Self-Supervised Learning Driven Cross-Domain Feature Fusion Network for Hyperspectral Image Classification

Qizhi FANG, Yubo ZHAO, Jingang WANG, Lili ZHANG

College of Electronic and Information Engineering, Shenyang Aerospace University, 110136, Shenyang, China

{arinc2006, 20052727}@sau.edu.cn, {zhaoyubo, wangjingang}@stu.sau.edu.cn

Submitted April 19, 2025 / Accepted June 25, 2025 / Online first July 21, 2025

Abstract. *Hyperspectral image (HSI) classification faces significant challenges due to the high cost of acquiring labeled samples. To mitigate this, we propose SSCF-Net, a novel self-supervised learning driven cross-domain feature fusion Network. SSCF-Net uniquely leverages readily available labeled natural images (source domain) to aid HSI (target domain) classification by transfer learning. Specifically, we employ rotation-based self-supervision in the source domain to learn transferable features, which are then transferred to the HSI domain. Within SSCF-Net, we effectively fuse local and global features: local features are extracted by a jointly trained module combining VGG and two-dimensional long short-term memory networks (TDLSTM) networks, while global features capturing long-range dependencies are learned via a Transformer model. Crucially, in the HSI domain, we further employ contrastive learning as a self-supervised strategy to maximally utilize the limited labeled data. Extensive experiments on three benchmark HSI datasets (Salinas, Indian Pines, WHU-HiLongKou) demonstrate that SSCF-Net significantly outperforms existing methods, validating its effectiveness in addressing the label scarcity problem. The code is available at https://github.com/6pangbo/SSCF-Net.*

Keywords

Hyperspectral image classification, self-supervised learning, transfer learning, feature fusion

1. Introduction

Hyperspectral image (HSI), acquired by airborne sensors, captures three-dimensional data containing rich spatial and spectral information across numerous contiguous wavelengths. This detailed reflectance signature enables precise identification of materials, making hyperspectral image classification (HSIC) a critical task with diverse applications. However, the complexity of land cover distributions makes acquiring sufficient labeled samples costly and time-consuming, posing a major challenge to effective classification [1].

Early HSIC research relied on algorithms like Support Vector Machine [2], Random Forest [3], and Neural Networks [4]. The advent of deep learning has revolutionized the field, with methods such as Stacked Autoencoders [5], Deep Belief Networks [6] and particularly Convolutional Neural Networks (CNN) demonstrating substantial improvements. CNN is widely used due to its powerful feature extraction ability [7]. Chen et al. [8] pioneered their use for extracting regularized deep features in HSIC. Considering the different structural scales in HSI, Ye et al. [9] proposed a lightweight multi-scale CNN, and Zhang et al. [10] constructed a multi-scale dense network to extract fusing features at different scales. Fang et al. [11] proposed FDEGCNet, which allows the CNN to dynamically focus on important features and capture cross-dimensional context. Currently, the graph convolutional network (GCN) treats the image pixels as graph nodes connected by edges representing spatial relationships, performing convolution directly on the graph structure to capture the interactions of the nodes. Ding et al. [12] proposed a diversity connection GCN to improve graph structure quality, while Liu et al. [13] proposed a comparison GCN with skip connection to solve the problem that the potentially important information is submerged in the iterative convolutional process. Additionally, other deep learning-based methods are also attempting to classify images, including Recurrent Neural Networks (RNN) [14] and Long Short-Term Memory (LSTM) [15].

While CNNs are powerful for local patterns, they often struggle with long-range dependencies. Transformer, built on self-attention mechanisms, addresses this limitation by globally weighting feature importance. Relevant researches in recent years include that Sun et al. [16] proposed a spectral-spatial tokenization Transformer for HSIC, while Mei et al. [17] proposed a hierarchical Transformer using local spatial-spectral attention. Although Vision Transformer (ViT) has shown some potential in image classification tasks [18], its performance is limited when dealing with large-scale datasets due to the lack of hierarchical feature extraction capability. To address this issue, Swin Transformer was introduced [19]. It combines local window self-attention and hierarchical feature representation to improve the processing ability for high-resolution images. Recog-

DOI: 10.13164/re.2025.0494

nizing the complementary strengths of CNNs (local detail) and Transformers (global context), recent research focuses on hybrid CNN-Transformer architectures. Qi et al. [20] proposed a method called global-local 3-D convolutional Transformer network. This network innovatively embedded three-dimensional convolution into a dual-branch Transformer structure to capture local-global correlation in spectral and spatial domains. Feng et al. [21] proposed a hybrid network based on multiple vision architectures-based hybrid network for HSIC. The framework consisted of a joint CNN and Transformer structure. It also included a GCN, which realized the integration of different methods and aimed to capture various types of feature information.

Despite these advances, a fundamental challenge persists: deep learning models typically demand large labeled datasets, which are scarce and costly to obtain for HSI. To alleviate this label dependency (only 1 to 5 labeled samples for training), researchers have turned to self-supervised learning [22]. Self-supervised learning leverages the intrinsic properties of data to learn effective feature representations before formal supervised learning. In HSIC, self-supervised learning aids in understanding data complexity and handling high-dimensional data. Bai et al. [23] proposed a hyperspectral classification method using masked self-supervised pretraining, which enables effective model training across datasets. Cao et al. [24] proposed an efficient hybrid self-supervised learning method that fully integrates the generative-based method and the contrastive-based method, and achieved high stability and strong reliability. Ye et al. [25] introduced a novel unsupervised approach called self-supervised learning with the multiscale densely connected network to make full use of unlabeled samples for HSIC. He et al. [26] trained the model by randomly masking image blocks and reconstructing them. The self-supervised learning model was then trained by minimizing the difference between the reconstructed data and the input data. Zhou et al. [27] proposed a new HSIC method called masked spectral-spatial feature prediction. This method help Transformer understand the complex spectral-spatial structure of unmarked HSI and further improve the classification performance.

Few-shot learning (FSL) [28] has emerged as another key solution for label dependency. Currently, many FSL-based HSIC methods focus on transferring meta-knowledge within HSI data. Li et al. [29] addressed both FSL and domain adaptation issues within an integrated framework resistant to domain shift. Wang et al. [30] enhanced model classification ability by learning transferable spatial structure and texture information from natural images. Zhang et al. [31] combined FSL with graph-based domain alignment and proposed a cross-domain FSL framework based on graph information aggregation. An FSL classification framework based on self-supervised learning is proposed by Li et al. [32], which integrated a spatial-spectral feature extraction network to achieve good classification results. Xiao et al. [33] developed an embedding feature extractor based on neural architecture search, which aggregated heterogeneous and homogeneous source data with a multi-source learning framework. Qin et al. [34] employed an orthogonal low-rank feature disentanglement method, which allowed the model to implicitly focus on the inherent knowledge.

Despite advances in FSL for HSIC, significant challenges persist under extreme label scarcity. Current methods predominantly rely on CNN for spatial-spectral fusion, largely overlooking the potential of Transformer architectures [30]. Moreover, the scarcity and high cost of obtaining labeled hyperspectral data severely limit model training [35]. Crucially, existing cross-domain approaches often struggle to adequately bridge the significant domain shift between the labeled images (source domain) and HSIs (target domain), resulting in compromised classification accuracy on the target data [29], [31].

To address these challenges-leveraging underutilized Transformers, maximizing labeled data, and bridging the domain gap, the Self-Supervised Learning Driven Cross-domain Feature fusion Network (SSCF-Net) is proposed. Our core strategy is to fully exploit abundant labeled natural images to boost HSIC under label scarcity, facilitated by tailored self-supervision in both domains. SSCF-Net operates on two synergistic paths: Source domain: We leverage a combination of VGG and two-dimensional LSTM (TD-LSTM). The goal of this combination is to take advantage of the powerful feature extraction capability of VGG for natural images and the sequence modeling ability of TD-LSTM. This collaborative approach generates weight transfers that are then used to improve performance. Target Domain: We design a novel hybrid branch integrating a local feature extractor (VGG-based) and a global context encoder based on Swin Transformer. This architecture uniquely combines the complementary strengths of CNNs for local patterns and Transformers for long-range dependencies within HSIs. Self-supervised learning is pivotal to our approach, enhancing feature robustness and mitigating the source-target domain discrepancy. Self-supervised learning in source domain: Natural images are augmented via rotation transformations, with rotation prediction as the pretext task. This encourages learning orientation-invariant features beneficial for diverse object appearances in HSIs. Self-supervised learning in target domain: Gaussian noise is added to HSI patches. The denoising task helps the model learn robust representations invariant to common sensor noise and environmental variations. Our principal contributions are:

- We introduce SSCF-Net, a novel framework integrating cross-domain knowledge transfer and dual-domain self-supervised learning to tackle HSIC under severe label constraints.
- In the source domain, a combination of VGG and TD-LSTM is designed for processing natural images. The VGG extracts features from natural images, which are then processed by TD-LSTM to capture contextual relationships. This method of feature extraction and se-

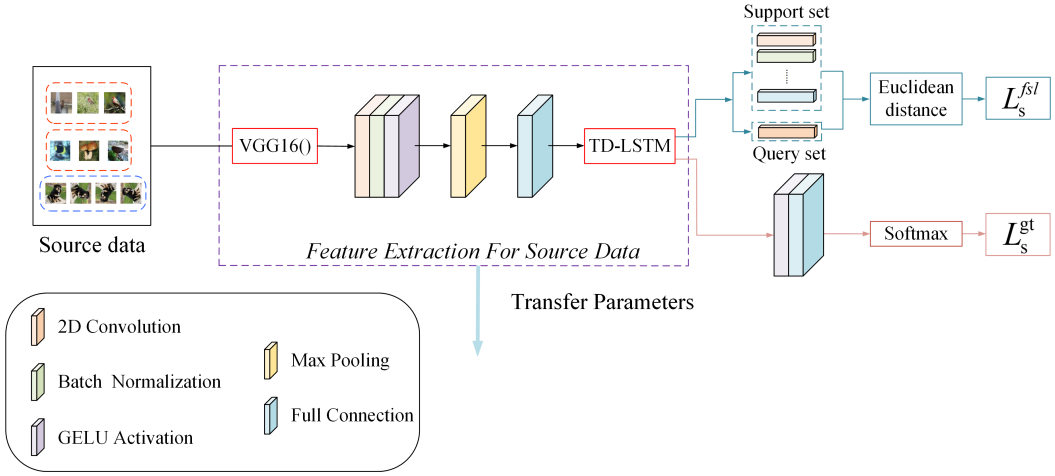

Fig. 1. The overall framework of the source domain.

quence modeling is transferred to hyperspectral data, which can effectively improve classification accuracy under conditions of scarce labels.

- In the target domain, a local-global feature extraction branch (LGFEB) is designed. LGFEB consists of three modules: local feature extraction, global feature extraction and feature fusion. This design utilizes the general features learned from the source domain. It also deeply explores the detailed and global information in the target domain, which achieves comprehensive learning of HSI data.
- Self-supervised strategies are used in both natural images and HSIs. Natural images are augmented with rotation transformations, while HSIs are added with random Gaussian noise. The former improves generalization ability to different perspectives, while the latter simulates sensor noise and environmental variations. This helps the model learn subtle differences.

The remainder of this paper is organized as follows. Section 2 presents a detailed overview of the proposed method. Section 3 describes the experimental datasets and results. Section 4 provides the conclusion of the study.

2. SSCF-Net for HSI

SSCF-Net consists of three core modules: the feature extraction module for the source domain, the LGFEB for the target domain, and the local-global feature fusion module.

2.1 Feature Extraction Module of the Source Domain

Due to complex spectral characteristics and high-dimensional structure, the process of annotating HSIs is more time-consuming and challenging than natural images. In contrast, natural images offer abundant labeled data. Thus, a feature extraction module for the source domain is designed. Figure 1 shows the overall framework of the source domain.

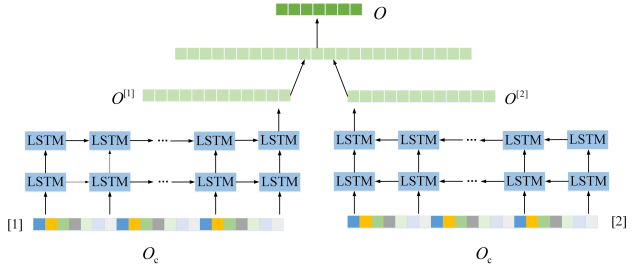

Fig. 2. TD-LSTM methodology framework.

First, the general features of the source domain are extracted through the VGG module. Then, operations such as 2D convolution, batch normalization, max pooling and full connection are applied to further process the features. Among them, the size of the 2D convolution kernel is 3×3 and the number of convolution kernel is set as 512. A batch normalization layer is added after the 2D convolutional layer, which not only effectively mitigates the vanishing gradient problem but also enhances the generalization ability of the model. Gaussian Error Linear Units (GELU) is an activation function commonly used in neural networks. It is based on Gaussian error function and helps to improve the convergence speed and performance of the training process. The 2D convolutional layer, batch normalization layer, and GELU activation function together form a basic unit. This unit is followed by a 2×2 max pooling layer to address the problem of redundancy. Finally, a full connection layer with 100 neurons is used to convert the feature maps into feature vectors.

On this basis, TD-LSTM network is proposed, which effectively handles the temporal dependencies in sequential data. When features vectors are flattened or arranged in a specific sequence, TD-LSTM processes information in both forward and backward directions to capture dependencies and contextual information. By combining VGG and TD-LSTM, the feature extraction ability of CNNs and the sequence modeling capability of LSTMs can be effectively leveraged. The TD-LSTM module is shown in Fig. 2.

The TD-LSTM consists of LSTM blocks and linear layers. The feature vectors generated by the full connection are sent into multiple LSTM units, as shown in [1] and [2] in the figure. These two layers of LSTM model the forward and backward correlations of features vectors $\mathbf{O}_c$, generating corresponding feature vector $\mathbf{O}^{[1]}$ and $\mathbf{O}^{[2]}$. Then, these two feature vectors are further concatenated into a linear layer for feature connection. Finally, the concatenated feature vectors are passed to the softmax layer for classification. Thus, the computation process of the TD-LSTM module is represented by the following formula:

$$\mathbf{O} = \text{Linear}\left(\text{Con}\left(\mathbf{O}^{[1]}, \mathbf{O}^{[2]}\right)\right) \tag{1}$$

where Con $(\cdot)$ denotes the concatenation of two feature vectors, and Linear $(\cdot)$ represents the linear layer, [1] denotes the forward RNN ,[2] denotes the backward RNN, $\mathbf{O}^{[1]}$ and $\mathbf{O}^{[2]}$ represent the outputs of the hidden layer. The output of the hidden layer is represented as follows:

$$s^{(t)} = g_{\text{f}}^{(t)} s^{(t-1)} + g_{\text{i}}^{(t)} \delta_{\text{s}}\left(Wh^{(t-1)}\right) + UX^{(t)} + b, \tag{2}$$

$$h^{(t)} = g_{\text{o}}^{(t)} \delta_{\text{h}}\left(s^{(t)}\right), \tag{3}$$

$$g_{\text{i}}^{(t)} = F_{\text{s}}\left(W_{\text{i}} h^{(t-1)} + U_{\text{i}} X^{(t)} + b_{\text{i}}\right), \tag{4}$$

$$g_{\text{f}}^{(t)} = F_{\text{s}}\left(W_{\text{f}} h^{(t-1)} + U_{\text{f}} X^{(t)} + b_{\text{f}}\right), \tag{5}$$

$$g_{\text{o}}^{(t)} = F_{\text{s}}\left(W_{\text{o}} h^{(t-1)} + U_{\text{o}} X^{(t)} + b_{\text{o}}\right) \tag{6}$$

where i, f, and o represent the input gate, forget gate, and output gate of the LSTM unit, respectively, h represents the system state, both δ_{s} and δ_{h} are activation functions for the system state and the hidden layer state, with the tanh activation function, b is the bias coefficient, g is the gating unit, W and U are weight coefficients, and F_{s} is the sigmoid function.

2.2 LGFEB of the Target Domain

CNN is renowned for its strength in extracting fine-grained local features. In contrast, the Transformer architecture excels at capturing long-range global dependencies via its self-attention mechanism. Fusing these complementary strengths presents a significant challenge in effectively mining and integrating discriminative feature information. To address this, we design the LGFEB that synergistically combines the capabilities of CNN and Transformer to comprehensively model both local details and global contexts. The overall framework of the proposed model is illustrated in Fig. 3.

2.2.1 The Local Feature Extraction Module of the Target Domain

To better represent and understand the details in the image, a local feature extraction module for the target domain is designed. A mapping layer is employed to address the issue of different channel numbers between the source domain and the target domain. Specifically, cubes of size 33×33×B (where B is the number of bands) are extracted from HSI and mapped to an image cube with 3 channels, that is, an 33×33×3 small cube. The weights from the first seven layers of the VGG and TD-LSTM models, trained on the source domain, are transferred to the local convolution layer in the target domain. Then, local features are extracted by a 2D convolutional, batch normalization, GELU activation function, max pooling layer, and full connection layers. The size of the 2D convolution kernel is 3 × 3 and the number of

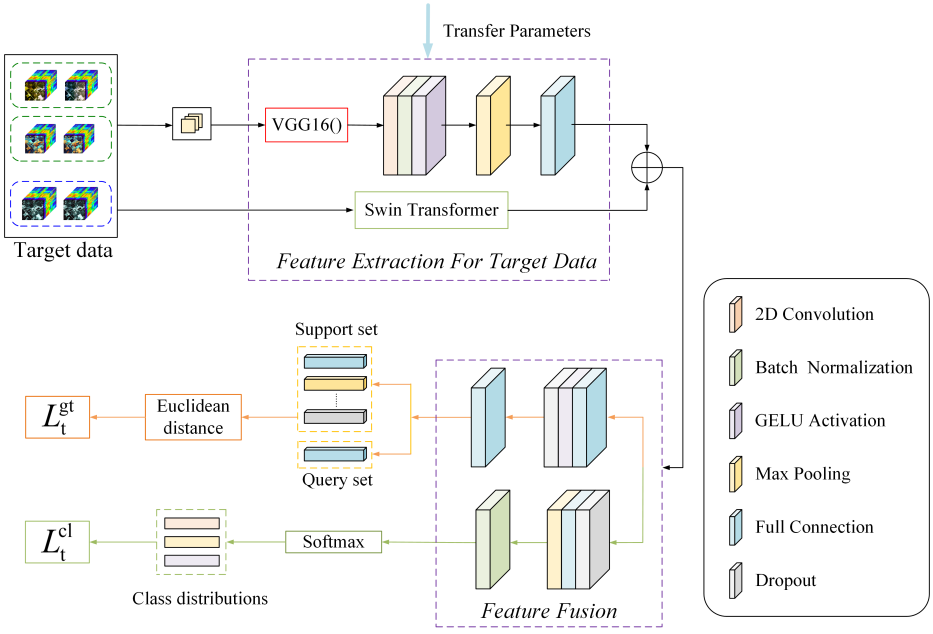

Fig. 3. The overall framework of the target domain.

the 2D convolution kernel is 512. Following the 2D convolution layer is the batch normalization layer, GELU activation functions, and max pooling. The size of the max pooling layer is 2×2, so as to deal with the problem of redundancy. Finally, an full connection layer with 100 neurons is added to generate feature vectors.

2.2.2 The Global Feature Extraction Module of the Target Domain

To fully capture the characteristics of HSI, a Swin Transformer is employed as the core algorithm for global feature extraction. The Swin Transformer is a computer vision model based on the ViT architecture. It introduces a sliding window mechanism that restricts the attention mechanism to a fixed window size, enabling the model to effectively learn features across different windows. The self-attention calculation formula for each divided window is:

$$\hat{z}^l = \mathrm{W-MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + \mathrm{MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + z^{l-1}, \quad (7)$$

$$z^{l+1} = \mathrm{SW-MSA}\left(\mathrm{LN}\left(z^{l}\right)\right) + \mathrm{MSA}\left(\mathrm{LN}\left(z^{l}\right)\right) + z^{l}. \quad (8)$$

In this stage, the input feature z^{l-1} undergoes layer normalization (LN), window multi-head self-attention (W-MSA), and a residual layer to obtain $\hat{z}^l$. After passing through LN and multilayer perceptron (MLP), it enters the second block of the multi-head self-attention shifted window (SW-MSA). Two successive Swin Transformer blocks are shown in Fig. 4.

The self-attention calculation formula for each partition window is as follows:

$$\mathrm{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}}\right)\mathbf{V} \quad (9)$$

here $\mathbf{V}$ represents the relative positional encoding of $\mathbf{Q}$ and $\mathbf{K}$. The dot product of $\mathbf{Q}$ and $\mathbf{K}$ indicates similarity, and a mask matrix is obtained through softmax normalization, with values ranging from 0 to 1. This mask matrix is multiplied by $\mathbf{V}$ to obtain the weighted $\mathbf{V}$ features. The HSI is sent into the Swin Transformer network, which efficiently processes HSIs and excels in global feature extraction.

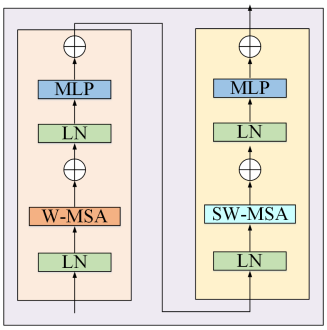

Fig. 4. Two successive Swin Transformer blocks.

2.2.3 The Feature Fusion Module for the Target Domain

The local-global feature fusion module is designed to fully utilize the local features and global features of HSI. The upper branch processes features through full connection layers, GELU activation functions, and a dropout layer. The unit number of full connection layer is set to 64. In addition, few-shot learning is performed by comparing the similarity between samples based on the corresponding feature vectors. The lower branch includes a dropout layer, GELU activation functions, full connection layers and batch normalization layers. Finally, contrastive self-supervised learning is achieved by evaluating the relationships between samples through their feature embeddings. The unit number of the second full connection layer is the class number of the HSI dataset. Through the collaboration of these two branches, the model can comprehensively consider local details and global context.

2.3 FSL

2.3.1 FSL in Source Domain

In the source domain, N_s classes are randomly selected from source domain D_s. For each selected class, K labeled samples are randomly chosen to form the support set S_s. Then, C samples are randomly selected from each class to form the query set Q_s. Throughout this, it is guaranteed that there is no overlap between the samples in the support set and those in the query set. The sample $x_\mathrm{s}^{\mathrm{que}}$ in the query set belonging to class m can be calculated as:

$$P\left(y_\mathrm{s}^{\mathrm{que},j} = m \mid x_\mathrm{s}^{\mathrm{que},j} \in Q_\mathrm{s}\right) = \frac{\exp\left(-d\left(F_\theta\left(x_\mathrm{s}^{\mathrm{que},j}\right), c_m\right)\right)}{\sum_{m=1}^{N_\mathrm{s}} \exp\left(-d\left(F_\theta\left(x_\mathrm{s}^{\mathrm{que},j}\right), c_m\right)\right)} \quad (10)$$

where c_m represents the embedded feature of the m-th class in the support set, and $d(\cdot)$ is a function of Euclidean distance, $x_\mathrm{s}^{\mathrm{que},j}$ and $y_\mathrm{s}^{\mathrm{que},j}$ represent the j-th sample in the query set and its corresponding label, respectively, and N_s is the number of classes.

2.3.2 FSL in Target Domain

In the target domain, FSL is applied to the HSI data D_l to extract discriminative features and individual knowledge to improve few-shot classification performance. First, classes N_t are randomly selected from D_l. Then, K labeled samples are selected from each class as the support set, and C samples are selected from the remaining samples in each class as the query set. To make the probability distribution of the predicted sample as close as possible to the probability distribution of the real sample, the cross-entropy loss function is used. The formula is as follows:

$$H\left(p, q\right) = -\sum_{i=1}^{n} p\left(x_i\right) \log\left(q\left(x_i\right)\right) \quad (11)$$

here $p(x_i)$ represents the true distribution of the sample, and $q(x_i)$ represents the distribution predicted by the model. Therefore, the training loss for each training set is:

$$L_{\mathrm{t}}^{\mathrm{fsl}} = -\sum E_{S_{\mathrm{t}},Q_{\mathrm{t}}}\left[\log p\left(y_{\mathrm{t}}^{\mathrm{que}} = n \mid x_{\mathrm{t}}^{\mathrm{que}}, \eta\right)\right] \tag{12}$$

here S_{t} is the support set generated by D_{l}, Q_{t} is the query set generated by D_{l}, and η is the parameter for feature extraction.

2.4 Self-supervised Learning

2.4.1 Self-supervised Learning in Source Domain

Self-supervised learning on the source domain aims to increase the model's diversity. This approach facilitates improved adaptation to various sample transformations and enhances the capacity to extract local features in the target domain. The images in the source domain are rotated with the transformation angle chosen as $R = \{90°, 180°, 270°\}$. The transformed image and the original image form a sample pair, which is sent to the feature extraction module of the source domain. The self-supervised loss for the source domain can be represented as:

$$L_{\mathrm{s}}^{\mathrm{gt}}(D_{\mathrm{s}}; \phi, \theta) = -E_{x_{\mathrm{s}}\ D_{\mathrm{s}}}\left[\sum_{\forall r \in R} \log R_{\phi}^{r}\left(F_{\theta}\left(x_{\mathrm{s}}^{r}\right)\right)\right] \tag{13}$$

where x_{s}^{r} represents the sample pair formed by the rotation transformation of the original image x_{s}, $F_{\theta}\left(x_{\mathrm{s}}^{r}\right)$ represents the features extracted from the rotated image, $R_{\phi}^{r}(\cdot)$ is the predicted score for rotation r.

2.4.2 Self-supervised Learning in Target Domain

Self-supervised learning in the target domain can extract high-quality and class-invariant features, so as to improve the classification accuracy in the target domain. First, a sample is selected and copied from D_{l}. Gaussian noise [30] is added to generate a noise-enhanced sample. Then, the original sample and the noise-added sample form a sample pair, which is passed to the global feature extraction network to extract the corresponding features. Afterwards, dropout is applied to both feature sets. Due to the randomness of dropout, two augmented matrices are obtained. The final output of self-supervised learning is two class distributions $z_{\mathrm{t}}^{\mathrm{cl},1}$ and $z_{\mathrm{t}}^{\mathrm{cl},2}$, and the self-supervised loss function is [36].

$$L_{\mathrm{t}}^{\mathrm{cl}}\left(z_{\mathrm{t}}^{\mathrm{cl}}, z_{\mathrm{t}}^{\mathrm{cl},2}; \gamma\right) = \frac{1}{2}\left(L\left(z_{\mathrm{t}}^{\mathrm{cl},1} \,||\, z_{\mathrm{t}}^{\mathrm{cl},2}\right) + L\left(z_{\mathrm{t}}^{\mathrm{cl},2} \,||\, z_{\mathrm{t}}^{\mathrm{cl},1}\right)\right) \tag{14}$$

where:

$$\begin{aligned} L\left(z_{\mathrm{t}}^{\mathrm{cl},1} \,||\, z_{\mathrm{t}}^{\mathrm{cl},2}\right) &= \frac{1}{B}\sum_{i=1}^{B} D_{\mathrm{kl}}\left(z_{\mathrm{t}}^{\mathrm{cl},i,1} \,||\, z_{\mathrm{t}}^{\mathrm{cl},i,2}\right) \\ &+ \frac{1}{B}\sum_{i=1}^{B}\left[\mathrm{H}\left(z_{\mathrm{t}}^{\mathrm{cl},i,1}\right) - \mathrm{H}\left(\frac{1}{B}\sum_{i=1}^{B} z_{\mathrm{t}}^{\mathrm{cl},i,1}\right)\right]. \end{aligned} \tag{15}$$

In the above equation, $D_{\mathrm{kl}}(\cdot||\cdot)$ represents the Kullback-Leibler divergence between two probability distributions. It is an asymmetric measure of the difference between two distributions and commonly used to evaluate the information loss of one distribution relative to another. $\mathrm{H}(\cdot)$ represents the entropy of a specific probability distribution, B is the batch size, and γ is the spectral space feature extraction parameter. The first component is referred to as the consistency term. By minimizing the consistency term, it ensures that different feature predictions from the same sample remain consistent. The second term is the sharpness term. For this term, the output distribution is regularized by minimizing the class distribution entropy for each sample. This process enhances the certainty of the output distribution and facilitates the assignment of distinct categories to each sample. As a result, features of the same class become more compact, improving feature discriminability. The third term is the diversity term. For this term, the entropy of the average distribution between different samples is maximized. This encourages the predictions of different samples to be distributed across new classes, which prevents the network from assigning all images to the same class.

Thus, the loss for the source or target domain can be represented as:

$$L^{\mathrm{total}} = L^{\mathrm{fsl}} + L^{\mathrm{ssl}} \tag{16}$$

here L^{fsl} represents the loss for FSL in the source or target domain, and L^{ssl} represents the loss for self-supervised learning in the source or target domain, L^{total} represents the total loss for the source or target domain.

3. Experiment Results and Analysis

3.1 Experimental Dataset

To evaluate the performance of SSCF-Net, mini-ImageNet [30] with a large amount of labeled data is selected as the source domain dataset. Three HSI datasets are used as target domain datasets, including Salinas (SA), Indian Pines (IP), and WHU-Hi-LongKou (LK) [37]. The number of labled samples per class are reported in Tab. 1. Figures 5–7 show the pseudocolor images and ground truth maps of SA, IP and LK.

The SA dataset was derived from the agricultural region of Salinas Valley in California, USA, which contains rich spectral information and is suitable for land cover classification tasks. The dataset includes 224 spectral bands, with 204 bands retained after discarding water absorption bands. The wavelength range spans from 400 nm to 2500 nm. The image size is 512 × 217 pixels, with a spatial resolution of 3.7 m. It includes 16 different land cover classes and provides high-detail ground truth information.

Datasets	SA		IP		LK	
Class	Name	Number	Name	Number	Name	Number
1	Brocoli-green-weeds-1	2009	Alfalfa	46	Corn	34511
2	Brocoli-green-weeds-2	3726	Corn-notill	1428	Cotton	8374
3	Fallow	1976	Corn-mintill	830	Sesame	3031
4	Fallow-rough-plow	1394	Corn	237	Broad-leaf soybean	63212
5	Fallow-smooth	2678	Grass-pasture	483	Narrow-leaf soybean	4151
6	Stubble	3959	Grass-tree	730	Rice	11854
7	Celery	3579	Grass-pasture-mowed	28	Water	67056
8	Grapes-untrained	11271	Hay-windrowed	478	Roads and houses	7124
9	Soil-vinyard-develop	6203	Oats	20	Mixed weed	5229
10	Corn senesced green weeds	3278	Soybean-notill	972		
11	Lettuce romaine-4wk	1068	Soybean-mintill	2455		
12	Lettuce romaine-5wk	1927	Soybean-clean	593		
13	Lettuce romaine-6wk	916	Wheat	205		
14	Lettuce romaine-7wk	1070	Woods	1265		
15	Vinyard untrained	7268	Buildings-Grass-Trees	386		
16	Vinyard vertical trellis	1807	Stone-Steel-Towers	93		
	Total	54129	**Total**	10249	**Total**	204542

Tab. 1. Number of samples per class for three datasets.

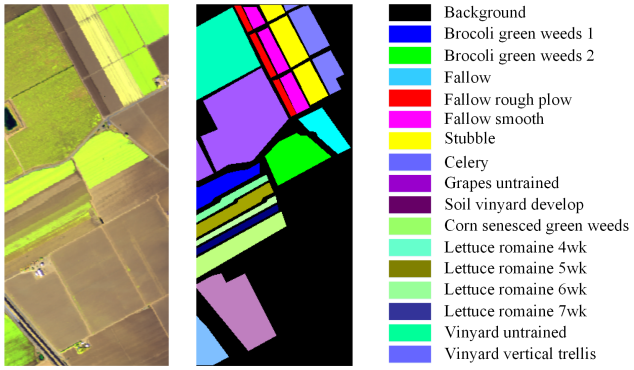

Fig. 5. Pseudocolor image and ground-truth map of SA. (a) Pseudocolor image of SA; (b) Ground-truth map of SA.

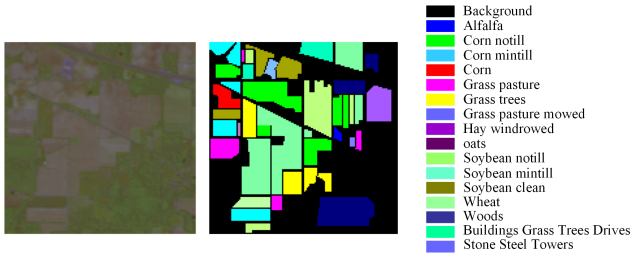

Fig. 6. Pseudocolor image and ground-truth map of IP. (a) Pseudocolor image of IP; (b) Ground-truth map of IP.

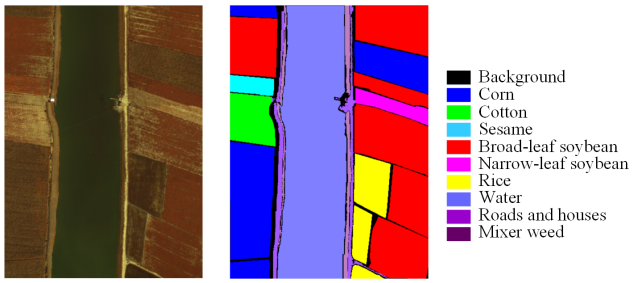

Fig. 7. Pseudocolor image and ground-truth map of LK. (a) Pseudocolor image of LK; (b) Ground-truth map of LK.

The IP dataset was collected from the agricultural area of Indian Pines in Indiana, USA. It includes a 145 × 145 pixel image, which provides hyperspectral data and corresponding ground truth labels. The dataset covers multiple bands from 400 nm to 2500 nm, with 200 spectral bands in total. It contains 16 different land cover classes, and the spatial resolution is 20 meters, which is suitable for fine-grained land cover analysis.

The LK dataset was captured from the LongKou region in Wuhan, Hubei Province, China, covering both urban and rural environments. The dataset includes 270 spectral bands, with a wavelength range from 400 nm to 2500 nm. The image size is 550 × 400 pixels and contains 9 land cover classes. The spatial resolution is 0.463 m, which can capture the ground information with rich details.

3.2 Experimental Setup

The configuration used in the experiment is an i7-10700F (3.7 GHz), 32 GB RAM, and an Nvidia GeForce RTX3060. The open-source software framework is PyTorch. For mini-ImageNet dataset, the number of episodes is set to 1500. For datasets from the target domain, it is set to 1000. The model is optimized by Adam optimizer with a learning rate of 0.001. The classification performance of the methods is evaluated by four metrics: F_1 score (F_1), overall accuracy (OA), average accuracy (AA), and kappa coefficient (Kappa).

The F_1 combines precision and recall to evaluate performance of the model. By accounting for both false positives and false negatives, it provides an effective metric for assessing the ability of the algorithm to capture class-specific details and distinguish between different land cover categories. F_1 can be calculated using the following formula, as derived from [38].

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{17}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{18}$$

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{19}$$

where TP, FP, TN, and FN stand for true positive, false positive, true negative, and false negative, respectively.

To evaluate the performance of SSCF-Net, SSCF-Net is compared with SpectralFormer [39], SSFTT [16], Gia-CFSL [31], DCFSL [29], HFSL [30], FSCF-SSL [32], HCFSL-NAS [33], and FDFSL [34]. For all comparison methods, 5 labeled samples per class are selected from the target domain dataset. The experiments are repeated 10 times to reduce randomness.

- The SpectralFormer, a Transformer-based backbone, enhances HSIC by learning spectral sequences, using cross-layer skip connections for improved feature extraction.
- The SSFTT systematically combines CNN network and Transformer structure to exploit spectral-spatial information in the HSI, with a Gaussian weighted feature tokenizer module making the samples more separable.
- The Gia-CFSL combines few-shot learning and domain alignment to address domain shift issues in cross-scene HSIC, improving performance with nonlocal relationship aggregation and graph-based alignment.
- The DCFSL is a method that addresses few-shot learning and domain adaptation in HSIC, using adversarial strategies to overcome domain shift and improve performance on target classes.
- The HFSL is a method for HSIC using few labeled samples, leveraging knowledge transfer from mini-ImageNet and a spectral-spatial fusion network to improve performance.
- The FSCF-SSL utilizes base class data from natural images to improve classification accuracy on novel HSI classes by transferring spatial meta-knowledge and learning discriminative features from limited data.
- The HCFSL-NAS utilizes neural architecture search for embedding feature extraction, multisource learning for data aggregation, and a combined loss function to improve classification performance with few labeled samples.
- The FDFSL employs feature disentanglement to reduce source bias, a multiorder spectral interaction block for data integration, and a self-distillation scheme to enhance feature diversity.

Compared with the above methods, SSCF-Net deeply integrates the advantages of convolution in local feature extraction with those of Swin Transformer in long-range dependency for HSI. Under the condition of limited labeled samples, with the help of self-supervised learning strategy, the method in this paper successfully facilitates cross-domain feature transfer between source and target domains.

3.3 Experimental Results

As shown in Tab. 2, on SA dataset, SSCF-Net outperforms other comparison methods. SSCF-Net performs excellently in F_1, OA and Kappa. Compared with FSCF-SSL, HCFSL-NAS, and FDFSL, F_1 is improved by 0.79%, 1.11%, and 2.06%, respectively. OA is improved by 0.54%, 2.76%, and 1.68%, respectively. Kappa is improved by 0.6%, 3.04%, and 1.87%, respectively. In AA, the top position is held by FSCF-SSL, which effectively utilizes the spectral-spatial feature information in HSI. The SA dataset has a higher number of bands, providing FSCF-SSL with more feature dimensions. It allows for better discrimination between multiple classes. This is particularly evident in the second class, Broccoli-green-weeds-2, and the fifteenth class, Vineyard untrained, where FSCF-SSL outperforms other methods.

From the data analysis in Tab. 3, SSCF-Net improves OA and Kappa by 0.4% and 0.57%, respectively, compared with the best cross-domain method, DCFSL, on the LK dataset. In F_1 and AA, SSCF-Net also outperforms the best cross-domain method, FSCF-SSL, with improvements of 1.04% and 0.99%. In addition, compared with the traditional deep learning method, SSCF-Net exhibits superior performance across the three indicators of OA, AA and Kappa, which are 4.41%, 8.8% and 5.72% higher than SSFTT, respectively. These significant performance improvements validate the effectiveness of SSCF-Net in cross-domain classification tasks. The outstanding performance of SSCF-Net is attributed to its self-supervised learning approach, which reduces the reliance on a large amount of labeled data. Additionally, the model improves its modeling ability in the target domain by incorporating LGFEB.

As shown in Tab. 4, classification performance of the SSCF-Net on the IP dataset is better than other comparative methods. Although HFSL and FSCF-SSL perform well, SSCF-Net outperforms FSCF-SSL and HFSL in F_1 by 1.26% and 2.26%, respectively; in OA by 0.74% and 2.01%, respectively; in AA by 0.10% and 1.72%, respectively; and in Kappa by 0.85% and 2.29%, respectively. In addition, compared with other methods, SSCF-Net achieved the best classification performance in eight classes, including Alfalfa, Corn-mintill, Corn, Grass-pasture-mowed, Hay-windrowed, Oats, Soybean-notill, and Soybean-mintill. Especially in the more challenging Corn and Soybean-mintill classes, SSCF-Net achieved classification accuracies of 93.25% and 76.20%, outperforming other methods by at least 2.22% and 4.65%, respectively. This highlights the exceptional performance of SSCF-Net in fine-grained classification.

Class	Spectral Former	SSFTT	Gia-CFSL	DCFSL	HFSL	FSCF-SSL	HCFSL-NAS	FDFSL	OURS
1	66.75	78.31	98.77	99.69	98.56	**99.96**	98.05	99.20	96.06
2	55.78	69.92	99.54	99.73	96.31	97.69	**99.97**	98.68	93.83
3	72.72	73.34	88.26	93.46	94.34	95.77	**95.99**	87.14	95.89
4	89.48	91.02	99.20	99.65	99.39	**99.73**	98.84	98.93	98.86
5	82.49	82.20	90.42	91.66	97.17	**97.13**	90.72	89.70	93.69
6	86.19	93.12	99.12	99.44	98.88	99.44	**99.89**	99.19	97.96
7	75.28	87.08	98.39	98.49	93.45	95.20	**99.82**	99.20	99.27
8	49.71	63.39	76.33	75.14	77.45	79.59	67.78	80.13	**83.24**
9	87.04	94.61	99.201	99.78	99.55	99.56	**99.72**	98.03	98.84
10	72.56	70.89	83.71	84.47	91.02	**94.52**	85.54	83.21	90.68
11	83.84	90.85	97.47	98.62	98.63	**98.95**	96.82	96.71	97.06
12	83.49	86.30	99.17	**99.47**	97.21	99.22	99.40	99.42	95.41
13	87.17	88.69	97.75	99.29	98.21	**99.64**	98.00	97.15	97.87
14	87.65	88.08	97.79	98.66	96.78	97.15	98.61	98.93	**99.52**
15	72.50	75.49	74.16	75.85	66.21	80.16	**87.83**	81.27	85.38
16	48.24	87.93	90.08	90.42	94.63	**95.46**	85.54	92.98	95.08
F_1	68.46	69.17	90.80	91.89	91.38	91.48	91.16	90.21	**92.27**
OA	70.52 ±3.21	78.88 ±3.24	88.85 ±2.25	89.37 ±2.14	88.55 ±1.38	91.61 ±1.69	89.39 ±1.29	90.47 ±1.54	**92.15 ±0.79**
AA	75.03 ±3.20	82.58 ±2.66	93.09 ±1.39	93.99 ±1.08	93.61 ±1.31	**95.57 ±1.34**	93.93 ±0.78	93.74 ±1.42	94.41 ±0.77
Kappa	67.48 ±3.48	76.67 ±3.54	87.62 ±2.49	88.20 ±2.35	87.26 ±1.54	90.68 ±1.88	88.24 ±1.42	89.41 ±1.71	**91.28 ±0.87**

Tab. 2. Classification performance [%] of different methods on SA dataset.

Class	Spectral Former	SSFTT	Gia-CFSL	DCFSL	HFSL	FSCF-SSL	HCFSL-NAS	FDFSL	OURS
1	74.14	93.42	98.03	98.91	97.33	98.53	88.59	96.95	**99.00**
2	59.68	79.13	85.37	89.40	95.72	**96.31**	78.78	83.38	96.24
3	53.36	92.73	89.00	89.28	86.49	88.56	91.05	**95.38**	91.17
4	54.84	90.62	91.00	92.11	85.75	88.53	**94.08**	91.76	91.47
5	57.47	90.63	89.85	92.80	93.71	95.32	69.16	92.22	**98.43**
6	57.77	84.74	91.90	93.68	92.07	93.40	**96.25**	94.58	95.31
7	89.99	96.15	99.83	99.83	99.88	**99.90**	99.11	99.71	97.94
8	63.81	63.88	84.68	82.03	**93.29**	90.65	77.71	77.98	88.46
9	54.33	81.94	72.19	75.87	92.68	92.39	73.03	83.79	**94.46**
F_1	67.39	79.30	89.96	85.00	85.20	90.17	84.14	88.95	**91.21**
OA	70.34 ±1.64	90.97 ±1.77	94.16 ±1.73	94.98 ±1.61	93.74 ±1.84	94.87 ±2.05	92.67 ±1.31	94.45 ±1.96	**95.38 ±2.04**
AA	62.84 ±2.81	85.92 ±1.72	89.09 ±1.33	90.43 ±2.76	92.99 ±1.58	93.73 ±1.09	85.31 ±3.08	90.64 ±2.82	**94.72 ±2.11**
Kappa	62.99 ±1.84	88.30 ±2.19	92.39 ±2.20	93.45 ±2.07	91.90 ±2.32	93.35 ±2.60	90.40 ±1.73	92.78 ±2.51	**94.02 ±2.60**

Tab. 3. Classification performance [%] of different methods on LK dataset.

3.4 Classification Result Visualization

To visually demonstrate the classification performance of various methods with 5 labeled samples per class in the target domain, the classification results on the IP dataset are presented and compared with the ground truth labels. The results are depicted in Figs. 8–10. The visual analysis indicates that the classification map produced by the SSCF-Net method demonstrates the highest degree of alignment with the ground truth. Additionally, SSCF-Net shows the least misclassification between categories compared with the other methods. This suggests that the SSCF-Net approach is more effective in distinguishing between different classes and produces results that are closer to the actual distribution of the data.

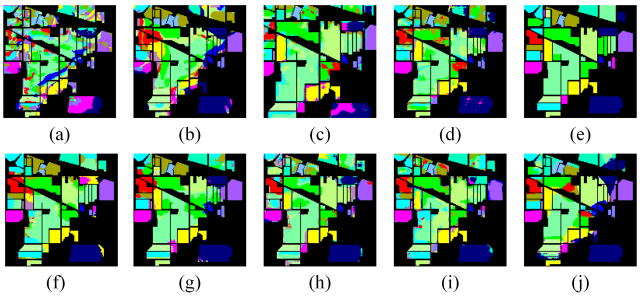

Fig. 8. Classification results of different methods on IP. (a) Spectral Former, (b) SSFTT, (c) Gia-CFSL, (d) DCFSL, (e) Ground truth, (f) HFSL, (g) FSCF-SSL, (h) HCFSL-NAS, (i) FDFSL, f (j) OURS.

Class	Spectral Former	SSFTT	Gia-CFSL	DCFSL	HFSL	FSCF-SSL	HCFSL-NAS	FDFSL	OURS
1	94.15	95.37	91.71	95.85	99.51	**100**	96.34	96.59	**100**
2	29.46	45.29	45.11	41.88	57.89	**62.30**	43.65	51.69	56.88
3	27.83	35.50	48.13	47.96	66.38	65.38	40.15	68.48	**69.82**
4	46.59	86.25	78.28	79.53	88.88	89.27	91.03	89.27	**93.25**
5	40.10	54.12	73.26	71.76	71.15	76.78	**80.67**	75.04	73.94
6	44.90	45.34	83.23	**84.72**	75.06	81.83	81.94	85.14	83.19
7	96.96	97.39	99.57	99.13	**100**	**100**	99.57	99.13	**100**
8	61.10	88.99	88.52	83.70	99.66	97.57	84.80	78.37	**99.96**
9	99.33	98.00	99.33	99.33	**100**	**100**	98.67	99.33	**100**
10	44.44	46.69	60.64	60.94	67.68	61.49	70.47	61.80	**72.70**
11	40.76	41.83	61.64	59.47	71.55	71.29	69.05	61.56	**76.20**
12	34.06	46.58	44.88	46.68	66.11	**67.65**	54.61	53.59	65.95
13	72.15	85.45	98.60	97.95	98.75	**99.10**	**99.10**	97.95	97.04
14	77.02	75.33	78.32	84.71	90.95	**93.99**	92.56	86.76	87.20
15	61.68	72.86	70.45	68.45	91.94	**97.87**	60.39	83.18	90.71
16	90.23	89.43	**98.75**	98.64	94.20	96.14	98.64	97.39	95.63
F_1	41.09	43.00	63.62	65.40	73.36	74.36	73.03	71.81	**75.62**
OA	46.19 ±1.75	53.58 ±3.66	64.59 ±4.02	64.29 ±2.76	74.59 ±3.03	75.86 ±4.15	68.81 ±2.23	69.27 ±2.46	**76.60 ±2.52**
AA	60.05 ±1.00	69.03 ±2.35	76.28 ±2.20	76.28 ±1.24	83.73 ±2.45	85.05 ±2.42	78.85 ±1.34	80.33 ±1.61	**85.15 ±1.87**
Kappa	40.47 ±1.65	48.89 ±3.76	60.04 ±4.25	59.87 ±2.82	71.27 ±3.38	72.71 ±4.55	64.77 ±2.45	65.50 ±2.55	**73.56 ±2.78**

Tab. 4. Classification performance [%] of different methods on IP dataset.

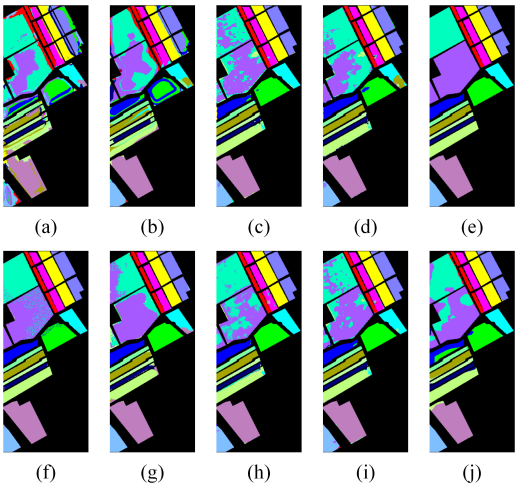

Fig. 9. Classification results of different methods on SA. (a) Spectral Former, (b) SSFTT, (c) Gia-CFSL, (d) DCFSL, (e) Ground truth, (f) HFSL, (g) FSCF-SSL, (h) HCFSL-NAS, (i) FDFSL, (j) OURS.

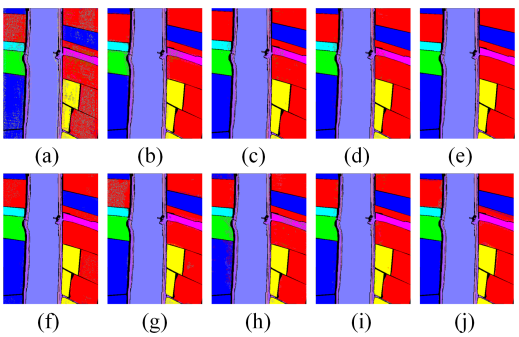

Fig. 10. Classification results of different methods on LK. (a) Spectral Former, (b) SSFTT, (c) Gia-CFSL, (d) DCFSL, (e) Ground truth, (f) HFSL, (g) FSCF-SSL, (h) HCFSL-NAS, (i) FDFSL, (j) OURS.

Datasets	TD-LSTM	RT,NA	Transformer	OA(%)
SA	×	×	✓	87.21
	✓	×	✓	87.65
	×	✓	✓	89.86
	✓	✓	✓	**92.15**
IP	×	×	✓	65.49
	✓	×	✓	72.28
	×	✓	✓	72.13
	✓	✓	✓	**76.60**
LK	×	×	✓	92.18
	✓	×	✓	92.88
	×	✓	✓	93.08
	✓	✓	✓	**95.38**

Tab. 5. Ablation comparison of each module in OA [%].

3.5 Ablation Experiment

Ablation experiments are conducted by evaluating four networks across two domains, further validating the effectiveness of the proposed SSCF-Net in both the source and target domains. The four networks are: 1) Network 1: Transformer, 2) Network 2: Transformer + TD-LSTM, 3) Network 3: Transformer + rotation transformation (RT) + noise addition (NA), and 4) Network 4: Transformer + TD-LSTM + RT + NA.

As shown in Tab. 5, the classification performance on the SA, IP, and LK datasets exhibited a progressive enhancement. This result demonstrates the significant gain from the three modules designed in HSIC tasks. Taking the SA dataset as an example: In Network 1, the base model consists of the Transformer for the target domain and the VGG network for the source domain, with an overall accuracy of 87.21%. This shows that relying only on these two modules is insufficient for the HSIC task. In Network 2, the overall accuracy is improved by 0.44% when the TD-LSTM module is added to

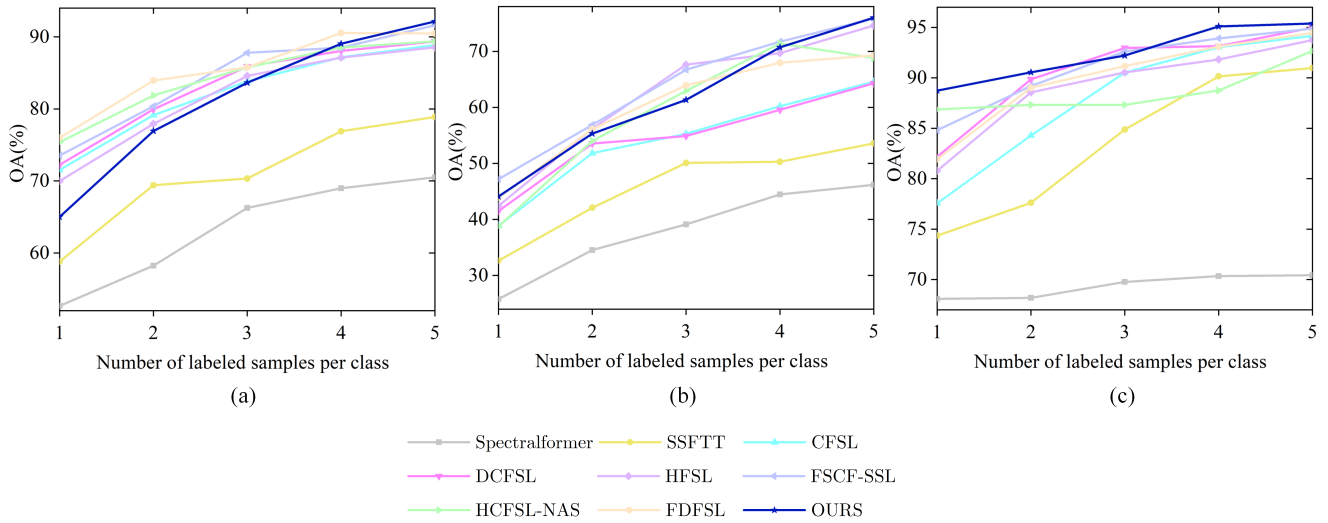

Fig. 11. The OA of all methods under different numbers of labeled samples. (a) SA, (b) IP, (c) LK.

the VGG network of the source domain. It shows that the addition of TD-LSTM enhances the transfer learning ability from the source domain to the target domain. It also improves the model's capability to extract local features. In Network 3, self-supervised learning is applied to both the source and target domains without adding TD-LSTM to the source domain. The OA increases by 2.65% compared with the base model, which indicates that self-supervised learning helps in recognizing subtle differences. This enhances the generalization ability of HSI. In Network 4, the combination of the above methods leads to an increase in overall accuracy from 89.89% to 92.15%.

3.6 Experiments With Different Numbers of Labeled Samples

To thoroughly evaluate the classification performance of SSCF-Net, different numbers of labeled samples in the target domain are utilized. Labeled samples from 1 to 5 per class are randomly selected from each category for model training and the remaining samples for testing. All classification results are based on the average of 10 experiments. As shown in Fig. 11, the OA of different methods under different number of labeled samples is compared across the SA, IP, and LK. The results show a positive trend in OA as the number of labeled samples increases. SSCF-Net demonstrates remarkable classification performance on the LK dataset, which ranks among the top methods. For the SA and IP datasets, when the number of labeled samples is less than 5, SSCF-Net does not achieve optimal performance. The possible reason is that SA and IP contain many fine-grained feature categories, which increases the difficulty of the classification task. Additionally, when the number of labeled training samples is small, the Transformer model used for global feature extraction does not perform optimally. This limitation affects the model's ability to learn useful knowledge for target domain classification, which ultimately impacts classification performance.

3.7 Complexity Analysis

To evaluate and demonstrate the computational efficiency of the various methods, as shown in Tab.6, the computational complexity of the model is evaluated based on training time, testing time, floating point operations (FLOPs), number of parameters and memory usage. For traditional deep learning methods (SpectralFormer and SSFTT), the training time only includes the time for training a single domain. In contrast, cross-domain FSL methods (Gia-CFSL, DCFSL, HFSL, and FSCF-SSL) include source domain training, transfer time, and target domain training, leading to higher computational costs but better classification performance. Compared with other cross-domain methods, SSCF-Net requires longer training time and testing time. This is due to the fact that SSCF-Net introduces Swin Transformer in the target domain, which is a module that has advantages in extracting global contextual information, and is able to better capture long-range dependencies in HSI through its hierarchical structure. However, the advantage of this structure is also accompanied by an increase in the number of FLOPs and parameters, which in turn leads to an increase in training time and testing time. In addition, on the same hardware platform (Nvidia GeForce RTX3060), based on GPU memory usage, SSCF-Net ranks first in terms of memory consumption among all methods.

3.8 2D Projection Features

To intuitively compare feature extraction performance, we visualize high-dimensional features using t-distributed Stochastic Neighbor Embedding (t-SNE) [34]. t-SNE is a nonlinear manifold learning algorithm that preserves local neighborhood structures in high-dimensional space by minimizing the KL divergence between probability distributions. This allows qualitative assessment of feature discriminability: well-separated and compact clusters indicate robust feature representation.

Datasets	Metric	Spectral Former	SSFTT	Gia-CFSL	DCFSL	HFSL	FSCF-SSL	HCFSL-NAS	FDFSL	OURS
SA	Training time	12185.1	5892.7	9334.8	4487.7	5515.5	87659.7	74.9	1468.3	122543.1
	Testing time	138.38	86.52	15.86	16.18	5.55	122.63	13.17	15.96	134.52
	FLOPs	0.52	0.29	8.90	8.87	3.21	3.21	0.0030	0.83	10.28
	#Params	0.16	0.04	2.21	0.17	3.31	3.31	0.06	0.18	30.80
	Memory usage	4.77	2.94	6.05	3.40	3.63	4.43	3.44	3.80	7.65
IP	Training time	7117.6	1010.3	9065.4	4329.2	5561.2	9955.1	74.4	1328.9	26427.8
	Testing time	10.87	5.07	2.94	3.04	4.42	4.65	2.52	2.39	10.10
	FLOPs	0.51	0.29	8.90	8.87	3.21	3.21	0.0030	0.83	10.28
	#Params	0.16	0.04	2.21	0.17	3.30	3.30	0.06	0.18	30.79
	Memory usage	4.79	3.00	5.92	3.43	3.80	4.41	3.44	3.84	7.26
LK	Training time	9051.5	9976.1	3984.3	2914.4	5295.6	75177.2	74.2	1033.4	94332.8
	Testing time	131.52	100.44	7.50	12.23	6.77	114.13	10.10	11.48	132.06
	FLOPs	0.69	0.37	8.90	8.87	3.22	3.22	0.0036	0.83	10.28
	#Params	0.16	0.04	2.21	0.12	3.41	3.41	0.07	0.18	30.89
	Memory usage	4.11	2.66	5.86	3.22	3.21	4.26	2.75	3.86	7.76

Tab. 6. Training time [s], testing time [s], FLOPs [G], #Params [M] and memory usage comparison [GB].

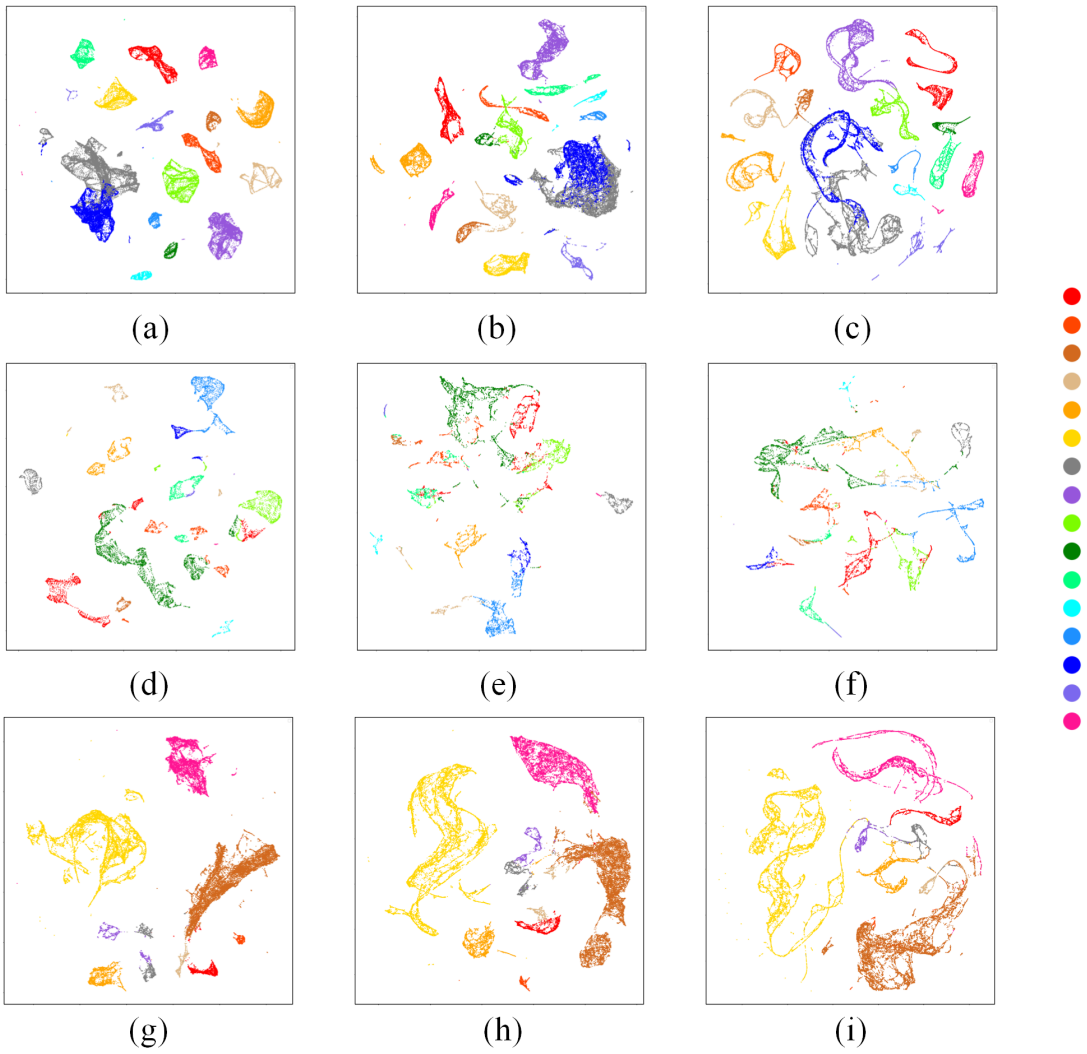

Fig. 12. Projection feature visualization on different datasets. (a) Feature visualization of FDFSL in SA, (b) Feature visualization of HDFSL-NAS in SA, (c) Feature visualization of SSCF-Net in SA, (d) Feature visualization of FDFSL in IP, (e) Feature visualization of HDFSL-NAS in IP, (f) Feature visualization of SSCF-Net in IP, (g) Feature visualization of FDFSL in LK, (h) Feature visualization of HDFSL-NAS in LK, (i) Feature visualization of SSCF-Net in LK.

As shown in Fig. 12, three cross-domain FSL methods (HCFSL-NAS, FDFSL, and SSCF-Net) are compared on SA, IP and LK. Each color represents a specific class, with the legend corresponding to the class indices listed in Tab.1. For instance, on SA, Classes 7 (Celery) and 14 (Lettuce romaine-7wk) are clearly separated in SSCF-Net, whereas they heavily overlap in FDFSL and HCFSL-NAS. The results indicate that SSCF-Net achieves the least feature category confusion and shows the clearest classification boundaries.

4. Conclusion

In this paper, an innovative method SSCF-Net is proposed. It aims to solve the problems of scarcity of labeled samples in HSIC. The classification method combines cross-domain feature transfer, self-supervised learning, and feature fusion technologies, which effectively transfer knowledge from different domains. They also make full use of the information from unlabeled data, which helps enhance the model's generalization ability. Through experimental validation on three hyperspectral datasets, the results show that SSCF-Net significantly improves in terms of accuracy compared with other comparative methods, especially in terms of OA. On SA, IP and LK datasets, SSCF-Net outperforms other state-of-the-art methods by 0.54%, 0.74% and 0.40%, respectively. SSCF-Net demonstrates stronger effectiveness than traditional methods and other few-shot learning methods.

Although the classification results on the three datasets validate the effectiveness of SSCF-Net, the complexity analysis indicates certain limitations in terms of training and testing time. In particular, the introduction of the self-attention mechanism from the Transformer in the target domain has created a bottleneck in computational complexity, leading to relatively low inference efficiency.

To address this issue, future research will explore various optimization strategies to improve the inference efficiency of the model. Lightweight network architectures, as an effective solution, will be used to reduce the computational load while maintaining high classification performance. Further research directions also include leveraging hardware acceleration technologies, such as customized hardware design and parallel computing capabilities of FPGA, which can significantly enhance the acceleration efficiency of high-complexity models.

References

[1] CHEN, W., OUYANG, S., YANG, J., et al. A framework for complex land cover classification using Gaofen-5 AHSI images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022, vol. 15, p. 1591–1603. DOI: 10.1109/JSTARS.2022.3144339

[2] PENG, J., ZHOU, Y., CHEN, C. L. P. Region-kernel-based support vector machines for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2015, vol. 53, no. 9, p. 4810–4824. DOI: 10.1109/TGRS.2015.2410991

[3] ZHANG, Y., CAO, G., LI, X., et al. Cascaded random forest for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018, vol. 11, no. 4, p. 1082–1094. DOI: 10.1109/JSTARS.2018.2809781

[4] YU, C., HUANG, J., SONG, M., et al. Edge-inferring graph neural network with dynamic task-guided self-diagnosis for few-shot hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, vol. 60, p. 1–13. DOI: 10.1109/TGRS.2022.3196311

[5] CHEN, Y., LIN, Z., ZHAO, X., et al. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2014, vol. 7, no. 6, p. 2094–2107. DOI: 10.1109/JSTARS.2014.2329330

[6] CHEN, Y., ZHAO, X., JIA, X., et al. Spectral–spatial classification of hyperspectral data based on deep belief network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2015, vol. 8, no. 6, p. 2381–2392. DOI: 10.1109/JSTARS.2015.2388577

[7] HU, W., HUANG, Y., WEI, L., et al. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors*, 2015, vol. 2015, no. 1, p. 1–13. DOI: 10.1155/2015/258619

[8] CHEN, Y., JIANG, H., LI, C., et al. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, vol. 54, no. 10, p. 6232–6251. DOI: 10.1109/TGRS.2016.2584107

[9] YE, Z., LI, C., LIU, Q., et al. Computationally lightweight hyperspectral image classification using a multiscale depthwise convolutional network with channel attention. *IEEE Geoscience and Remote Sensing Letters*, 2023, vol. 20, p. 1–5. DOI: 10.1109/LGRS.2023.3285208

[10] ZHANG, C., LI, G., DU, S., et al. Multi-scale dense networks for hyperspectral remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, vol. 57, no. 11, p. 9201–9222. DOI: 10.1109/TGRS.2019.2925615

[11] FANG, Q., GU, S., WANG, J., et al. A feature dynamic enhancement and global collaboration guidance network for remote sensing image compression. *Radioengineering*, 2025, vol. 34, no. 2, p. 324–341. DOI: 10.13164/re.2025.0324

[12] DING, Y., CHONG, Y., PAN, S., et al. Diversity-connected graph convolutional network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–18. DOI: 10.1109/TGRS.2023.3298848

[13] LIU, C., DONG, A., DONG, D., et al. Contrastive graph convolution network with skip connections for few-shot hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 2024, vol. 21, p. 1–5. DOI: 10.1109/LGRS.2024.3355147

[14] ZHOU, W., KAMATA, S.-I., WANG, H., et al. Multiscanning-based RNN-transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–19. DOI: 10.1109/TGRS.2023.3277014

[15] MEI, S., LI, X., LIU, X., et al. Hyperspectral image classification using attention-based bidirectional long short-term memory network. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, vol. 60, p. 1–12. DOI: 10.1109/TGRS.2021.3102034

[16] SUN, L., ZHAO, G., ZHENG, Y., et al. Spectral-spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, vol. 60, p. 1–14. DOI: 10.1109/TGRS.2022.3144158

[17] MEI, S., SONG, C., MA, M., et al. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, vol. 60, p. 1–14. DOI: 10.1109/TGRS.2022.3207933

[18] ZHAO, Z., XU, X., LI, S., et al. Hyperspectral image classification using groupwise separable convolutional vision transformer network. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 61, p. 1–17. DOI: 10.1109/TGRS.2024.3377610

[19] LONG, Y., WANG, X., XU, M., et al. Dual self-attention Swin transformer for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–12. DOI: 10.1109/TGRS.2023.3275146

[20] QI, W., HUANG, C., WANG, Y., et al. Global-local 3-D convolutional transformer network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–20. DOI: 10.1109/TGRS.2023.3272885

[21] ZHAO, F., ZHANG, J., MENG, Z., et al. Multiple vision architectures-based hybrid network for hyperspectral image classification. *Expert Systems with Applications*, 2023, vol. 234, p. 1–16. DOI: 10.1016/j.eswa.2023.121032

[22] WANG, Y., MEI, J., ZHANG, L., et al. Self-supervised feature learning with CRF embedding for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, vol. 57, no. 5, p. 2628–2642. DOI: 10.1109/TGRS.2018.2875943

[23] BAI, J., ZHOU, Z., CHEN, Z., et al. Cross-dataset model training for hyperspectral image classification using self-supervised learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 62, p. 1–17. DOI: 10.1109/TGRS.2024.3493969

[24] CAO, X., YU, J., XU, R., et al. Mask-enhanced contrastive learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 62, p. 1–15. DOI: 10.1109/TGRS.2024.3479220

[25] YE, Z., CAO, Z., LIU, H., et al. Self-supervised learning with multiscale densely connected network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 62, p. 1–15. DOI: 10.1109/TGRS.2024.3424394

[26] HE, K., CHEN, X., XIE, S., et al. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans (USA), 2022, p. 16000–16009. DOI: 10.1109/CVPR52688.2022.01553

[27] ZHOU, F., XU, C., YANG, G., et al. Masked spectral-spatial feature prediction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 62, p. 1–13. DOI: 10.1109/TGRS.2023.3344782

[28] CAO, M., ZHANG, X., CHENG, J., et al. Spatial-spectral-semantic cross-domain few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 62, p. 1–19. DOI: 10.1109/TGRS.2024.3434484

[29] LI, Z., LIU, M., CHEN, Y., et al. Deep cross-domain few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, vol. 60, p. 1–18. DOI: 10.1109/TGRS.2021.3057066

[30] WANG, Y., LIU, M., YANG, Y., et al. Heterogeneous few-shot learning for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 2021, vol. 19, p. 1–5. DOI: 10.1109/LGRS.2021.3117577

[31] ZHANG, Y., LI, W., ZHANG, M., et al. Graph information aggregation cross-domain few-shot learning for hyperspectral image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, vol. 35, no. 2, p. 1912–1925. DOI: 10.1109/TNNLS.2022.3185795

[32] LI, Z., GUO, H., CHEN, Y., et al. Few-shot hyperspectral image classification with self-supervised learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–17. DOI: 10.1109/TGRS.2023.3298851

[33] XIAO, F., HAN, X., CAO, C., et al. Neural architecture search-based few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 61, p. 1–15. DOI: 10.1109/TGRS.2024.3385478

[34] QIN, B., FENG, S., ZHAO, C., et al. Cross-domain few-shot learning based on feature disentanglement for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, vol. 62, p. 1–15. DOI: 10.1109/TGRS.2024.3386256

[35] YE, Z., WANG, J., LIU, H., et al. Adaptive domain-adversarial few-shot learning for cross-domain hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–17. DOI: 10.1109/TGRS.2023.3334289

[36] WANG, F., KONG, T., ZHANG, R., et al. Self-supervised learning by estimating twin class distribution. *IEEE Transactions on Image Processing*, 2023, vol. 32, p. 2228–2236. DOI: 10.1109/TIP.2023.3266169

[37] GRANA, M., VEGANZONS, M. A., AYERDI, B. *Hyperspectral Remote Sensing Scenes*. Available online: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[38] DONG, Z., LIU, T., GU, Y. Spatial and semantic consistency contrastive learning for self-supervised semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, vol. 61, p. 1–12. DOI: 10.1109/TGRS.2023.3317016

[39] HONG, D., HAN, Z., YAO, J., et al. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, vol. 60, p. 1–15. DOI: 10.1109/TGRS.2021.3130716

About the Authors . . .

Qizhi FANG received a B.E. degree from Shenyang Aerospace University, Shenyang, China, in 2002 and received a M.E. degree from Northeastern University, Shenyang, China, in 2008. He is currently an Associate Professor at the College of Electronic Information Engineering, Shenyang Aerospace University and Liaoning General Aviation Academy, Shenyang, China. His research interests include image compression, image classification, radar-based human activity classification, and SAR-based ship detection.

Yubo ZHAO received a B.E. degree from Shenyang Aerospace University, Shenyang, China, in 2023. He is currently a postgraduate student at the College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China. His current research interests include deep learning and image classification.

Jingang WANG received a B.E. degree from Shenyang Aerospace University, Shenyang, China, in 2023. He is currently a postgraduate student at the College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China. His current research interests include deep learning and multispectral image compression.

Lili ZHANG received her B.E., M.E., and Ph.D. degrees from Jilin University, Changchun, China, in 2002, 2005, and 2012. She is currently an Associate Professor at the College of Electronic Information Engineering, Shenyang Aerospace University, Shenyang, China. Her current research interests include image compression, radar-based human activity classification, and SAR-based ship detection.