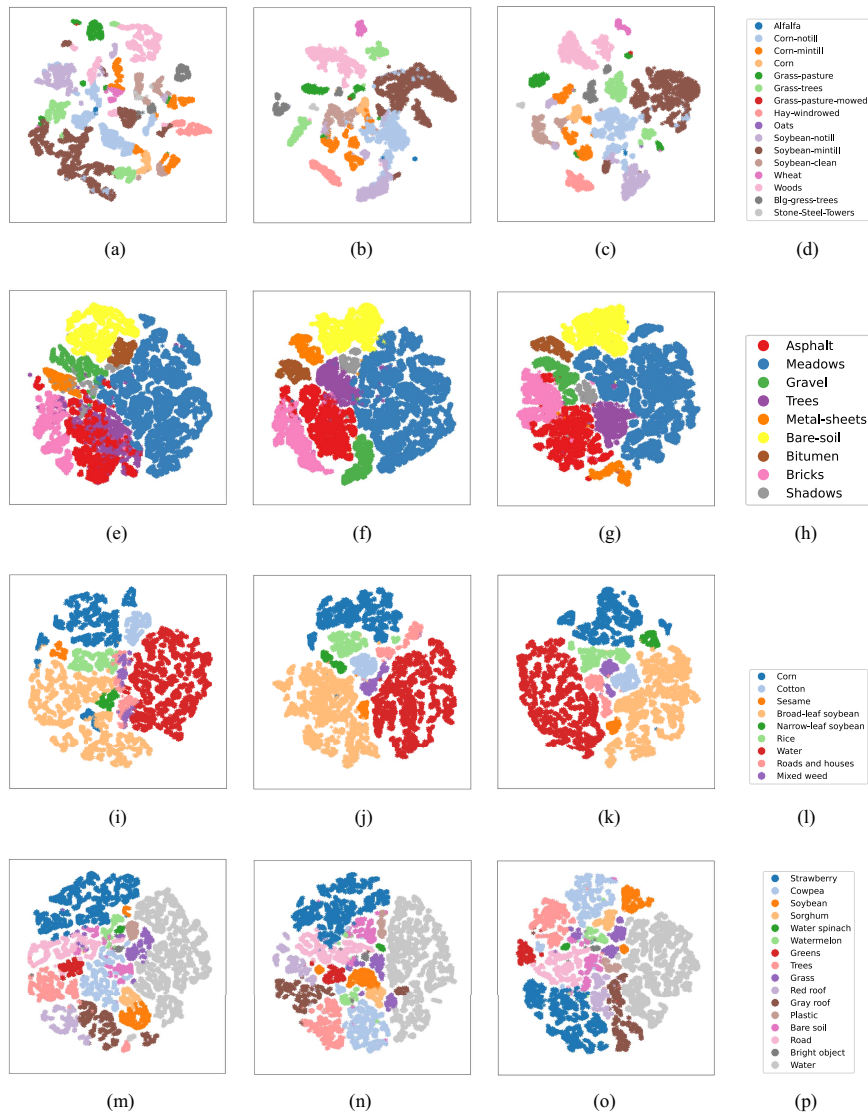


Fig. 8. Feature visualization of encoders for HSI on four datasets. (a), (b), (c), (d) Results on the Indian Pine dataset. (e), (f), (g), (h) Results on the Pavia University dataset. (i), (j), (k), (l) Results on the WHU-Hi-LongKou dataset. (m), (n), (o), (p) Results on the WHU-Hi-HanChuan dataset. The first column denotes features that only explore spatial diversity, the second column denotes features that explore spectral-spatial diversity, the third column denotes features obtained by incorporating multilevel fusion modules on the basis of exploring spectral-spatial diversity, and the last column shows the color represented by each class.

each class. It should be noted that if the sample quantity exceeds the total number, we took 15 samples.

As the number of training samples decreases, the accuracy of the nine classification algorithms shows varying degrees of decline on the four datasets. On the Indian Pine dataset, our proposed algorithm exhibits a relatively small decrease in

accuracy compared to other algorithms. Specifically, spectral-spatial residual blocks networks (SSRN) and 3-D convolutional neural network (3DCNN) algorithms experience a significant decrease in OA, while our algorithm is the only one with accuracy above 80%. DCFSL is sensitive to the number of samples, and as the sample quantity increases, the accuracy tends to

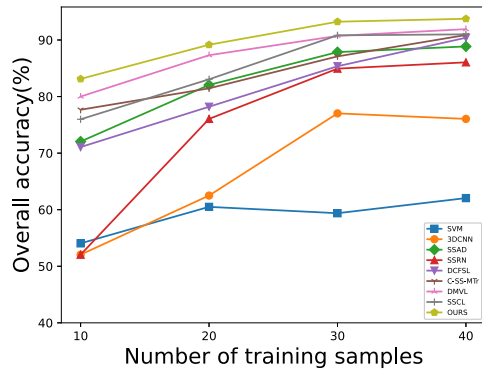


Fig. 9. Classification results in terms of OA values under different training samples on Indian Pines dataset.

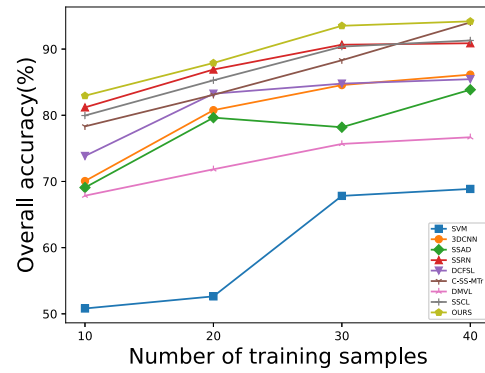


Fig. 12. Classification results in terms of OA values under different training samples on WHU\_Hi\_HanChuan dataset.

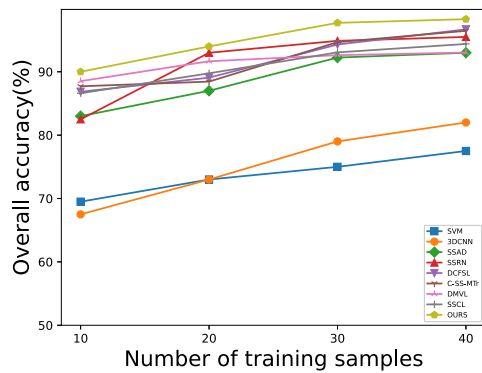


Fig. 10. Classification results in terms of OA values under different training samples on Pavia University dataset.

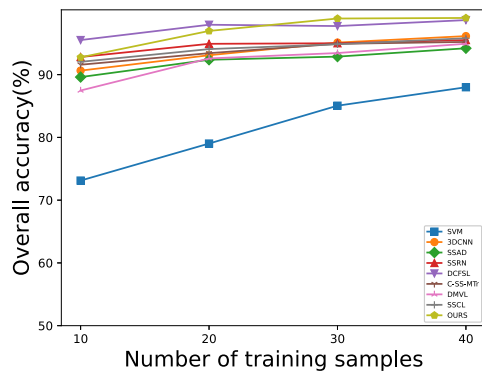


Fig. 11. Classification results in terms of OA values under different training samples on WHU\_Hi\_LongKou dataset.

exhibit nearly linear growth. On the Pavia University dataset, the data sensitivity of various algorithms is similar to that of the Indian Pines dataset. Our algorithm achieves high levels of classification accuracy for 20, 30, and 40 training samples. On the LongKou dataset, due to a significant overlap in the sample space between the Chikusei and LongKou datasets, the transfer effectiveness of DFSQL is quite remarkable, exhibiting the best classification performance when utilizing 10 and 20 samples. Our approach demonstrated superior performance in subsequent stages, delivering the best results with 30 and 40 samples. On

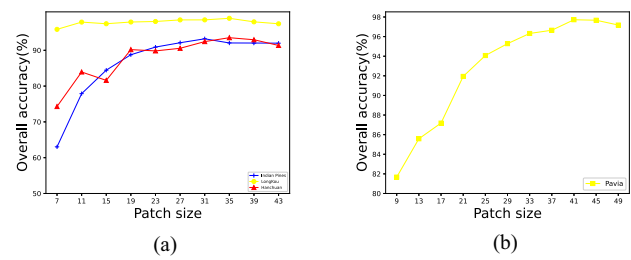


Fig. 13. Parameter analysis of patch size. (a) Performance of different patch size in Indian Pines, LongKou, and Hanchuan datasets. (b) Performance of different patch size in Pavia University dataset.

the HanChuan dataset, the accuracy of C-SS-MTr increases nearly linearly. Overall, our proposed algorithm shows insensitivity to samples and achieves the best performance across different datasets and various numbers of training samples. This also verifies that the contrastive representation learning method and the pretraining method employed in this study effectively enhance the feature extraction capability of the network under small-sample conditions and improve the overall performance of the fusion network.

### E. Analysis of Patch Size

As demonstrated in Fig. 13, the impact of spatial block size on the classification accuracy of various datasets is evident. A larger spatial block size allows for the utilization of more spatial correlation information, but simultaneously, it could introduce interference to the classification from pixels within the spatial block that does not belong to the same class as the target pixel. Moreover, the complex structure requires a larger number of samples for training parameter fitting. Our model harnesses joint spatial-spectral feature information to mitigate the influence of nontarget pixel classes within the spatial block. In addition, we employ contrastive learning, enabling the utilization of a substantial amount of unlabeled samples for model training. The optimal patch size is 35 for the LongKou and Hanchuan datasets, 31 for the IP dataset, and 41 for the PU dataset.

TABLE VIII  
CLASSIFICATION PERFORMANCE OF DIFFERENT MODELS ON THE INDIAN PINES DATASET

| Class No. | RBF-SVM      | 3DCNN        | SSAD                | SSRN                | DCFSL        | C-SS-MTr     | DMVL                | SSCL                | Ours                |
|-----------|--------------|--------------|---------------------|---------------------|--------------|--------------|---------------------|---------------------|---------------------|
| 1         | 81.25 ± 0.12 | 100.00 ± 0.0 | 93.75 ± 0.29        | 100.00 ± 0.0        | 100.00 ± 0.0 | 99.37 ± 1.88 | 57.13 ± 14.58       | 100.00 ± 0.0        | <b>100.00 ± 0.0</b> |
| 2         | 58.94 ± 1.89 | 46.20 ± 3.23 | 76.96 ± 0.28        | 91.18 ± 0.44        | 79.61 ± 4.09 | 77.84 ± 7.21 | <b>92.75 ± 3.32</b> | 84.34 ± 2.71        | 88.68 ± 0.81        |
| 3         | 63.62 ± 0.45 | 55.62 ± 2.00 | 86.00 ± 0.21        | 71.32 ± 0.78        | 82.53 ± 3.26 | 84.56 ± 3.68 | <b>94.32 ± 4.46</b> | 88.12 ± 5.33        | 93.40 ± 0.33        |
| 4         | 78.74 ± 0.61 | 71.98 ± 0.34 | 99.03 ± 0.01        | <b>100.00 ± 0.0</b> | 96.91 ± 2.13 | 98.21 ± 2.05 | 90.65 ± 4.86        | 100.00 ± 0.0        | 98.36 ± 2.57        |
| 5         | 89.62 ± 0.22 | 74.83 ± 0.45 | 88.30 ± 0.54        | 83.85 ± 0.45        | 92.63 ± 5.37 | 87.26 ± 4.26 | 88.21 ± 4.07        | <b>91.00 ± 1.81</b> | 87.28 ± 2.19        |
| 6         | 93.28 ± 0.12 | 90.28 ± 0.60 | <b>99.85 ± 0.79</b> | 88.52 ± 0.63        | 96.71 ± 2.32 | 94.89 ± 4.57 | 86.67 ± 6.35        | 93.61 ± 0.44        | 97.94 ± 1.26        |
| 7         | 76.92 ± 0.43 | 100.00 ± 0.0 | 100.00 ± 0.0        | 0.00 ± 0.0          | 100.00 ± 0.0 | 100.00 ± 0.0 | 57.23 ± 13.19       | 100.00 ± 0.0        | <b>100.00 ± 0.0</b> |
| 8         | 86.60 ± 0.46 | 95.31 ± 0.64 | 100.00 ± 0.0        | 100.00 ± 0.0        | 98.30 ± 1.66 | 99.86 ± 0.40 | 98.10 ± 2.84        | 100.00 ± 0.0        | <b>100.00 ± 0.0</b> |
| 9         | 100.00 ± 0.0 | 100.00 ± 0.0 | 100.00 ± 0.0        | 0.00 ± 0.0          | 100.00 ± 0.0 | 98.00 ± 6.00 | 41.17 ± 10.28       | 100.00 ± 0.0        | <b>100.00 ± 0.0</b> |
| 10        | 46.49 ± 3.67 | 66.66 ± 0.53 | 88.32 ± 0.74        | 85.90 ± 0.34        | 84.03 ± 3.02 | 85.80 ± 4.37 | 86.60 ± 5.74        | 84.36 ± 3.10        | <b>92.97 ± 1.67</b> |
| 11        | 45.16 ± 2.11 | 59.38 ± 0.09 | 83.58 ± 0.33        | 74.55 ± 0.03        | 74.52 ± 1.59 | 83.39 ± 4.23 | <b>94.66 ± 2.23</b> | 88.73 ± 1.22        | 89.71 ± 0.81        |
| 12        | 37.19 ± 0.34 | 60.56 ± 0.53 | 79.75 ± 2.43        | 65.59 ± 0.83        | 85.29 ± 6.25 | 79.15 ± 5.30 | <b>94.26 ± 2.76</b> | 93.46 ± 1.81        | 92.01 ± 2.10        |
| 13        | 94.73 ± 0.67 | 96.00 ± 0.45 | 100.00 ± 0.0        | 100.00 ± 0.0        | 99.89 ± 0.23 | 99.31 ± 1.02 | 90.24 ± 4.94        | 100.00 ± 0.0        | <b>100.00 ± 0.0</b> |
| 14        | 61.52 ± 3.45 | 85.99 ± 0.29 | 97.24 ± 0.11        | <b>99.68 ± 0.10</b> | 93.91 ± 2.97 | 95.41 ± 2.16 | 95.26 ± 3.33        | 96.98 ± 1.53        | 99.06 ± 0.38        |
| 15        | 32.88 ± 1.59 | 74.15 ± 1.84 | 88.76 ± 0.34        | <b>100.00 ± 0.0</b> | 97.02 ± 1.59 | 94.89 ± 3.96 | 87.09 ± 7.20        | 94.20 ± 1.55        | 98.65 ± 0.86        |
| 16        | 93.58 ± 0.29 | 90.47 ± 0.43 | 100.00 ± 0.0        | 100.00 ± 0.0        | 100.00 ± 0.0 | 95.84 ± 2.03 | 58.68 ± 6.45        | 100.00 ± 0.0        | <b>100.00 ± 0.0</b> |
| OA        | 59.38 ± 0.40 | 77.05 ± 0.63 | 87.86 ± 0.32        | 84.94 ± 0.79        | 85.36 ± 0.88 | 87.10 ± 2.32 | 90.73 ± 0.94        | 90.82 ± 0.99        | <b>93.23 ± 0.23</b> |
| AA        | 71.28 ± 1.64 | 81.44 ± 1.10 | 92.59 ± 0.12        | 78.78 ± 0.72        | 92.58 ± 1.32 | 92.11 ± 1.45 | 82.07 ± 1.67        | 94.68 ± 0.58        | <b>96.13 ± 0.34</b> |
| Kappa     | 54.51 ± 0.43 | 70.47 ± 0.77 | 86.11 ± 1.23        | 82.98 ± 0.85        | 83.36 ± 0.98 | 85.35 ± 2.60 | 89.49 ± 1.05        | 89.59 ± 1.13        | <b>92.28 ± 0.26</b> |

### F. Comparison of Classification Results

In this section, we compare our proposed model with RBF-SVM [63], 3DCNN [64], SSAD [65], SSRN [66], DCFSL [47], C-SS-MTr [67], DMVL [68], and SSCL [69] under the same experimental conditions. Among these approaches, SVM is a classic machine learning classification method, and 3DCNN and SSRN are deep learning methodologies that capitalize on the combined spatial and spectral information. SSAD represents an excellent semisupervised HSI classification method, and DCFSL is a transfer learning method. C-SS-MTr, DMVL, and SSCL are HSI classification methods based on contrastive learning. C-SS-MTr is a contrastive learning method based on the transformer architecture. DMVL is a contrastive learning method that explores spectral diversity, while SSCL is a contrastive learning method that explores spatial diversity.

- 1) *RBF-SVM*: Employing a nonlinear Support Vector Machine based on Gaussian radial basis kernel functions for the classification of hyperspectral features.
- 2) *3DCNN*: In consideration of the characteristics unique to hyperspectral data, a 3DCNN architecture has been devised, accompanied by the implementation of regularization strategies to mitigate overfitting concerns.
- 3) *SSAD*: Drawing inspiration from active learning principles, this approach involves iteratively selecting samples to construct a valuable training set, which is subsequently utilized for feature extraction using convolutional networks.
- 4) *SSRN*: Formulating an end-to-end approach involving SSRN designed to independently learn more distinctive spectral and spatial features.
- 5) *DCFSL*: Using meta-learning in both the source and target domains enables the encoder to quickly adapt to new tasks. Subsequently, employing conditional adversarial techniques achieves domain alignment, reducing the difficulty of transfer.

- 6) *C-SS-MTr*: Utilizing HSI patch masking as an augmentation view of the original patch for contrastive learning to obtain instance discrimination information. In addition, it employs a transformer for mask block reconstruction to obtain local semantic information.
- 7) *DMVL*: Treating different spectral bands of HSI as different perspectives of the image allows for the exploration of spectral diversity.
- 8) *SSCL*: Implementing spatial feature augmentation to generate pairs of samples with enhanced spatial features, subsequently employed for contrastive learning within the same class.
  - 1) *Indian Pines*: The classification results of the IP dataset are presented in Table VIII, encompassing the average classification accuracies and standard deviations for each class across five independent experimental runs. In addition, OA, AA, and Kappa coefficient averages and standard deviations over the five experiments are documented. The performance of the RBF-SVM method, which only employs spectral information, overlooking the spatial information, is notably inferior to the 3DCNN. The SSRN model, incorporating residual blocks, effectively retains feature information in deep networks, mitigating accuracy degradation and improving OA by 7.89% compared to the 3DCNN method. However, due to the class imbalance issue within the IP dataset, instances, such as the seventh and ninth classes exhibit terrible classification results in SSRN, owing to their limited sample sizes. Approaches utilizing pretraining on unlabeled samples, such as SSAD and SSCL, demonstrate significant advancements over previous methodologies. C-SS-MTr, built upon the transformer architecture, necessitates a substantial volume of samples for parameter fitting. Consequently, it exhibits a 3.72% lower performance in terms of OA compared to SSCL. In contrast to the SSCL method, which focuses solely on spatial diversity, our proposed model harnesses the diversity of spectral information, leading to improvements of 2.41%, 1.45%, and 2.69% in terms of OA, AA, and Kappa, respectively.

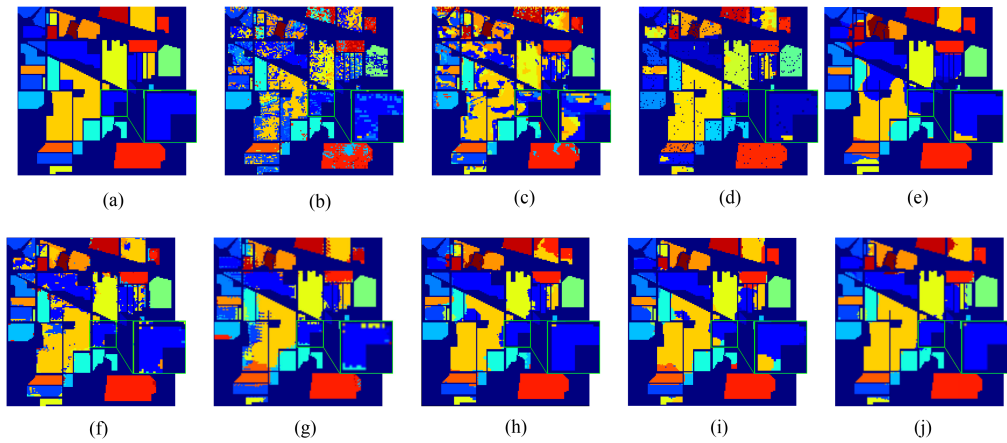


Fig. 14. Classification maps of each algorithm on the Indian Pines dataset. (a) Ground Truth. (b) RBF-SVM (59.38%). (c) 3DCNN (77.05%). (d) SSAD (87.86%). (e) SSRN (84.94%). (f) DCFSL (85.36%). (g) C-SS-MTr (87.10%). (h) DMVL (90.73%). (i) SSCL (90.82%). (j) Proposed method (93.23%).

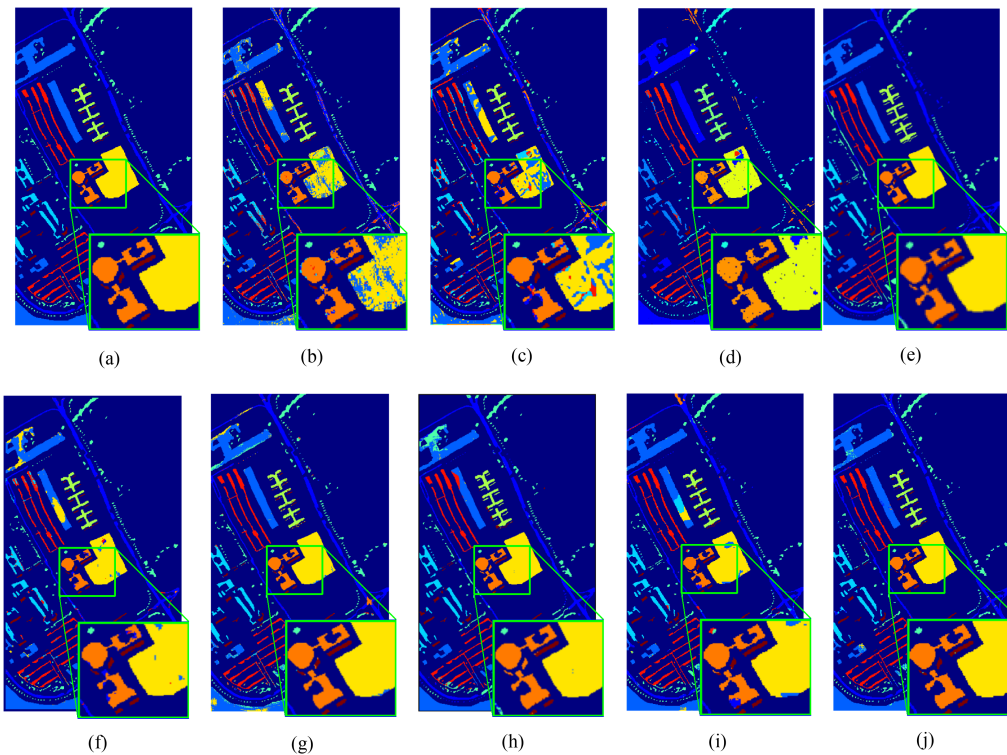


Fig. 15. Classification maps of each algorithm on the Pavia University dataset. (a) Ground Truth. (b) RBF-SVM (79.32%). (c) 3DCNN (79.01%). (d) SSAD (92.24%). (e) SSRN (94.88%). (f) DCFSL (94.30%). (g) C-SS-MTr (94.67%). (h) DMVL (92.63%). (i) SSCL (93.06%). (j) Proposed method (97.72%).

This illustrates that harnessing the diversity within spectral can enhance the performance of models in downstream classification tasks. Moreover, our model attains superior performance across all seven classes, including precise classification of the severely undersampled seventh and ninth classes.

Fig. 14 presents the classification result maps of various algorithms, where our model demonstrates the best performance. Specifically, the utilization of RBF-SVM solely based on spectral information has resulted in the presence of numerous point noises. The SSAD method demonstrates good classification performance at category edges but suffers from numerous misclassified points within regions. The enlarged region typically

consists of a homogeneous medium where samples mostly belong to the same class but with significant spectral variations. Therefore, our approach, along with the spectral diversity mining contrastive method DMVL, has demonstrated promising results. While SSCL focuses solely on exploring spatial diversity, it encounters interference from other classes in the edge regions. Our approach, on the other hand, combines spectral features with spatial information not only preserves edges effectively but also exhibits fewer noise artefacts within regions, outperforming other methods.

2) *Pavia University*: Table IX displays the classification accuracy and evaluation metrics of different algorithms on the

TABLE IX  
CLASSIFICATION PERFORMANCE OF DIFFERENT MODELS ON THE PAVIA UNIVERSITY DATASET

| Class No. | RBF-SVM             | 3DCNN        | SSAD         | SSRN                | DCFSL               | C-SS-MTr            | DMVL                | SSCL         | Ours                |
|-----------|---------------------|--------------|--------------|---------------------|---------------------|---------------------|---------------------|--------------|---------------------|
| 1         | 75.24 ± 1.44        | 75.09 ± 0.96 | 88.44 ± 0.29 | 89.35 ± 0.54        | 93.56 ± 0.85        | <b>96.80 ± 1.05</b> | 92.10 ± 2.50        | 92.68 ± 1.73 | 95.86 ± 0.72        |
| 2         | 85.62 ± 0.94        | 81.88 ± 0.44 | 95.71 ± 0.30 | 97.90 ± 0.09        | 95.37 ± 1.85        | 92.21 ± 3.57        | <b>99.44 ± 0.31</b> | 92.36 ± 1.58 | 97.66 ± 0.59        |
| 3         | 82.45 ± 0.32        | 66.45 ± 2.24 | 96.76 ± 0.24 | <b>98.38 ± 0.07</b> | 87.55 ± 1.47        | 96.19 ± 2.13        | 90.57 ± 1.59        | 94.42 ± 2.25 | 96.72 ± 0.75        |
| 4         | 88.79 ± 0.34        | 91.10 ± 0.32 | 95.25 ± 0.45 | 71.36 ± 1.74        | <b>97.92 ± 0.99</b> | 91.98 ± 1.99        | 66.96 ± 4.80        | 84.66 ± 4.08 | 95.78 ± 0.92        |
| 5         | 99.23 ± 0.12        | 99.46 ± 0.03 | 100.00 ± 0.0 | <b>100.00 ± 0.0</b> | 99.92 ± 0.08        | 97.41 ± 1.95        | 86.68 ± 0.80        | 95.16 ± 1.35 | 99.60 ± 0.11        |
| 6         | 72.41 ± 1.60        | 65.77 ± 1.62 | 91.29 ± 0.16 | 100.00 ± 0.0        | 95.52 ± 0.25        | 98.94 ± 1.21        | 96.14 ± 1.18        | 96.73 ± 1.86 | <b>100.00 ± 0.0</b> |
| 7         | 92.38 ± 0.49        | 85.61 ± 0.90 | 96.07 ± 0.43 | <b>100.00 ± 0.0</b> | 90.79 ± 2.42        | 99.91 ± 0.18        | 95.13 ± 0.85        | 96.33 ± 1.96 | 99.98 ± 0.04        |
| 8         | 84.00 ± 0.33        | 71.96 ± 0.54 | 94.85 ± 0.33 | 97.85 ± 0.21        | 87.22 ± 2.18        | 97.31 ± 1.03        | 89.98 ± 2.73        | 96.27 ± 1.65 | <b>98.36 ± 0.39</b> |
| 9         | <b>99.89 ± 0.01</b> | 98.69 ± 0.09 | 99.78 ± 0.04 | 69.48 ± 1.63        | 99.64 ± 0.22        | 89.64 ± 5.86        | 75.16 ± 10.29       | 93.75 ± 1.47 | 99.40 ± 0.28        |
| OA        | 79.32 ± 0.67        | 79.01 ± 1.27 | 92.24 ± 0.11 | 94.88 ± 0.54        | 94.30 ± 0.38        | 94.67 ± 2.03        | 92.63 ± 0.90        | 93.06 ± 0.29 | <b>97.72 ± 0.19</b> |
| AA        | 83.67 ± 0.26        | 81.78 ± 0.79 | 95.35 ± 0.32 | 91.70 ± 0.57        | 94.17 ± 0.41        | 95.60 ± 1.40        | 88.02 ± 1.51        | 93.60 ± 0.53 | <b>98.16 ± 0.08</b> |
| Kappa     | 73.22 ± 0.80        | 72.75 ± 1.46 | 92.39 ± 0.17 | 93.21 ± 0.95        | 92.49 ± 0.47        | 93.04 ± 2.60        | 90.34 ± 1.14        | 90.92 ± 0.36 | <b>96.96 ± 0.24</b> |

TABLE X  
CLASSIFICATION PERFORMANCE OF DIFFERENT MODELS ON THE WHU-HI-LONGKOU DATASET

| Class No. | RBF-SVM             | 3DCNN               | SSAD          | SSRN                | DCFSL        | C-SS-MTr     | DMVL          | SSCL         | Ours                |
|-----------|---------------------|---------------------|---------------|---------------------|--------------|--------------|---------------|--------------|---------------------|
| 1         | 91.65 ± 0.93        | 97.82 ± 1.97        | 99.33 ± 0.48  | 98.75 ± 0.81        | 99.62 ± 0.19 | 97.92 ± 1.45 | 95.89 ± 2.63  | 97.06 ± 0.39 | <b>99.70 ± 0.08</b> |
| 2         | 20.65 ± 0.15        | 80.61 ± 5.00        | 91.60 ± 9.86  | 98.16 ± 1.28        | 97.94 ± 0.89 | 92.28 ± 1.63 | 93.32 ± 2.23  | 97.50 ± 0.91 | <b>99.85 ± 0.10</b> |
| 3         | 3.33 ± 0.05         | 65.87 ± 6.94        | 77.48 ± 20.82 | <b>99.89 ± 0.08</b> | 97.93 ± 1.31 | 99.25 ± 0.48 | 73.74 ± 8.90  | 99.32 ± 0.82 | 99.79 ± 0.15        |
| 4         | <b>99.13 ± 0.14</b> | 96.81 ± 0.47        | 93.75 ± 3.57  | 92.62 ± 1.74        | 96.49 ± 1.31 | 92.72 ± 2.41 | 96.47 ± 3.24  | 94.62 ± 0.30 | 97.44 ± 0.51        |
| 5         | 12.19 ± 16.31       | 68.89 ± 9.82        | 60.68 ± 38.07 | 99.92 ± 0.06        | 98.90 ± 0.48 | 94.32 ± 3.63 | 70.19 ± 9.13  | 99.75 ± 0.40 | <b>100.00 ± 0.0</b> |
| 6         | 72.72 ± 2.96        | 91.52 ± 5.12        | 95.03 ± 4.65  | 97.50 ± 1.11        | 98.88 ± 0.62 | 99.12 ± 0.72 | 91.40 ± 2.24  | 91.80 ± 2.87 | <b>99.90 ± 0.03</b> |
| 7         | 99.25 ± 0.31        | <b>99.99 ± 0.01</b> | 98.38 ± 0.85  | 96.24 ± 0.75        | 99.80 ± 0.16 | 95.58 ± 2.10 | 99.81 ± 0.17  | 95.89 ± 0.93 | 99.90 ± 0.02        |
| 8         | 39.39 ± 33.35       | 91.74 ± 3.80        | 89.58 ± 1.65  | 73.74 ± 4.94        | 92.84 ± 2.74 | 85.23 ± 4.91 | 73.87 ± 7.50  | 76.46 ± 6.72 | <b>96.60 ± 0.55</b> |
| 9         | 4.88 ± 6.89         | 68.34 ± 11.08       | 71.34 ± 8.18  | 94.55 ± 1.23        | 97.50 ± 1.19 | 94.15 ± 2.66 | 66.03 ± 14.21 | 89.30 ± 4.11 | <b>97.86 ± 0.34</b> |
| OA        | 85.05 ± 1.18        | 95.11 ± 0.51        | 92.87 ± 1.17  | 95.00 ± 0.32        | 98.27 ± 0.39 | 94.91 ± 0.70 | 93.44 ± 1.26  | 97.84 ± 0.50 | <b>98.94 ± 0.16</b> |
| AA        | 48.05 ± 5.80        | 84.59 ± 1.97        | 83.01 ± 6.86  | 94.60 ± 0.24        | 97.77 ± 0.31 | 94.84 ± 0.48 | 84.53 ± 1.72  | 93.61 ± 0.88 | <b>98.91 ± 0.12</b> |
| Kappa     | 79.40 ± 1.75        | 93.55 ± 0.67        | 90.62 ± 1.44  | 93.51 ± 0.41        | 97.74 ± 0.50 | 93.40 ± 0.89 | 91.50 ± 1.62  | 93.29 ± 0.64 | <b>98.61 ± 0.20</b> |

TABLE XI  
CLASSIFICATION PERFORMANCE OF DIFFERENT MODELS ON THE WHU-HI-HANCHUAN DATASET

| Class No. | RBF-SVM             | 3DCNN               | SSAD          | SSRN                | DCFSL        | C-SS-MTr     | DMVL                | SSCL                | Ours                |
|-----------|---------------------|---------------------|---------------|---------------------|--------------|--------------|---------------------|---------------------|---------------------|
| 1         | 90.21 ± 1.69        | <b>96.36 ± 0.81</b> | 92.79 ± 2.73  | 96.15 ± 0.15        | 86.10 ± 3.21 | 91.09 ± 4.42 | 91.30 ± 4.59        | 90.39 ± 0.37        | 93.89 ± 1.14        |
| 2         | 60.71 ± 13.44       | 83.93 ± 1.13        | 63.66 ± 7.83  | 80.69 ± 3.38        | 64.97 ± 5.43 | 85.61 ± 1.66 | <b>91.69 ± 3.39</b> | 84.52 ± 1.36        | 91.04 ± 1.22        |
| 3         | 26.22 ± 12.67       | 50.24 ± 5.94        | 67.11 ± 12.73 | 95.96 ± 1.81        | 84.04 ± 2.77 | 88.91 ± 2.21 | 87.52 ± 4.01        | 96.27 ± 0.47        | <b>96.66 ± 0.92</b> |
| 4         | 11.81 ± 14.69       | 97.89 ± 0.63        | 80.21 ± 9.68  | <b>99.91 ± 0.15</b> | 95.62 ± 2.56 | 97.35 ± 0.66 | 83.14 ± 4.47        | 93.31 ± 3.61        | 86.32 ± 8.76        |
| 5         | 0.02 ± 0.03         | 13.66 ± 10.20       | 13.95 ± 17.20 | 85.31 ± 24.87       | 97.18 ± 0.99 | 98.43 ± 2.39 | 81.46 ± 9.67        | <b>99.94 ± 0.06</b> | 99.87 ± 0.23        |
| 6         | 0.02 ± 0.03         | 37.43 ± 1.33        | 34.84 ± 8.67  | 82.65 ± 10.72       | 63.08 ± 1.35 | 75.72 ± 5.74 | 74.97 ± 5.45        | 81.78 ± 3.58        | <b>99.09 ± 2.76</b> |
| 7         | 20.33 ± 15.65       | 68.81 ± 3.42        | 49.79 ± 11.78 | <b>96.02 ± 4.36</b> | 92.74 ± 3.22 | 91.09 ± 3.51 | 26.50 ± 3.12        | 95.18 ± 2.27        | 95.27 ± 1.61        |
| 8         | 45.12 ± 25.70       | 76.07 ± 1.90        | 58.63 ± 3.29  | 87.20 ± 3.81        | 66.96 ± 3.06 | 79.91 ± 4.87 | 86.30 ± 5.59        | 84.27 ± 0.62        | <b>88.21 ± 2.18</b> |
| 9         | 3.74 ± 4.82         | 57.43 ± 4.96        | 52.81 ± 14.94 | 91.96 ± 2.49        | 71.92 ± 5.32 | 78.79 ± 8.05 | 58.92 ± 5.55        | 77.88 ± 2.33        | <b>93.01 ± 1.71</b> |
| 10        | 13.88 ± 8.97        | 93.76 ± 6.35        | 95.83 ± 1.92  | 96.44 ± 3.13        | 90.47 ± 5.01 | 79.34 ± 6.81 | 58.09 ± 7.18        | 94.37 ± 0.25        | <b>96.87 ± 0.93</b> |
| 11        | 31.15 ± 10.46       | <b>91.37 ± 1.64</b> | 59.33 ± 8.92  | 96.08 ± 1.27        | 88.64 ± 1.70 | 79.66 ± 5.68 | 66.48 ± 8.34        | 86.08 ± 3.49        | 90.70 ± 3.77        |
| 12        | 0.00 ± 0.0          | 23.70 ± 4.73        | 13.49 ± 16.52 | 89.67 ± 5.91        | 82.63 ± 2.68 | 96.38 ± 1.80 | 78.49 ± 6.10        | 97.39 ± 0.55        | <b>99.84 ± 0.19</b> |
| 13        | 3.64 ± 5.66         | 48.25 ± 3.36        | 42.57 ± 12.91 | 49.32 ± 4.45        | 65.99 ± 0.71 | 78.12 ± 5.19 | 50.77 ± 6.08        | 79.72 ± 1.17        | <b>83.90 ± 1.86</b> |
| 14        | 53.27 ± 4.73        | 81.76 ± 2.53        | 69.85 ± 3.83  | 76.89 ± 2.79        | 80.07 ± 2.95 | 83.23 ± 4.96 | 61.77 ± 6.98        | 87.36 ± 2.11        | <b>89.74 ± 1.87</b> |
| 15        | 0.04 ± 0.07         | 15.31 ± 5.47        | 75.65 ± 20.50 | 79.95 ± 4.80        | 91.35 ± 3.41 | 88.63 ± 7.58 | 54.04 ± 12.48       | <b>94.48 ± 0.44</b> | 91.29 ± 0.67        |
| 16        | <b>99.14 ± 0.26</b> | 98.05 ± 0.23        | 96.72 ± 1.39  | 95.70 ± 2.02        | 97.43 ± 0.60 | 95.58 ± 1.58 | 91.35 ± 3.55        | 96.13 ± 0.51        | 97.53 ± 1.07        |
| OA        | 67.82 ± 11.74       | 84.56 ± 0.28        | 78.19 ± 1.67  | 90.66 ± 0.50        | 84.78 ± 0.63 | 88.31 ± 1.05 | 75.67 ± 1.60        | 90.37 ± 0.23        | <b>93.52 ± 0.51</b> |
| AA        | 29.10 ± 1.75        | 64.65 ± 1.26        | 60.45 ± 1.89  | 87.98 ± 1.09        | 82.45 ± 0.91 | 86.74 ± 1.24 | 71.43 ± 1.54        | 89.94 ± 0.68        | <b>92.67 ± 0.87</b> |
| Kappa     | 53.37 ± 1.61        | 81.88 ± 0.33        | 74.33 ± 1.98  | 89.12 ± 0.58        | 82.32 ± 0.71 | 86.40 ± 1.19 | 72.18 ± 1.68        | 88.79 ± 0.27        | <b>92.44 ± 0.60</b> |

PU dataset. The PU dataset exhibits a scattered sample distribution and contains numerous boundary regions. When relying on spatial information and utilizing image patches, there might be interference from incongruent information with the target pixel, potentially affecting classification. SSAD, through its semisupervised training approach, achieves better results for RBF-SVM and 3DCNN by selecting samples that contribute to classification. In contrast, SSCL primarily explores spatial diversity, disregarding spectral diversity. Consequently, its performance is relatively inferior when compared to the SSRN

method, which leverages a dual-branch network approach incorporating spatial and spectral information. Compared to SSRN, our proposed method achieves notable improvements of 2.84% in AA, 6.46% in OA, and 3.75% in the Kappa coefficient.

Fig. 15 displays the classification result maps of various algorithms on the Pavia University dataset. The enlarged region encompasses multiple different classes, and because the spectral of different classes may be similar, using SVM methods that rely solely on spectral information results in significant point noise. SSCL, due to interference from nontarget pixel classes at

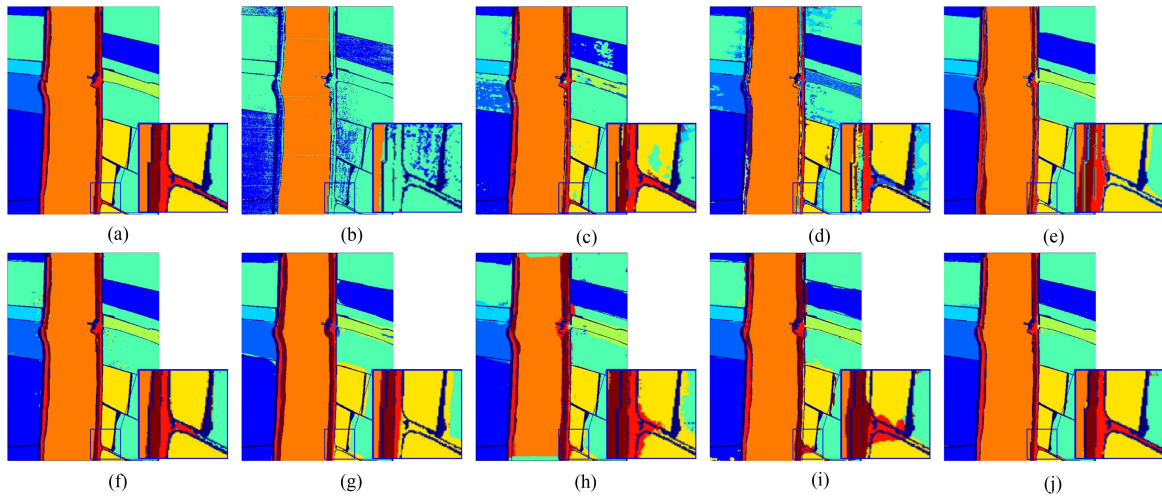


Fig. 16. Classification maps of each algorithm on the WHU-Hi-LongKou dataset. (a) Ground Truth. (b) RBF-SVM (85.05%). (c) 3DCNN (95.11%). (d) SSAD (92.87%). (e) SSRN (95.00%). (f) DCFSL (98.27%). (g) C-SS-MTr (94.91%). (h) DMVL (93.44%). (i) SSCL (97.84%). (j) Proposed method (98.94%).

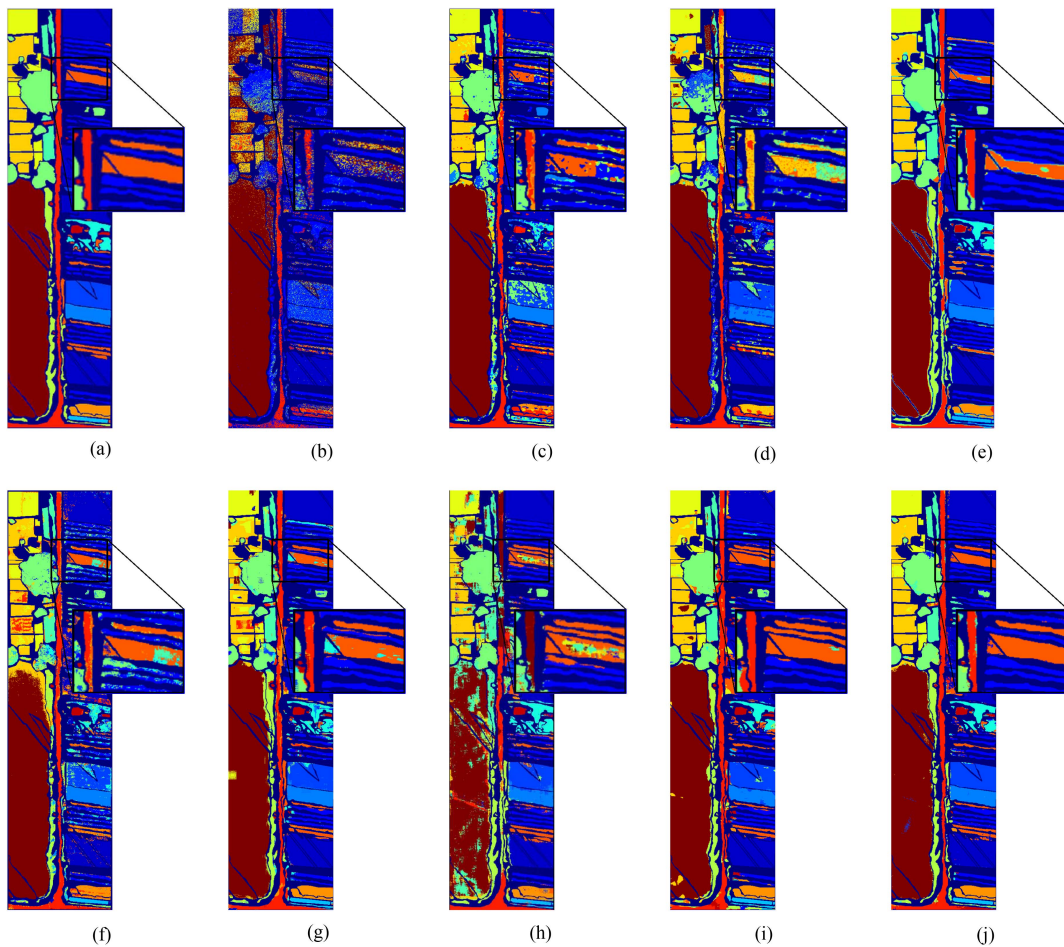


Fig. 17. Classification maps of each algorithm on the WHU-Hi-HanChuan dataset. (a) Ground Truth. (b) RBF-SVM (67.82%). (c) 3DCNN (84.56%). (d) SSAD (78.19%). (e) SSRN (90.66%). (f) DCFSL (84.78%). (g) C-SS-MTr (88.31%). (h) DMVL (75.67%). (i) SSCL (90.37%). (j) Proposed method (93.52%).

category edges, exhibits suboptimal classification performance. In contrast, DMVL only exhibits sporadic misclassification points within the regions. Our approach, however, demonstrates the best results in addressing these challenges.

3) *WHU-Hi-LongKou*: Table X presents the classification results of various methods on the LongKou dataset. The LongKou dataset has fewer boundary regions and more homogeneous areas, making SSCL methods that explore spatial diversity

superior to DMVL methods, resulting in a relative accuracy improvement of 4.4%. Regarding the DCFSL, the sample space distribution between the Chikusei dataset and the LongKou dataset is similar, and the sample space deviation is relatively small, making the transfer relatively easier, the classification accuracy reached the second best. However, our method still outperformed it by 0.67% in OA, 1.14% in AA, and 0.87% in Kappa, our method achieves the highest accuracy among the six classes.

Fig. 16 presents the classification result maps of various algorithms on the LongKou dataset. Similar to the previous datasets, our method achieves the best classification results in the enlarged region by combining spatial and spectral features. The SSCL, which explores spatial diversity, is influenced by pixels from different classes within the same image block, leading to classification errors in boundary areas. Meanwhile, due to the characteristic of hyperspectral data, where different materials may have similar spectral, DMVL exhibits some misclassification points within regions.

4) *WHU-Hi-HanChuan*: The classification results for the HanChuan dataset are presented in Table XI. Due to the influence of sunlight angles on dataset acquisition, resulting in a significant amount of noise in the dataset, the DMVL method, which aims to extract spectral diversity, performs poorly on this dataset. Compared to SSCL, our proposed method achieves better improvements of 3.15% in OA, 2.73% in AA, and 3.65% in Kappa coefficient. Among these 16 classes, 8 classes achieved the best classification performance.

Fig. 17 illustrates the classification maps for different methods on the HanChuan dataset. Regarding the SSAD, it can be observed in the zoomed-in regions that suboptimal sample selection leads to error accumulation, resulting in nearly the entire area being misclassified in certain cases. Our method exhibits the best classification performance both within the region and along the boundaries.

#### IV. CONCLUSION

In this article, we propose a two-stream contrastive learning network to address the challenge of few-shot HSI classification. Building upon the existing contrastive learning framework that primarily explores spatial feature diversity within surface samples, we introduce an additional stream dedicated to exploiting the spectral diversity of target points. Meanwhile, we devise four distinct data augmentation methods to enrich spectral diversity. Moreover, to mitigate the issue of inconsistency between upstream and downstream tasks, we architect a multilevel network structure, this structure fuses information from varying network levels, rendering the model more generalized. Extensive experimental results and visualization of feature maps indicate that our proposed approach can uncover discriminative features that are more conducive to generalization for downstream tasks. It can improve the classification accuracy on four public datasets. Our approach delves into both spatial and spectral aspects of contrastive learning, ignoring mutual interaction between the two types of features during the pretraining phase.

In the future, we aspire to leverage deep learning networks, such as GAN-based models, to design more effective data augmentation techniques. In addition, we intend to craft a

module that facilitates interaction between spatial and spectral information.

#### REFERENCES

- [1] T. H. Kurz, S. J. Buckley, and J. A. Howell, "Close-range hyperspectral imaging for geological field studies: Workflow and methods," *Int. J. Remote Sens.*, vol. 34, no. 5, pp. 1798–1822, 2013.
- [2] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, 2014, Art. no. 010901.
- [3] E. K. Hege, D. O'Connell, W. Johnson, S. Basty, and E. L. Dereniak, "Hyperspectral imaging for astronomy and space surveillance," in *Proc. Conf. SPIE Imag. Spectrometry IX*, 2004, vol. 5159, pp. 380–391.
- [4] Y.-Z. Feng and D.-W. Sun, "Application of hyperspectral imaging in food safety inspection and control: A review," *Crit. Rev. Food Sci. Nutr.*, vol. 52, no. 11, pp. 1039–1058, 2012.
- [5] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas, "Modern trends in hyperspectral image analysis: A review," *IEEE Access*, vol. 6, pp. 14118–14129, 2018.
- [6] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.
- [7] G. Mercier and M. Lennon, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2003, vol. 1, pp. 288–290.
- [8] P. Quesada-Barriuso, F. Argüello, and D. B. Heras, "Spectral–spatial classification of hyperspectral images using wavelets and extended morphological profiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1177–1185, Apr. 2014.
- [9] W. Duan, S. Li, and L. Fang, "Superpixel-based composite kernel for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1698–1701.
- [10] J. C.-W. Chan and D. Paelinckx, "Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2999–3011, 2008.
- [11] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image classification using soft sparse multinomial logistic regression," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 2, pp. 318–322, Mar. 2012.
- [12] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.
- [13] 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.
- [14] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
- [15] W. Wang, X. Ma, L. Leng, Y. Wang, B. Liu, and J. Sun, "A hybrid CNN based on global reasoning for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6012605.
- [16] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
- [17] H. Xu, W. Yao, L. Cheng, and B. Li, "Multiple spectral resolution 3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 7, 2021, Art. no. 1248.
- [18] D. Hong et al., "Spectralformer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5518615.
- [19] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5514715.
- [20] 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.
- [21] J. Xu, J. Zhao, and C. Liu, "An effective hyperspectral image classification approach based on discrete wavelet transform and dense CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6011705.
- [22] L. Liang, S. Zhang, and J. Li, "Multiscale densenet meets with Bi-RNN for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5401–5415, 2022.



- [23] Z. Xiaojin, "Semi-supervised learning literature survey," Dept. of Comput. Sci., Univ. of Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1530, 2006.
- [24] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," 2020, *arXiv:2006.05278*.
- [25] M. Huisman, J. N. Van Rijn, and A. Plaata, "A survey of deep meta-learning," *Artif. Intell. Rev.*, vol. 54, no. 6, pp. 4483–4541, 2021.
- [26] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.
- [27] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," 2017, *arXiv:1702.05374*.
- [28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
- [29] W. Ying, Y. Zhang, J. Huang, and Q. Yang, "Transfer learning via learning to transfer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5085–5094.
- [30] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
- [31] T. S. Reddy and J. Hariikiran, "A semi-supervised cycle-GAN neural network for hyperspectral image classification with minimum noise fraction," *J. Spectral Imag.*, vol. 11, 2022, Art. no. a2.
- [32] X. Zheng et al., "Hyperspectral image classification with imbalanced data based on semi-supervised learning," *Appl. Sci.*, vol. 12, no. 8, 2022, Art. no. 3943.
- [33] Z. Wang and B. Du, "Unified active and semi-supervised learning for hyperspectral image classification," *GeoInformatica*, vol. 27, no. 1, pp. 23–38, 2023.
- [34] S. S. Sawant and M. Prabukumar, "A review on graph-based semi-supervised learning methods for hyperspectral image classification," *Egyptian J. Remote Sens. Space Sci.*, vol. 23, no. 2, pp. 243–248, 2020.
- [35] Z. He, K. Xia, P. Ghamisi, Y. Hu, S. Fan, and B. Zu, "HypervitGAN: Semisupervised generative adversarial network with transformer for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6053–6068, 2022.
- [36] H. Pan, M. Liu, H. Ge, and S. Chen, "Semi-supervised spatial-spectral classification for hyperspectral image based on three-dimensional gabor and co-selection self-training," *J. Appl. Remote Sens.*, vol. 16, no. 2, pp. 028501–028501, 2022.
- [37] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Collaborative learning of lightweight convolutional neural network and deep clustering for hyperspectral image semi-supervised classification with limited training samples," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 164–178, 2020.
- [38] Y. Yang et al., "Semi-supervised multiscale dynamic graph convolution network for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022, to be published, doi: [10.1109/TNNLS.2022.3212985](https://doi.org/10.1109/TNNLS.2022.3212985).
- [39] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang, "Deep few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2290–2304, Apr. 2019.
- [40] D. Pal, V. Bunde, B. Banerjee, and Y. Jeppu, "SPN: Stable prototypical network for few-shot learning-based hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 5506905.
- [41] D. AL-Alimi, M. A. Al-qaness, Z. Cai, A. Dahou, Y. Shao, and S. Issaka, "Meta-learner hybrid models to classify hyperspectral images," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 1038.
- [42] K. Gao, B. Liu, X. Yu, P. Zhang, X. Tan, and Y. Sun, "Small sample classification of hyperspectral image using model-agnostic meta-learning algorithm and convolutional neural network," *Int. J. Remote Sens.*, vol. 42, no. 8, pp. 3090–3122, 2021.
- [43] H. Wu, M. Li, and A. Wang, "A novel meta-learning-based hyperspectral image classification algorithm," *Front. Phys.*, vol. 11, 2023, Art. no. 1163555.
- [44] Y. Wang, X. Chen, F. Wang, M. Song, and C. Yu, "Meta-learning based hyperspectral target detection using Siamese network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5527913.
- [45] W. Wang, F. Liu, J. Liu, and L. Xiao, "Cross-domain few-shot hyperspectral image classification with class-wise attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5502418.
- [46] H. Lee, S. Eum, and H. Kwon, "Exploring cross-domain pretrained model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5526812.
- [47] S. Zhang, Z. Chen, D. Wang, and Z. J. Wang, "Cross-domain few-shot contrastive learning for hyperspectral images classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5514505.
- [48] Y. Wang, C. M. Albrecht, N. A. A. Braham, L. Mou, and X. X. Zhu, "Self-supervised learning in remote sensing: A review," 2022, *arXiv:2206.13188*.
- [49] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1422–1430.
- [50] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.
- [51] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
- [52] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.
- [53] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
- [54] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
- [55] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3733–3742.
- [56] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 776–794.
- [57] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.
- [58] B. Deng, S. Jia, and D. Shi, "Deep metric learning-based feature embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1422–1435, Feb. 2020.
- [59] W. Song, Y. Dai, Z. Gao, L. Fang, and Y. Zhang, "Hashing-based deep metric learning for the classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5704513.
- [60] J. Li, X. Li, Z. Cao, and L. Zhao, "Robyol: Random-occlusion-based byol for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6014405.
- [61] Z. Cao, X. Li, Y. Feng, S. Chen, C. Xia, and L. Zhao, "Contrastnet: Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral imagery classification," *Neurocomputing*, vol. 460, pp. 71–83, 2021.
- [62] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] L. Hu, X. Luo, and Y. Wei, "Hyperspectral image classification of convolutional neural network combined with valuable samples," in *Proc. J. Phys.: Conf. Ser.*, 2020, vol. 1549, Art. no. 052011.
- [67] L. Huang, Y. Chen, and X. He, "Spectral-spatial masked transformer with supervised and contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508718.
- [68] 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.
- [69] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5521213.

**Shuxiang Xia** received the B.S. degree in computer science and technology from the Xi'an University of Technology, Xi'an, China, where he is currently working toward the master's degree with Xidian University, Xi'an, China. His main research interests include deep learning, computer vision, and hyperspectral image classification.

**Xiaohua Zhang** (Member, IEEE) received the B.S. and M.S. degrees in applied mathematics from Northwest University, Xi'an, China, in 2000 and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2004.

In 2009, he was a Visiting Scholar with the Georgia Institute of Technology, Atlanta, GA, USA. He is currently an Associate Professor with the School of Artificial Intelligence, Xidian University, and a Member of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. His research interests include image processing, deep learning, computer vision, and remote sensing image processing.

Dr. Zhang is a Senior Member of the Chinese Institute of Electronics and the China Computer Federation.

**Hongyun Meng** received the Ph.D. degree in operating research and control from the School of Mathematics and Statistics, Xidian University, Xi'an, China, in 2004.

She is currently an Associate Professor with the School of Mathematics and Statistics, Xidian University. Her research interests include intelligent information processing, natural computing, and multiobjective optimization.

**Jiaxin Fan** is currently working toward the master's degree with the University of New South Wales, Business School, Kensington, NSW, Australia.

His main research interests include deep learning, computer vision, and hyperspectral image classification.

**Licheng Jiao** (Fellow, IEEE) received the B.S. degree in electrical and computer science from Shanghai Jiao Tong University, Shanghai, China, in 1982 and the M.S. and Ph.D. degrees in theoretical electrician from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, Xi'an, China, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. His research interests include image processing, natural computation, machine learning, and intelligent information processing.

Dr. Jiao is a Foreign Member of the Academia Europaea and the Russian Academy of Natural Sciences. He is a Fellow of the Institution of Engineering and Technology, Chinese Association for Artificial Intelligence, Chinese Institute of Electronics, China Computer Federation, and the Chinese Association of Automation. He is a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, and an Expert of the Academic Degrees Committee of the State Council. He is the Chairman of the Awards and Recognition Committee and the Vice Board Chairperson of the Chinese Association of Artificial Intelligence.