



Article

Enhancing Semi-Supervised Few-Shot Hyperspectral Image Classification via Progressive Sample Selection

Jiaguo Zhao ¹, Junjie Zhang ^{1,*} , Huaxi Huang ² and Jian Zhang ³

¹ Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute of Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China; jiaguo_zhao@shu.edu.cn

² Commonwealth Scientific and Industrial Research Organisation, Sydney, NSW 2601, Australia; huaxi.huang@csiro.au

³ Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia; jian.zhang@uts.edu.au

* Correspondence: junjie_zhang@shu.edu.cn

Abstract: Hyperspectral images (HSIs) provide valuable spatial–spectral information for ground analysis. However, in few-shot (FS) scenarios, the limited availability of training samples poses significant challenges in capturing the sample distribution under diverse environmental conditions. Semi-supervised learning has shown promise in exploring the distribution of unlabeled samples through pseudo-labels. Nonetheless, FS HSI classification encounters the issue of high intra-class spectral variability and inter-class spectral similarity, which often lead to the diffusion of unreliable pseudo-labels during the iterative process. In this paper, we propose a simple yet effective progressive pseudo-label selection strategy that leverages the spatial–spectral consistency of HSI pixel samples. By leveraging spatially aligned ground materials as connected regions with the same semantic and similar spectrum, pseudo-labeled samples were selected based on round-wise confidence scores. Samples within both spatially and semantically connected regions of FS samples were assigned pseudo-labels and joined subsequent training rounds. Moreover, considering the spatial positions of FS samples that may appear in diverse patterns, to fully utilize unlabeled samples that fall outside the neighborhood of FS samples but still belong to certain connected regions, we designed a matching active learning approach for expert annotation based on the temporal confidence difference. We identified samples with the highest training value in specific regions, utilizing the consistency between predictive labels and expert labels to decide whether to include the region or the sample itself in the subsequent semi-supervised iteration. Experiments on both classic and more recent HSI datasets demonstrated that the proposed base model achieved SOTA performance even with extremely rare labeled samples. Moreover, the extended version with active learning further enhances performance by involving limited additional annotation.

Keywords: hyperspectral image classification; semi-supervised learning; pseudo-label selection; active learning



Citation: Zhao, J.; Zhang, J.; Huang, H.; Zhang, J. Enhancing Semi-Supervised Few-Shot Hyperspectral Image Classification via Progressive Sample Selection. *Remote Sens.* **2024**, *16*, 1747. <https://doi.org/10.3390/rs16101747>

Academic Editor: Pedro Melo-Pinto

Received: 19 March 2024

Revised: 7 May 2024

Accepted: 13 May 2024

Published: 15 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Hyperspectral imaging technology has found extensive applications in various domains, including geological mapping, environmental monitoring, vegetation analysis, atmospheric characterization, biochemical detection, and disaster assessment. The pixel-level classification of hyperspectral images (HSIs) based on spectral information, known as HSI classification, is a prominent research focus due to its significant academic, civilian, and military value. Despite the remarkable achievements of deep learning-based HSI classification methods in recent years [1–5], it is evident that these methods often rely on a large number of high-quality labeled samples, which necessitates time-consuming and labor-intensive real-time investigations, along with expert interpretation of pixel-level

images. Therefore, conducting research on HSI classification in few-shot (FS) scenarios holds immense practical significance for advancing and applying hyperspectral technology.

FS HSI classification poses challenges related to the curse of dimensionality (Hughes phenomenon), scarcity of training samples, and domain adaptation, leading to subpar performance. To address these challenges, researchers have explored various data pre-processing techniques. One common approach involves data augmentation methods such as spectral data enhancement [6], pixel-block pair augmentation [7], random occlusion [8], and expanded pixel neighborhood information [9]. These techniques aim to increase the quantity and diversity of training samples. However, it is important to note that, in FS scenarios, these methods often result in a significant increase in computational complexity while providing only marginal improvements in classification performance.

Apart from pre-processing, existing works also focus on addressing FS classification issues through the design of training strategies, which can be roughly categorized as strict Few-Shot Learning (FSL) [10,11], semi-supervised learning (SSL) [12,13], and active learning (AL) [14,15]. FSL models often rely on meta-learning training strategies that utilize a limited number of labeled samples and an auxiliary set to construct multiple meta-learning tasks [16], enabling the acquisition of an initialization model with strong generalization capabilities. These models typically extract spatial–spectral features as a first step, followed by classification using methods such as measuring cosine similarities between queries and prototypes or predicting pairwise relation scores [17]. In contrast, SSL methods consider a large number of unlabeled samples in conjunction with extremely scarce labeled samples to guide model construction. Representative approaches include self-training via pseudo-labeling [18], the combination of deep clustering [19,20], and label propagation through graph networks [21], among others. Semi-supervised methods demonstrate particular effectiveness when only a few labeled samples are available, often outperforming their FS-supervised counterparts. Unlike FSL and SSL approaches, AL models selectively perform expert labeling by assessing the information of unlabeled samples. These models evaluate the unlabeled samples using techniques such as the large margin theorem [22], posterior probability [23], and committee-based methods [24], etc. AL techniques offer a different perspective by actively involving experts in the labeling process to enhance classification performance.

Given its generalization ability and the natural usage of unlabeled samples, we explore the FS HSI classification under a semi-supervised setting, and establish a self-training-based model. Though existing training strategies serve to mitigate the issue of the limited generalization of FS samples and enhance classification accuracy to a certain extent, most of them primarily stem from techniques developed for generic optical images, often overlooking the unique characteristics of HSIs. The high-dimensional yet redundant spectral channels, and anabatic high intra-class spectral and inter-class spectral similarity largely affect the process of pseudo-labeling with FS-labeled samples, which induces the risk of diffusing unreliable pseudo-labels during the iterative process.

To address the aforementioned issues, we first performed a dimension reduction in conjunction with a hybrid feature extraction backbone. This integration combines convolutional operations and self-attention mechanisms, enabling the comprehensive extraction of spatial, spectral, and global contextual correlations, facilitating multilevel data comprehension. Taking into account the abundant information from unlabeled samples in HSI, we developed a simple yet effective pseudo-label selection strategy that leverages the nuanced spatial–spectral consistency of HSI, exploring valuable unlabeled samples within both semantically and spatially connected regions, distinguishing our model from previous semi-supervised approaches. As illustrated in Figure 1, from the aerial perspective of the HSI, ground materials are represented as high-dimensional pixels, with materials of the same type naturally appearing as spatially connected regions with similar spectral distributions.

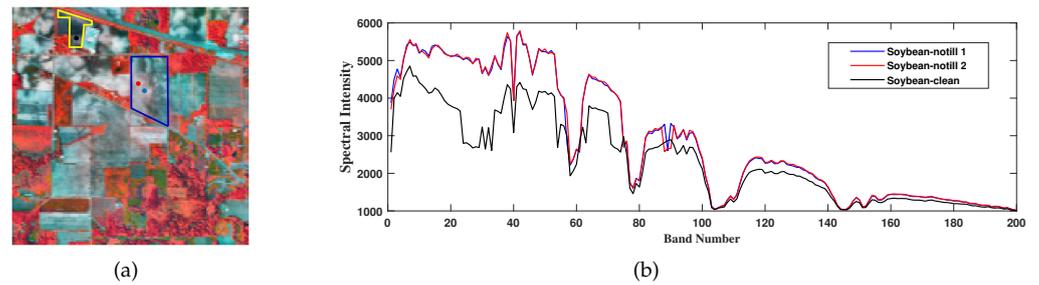


Figure 1. The The spatial–spectral consistency on Indian Pines dataset. Two regions with different types of ground materials (soybean-notill and soybean-clean) are noted in blue and yellow, and the spectral intensity of three pixels are sampled from them. The spatially adjacent pixels with the same class label are present in similar spectral distributions (blue and red), while the distant black pixel with another label is much different. This figure is better viewed in color. (a) Sampling regions, (b) Spectral curves of sampled samples.

Leveraging the robust feature extractor that integrates the capabilities of convolutional networks and vision transformers, our proposed approach selects pseudo-labeled samples based on connected regions derived from iterative confidence scores. The overall framework is illustrated in Figure 2. Initially, using the available FS training samples, we identify unlabeled samples with confident predictions and link them to connected regions determined from confidence scores. Unlabeled samples situated in the same semantic regions as the FS samples are deemed reliable and assigned pseudo-labels for inclusion in subsequent training rounds, thus selecting pseudo-labeled samples in a stepwise manner. This self-training process iterates to enhance the model’s performance. Additionally, considering the varied spatial positions of FS training samples, to fully utilize unlabeled samples outside the semantic neighborhood of FS samples but still within confidently connected regions, we identify the highest value training samples in a given region by evaluating their confidence variations across training rounds and request expert annotations. The agreement between predicted and expert labels determines whether to include the region or the sample itself in the subsequent round. This approach facilitates the global expansion of pseudo-labeled samples, enriching the feature distribution and providing additional benefits when integrated with pseudo-labeling techniques.

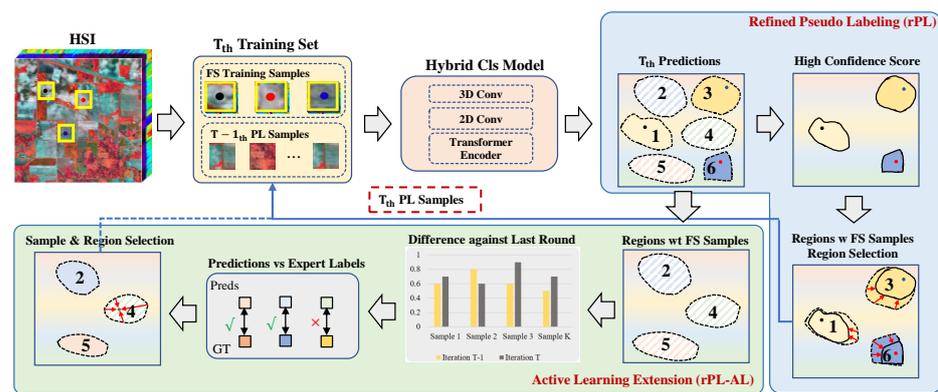


Figure 2. The overall pseudo-labeling framework. FS training samples are passed through the hybrid classification model to obtain initial predictions. Samples from the connected regions that contain FS samples are selected for the next round of training based on the proposed refined pseudo-labeling. Conversely, active learning is employed to leverage the connected regions without FS samples, determining whether to include the entire region or specific samples for the next iteration.

The main contributions can be summarized as follows:

- We identified the key challenges in FS HSI classification and performed a comprehensive analysis of the spatial–spectral consistency presented in HSI data. We thus designed an efficient pseudo-label selection strategy that fully utilizes this property and incorporates the feature distribution of abundant unlabeled samples.
- In addition to the proposed pseudo-labeling model, we propose an AL approach for expert annotation selection based on the temporal spatial–spectral confidence difference to mitigate the randomness of initial FS training samples, thereby expanding the utilization of unlabeled samples.
- The extensive experimental results demonstrate that our base model manages to achieve state-of-the-art performance without any bells and whistles, even when provided with extremely limited labeled samples. Furthermore, the incorporation of the AL approach enhances performance at the cost of limited expert annotations.

2. Related Work

2.1. Semi-Supervised Learning Methods

In the field of HSI classification, deep learning models such as 2D-CNN [25], 3D-CNN [26], and Long Short-Term Memory Network (LSTM) [27] frequently demonstrate limited classification performance when labeled samples are scarce. Consequently, researchers have turned to semi-supervised learning methods to improve the classification performance of these models by utilizing both labeled and unlabeled samples. Commonly employed semi-supervised learning methods in HSI classification encompass self-training models [28], generative models [29], and graph models [30–33].

For instance, Chen et al. [34] introduced a semi-supervised approach to alleviate overfitting by employing the soft pseudo-labeling of auxiliary unlabeled samples. This method aids in training the feature extractor with a limited number of labeled samples. Each labeled sample serves as a reference, and soft pseudo-labels are assigned by computing the distance between the unlabeled samples and the labeled sample. Similarly, Cui et al. [35] proposed a sparse representation-based (SRSPL) pseudo-labeling method for HSI classification. This method utilizes sparse representation to select the most reliable pseudo-labeled samples for extending the training set. Yue et al. [36] designed a joint loss function that simultaneously considers label loss for labeled samples, soft label loss for unlabeled samples, and self-supervised loss functions. Zhao et al. [12] introduced a novel hyperspectral-specific augmentation module to generate sample pairs. They employed self-supervised models to extract features from these pairs through contrastive learning. Feng et al. [13] proposed constructing specific branches consisting of multiple sets of generators and discriminators based on clustering partitions. This adaptation helps mitigate the class and mode collapse problem when dealing with a large number of samples. Tong et al. [37] proposed a graph convolution model that utilizes attention mechanism weights to exploit the correlation among unlabeled samples for category inference. Building upon conventional classification networks, Li et al. [38] incorporated the constraints of self-supervision through a metric learning branch, utilizing compositional sample pairs.

2.2. Active Learning Methods

Active learning (AL) is a sample partitioning method that iteratively selects high-value annotated samples using a sampling strategy to identify samples with significant information content in each round, which are then provided to experts for annotation. Unlike traditional machine learning's sample selection strategy, AL selectively chooses samples from a candidate sample set based on their inherent properties that align with the specific sampling strategy.

With the advancement of deep learning, the integration of active learning (AL) with deep neural networks has become increasingly prominent. In HSI processing, Liu et al. [39] addressed the spatial characteristics of HSI by integrating multi-scale residual networks as classifiers into AL frameworks. This approach fully leverages the contextual features of the spatial dimension of HSI and devises a sampling strategy based on BreakingTies, reducing

the sample requirement to one-third of the original. YANG et al. [40] proposed a multi-path residual pairwise Siamese network integrated with AL. Initially, they developed a Siamese network with a relatively low requirement for sample data and integrated the AL strategy to select more representative samples, thereby enhancing the discriminative ability of the model. Additionally, et al. [41] established a novel active inference transfer convolutional fusion network (AI-TFNet) for HSI classification. Furthermore, they constructed an active inference (AI) pseudo-label propagation algorithm for spatially chi-squared samples by leveraging the chi-squared pre-segmentation of TFNet. Ding et al. [15] proposed a heuristic AL approach based on clustering constraints to select informative and diverse samples for expert labeling. In contrast, Liu et al. [42] introduced the concept of feature-driven AL, wherein sample selection is conducted in an optimized feature space, and its efficacy is evaluated based on the overall probability of error.

3. Proposed Method

3.1. Generic Pseudo-Labeling FS HSI Classification and Limitations

Building upon prior few-shot HSI classification works [21], we established notations for this task. We defined the few-shot labeled training set \mathcal{DL} , which consists of samples from C classes; K is the number of samples in each class. In the stage of prediction, given a sample from the unlabeled set \mathcal{DQ} , the objective is to assign the sample to one of the C classes. The entire dataset, representing the complete HSI, is denoted as follows:

$$\mathcal{D} = \left\{ \mathcal{DL} = \left\{ (x_l, y_l)_{l=1}^{K \times C} \right\} \cup \left\{ \mathcal{DQ} = \{x_q\}_{q=1}^N \right\} \right\}, \quad (1)$$

$$y_l \in 1, 2, \dots, C, N \gg K \times C$$

Here, $x_l \in \mathbb{R}^{d \times 1}$ denotes the l -th d -dimensional FS-labeled sample, y_l denotes its corresponding class label, $x_q \in \mathbb{R}^{d \times 1}$ is the q -th unlabeled sample, and N represents the number of unlabeled samples. It is important to note that the training and testing sets share the same label space. For FS HSI classification, the labeled samples \mathcal{DL} were used for model training. In the testing phase, all the remaining unlabeled samples \mathcal{DQ} were treated as the whole test set.

Pseudo-labeling is a classic semi-supervised framework used to improve the performance of supervised learning models by leveraging unlabeled data. The process involves training an initial model using a limited set of labeled data, which is then used to generate pseudo-labels for the unlabeled dataset based on the model's predictions. In line with this concept, given an unlabeled sample x_q , the pseudo-label \hat{y}_q can be obtained as follows:

$$\hat{y}_q = \operatorname{argmax}_c f_\theta(x_q)_c \quad (2)$$

where $f_\theta(\cdot)$ represents the current model, $f_\theta(x_q)$ is a vector of category probabilities, where the length of the vector is the number of classes in the dataset, and each value in the vector represents the probability value that the sample belongs to that class, and $f_\theta(x_q)_c$ is the predicted probability of sample x_q belonging to class C . The $\operatorname{argmax}_c f_\theta(x_q)_c$ denotes the class corresponding to the number with the largest value in the probability vector as the pseudo-label for that sample. The labeled data are augmented with the pseudo-labeled data to create an enlarged training set. This augmented dataset is then used to retrain the model, incorporating the new information from the pseudo-labeled samples. The process of pseudo-label generation and model retraining is iteratively repeated, refining the model's predictions on the unlabeled data and potentially improving its overall performance.

While generic pseudo-labeling has shown promising improvements in classification performance when the amount of labeled data is limited, directly applying it to FS HSI classification introduces specific challenges. FS HSI classification is often characterized by the severe Hughes phenomenon, sample scarcity, and domain adaptation. These challenges may increase the risks of diffusing unreliable pseudo-labels during the rounding process. We evaluated the percentage of wrongly annotated pseudo-labels by comparing them

against the ground-truth labels. The results of different training rounds are illustrated in Figure 3. It is evident that, without the careful selection of reliable pseudo-labels, it is difficult to reduce the error rate as the rounds increase. However, with our proposed progressive pseudo-label selection strategy, the accuracy is significantly improved. In the following sections, we provide detailed explanations of our proposed method.

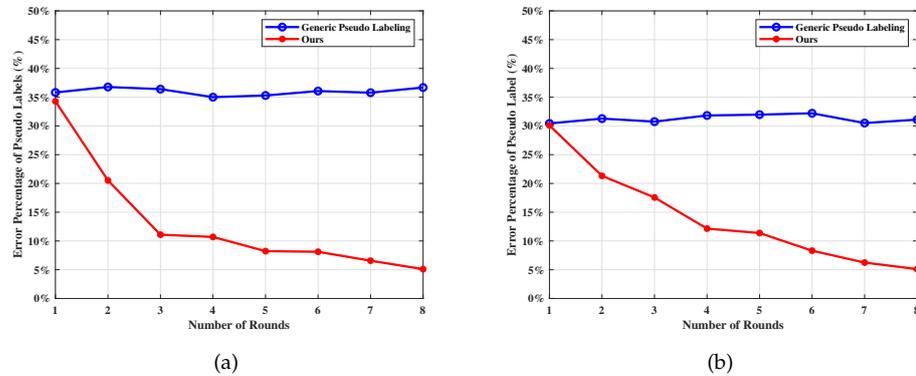


Figure 3. The error rates of wrong pseudo-labels during different rounds with two sampling strategies: (a) Indian Pines, (b) Pavia University.

3.2. Progressive Pseudo-Label Selection Guided by Spatial–Spectral Consistency

The key to enhancing the performance of the pseudo-labeling process lies in the selection of the most reliable samples for subsequent training. In order to tackle this challenge, we proposed leveraging the inherent spatial–spectral consistency found in HSI to allocate the pseudo-labels. In HSI, ground materials of the same type often exhibit spatial proximity, appearing as pixels within a certain neighborhood range with similar spectral signatures. Building on this observation, we introduced a hybrid classification framework; this framework enables the extraction of fine-grained spatial–spectral features from pixel patches centered around the FS training samples. The original training samples are shown in Figure 4a; they were used to train the framework. At the initial training rounds, after obtaining predicted confidence scores from the test patches, we generated two maps, namely the score map $P(i, j)$ Figure 4b and the label map $\hat{Y}(i, j)$ Figure 4c, which possess the same shape $[H, W]$ as the original HSI. In these maps, each spatial position (i, j) corresponds to the score p_{ij} and the predicted label \hat{y}_{ij} of the test patch centered around the corresponding pixel x_{ij} , respectively. The FS training samples were naturally included in the next training round. To filter out unreliable predictions, we masked out positions with confidence scores lower than the threshold τ using the following equations:

$$p_{ij} = \max(f_{\theta}^t(x_{ij})) \quad (3)$$

$$\hat{y}_{ij} = \operatorname{argmax}_c f_{\theta}^t(x_{ij})_c \quad (4)$$

$$P(i, j) = \begin{cases} p_{ij} & p_{ij} > \tau \\ 0; & \text{Otherwise} \end{cases} \quad (5)$$

$$\hat{Y}(i, j) = \begin{cases} \hat{y}_{ij} & p_{ij} > \tau \\ y_{ij}; & x_{ij} \in \mathcal{DL} \\ -1; & \text{Otherwise} \end{cases} \quad (6)$$

Here, for the determination of the threshold τ , we made a reasonable choice through the following experiments. y_{ij} is the ground-truth label for x_{ij} sample; $i \in [1, H]$ and $j \in [1, W]$.

For simplicity, we omitted the spectral dimension of the sample x_{ij} . The index t represents the current training round.

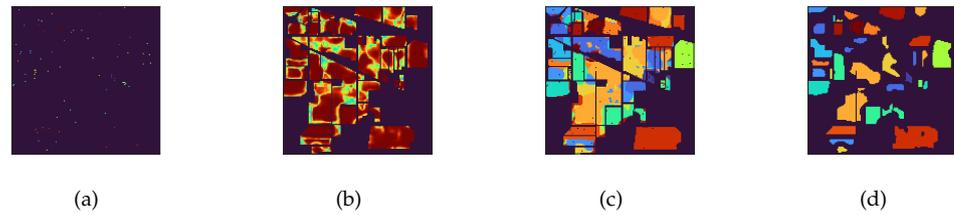


Figure 4. Progressive pseudo-label selection strategy. (a) Original few-shot training samples. (b) Predicted confidence score map. (c) Predicted label map. (d) Selected pseudo-label map. The brownish-red color in (b) represents higher confidence scores. Different colors in (c,d) represent different classes.

Considering the regional distribution of HSI pixels, as depicted in Figure 1, there exists a substantial likelihood that ground materials within a specific range of spatial neighborhoods belong to the same class. Initially, each pixel in the HSI is transformed into spatial-spectral features by the classification network in patch form. We denote the category information of the pixel in terms of confidence score, wherein pixels with low confidence, indicating difficulty in classification, are filtered out using a thresholding approach. Subsequently, the resulting predicted label map $\hat{Y}(i, j)$ naturally comprises multiple connected regions, signifying similarity in the predicted label of pixels within these regions. Among them, we prioritize those encompassing FS training samples as the most dependable. This preference is attributed to the spatial proximity of the samples within these regions to the training samples, consistent with the inherent distribution of the corresponding ground materials. Finally, we utilize these regions to select samples whose predicted labels align with those of the training samples, designating them as pseudo-labeled samples, as shown in Figure 4d. These pseudo-labeled samples are then integrated into subsequent training iterations using the following formula:

$$\mathcal{G} = \{g_i; \text{ if } x_l \in g_i\} \quad (7)$$

$$\hat{\mathcal{D}}\mathcal{L}^{t+1} = \hat{\mathcal{D}}\mathcal{L}^t \cup \{x_q; \text{ if } x_q \in \mathcal{G}\} \quad (8)$$

Here, g_i represents the i -th connected region, identified using an eight-neighborhood approach. The set \mathcal{G} denotes the selected regions that contain FS samples. $\hat{\mathcal{D}}\mathcal{L}^{t+1}$ represents the training set that incorporates the newly added pseudo-labels for the $(t + 1)$ -th iteration.

3.3. Incorporation of Active Learning

The semi-supervised sample selection strategy described above shows good results in exploiting the potential of unlabeled samples within the neighborhood of FS training samples. However, the limitation of this approach is that unlabeled samples can only be utilized based on the initial position of the FS training samples. To overcome this limitation, we designed a new active learning (AL) approach by additionally selecting some highly informative test samples for expert annotation within the connected regions that do not contain the original FS training samples, as depicted in Figure 5. This compensates for the shortcomings of the previous strategy, i.e., ignoring a certain number of high-confidence test samples in subsequent rounds.

In AL, the choice of a query strategy is pivotal. Our objective was to identify and select samples that provided the highest value for model training. As the connected regions no longer have the support of FS samples with ground-truth labels, it became crucial to maximize the utilization of these regions. Hence, our aim was to find test samples that, through expert annotations, could absorb as many confident pseudo-labels as possible for the next round. To achieve this, we adopted a two-fold perspective in identifying the most valuable queries: (1) As the first step of Figure 5, we prioritize regions that contain a significant number of confident predictions. These regions are more likely to

provide reliable information for further training. (2) Within these regions, we identify the sample with the lowest confidence by the maximum confidence difference of the score maps of neighboring rounds and compare their predictions with the expert labels. This comparison helps us determine whether the particular sample or the entire region should be included in subsequent training. By employing this approach, we ensured that informed decisions were made regarding the inclusion of specific samples and regions in the training process. This optimization of the selection strategy enhanced the overall performance of the classification task.

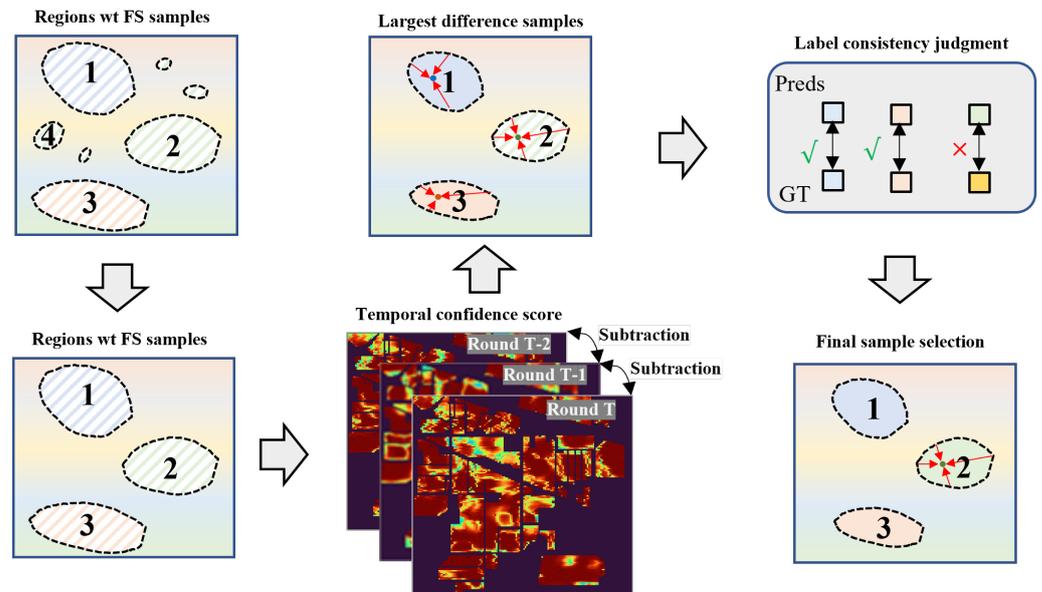


Figure 5. The proposed active learning framework.

More specifically, firstly, we consider the presence of outliers in the remaining connected regions that deviate significantly from the norm. These outliers may have limited suitability for sample selection or differentiation. In order to address this issue, we introduce a sample size threshold N_τ to identify and discard a small number of outliers, if any, from the remaining connected regions. This threshold ensures that the selection process focuses on the majority of samples while disregarding a few potential outliers:

$$\hat{\mathcal{G}} = \{g_i; \text{ if } x_l \notin g_i \& |g_i| > N_\tau\} \tag{9}$$

Here, $\hat{\mathcal{G}}$ denotes the set of regions that satisfy our selection criteria, and $|g_i|$ is the number of samples in region g_i .

Next, we proceed to identify a predicted sample in each region g_i in $\hat{\mathcal{G}}$. Firstly, we select the bottom n confident samples from g_i . For each selected sample, we calculate the difference between its confidence score from the current and the previous iteration. The sample with the largest difference indicates the most informative and valuable candidate for expert annotation. We consult the expert to obtain the ground-truth label for this sample. If the predicted label aligns with the expert one, we incorporate all samples within this region for the subsequent round of training. Otherwise, we include only the selected sample itself. This process can be expressed as follows:

$$v = \operatorname{argmax}_j |p^t(x_j) - p^{t-1}(x_j)|_j; \quad \forall x_j \in \text{Btm}(g_i) \tag{10}$$

$$\hat{y}_v = \operatorname{argmax}_c f_\theta^t(x_v)_c \tag{11}$$

$$\hat{\mathcal{D}}\mathcal{L}^{t+1} = \begin{cases} \hat{\mathcal{D}}\mathcal{L}^t \cup \{g_i\}; & \text{if } \hat{y}_v = y_v \\ \hat{\mathcal{D}}\mathcal{L}^t \cup \{x_v\}; & \text{Otherwise} \end{cases} \tag{12}$$

Here, x_v represents the unlabeled sample that undergoes the maximum change between two consecutive training rounds. $\text{Btm}(g_i)$ denotes the set of bottom n confident samples within g_i . \hat{y}_v represents the predicted label assigned by the current classifier, and y_v denotes the expert annotation for that particular sample.

It is worth noting that the number of expert-labeled samples in each round of the AL stage is less than 10. Therefore, our proposed AL method provides a complementary approach for sample selection, which can ensure that the pseudo-labeled samples cover the distribution of the entire dataset as comprehensively as possible by shifting from local to global patterns. Full details of the training process are given in Algorithm 1.

Algorithm 1 Expanding reliable training set via progressive sample selection

Input: FS-labeled set \mathcal{DL} ; unlabeled set \mathcal{DQ} ; max training rounds R ; confidence threshold τ ; connected region sample size threshold N_τ ; initialized model f_θ^0 ;

Output: Model f_θ^{Iter}

- 1: Training initialized model f_θ^0 with \mathcal{DL}
 - 2: **for** $t = 0$ to R **do**
 - 3: Get confidence score p_{ij} and label \hat{y}_{ij} of \mathcal{DQ} via Equations (2) and (3)
 - 4: Generate two maps $P(i, j)$ and $\hat{Y}(i, j)$ via Equations (4) and (5)
 - 5: Obtain connected regions of \hat{Y} by eight-neighborhood method
 - 6: Select regions \mathcal{G} that contain FS training samples as Equation (6)
 - 7: Merge pseudo-labeled samples as $\hat{\mathcal{DL}}^t$ for next iteration as Equation (7)
 - 8: Density-weighted selection is adopted for remaining regions $\hat{\mathcal{G}}$ via Equation (8)
 - 9: Get predicted label for samples in $\text{Btm}(g_i)$ via Equations (9) and (10)
 - 10: Augment $\hat{\mathcal{DL}}^t$ for next round as Equation (11) Re-train f_θ^t using $\hat{\mathcal{DL}}^{t+1}$ as training set
 - 11: **end for**
-

3.4. Hybrid Classification Framework

Considering the high spectral dimension of HSI pixels and their spatial-spectral consistency, we adopted a hybrid classification framework that leverages the benefits of both CNNs and vision transformers, as shown in Figure 6. CNNs are renowned for their robustness to local and translation invariance, as well as their ability to extract hierarchical features. Conversely, vision transformers excel at capturing global context information. Building upon the inspiration from previous work [43], we constructed our framework as follows.

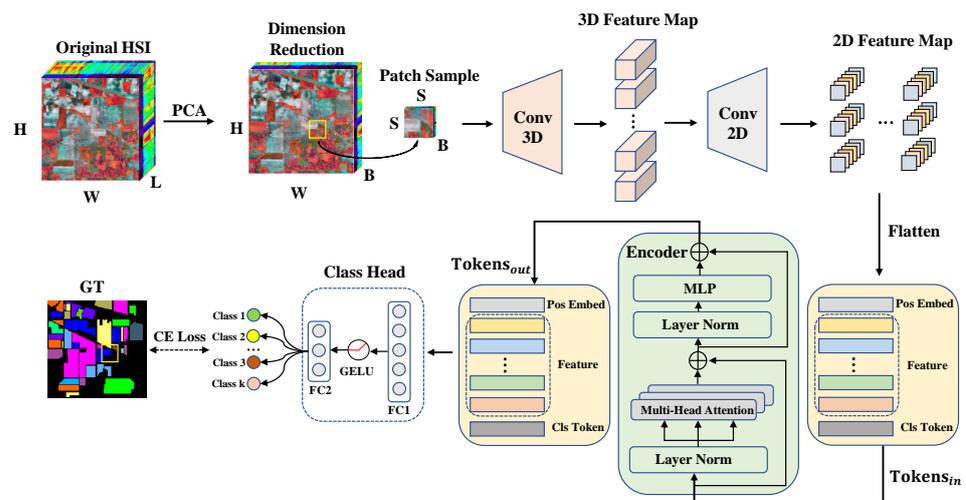


Figure 6. Hybrid classification framework. The patch samples are sequentially passed through the 3D-CNN and 2D-CNN blocks to obtain feature maps that combine spectral and spatial information. These features are then fed to the transformer encoder to capture the contextual information. In the classification head, the model update is constrained by the cross-entropy loss.

3.4.1. Conv Block

We begin by applying PCA dimensionality reduction to the high-dimensional raw data. Our convolutional structure consists of a 3D convolutional (Conv) layer and a 2D Conv layer, which handle spectral and spatial features, respectively. The 3D Conv layer incorporates information from the data to extract spatial–spectral semantic features, and the 2D Conv layer specializes in capturing fine-grained spatial features.

To process the data, we sample each patch $x_i \in \mathbb{R}^{s \times s \times B}$ from the original HSI, centered around a specific pixel using a window size of s . Here, B represents the reduced spectral dimension size. The semantic category of each patch is determined by the label of its center pixel. To account for edge pixels, we apply padding to generate corresponding patches. Instead of using single pixels, we treat patches as individual samples for both training and testing. The 3D Conv layer takes the training patch as input and extracts feature maps accordingly. These maps are reshaped and subsequently fed into the 2D Conv layer. Following standard practice, activation layers are applied after each Conv layer.

3.4.2. Encoder Block

Utilizing the Conv features that capture local information, we incorporate a transformer encoder block to fully explore the correlations within these features. To prepare the 2D feature maps for input into the transformer encoder, we reshape them into a sequence of tokenized data. The length of the sequence is determined by the spatial size of the feature maps, with an additional class token included. To enable the transformer to consider the spatial relationship between tokens, we apply positional embedding to the feature sequences. These sequences are then used as the query (Q), key (K), and value (V) inputs for the multi-head self-attention mechanism. By attending to the class token, we obtain an output that represents the classification probability. This output is processed through a fully connected layer followed by a SoftMax operation. During self-training, the cross-entropy loss is computed using the corresponding ground-truth or pseudo-labels. This loss serves as the objective function to guide the training process and update the model parameters accordingly.

3.4.3. Inference

In the final stage of prediction, the test patches undergo processing through the 3D–2D Conv block, followed by the transformer encoder. The predicted label for each patch is determined based on the category assigned to the center pixel of that particular patch. This approach ensures that the classification decision is made with respect to the central information within each patch, providing a reliable prediction for the HSI classification task.

4. Experimental Analysis

4.1. Datasets

To validate the effectiveness of our proposed method and enable fair comparison with existing approaches, we conducted performance evaluations on four widely recognized public datasets: Indian Pines (IP), Pavia University (PU), WHU-Hi-HanChuan (HC), and WHU-Hi-HongHu (HH).

4.1.1. IP Dataset

The Indian Pines (IP) dataset was acquired using the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over northwestern Indiana, USA. The original dataset consists of 224 spectral bands spanning the range of 0.4 to 2.5 μm . It comprises a total of 145×145 pixels with a spatial resolution of 20m and encompasses 16 distinct land cover classes. For visual reference, Figure 7 presents the false-color map and ground-truth map of the IP dataset.

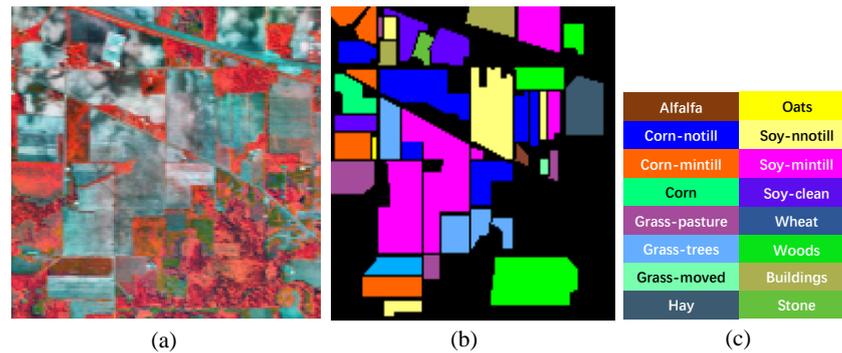


Figure 7. Indian Pines (IP) dataset. (a) False-color map. (b) Ground-truth map. (c) Classes by colors.

4.1.2. PU Dataset

The University of Pavia (PU) dataset was obtained using the Reflection Optical System Imaging Spectrometer (ROSIS) sensor, and provides valuable data captured over the University of Pavia in northern Italy. The original dataset consists of 115 spectral bands, covering a wavelength range of 0.43 to 0.86 μm . The image itself spans 610×340 pixels and has a spatial resolution of 1.3 m. It encompasses information related to nine distinct land cover classes. For a visual representation of the dataset, please refer to Figure 8, which showcases both the false-color map and the ground-truth map associated with the PU dataset.

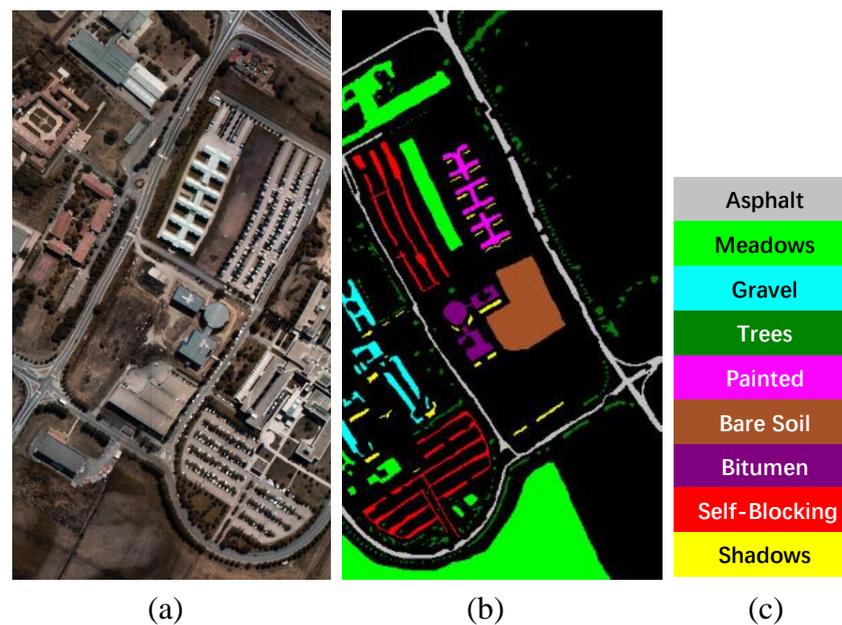


Figure 8. Pavia University (PU) dataset. (a) False-color map. (b) Ground-truth map. (c) Classes by colors.

4.1.3. HC Dataset

The WHU-Hi-HanChuan (HC) dataset was acquired in June 2016 in Hanchuan, Hubei province, China, using a 17 mm focal length Headwall Nano-Hyperspec imaging sensor mounted on a Leica Aibot X6 UAV V1 platform. The imagery has dimensions of 1217×303 pixels with 274 bands covering the range from 400 to 1000 nm. The spatial resolution of the UAV-borne hyperspectral imagery is approximately 0.109 m. Figure 9 is a visual representation of the dataset, which includes both the false-color map and the ground-truth map for the HC dataset.

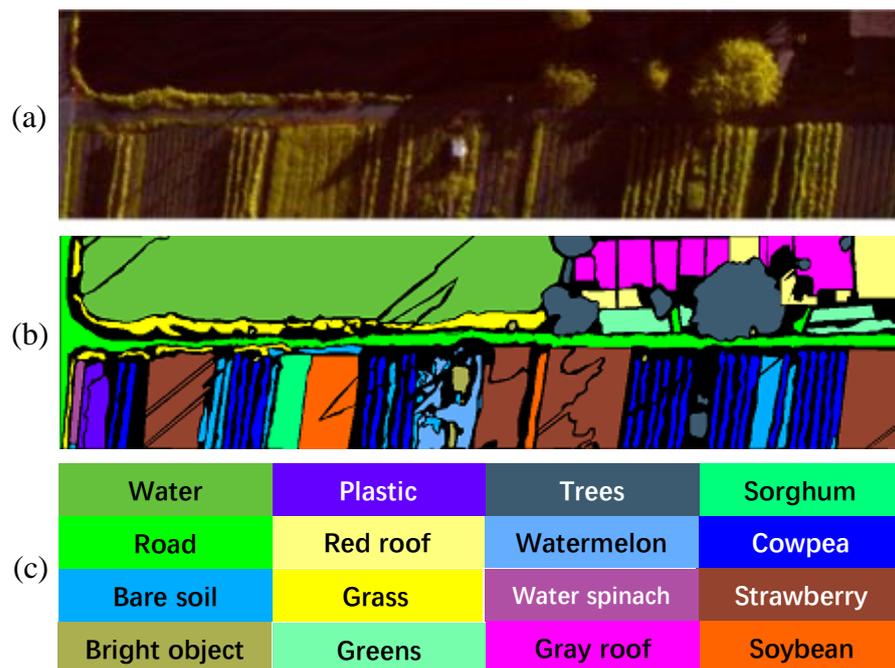


Figure 9. WHU-Hi-HanChuan (HC) dataset. (a) False-color map. (b) Ground-truth map. (c) Classes by colors.

4.1.4. HH Dataset

The WHU-Hi-HongHu (HH) dataset was acquired in November 2017 in Honghu City, Hubei province, China, using a 17 mm focal length Headwall Nano-Hyperspec imaging sensor mounted on a DJI Matrice 600 Pro UAV platform. The imagery has dimensions of 940×475 pixels with 270 bands covering the range from 400 to 1000 nm. The spatial resolution of the UAV-borne hyperspectral imagery is approximately 0.043 m. Figure 10 is a visual representation, which includes both the false-color composite image and the available ground-truth map for the HH dataset.

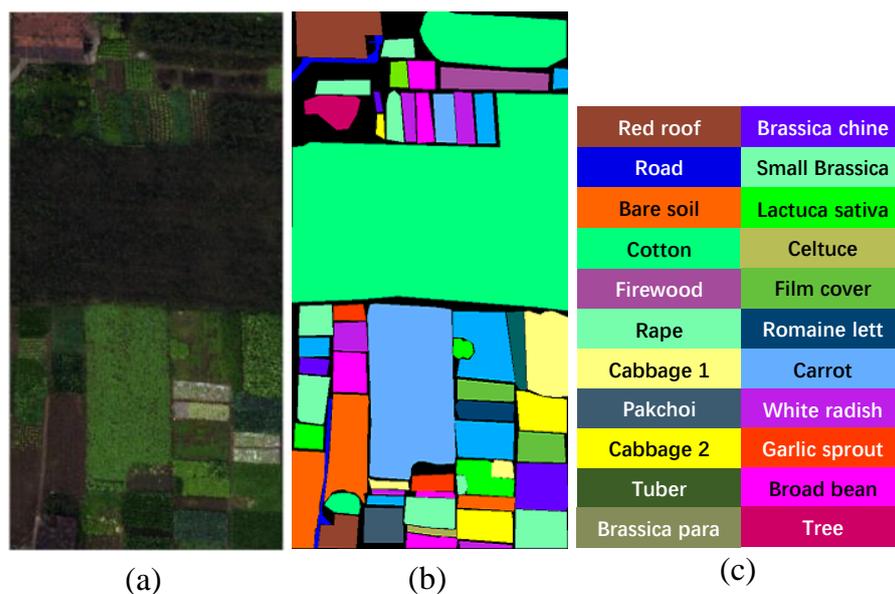


Figure 10. WHU-Hi-HongHu (HH) dataset. (a) False-color map. (b) Ground-truth map. (c) Classes by colors.

4.2. Evaluation Metrics

To comprehensively assess the effectiveness of our proposed method and enable a thorough comparison with other approaches, we employed four quantitative evaluation metrics for HSI classification. These metrics include overall accuracy (OA), average accuracy (AA), and kappa coefficient (Kappa). Additionally, we reported the classification accuracy for each specific land cover category. Higher values for these metrics indicate superior classification performance. In addition to quantitative evaluation, we also presented visual representations of the classification maps generated by different models.

4.3. Implementation Details

We utilized the conventional cross-entropy loss function for classification. Specifically, assuming that there are N samples, the output confidence of the model is noted as y_i and the corresponding ground-truth label is t_i , the cross-entropy loss can be expressed as follows:

$$L_{cross-entropy} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C t_{ij} \log y_{ij} \quad (13)$$

where C denotes the number of categories, t_{ij} indicates whether the i_{th} sample belongs to the j_{th} category, and y_{ij} is the predicted corresponding probability. A smaller cross-entropy value signifies a closer alignment between the model's predicted output and the true label, which correlates with improved performance.

In the experiments, we randomly selected five samples per class as initial samples (FS) and the number of annotations in each round of active learning was limited to no more than 10. We utilized the Adam optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 5×10^{-4} to train our models. Each training round was performed using a mini-batch size of 32 for a total of 50 epochs. We set the maximum number of rounds to eight. Empirically, we selected $\tau = 0.6$ as the threshold and initially set N_τ to 100. The value of N_τ was halved every two training rounds. Furthermore, we set the bottom sample size n_{btm} to 10.

For the hybrid classification framework, by following the parameter settings in the paper [44], the number of principal components B was set to 30, and the input patch sample width and height were set to 13×13 . The patch size was selected by referring to [43] and experimental comparisons. We observed that larger patches (15×15 and 17×17) showed only marginal impacts on results with heavier computational burden. To strike a balance between effectiveness and efficiency, we adopted a patch size of 13 in our final models. For the transformer backbone, we followed the standard ViT encoder and set the number of heads to eight with a token size of 64. In order to account for sample selection variability, each setting was executed 10 times on each dataset using randomly labeled samples each time. The average accuracy and standard deviation of these 10 runs were reported. All models were trained and tested using the PyTorch framework and executed on a single NVIDIA GeForce RTX 3090 GPU.

To demonstrate the superiority of our model, we conducted a comprehensive comparison with several well-established supervised and semi-supervised HSI classification methods. The comparative analysis included the following approaches: SSFTT [43], S3Net [45], DM-MRN [46], FPGA [47], Fusion of Spectral-Spatial Classifiers (FSS) [48], as well as Superpixel-guided Training Sample Enlargement with Distance-Weighted Linear Regression (STSE-DWLR) [49], and the combination of Superpixel Graph and Poisson Learning (DSSPL) [21]. Moreover, in the HC and HH datasets, we added MDMC [50], SSRN [51], and A2S2K-ResNet [52] to compare. To ensure fair and consistent comparisons, only five original samples of each class were randomly selected for training for all methods. Our proposed base model is referred to as refined Pseudo-Labeling (rPL), and its full version with AL is denoted as rPL-AL.

4.4. Results and Analysis

Tables 1–4 present the classification results of the compared methods on four datasets. The best and second-best performances are highlighted in bold and italic, respectively. Figures 11–14 provide visualizations of the predicted categories, with each category represented by a different color. Our rPL model demonstrates superior performance across all evaluation metrics. Additionally, the incorporation of AL techniques further enhances the overall performance by effectively leveraging limited expert annotations. These conclusions are reinforced by the visualized classification diagrams, confirming the effectiveness and superiority of our proposed approach.

Table 1. Per-class accuracies, OA, AA, and Kappa metrics of different methods on the IP dataset. The best and second-best scores are noted in bold and italic.

Class	SSFTT	S3Net	DM-MRN	FPGA	FSS	STSE-DWLR	DSSPL	Ours: rPL	Ours: rPL-AL
1	100.00 ± 0.00	100.00 ± 0.00	98.76 ± 1.21	100.00 ± 0.00	77.15 ± 2.62	98.48 ± 1.79	99.13 ± 0.52	100.00 ± 0.00	100.00 ± 0.00
2	57.09 ± 4.23	55.77 ± 6.32	78.50 ± 3.23	57.09 ± 3.42	83.88 ± 1.67	47.91 ± 1.93	70.78 ± 1.2	74.52 ± 6.11	91.56 ± 0.78
3	59.07 ± 6.21	68.40 ± 5.22	66.79 ± 7.72	54.72 ± 4.42	77.59 ± 1.93	64.25 ± 1.05	78.29 ± 1.37	81.26 ± 2.53	99.58 ± 0.23
4	98.00 ± 0.34	97.89 ± 1.10	95.69 ± 1.03	99.20 ± 0.21	79.65 ± 2.23	83.84 ± 1.18	85.40 ± 1.92	99.95 ± 0.01	99.96 ± 0.13
5	87.05 ± 4.11	98.11 ± 0.87	87.66 ± 3.43	86.84 ± 3.22	69.07 ± 3.30	81.70 ± 0.94	81.06 ± 1.21	88.20 ± 2.21	99.44 ± 0.23
6	94.35 ± 0.92	98.10 ± 0.41	97.24 ± 1.22	95.31 ± 2.32	81.38 ± 2.67	91.16 ± 0.10	89.40 ± 5.93	99.05 ± 0.32	99.97 ± 0.01
7	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	72.64 ± 3.08	97.86 ± 0.18	98.21 ± 1.98	100.00 ± 0.00	100.00 ± 0.00
8	95.14 ± 0.98	99.76 ± 0.10	100.00 ± 0.00	100.00 ± 0.00	91.29 ± 1.15	99.94 ± 0.01	88.89 ± 3.32	100.00 ± 0.00	100.00 ± 0.00
9	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	66.10 ± 4.34	99.00 ± 0.32	99.00 ± 3.16	100.00 ± 0.00	100.00 ± 0.00
10	52.47 ± 5.61	80.39 ± 3.65	84.39 ± 2.70	55.06 ± 4.91	75.08 ± 1.86	69.89 ± 2.31	81.48 ± 2.31	85.52 ± 2.19	100.00 ± 0.00
11	67.89 ± 5.40	70.56 ± 3.14	63.22 ± 4.54	77.47 ± 2.81	90.42 ± 0.50	83.29 ± 1.92	82.90 ± 2.13	82.93 ± 4.31	99.30 ± 0.23
12	41.59 ± 2.43	63.84 ± 4.54	83.67 ± 5.34	35.47 ± 4.74	79.21 ± 2.15	77.64 ± 1.53	72.48 ± 6.32	70.52 ± 4.45	80.00 ± 3.79
13	98.00 ± 3.42	99.60 ± 1.11	98.13 ± 0.43	99.87 ± 0.11	88.55 ± 1.82	99.51 ± 0.10	99.03 ± 0.21	100.00 ± 0.00	100.00 ± 0.00
14	89.05 ± 2.98	94.24 ± 3.2	94.42 ± 1.38	84.37 ± 3.53	74.38 ± 2.94	83.10 ± 3.51	87.12 ± 1.87	95.44 ± 1.10	99.99 ± 0.00
15	90.57 ± 2.21	91.97 ± 3.41	91.91 ± 1.02	73.66 ± 3.21	64.80 ± 3.55	87.28 ± 2.34	88.06 ± 4.17	94.69 ± 1.40	99.55 ± 0.02
16	83.14 ± 4.10	99.56 ± 0.10	99.56 ± 0.21	89.88 ± 2.10	89.63 ± 1.41	95.14 ± 0.45	90.03 ± 0.34	96.46 ± 2.56	97.35 ± 1.23
OA	71.93 ± 3.56	78.50 ± 1.91	80.54 ± 2.52	72.92 ± 3.92	81.42 ± 1.12	80.52 ± 5.98	85.91 ± 4.84	86.25 ± 3.81	97.41 ± 0.85
AA	82.25 ± 2.76	88.01 ± 0.75	88.89 ± 2.03	81.79 ± 2.34	78.83 ± 2.07	85.76 ± 3.45	87.02 ± 3.12	91.35 ± 1.83	97.95 ± 0.65
Kappa	68.42 ± 3.10	75.83 ± 2.01	78.14 ± 2.08	69.21 ± 4.12	80.28 ± 1.27	78.68 ± 6.41	83.32 ± 5.35	84.30 ± 4.41	97.04 ± 0.97

Table 2. Per-class accuracies, OA, AA, and Kappa metrics of different methods on the PU dataset. The best and second-best scores are noted in bold and italic.

Class	SSFTT	S3Net	DM-MRN	FPGA	FSS	STSE-DWLR	DSSPL	Ours: rPL	Ours: rPL-AL
1	63.93 ± 5.34	81.30 ± 2.41	82.17 ± 4.21	66.98 ± 6.42	70.00 ± 1.78	77.69 ± 4.41	80.72 ± 7.49	98.17 ± 0.41	100.00 ± 0.00
2	78.12 ± 3.22	73.59 ± 0.87	86.15 ± 1.13	74.41 ± 4.32	79.47 ± 8.26	82.25 ± 1.32	92.73 ± 5.09	93.49 ± 1.43	100.00 ± 0.00
3	18.53 ± 4.31	70.67 ± 3.31	84.13 ± 2.91	94.12 ± 2.11	95.06 ± 1.74	98.12 ± 2.43	95.24 ± 3.24	83.15 ± 1.54	99.80 ± 0.01
4	84.6 ± 2.09	66.67 ± 2.34	69.40 ± 3.32	95.02 ± 1.32	32.85 ± 2.00	59.44 ± 6.24	82.83 ± 6.1	77.15 ± 2.31	96.55 ± 1.80
5	99.63 ± 0.95	99.4 ± 0.20	99.01 ± 0.51	100.00 ± 0.00	95.47 ± 7.49	98.54 ± 0.50	96.31 ± 3.00	100.00 ± 0.00	99.81 ± 0.04
6	67.59 ± 4.31	89.27 ± 1.59	54.87 ± 3.67	78.01 ± 4.56	97.38 ± 1.94	97.74 ± 2.72	99.29 ± 0.03	100.00 ± 0.00	100.00 ± 0.00
7	98.40 ± 3.78	99.01 ± 0.69	94.10 ± 5.22	99.84 ± 0.24	99.67 ± 0.70	88.42 ± 5.8	88.06 ± 3.1	100.00 ± 0.00	100.00 ± 0.00
8	95.11 ± 3.54	62.53 ± 8.10	86.31 ± 6.10	57.49 ± 13.04	87.72 ± 2.76	75.12 ± 1.40	83.29 ± 1.57	95.52 ± 1.43	100.00 ± 0.00
9	93.94 ± 2.54	96.3 ± 2.40	88.72 ± 4.31	94.57 ± 2.11	30.41 ± 1.80	99.89 ± 0.00	99.89 ± 0.00	92.15 ± 2.11	99.50 ± 0.42
OA	75.42 ± 3.56	77.1 ± 1.97	81.31 ± 1.66	76.70 ± 2.28	83.78 ± 3.30	83.02 ± 3.80	86.74 ± 2.31	93.00 ± 1.23	99.52 ± 0.23
AA	77.86 ± 5.60	82.01 ± 3.21	82.88 ± 2.29	84.50 ± 1.34	83.78 ± 3.30	87.69 ± 4.03	89.37 ± 5.01	93.23 ± 1.98	99.74 ± 1.12
Kappa	68.69 ± 4.81	71.1 ± 2.55	75.3 ± 1.89	70.73 ± 0.79	77.74 ± 5.40	78.84 ± 4.50	84.05 ± 2.86	91.12 ± 2.5	99.64 ± 0.16

Table 3. Per-class accuracies, OA, AA, and Kappa metrics of different methods on the HC dataset. The best and second-best scores are noted in bold and italic.

Class	SSFTT	S3Net	DM-MRN	MDMC	FPGA	SSRN	A2S2K-ResNet	Ours: rPL	Ours: rPL-AL
1	80.60 ± 5.21	53.92 ± 7.21	68.86 ± 3.36	70.74 ± 4.32	76.45 ± 4.21	62.45 ± 3.11	72.51 ± 6.65	91.65 ± 2.11	97.99 ± 1.17
2	81.26 ± 4.15	82.63 ± 3.90	78.22 ± 3.21	70.60 ± 6.83	76.70 ± 5.32	35.63 ± 8.82	52.11 ± 4.56	87.49 ± 2.65	95.78 ± 1.43
3	84.82 ± 2.32	89.63 ± 4.12	92.29 ± 2.90	94.89 ± 1.54	94.72 ± 2.59	84.54 ± 3.53	90.74 ± 3.78	93.99 ± 2.11	93.58 ± 2.54
4	98.35 ± 1.88	98.99 ± 0.58	99.23 ± 0.95	96.86 ± 1.80	95.04 ± 2.13	95.02 ± 1.34	98.20 ± 2.35	98.99 ± 1.13	99.85 ± 0.26
5	94.89 ± 2.34	66.52 ± 6.98	95.89 ± 2.45	93.82 ± 3.42	99.70 ± 0.38	66.58 ± 6.43	99.91 ± 0.21	100.00 ± 0.00	100.00 ± 0.00
6	52.85 ± 5.34	60.53 ± 4.65	57.31 ± 6.43	66.91 ± 8.54	65.30 ± 4.65	61.34 ± 5.76	70.64 ± 3.54	77.57 ± 3.50	79.53 ± 2.75
7	94.46 ± 2.95	60.80 ± 5.31	74.56 ± 4.39	82.84 ± 2.86	81.11 ± 3.65	40.17 ± 10.54	46.76 ± 8.11	89.59 ± 2.69	96.25 ± 1.56
8	79.15 ± 3.25	65.23 ± 3.65	71.60 ± 5.32	76.45 ± 3.48	76.29 ± 2.65	52.66 ± 4.70	51.08 ± 3.65	79.51 ± 4.52	93.91 ± 2.11
9	21.65 ± 3.25	43.85 ± 2.84	66.91 ± 3.43	34.72 ± 6.47	62.25 ± 3.67	39.92 ± 2.57	32.78 ± 6.71	69.79 ± 3.49	83.26 ± 2.76
10	55.61 ± 7.37	86.63 ± 4.20	87.67 ± 3.10	87.70 ± 3.50	92.06 ± 2.49	93.66 ± 2.40	85.83 ± 2.43	93.72 ± 2.48	91.47 ± 2.43
11	69.14 ± 4.40	54.06 ± 6.33	67.18 ± 5.87	70.4 ± 4.70	87.45 ± 3.51	67.41 ± 8.55	73.28 ± 4.58	85.37 ± 3.61	97.04 ± 1.18
12	93.14 ± 2.41	90.17 ± 2.55	96.24 ± 1.69	81.00 ± 4.06	98.45 ± 0.64	39.99 ± 4.76	65.75 ± 6.70	97.97 ± 1.42	99.92 ± 0.01
13	35.15 ± 7.51	43.62 ± 5.08	50.19 ± 8.53	54.46 ± 6.48	49.95 ± 7.86	70.97 ± 4.54	43.03 ± 6.70	75.80 ± 4.82	76.31 ± 3.54
14	38.98 ± 4.60	43.63 ± 3.90	52.80 ± 6.70	50.46 ± 4.54	63.45 ± 3.41	34.68 ± 3.89	36.93 ± 6.54	70.79 ± 4.50	91.84 ± 2.54
15	96.81 ± 1.45	89.63 ± 3.02	94.00 ± 2.54	77.27 ± 3.57	92.04 ± 2.65	63.98 ± 5.02	88.34 ± 3.54	93.07 ± 2.52	96.51 ± 2.11
16	85.43 ± 2.43	75.63 ± 3.54	88.6 ± 2.65	95.42 ± 1.55	88.40 ± 2.03	70.41 ± 3.83	94.85 ± 2.43	93.71 ± 2.43	99.18 ± 0.87
OA	74.10 ± 3.42	66.87 ± 3.3	76.54 ± 3.82	77.69 ± 1.35	80.84 ± 2.40	61.53 ± 1.35	72.02 ± 2.55	87.41 ± 1.21	95.30 ± 1.19
AA	72.64 ± 2.5	69.31 ± 2.50	77.61 ± 1.73	75.43 ± 1.26	81.89 ± 1.98	61.21 ± 1.15	68.93 ± 2.02	86.7 ± 0.91	92.09 ± 1.03
Kappa	70.23 ± 2.84	62.38 ± 3.7	73.02 ± 3.86	74.17 ± 1.31	77.93 ± 2.64	56.65 ± 1.36	67.59 ± 1.92	85.37 ± 1.34	95.49 ± 1.20

Table 4. Per-class accuracies, OA, AA, and Kappa metrics of different methods on the HH dataset. The best and second-best scores are noted in bold and italic.

Class	SSFTT	S3Net	DM-MRN	MDMC	FPGA	SSRN	A252K-ResNet	Ours: rPL	Ours: rPL-AL
1	92.81 ± 2.80	58.02 ± 4.87	76.15 ± 3.80	88.95 ± 3.32	67.49 ± 3.54	70.08 ± 5.54	91.87 ± 2.55	97.74 ± 0.93	99.16 ± 0.36
2	45.23 ± 9.23	57.38 ± 5.62	46.65 ± 4.07	38.55 ± 5.91	55.94 ± 6.23	58.97 ± 4.43	61.51 ± 3.90	70.60 ± 3.94	84.43 ± 3.08
3	75.57 ± 3.94	26.76 ± 5.84	82.84 ± 3.90	89.22 ± 3.65	86.80 ± 4.75	75.51 ± 2.56	85.17 ± 4.32	93.43 ± 2.97	96.15 ± 2.55
4	83.14 ± 5.54	92.13 ± 3.81	90.74 ± 2.95	91.76 ± 3.10	94.20 ± 2.65	86.71 ± 3.31	94.34 ± 1.95	99.14 ± 0.54	99.80 ± 0.24
5	82.09 ± 3.36	82.18 ± 5.32	83.24 ± 4.87	89.25 ± 3.06	71.22 ± 5.80	91.75 ± 4.65	80.28 ± 5.43	85.82 ± 4.10	92.63 ± 3.59
6	86.59 ± 5.49	90.51 ± 3.89	87.59 ± 4.09	87.97 ± 5.84	93.95 ± 3.97	20.50 ± 7.32	90.55 ± 3.89	96.42 ± 2.54	99.10 ± 0.89
7	32.93 ± 8.64	18.75 ± 4.65	50.44 ± 6.54	30.86 ± 7.43	57.03 ± 6.67	39.96 ± 3.65	14.25 ± 6.04	64.95 ± 5.64	90.74 ± 3.45
8	58.28 ± 4.42	15.92 ± 5.53	45.09 ± 4.11	35.84 ± 3.40	60.04 ± 3.54	23.25 ± 2.86	47.96 ± 4.85	55.59 ± 5.33	84.10 ± 3.51
9	88.37 ± 3.21	87.88 ± 4.09	96.11 ± 2.15	89.46 ± 2.80	85.37 ± 3.69	84.69 ± 3.50	81.95 ± 5.33	96.93 ± 2.66	99.35 ± 1.05
10	33.92 ± 5.65	49.85 ± 5.33	68.25 ± 4.68	65.21 ± 7.41	69.86 ± 6.21	45.56 ± 5.98	35.08 ± 5.32	68.25 ± 6.34	85.79 ± 3.43
11	48.70 ± 8.53	21.76 ± 6.90	41.75 ± 5.32	70.68 ± 5.30	34.82 ± 5.96	49.51 ± 4.87	64.63 ± 5.98	66.48 ± 6.31	76.02 ± 4.86
12	52.36 ± 4.42	39.13 ± 5.42	55.19 ± 6.12	36.59 ± 7.53	60.39 ± 3.21	60.24 ± 5.06	46.87 ± 7.92	71.76 ± 5.54	85.23 ± 4.65
13	56.83 ± 5.11	37.17 ± 4.32	57.90 ± 5.21	56.37 ± 4.64	56.98 ± 6.10	27.67 ± 8.04	35.61 ± 6.32	59.59 ± 4.32	79.07 ± 2.11
14	50.25 ± 5.70	48.16 ± 5.30	70.22 ± 4.63	77.56 ± 3.57	77.63 ± 4.41	55.53 ± 7.39	86.55 ± 3.59	97.36 ± 1.77	99.72 ± 0.43
15	99.60 ± 0.58	97.70 ± 1.06	99.40 ± 0.73	98.80 ± 0.97	99.10 ± 0.79	97.40 ± 1.69	97.99 ± 1.16	99.50 ± 0.89	100.00 ± 0.00
16	91.46 ± 3.11	91.63 ± 3.65	86.85 ± 3.19	99.44 ± 0.69	68.98 ± 5.49	88.06 ± 3.20	85.89 ± 2.65	95.97 ± 1.33	97.82 ± 1.18
17	81.32 ± 5.28	71.94 ± 4.76	99.04 ± 1.10	86.02 ± 2.04	65.02 ± 5.72	77.83 ± 4.19	73.81 ± 4.94	76.94 ± 6.49	88.20 ± 3.19
18	91.72 ± 3.14	97.76 ± 2.11	91.28 ± 3.12	92.81 ± 2.98	93.28 ± 3.94	83.70 ± 3.85	94.36 ± 2.48	98.19 ± 1.10	97.82 ± 1.58
19	71.87 ± 4.87	51.83 ± 5.96	80.84 ± 3.7	82.74 ± 3.95	75.95 ± 3.70	57.14 ± 5.96	82.83 ± 3.54	83.98 ± 4.04	93.97 ± 2.67
20	54.00 ± 4.98	41.25 ± 8.54	70.78 ± 5.32	80.18 ± 3.46	62.77 ± 1.43	59.58 ± 8.33	60.32 ± 4.57	55.90 ± 5.48	75.04 ± 4.54
21	90.03 ± 3.43	90.12 ± 2.87	90.93 ± 3.12	82.46 ± 2.11	96.44 ± 2.15	93.89 ± 1.97	94.63 ± 2.08	98.64 ± 1.09	99.54 ± 0.86
22	98.54 ± 1.12	96.85 ± 2.60	77.72 ± 3.27	84.01 ± 3.62	82.31 ± 3.59	79.07 ± 4.78	94.62 ± 2.11	99.16 ± 0.73	99.63 ± 0.58
OA	70.43 ± 4.54	69.86 ± 2.40	80.37 ± 3.42	80.75 ± 3.21	81.9 ± 2.86	66.01 ± 1.21	78.67 ± 2.08	88.11 ± 1.56	95.16 ± 1.40
AA	71.62 ± 3.49	61.12 ± 2.65	75.41 ± 2.87	75.22 ± 2.53	73.43 ± 2.10	64.84 ± 1.56	72.78 ± 1.45	81.93 ± 1.65	91.06 ± 1.13
Kappa	68.85 ± 3.87	62.75 ± 1.46	75.66 ± 2.66	76.08 ± 2.21	77.31 ± 1.98	58.37 ± 1.43	73.39 ± 1.36	84.98 ± 1.76	93.88 ± 0.98

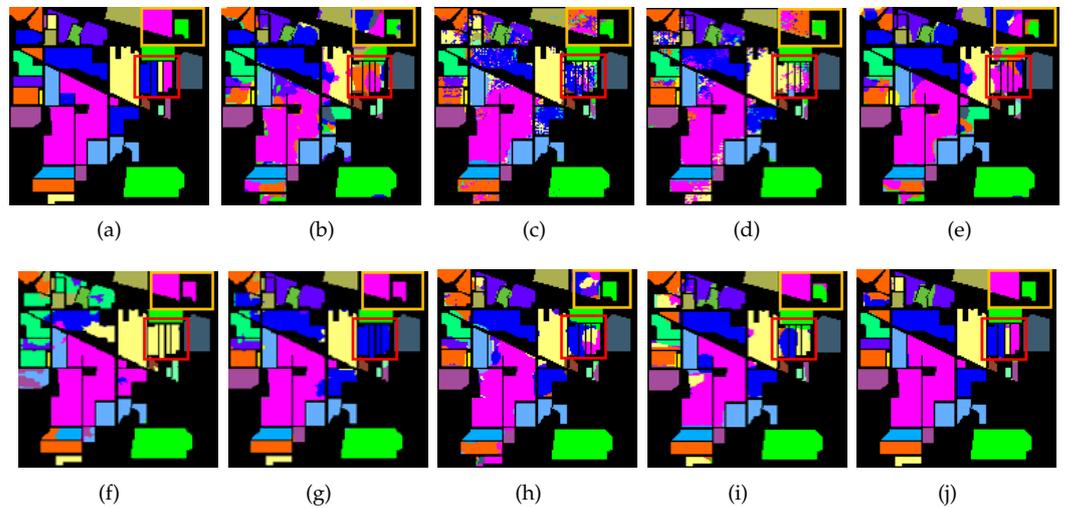


Figure 11. GT and predicted class maps of compared methods on 5-shot IP dataset: (a) GT, (b) SSFTT, (c) S3Net, (d) DM-MRN, (e) FPGA, (f) FSS, (g) STSE-DWLR, (h) DSSPL, (i) rPL, and (j) rPL-AL. The areas with significant effects are marked with yellow and red boxes.

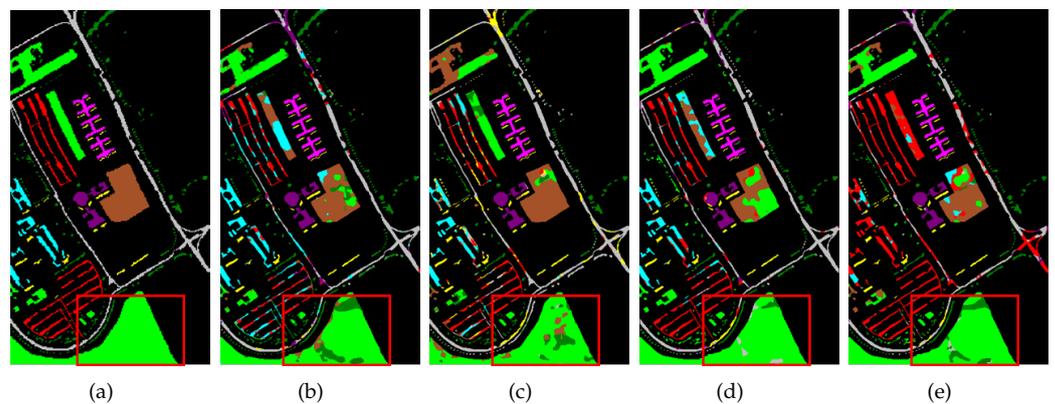


Figure 12. Cont.

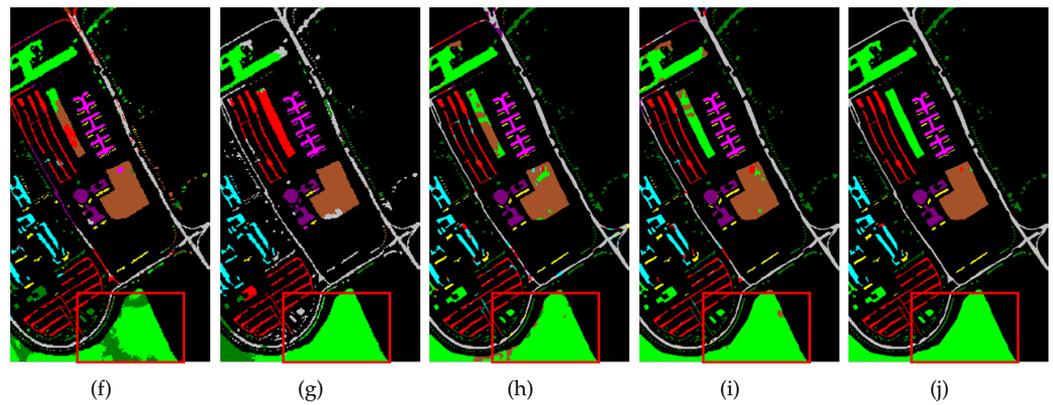


Figure 12. GT and predicted class maps of compared methods on 5-shot PU dataset: (a) GT, (b) SSFTT, (c) S3Net, (d) DM-MRN, (e) FPGA, (f) FSS, (g) STSE-DWLR, (h) DSSPL, (i) rPL, and (j) rPL-AL. The areas with significant effects are marked with red boxes.

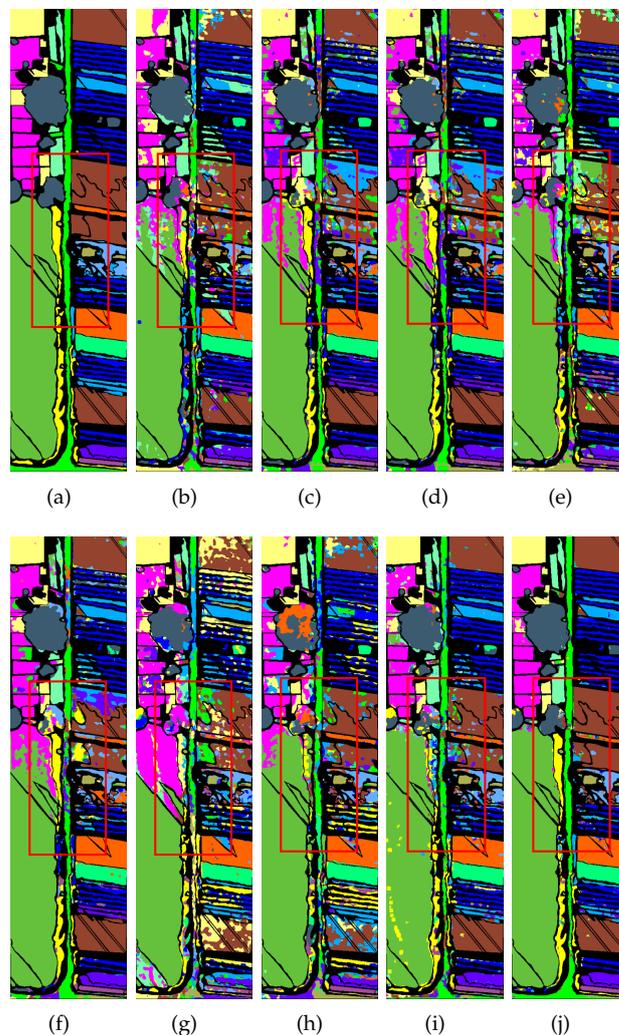


Figure 13. GT and predicted class maps of compared methods on 5-shot HC dataset: (a) GT, (b) SSFTT, (c) S3Net, (d) DM-MRN, (e) MDMC, (f) FPGA, (g) SSRN, (h) A2S2K-ResNet, (i) rPL, and (j) rPL-AL. The areas with significant effects are marked with red boxes.

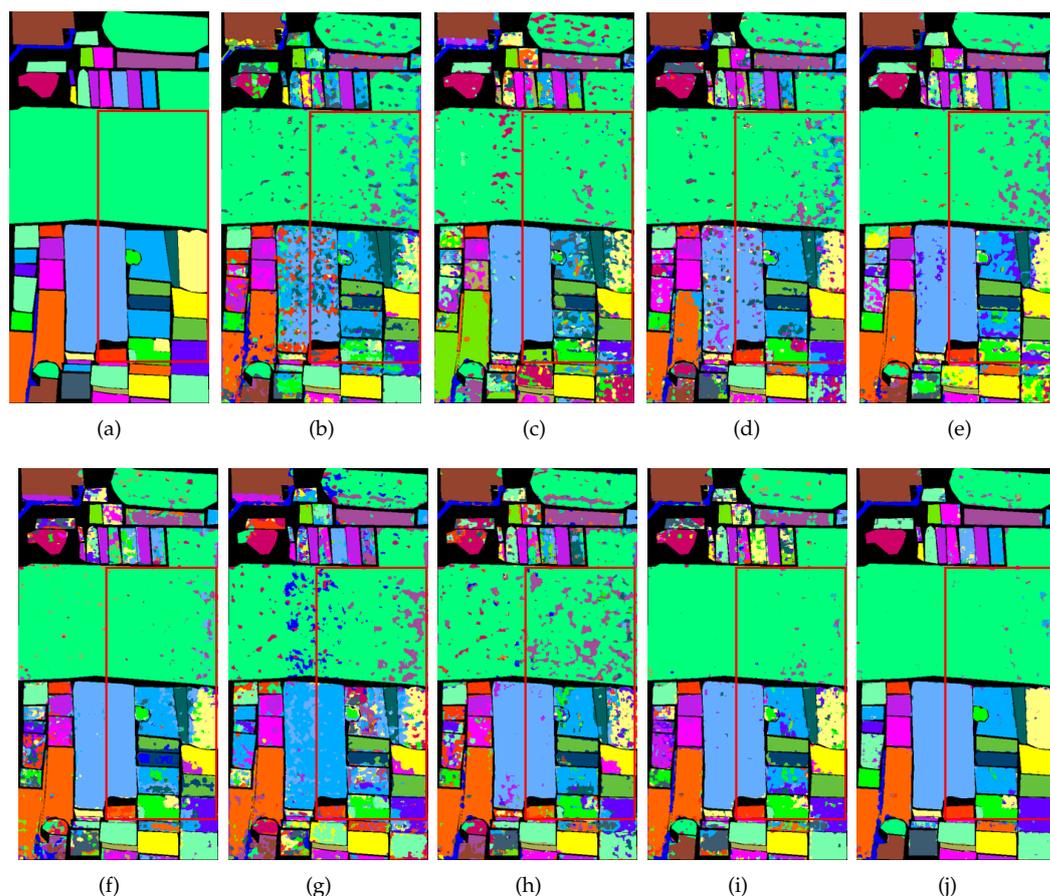


Figure 14. GT and predicted class maps of compared methods on 5-shot HH dataset: (a) GT, (b) SSFTT, (c) S3Net, (d) DM-MRN, (e) MDMC, (f) FPGA, (g) SSRN, (h) A2S2K-ResNet, (i) rPL, and (j) rPL-AL. The areas with significant effects are marked with red boxes.

4.4.1. IP Dataset Result

In the results of the IP dataset presented in Table 1, our rPL method achieves percentages of 86.25%, 91.35%, and 84.3% for three evaluation metrics, respectively. These results outperform those of the compared FS-supervised and semi-supervised methods. Specifically, our method demonstrates improvements of 0.4%, 4.97%, and 1.18% compared to the second-best performing method, DSSPL. Furthermore, our method achieves the highest accuracy in 11 out of the 16 land cover categories, highlighting its strong performance across a wide range of specific ground types.

On the other hand, the performance of the other compared methods display less stability when dealing with different categories. Classic SSFTT and FPGA methods struggle to accurately delineate the boundaries between features in inter-class samples, especially in cases where HSIs exhibit high spectral similarity. Graph-based models, which typically employ the superpixel approach to represent HSIs as weighted graphs, demonstrate improved performance compared to the classic methods. However, they are still influenced by the quality of the coarse superpixel segmentation, particularly in scenarios where both spatial adjacency and high spectral similarity coexist. For instance, in the categories “corn-note” and “corn-mint”, these models tend to produce subpar results. In contrast, our models notably excel in these challenging areas, while the comparison methods exhibit a significant “salt and pepper” artifact, emphasizing the effectiveness of our approach in managing intricate feature distributions.

In particular, when examining the local details framed by the yellow and red boxes, we observe multiple feature classes within a small area, posing a challenge to the model’s ability to handle such local intricacies. The classification outcomes of SSFTT, S3Net, DM-MRN,

FPGA, and DSSPL methods in these regions often exhibit confusion and a considerable number of misclassifications. Additionally, both FSS and STSE-DWLR methods tend to misclassify multiple features from different classes as the same class within this region. In contrast, our rPL and rPL-AL methods demonstrate better performance in handling such details within these regions. By leveraging the spatial-spectral consistency, our semi-supervised strategy effectively improves the reliability of pseudo-labeled samples during training.

Furthermore, by incorporating limited expert annotations so that the amount of labeled samples does not exceed 10 in each round, our rPL-AL model achieves respective increases of 12.9%, 7.22%, and 15.1% in terms of OA, AA, and Kappa metrics compared to the original rPL method. The integration of AL techniques has proven instrumental in effectively utilizing the unlabeled samples and extracting valuable information from the remaining regions. This strategic utilization of AL has resulted in significant improvements in performance.

4.4.2. PU Dataset Result

Compared to the IP dataset, the spatial distribution of land features in the PU dataset exhibits more irregular patterns, and the sample distribution is more dispersed. Despite this challenge, our method demonstrated excellent classification performance, emphasizing the generalization capability of our training strategy. As shown in Table 2, our rPL method achieves over 90% accuracy in terms of OA, AA, and Kappa metrics, surpassing the second-best method by 7.21%, 6.99%, and 8.42%, respectively. Our rPL also achieves the best performance in six out of nine categories, demonstrating its effectiveness in diverse land cover types. Furthermore, the incorporation of AL techniques significantly enhances the performance of our model. The AL-aided rPL model achieves gains of 7.01%, 6.98%, and 9.35% in OA, AA, and Kappa coefficients, respectively. This advancement results from the integration of semi-supervised pseudo-labeling and active learning, enhancing the model's adaptability across various spatial distributions of ground materials, especially in dispersed cases.

4.4.3. HC Dataset Result

The HC dataset possesses a larger sample size and a higher annotation rate compared to the IP and PU datasets, posing new challenges for classification. As shown in Table 3, our rPL model demonstrates improvements of 8.12%, 5.87%, and 9.54% compared to the second-best performing method, FPGA. Furthermore, it achieves the best performance in 10 out of 16 categories compared to other methods. Additionally, the AL-assisted rPL model achieves improvements of 8.71%, 6.21%, and 11.85% for three metrics. Moreover, integrating semi-supervised pseudo-labeling and active learning enhances the model's adaptability to the spatial distribution of samples. Our proposed strategy remains effective for datasets with both dense and sparse sample distributions.

4.4.4. HH Dataset Result

Like the HC dataset, the HH dataset contains more diverse annotated samples and categories, as depicted in Table 4. Both our rPL and rPL-AL models achieve the highest classification performance. Specifically, the rPL method outperforms the second-best method by 6.21%, 11.5%, and 9.92% across three metrics. Additionally, our rPL demonstrates the highest performance in 14 out of 22 categories, showcasing its effectiveness across diverse land cover types. Furthermore, the introduction of our proposed AL strategy leads to a further improvement in overall experimental performance by 8.0%, 11.1%, and 10.47%. These results demonstrate that our model maintains consistent performance with datasets containing a large amount of samples.

Moreover, we also compared the running time with other methods, and, as shown in Tables 1–4, the rPL and rPL AL show relative advantages. Compared to rPL, the running time of rPL AL slightly increases, indicating that incorporating AL into deep learning

frameworks has a negligible impact on running time. We emphasize that our method achieves an appropriate balance between effectiveness and efficiency through classification accuracy and operational complexity.

4.5. Hyperparameter Sensitivity

4.5.1. Convergence Analysis

Figure 15a illustrates the convergence curves of our proposed rPL and rPL-AL methods across all datasets used in this study. The x-axis represents the number of rounds, while the y-axis indicates accuracy values from 60 to 100. These curves portray the overall accuracy per round. A consistent trend spans all four datasets: initial increases in accuracy metrics followed by stabilization, signifying model convergence. Our models converge in a relatively brief span, typically within six to eight rounds, emphasizing the efficiency of rPL and rPL-AL in achieving stable, high-quality classification results swiftly. Additionally, Figure 15b depicts the loss during the semi-supervised training process, each round consisting of 50 epochs. This illustrates that our model converges after each round, with subsequent rounds exhibiting a gradual decrease in training loss.

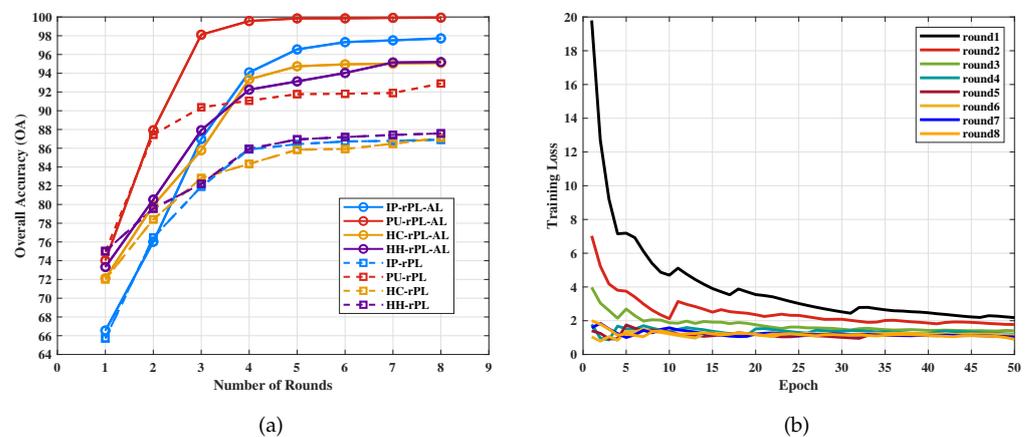


Figure 15. (a) Convergence of the proposed rPL and rPL-AL on four datasets with 5-shot setting. (b) Training loss per round on IP dataset.

4.5.2. Confidence Threshold τ

The selection of the threshold τ in the experiment played a crucial role in determining the initial screening rule for pseudo-labeling. To investigate the impact of different threshold settings on the experimental results, we considered τ values of 0.1, 0.4, 0.6, and 0.9, respectively. Figure 16a displays the classification performance corresponding to each threshold on the IP dataset. More comprehensive experiments on other datasets showed that both low (0.1) and high thresholds (0.9) lead to lower classification performance compared to the selected threshold of 0.6. These three metrics exhibit a pattern of initially increasing and then decreasing as the threshold increases. Therefore, we empirically chose the suitable threshold for pseudo-labeling in order to strike a balance between including enough confident pseudo-labeled samples for training and avoiding the inclusion of samples with uncertain labels that could negatively impact the classification performance.

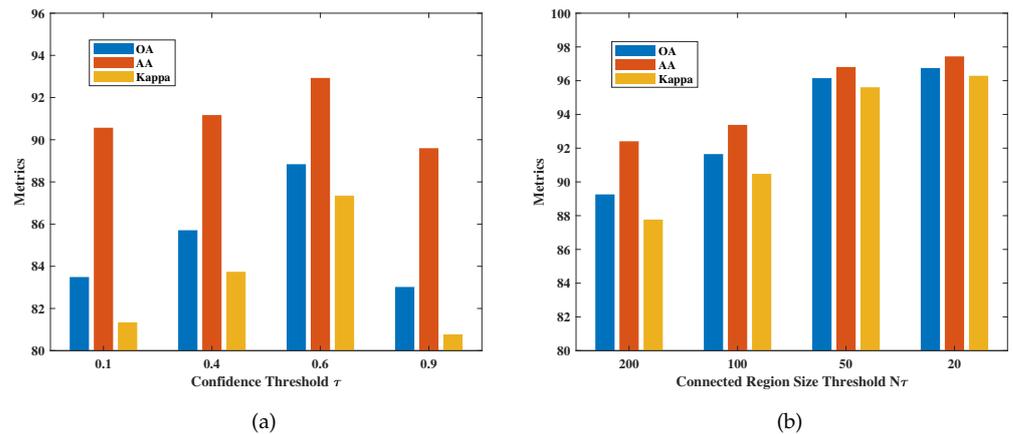


Figure 16. (a) Metrics with different confidence thresholds τ on 5-shot IP dataset. (b) Metrics with different region sizes N_τ on 5-shot IP dataset.

4.5.3. Connected Region Size Threshold N_τ

The decision of whether pixels within a connected region are “discarded” during AL is based on the size of the connected region, denoted as N_τ . To thoroughly investigate the influence of N_τ , we conducted a series of experiments with four datasets where N_τ was fixed for the whole training phase with varying values. Additionally, we explored the dynamic setting where N_τ was nearly halved every two rounds.

As depicted in Figure 16b, decreasing the value of N_τ results in an upward trend in performance, albeit with a longer convergence time. Our dynamic strategy achieves the best overall performance and efficiency. In the initial rounds of training, where instability is more prominent, we opted to reserve a small number of regions to ensure the reliability of the pseudo-labeled samples for rapid convergence. As the training progressed, we gradually relaxed the selection criteria for the regions, adapting to the changing dynamics of the learning process.

4.6. Ablation Study

4.6.1. Pseudo-Label Selection Strategy

To validate the effectiveness of our pseudo-label selection strategy, we conducted a comprehensive comparison with several alternative approaches. These include a 5-shot fully supervised baseline, selecting only eight spatially adjacent (EA) samples, randomly selecting a certain number of samples above a threshold (including 200, 1000, and all), and our rPL model. The classification results after one training iteration are presented in Figure 17a. Our proposed selection strategy showcases a substantial improvement over the alternative approaches. This outcome emphasizes the significance of leveraging the spatial-spectral consistency of HSI pixels to identify reliable samples. In addition, we also conducted experiments on the remaining two datasets, and our rPL consistently achieved the best results. This robust performance strongly supports the effectiveness of our proposed pseudo-labeled sample selection strategy.

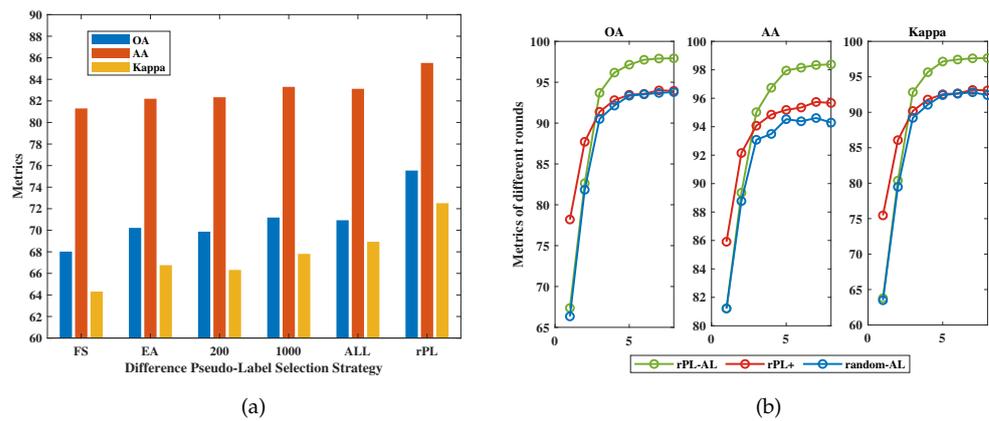


Figure 17. (a) Metrics with different pseudo-label selection strategies on 5-shot IP dataset. (b) Metrics with different AL strategies and rPL+ with similar amount of labeled samples on IP dataset.

4.6.2. AL Strategy

It is evident that rPL-AL demonstrates a significant improvement in classification accuracy compared to rPL. To validate the effectiveness of our AL strategy, we conducted a comparison with two alternative approaches. First, during the final round of training, we observed the classification performance of rPL with more labeled samples (approximated with rPL-AL), noted as rPL+. For rPL+, we used 10 samples per class for semi-supervised training. Secondly, for rPL-AL, we selected five samples per class as initial samples, and the number of labeled samples did not exceed ten in each round of AL. Finally, we replaced our AL strategy with randomly selected samples for expert annotation, without considering the connected regions, referred to as random-AL.

From Figure 17b, it is clear that rPL-AL outperforms rPL+, highlighting that the advantage of AL extends beyond simply increasing the number of labeled samples during pseudo-labeling. Furthermore, when compared to random-AL, our results emphasize the importance of carefully selecting suitable samples for expert annotation. This approach enables the effective utilization of unlabeled samples while minimizing the labeling cost. In addition, we conducted experiments on the remaining datasets. While rPL+ and random-rPL methods independently exhibit commendable results, rPL-AL consistently demonstrates the overall best performance across all four datasets. The superior performance of rPL-AL confirms the significance of our AL strategy in maximizing the potential of unlabeled samples while optimizing the classification process.

5. Conclusions

In this paper, we present a novel approach for FS HSI classification. The proposed method addresses the challenges of limited training samples and unreliable pseudo-label propagation by leveraging the spatial–spectral consistency of HSI pixel samples. By considering the spatially aligned ground materials as connected regions with the same semantic label and similar spectrum, we enhanced our hybrid classification framework by selecting confident pseudo-labeled samples. These samples were assigned pseudo-labels and used in the iterative process to improve classification accuracy. Furthermore, we introduced an active learning strategy to maximize the utilization of unlabeled samples. This strategy identifies the least confident sample within a region and requests expert annotation. The agreement between the predicted and expert labels is used to determine the inclusion of the region or the individual sample in the next iteration. The effectiveness of our approach was demonstrated through experimental results on benchmark datasets. The proposed method achieved state-of-the-art performance even with extremely limited labeled samples. Moreover, the active learning approach enhanced classification accuracy by involving a minimal amount of additional annotation. In the future, we will continue to explore HSI tasks in FS scenarios, mainly focusing on domain adaptation, and multi-source information fusion.

Author Contributions: Conceptualization, J.Z. (Jianguo Zhao) and J.Z. (Junjie Zhang); methodology, J.Z. (Jianguo Zhao) and J.Z. (Junjie Zhang); software, J.Z. (Jianguo Zhao) and J.Z. (Jian Zhang); validation, J.Z. (Jianguo Zhao) and J.Z. (Jian Zhang); formal analysis, J.Z. (Jianguo Zhao) and J.Z. (Junjie Zhang); resources, J.Z. (Jianguo Zhao) and J.Z. (Junjie Zhang); writing—original draft, J.Z. (Jian Zhang); writing, review and editing, H.H. and J.Z. (Junjie Zhang); supervision, J.Z. (Jian Zhang); project administration, H.H.; funding acquisition, J.Z. (Junjie Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 62202283).

Data Availability Statement: All the datasets are available at following URLs: https://www.ehu.es/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (accessed on 10 October 2023) and http://rsidea.whu.edu.cn/resource_WHUHi_sharing.htm (accessed on 15 October 2023), respectively.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaudhuri, B.B. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 277–281. [CrossRef]
- Santara, A.; Mani, K.; Hatwar, P.; Singh, A.; Garg, A.; Padia, K.; Mitra, P. BASS Net: Band-Adaptive Spectral-Spatial Feature Learning Neural Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5293–5301. [CrossRef]
- Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5518615. [CrossRef]
- Sun, H.; Zheng, X.; Lu, X.; Wu, S. Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3232–3245. [CrossRef]
- Sun, H.; Zheng, X.; Lu, X. A Supervised Segmentation Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 2810–2825. [CrossRef]
- Nalepa, J.; Myller, M.; Kawulok, M. Hyperspectral Data Augmentation. *arXiv* **2019**, arXiv:1903.05580.
- Li, W.; Chen, C.; Zhang, M.; Li, H.; Du, Q. Data Augmentation for Hyperspectral Image Classification with Deep CNN. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 593–597. [CrossRef]
- Haut, J.M.; Paoletti, M.E.; Plaza, J.; Plaza, A.; Li, J. Hyperspectral Image Classification Using Random Occlusion Data Augmentation. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1751–1755. [CrossRef]
- Li, W.; Wu, G.; Zhang, F.; Du, Q. Hyperspectral Image Classification Using Deep Pixel-Pair Features. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 844–853. [CrossRef]
- Zhang, C.; Yue, J.; Qin, Q. Global Prototypical Network for Few-Shot Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4748–4759. [CrossRef]
- Li, Z.; Liu, M.; Chen, Y.; Xu, Y.; Li, W.; Du, Q. Deep Cross-Domain Few-Shot Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5501618. [CrossRef]
- Zhao, L.; Luo, W.; Liao, Q.; Chen, S.; Wu, J. Hyperspectral Image Classification with Contrastive Self-Supervised Learning Under Limited Labeled Samples. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6008205. [CrossRef]
- Feng, J.; Zhao, N.; Shang, R.; Zhang, X.; Jiao, L. Self-Supervised Divide-and-Conquer Generative Adversarial Network for Classification of Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5536517. [CrossRef]
- Li, X.; Cao, Z.; Zhao, L.; Jiang, J. ALPN: Active-Learning-Based Prototypical Network for Few-Shot Hyperspectral Imagery Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5508305. [CrossRef]
- Ding, C.; Zheng, M.; Chen, F.; Zhang, Y.; Zhuang, X.; Fan, E.; Wen, D.; Zhang, L.; Wei, W.; Zhang, Y. Hyperspectral Image Classification Promotion Using Clustering Inspired Active Learning. *Remote Sens.* **2022**, *14*, 596. [CrossRef]
- Zhang, J.; Liu, L.; Zhao, R.; Shi, Z. A Bayesian Meta-Learning-Based Method for Few-Shot Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5500613. [CrossRef]
- Tang, H.; Li, Y.; Han, X.; Huang, Q.; Xie, W. A Spatial–Spectral Prototypical Network for Hyperspectral Remote Sensing Image. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 167–171. [CrossRef]
- Seydgar, M.; Rahnamayan, S.; Ghamisi, P.; Bidgoli, A.A. Semisupervised Hyperspectral Image Classification Using a Probabilistic Pseudo-Label Generation Framework. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5535218. [CrossRef]
- Yao, W.; Lian, C.; Bruzzone, L. ClusterCNN: Clustering-Based Feature Learning for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1991–1995. [CrossRef]
- Zhang, L.; Xu, J.; Zhang, J.; Gong, Y. Information enhancement for travelogues via a hybrid clustering model. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, ACT, Australia, 10–13 December 2018; pp. 1–8.
- Zhong, S.; Zhou, T.; Wan, S.; Yang, J.; Gong, C. Dynamic Spectral–Spatial Poisson Learning for Hyperspectral Image Classification With Extremely Scarce Labels. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5517615. [CrossRef]

22. Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.F.; Emery, W.J. Active Learning Methods for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2218–2232. [[CrossRef](#)]
23. Li, J.; Bioucas-Dias, J.M.; Plaza, A. Supervised hyperspectral image segmentation using active learning. In Proceedings of the 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, Reykjavik, Iceland, 14–16 June 2010; pp. 1–4. [[CrossRef](#)]
24. Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; Munoz-Mari, J. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 606–617. [[CrossRef](#)]
25. Roy, S.K.; Mondal, R.; Paoletti, M.E.; Haut, J.M.; Plaza, A. Morphological Convolutional Neural Networks for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8689–8702. [[CrossRef](#)]
26. Li, Y.; Zhang, H.; Shen, Q. Spectral–Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sens.* **2017**, *9*, 67. [[CrossRef](#)]
27. Zhou, F.; Hang, R.; Liu, Q.; Yuan, X. Hyperspectral image classification using spectral-spatial LSTMs. *Neurocomputing* **2019**, *328*, 39–47. [[CrossRef](#)]
28. Tan, K.; Zhou, S.; Du, Q. Semisupervised Discriminant Analysis for Hyperspectral Imagery With Block-Sparse Graph. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1765–1769. [[CrossRef](#)]
29. Krishnapuram, B.; Carin, L.; Figueiredo, M.A.T.; Hartemink, A.J. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 957–968. [[CrossRef](#)] [[PubMed](#)]
30. Gong, Z.; Tong, L.; Zhou, J.; Qian, B.; Duan, L.; Xiao, C. Superpixel Spectral–Spatial Feature Fusion Graph Convolution Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5536216. [[CrossRef](#)]
31. Zhang, H.; Zou, J.; Zhang, L. EMS-GCN: An End-to-End Mixhop Superpixel-Based Graph Convolutional Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5526116. [[CrossRef](#)]
32. Wang, W.; Liu, F.; Liao, W.; Xiao, L. Cross-Modal Graph Knowledge Representation and Distillation Learning for Land Cover Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5520318. [[CrossRef](#)]
33. Liu, Q.; Xiao, L.; Yang, J.; Wei, Z. CNN-Enhanced Graph Convolutional Network with Pixel- and Superpixel-Level Feature Fusion for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 8657–8671. [[CrossRef](#)]
34. Ding, C.; Li, Y.; Wen, Y.; Zheng, M.; Zhang, L.; Wei, W.; Zhang, Y. Boosting few-shot hyperspectral image classification using pseudo-label learning. *Remote Sens.* **2021**, *13*, 3539. [[CrossRef](#)]
35. Cui, B.; Cui, J.; Lu, Y.; Guo, N.; Gong, M. A sparse representation-based sample pseudo-labeling method for hyperspectral image classification. *Remote Sens.* **2020**, *12*, 664. [[CrossRef](#)]
36. Yue, J.; Fang, L.; Rahmani, H.; Ghamisi, P. Self-Supervised Learning with Adaptive Distillation for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5501813. [[CrossRef](#)]
37. Tong, X.; Yin, J.; Han, B.; Qv, H. Few-Shot Learning With Attention-Weighted Graph Convolutional Networks For Hyperspectral Image Classification. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1686–1690. [[CrossRef](#)]
38. Li, Y.; Zhang, L.; Wei, W.; Zhang, Y. Deep Self-Supervised Learning for Few-Shot Hyperspectral Image Classification. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September 2020–2 October 2020; pp. 501–504. [[CrossRef](#)]
39. Liu, S.; Luo, H.; Tu, Y.; He, Z.; Li, J. Wide Contextual Residual Network with Active Learning for Remote Sensing Image Classification. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 7145–7148. [[CrossRef](#)]
40. Yang, J.; Qin, J.; Qian, J.; Li, A.; Wang, L. AL-MRIS: An Active Learning-Based Multipath Residual Involution Siamese Network for Few-Shot Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 990. [[CrossRef](#)]
41. Wang, J.; Li, L.; Liu, Y.; Hu, J.; Xiao, X.; Liu, B. AI-TFNet: Active Inference Transfer Convolutional Fusion Network for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 1292. [[CrossRef](#)]
42. Liu, C.; He, L.; Li, Z.; Li, J. Feature-Driven Active Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 341–354. [[CrossRef](#)]
43. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [[CrossRef](#)]
44. Song, W.; Li, S.; Fang, L.; Lu, T. Hyperspectral Image Classification with Deep Feature Fusion Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3173–3184. [[CrossRef](#)]
45. Xue, Z.; Zhou, Y.; Du, P. S3Net: Spectral–Spatial Siamese Network for Few-Shot Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5531219. [[CrossRef](#)]
46. Zeng, J.; Xue, Z.; Zhang, L.; Lan, Q.; Zhang, M. Multistage Relation Network with Dual-Metric for Few-Shot Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5510017. [[CrossRef](#)]
47. Zheng, Z.; Zhong, Y.; Ma, A.; Zhang, L. FPGA: Fast Patch-Free Global Learning Framework for Fully End-to-End Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5612–5626. [[CrossRef](#)]
48. Zhong, S.; Chen, S.; Chang, C.I.; Zhang, Y. Fusion of Spectral–Spatial Classifiers for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5008–5027. [[CrossRef](#)]

49. Zheng, C.; Wang, N.; Cui, J. Hyperspectral Image Classification with Small Training Sample Size Using Superpixel-Guided Training Sample Enlargement. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7307–7316. [[CrossRef](#)]
50. Hu, L.; He, W.; Zhang, L.; Zhang, H. Cross-Domain Meta-Learning Under Dual-Adjustment Mode for Few-Shot Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5526416. [[CrossRef](#)]
51. Zhong, Z.; Li, J.; Luo, Z.; Chapman, M. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 847–858. [[CrossRef](#)]
52. Roy, S.K.; Manna, S.; Song, T.; Bruzzone, L. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7831–7843. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.