Prototype-Guided Class-Balanced Active Domain Adaptation for Hyperspectral Image Classification

Haiyang Luo, Shengwei Zhong, *Member, IEEE*, and Chen Gong, *Senior Member, IEEE*

***Abstract*— The high cost of data annotation has become a major factor restricting the hyperspectral image (HSI) classification task. To address this issue, domain adaptation (DA) techniques have been developed to adapt models trained on abundantly labeled HSIs to those with scarce labels. As a novel DA paradigm, active DA (ADA) seeks to selectively annotate informative examples using active learning (AL) techniques under domain shift scenarios, ultimately enhancing model adaptation performance. However, current ADA methods require annotating a relatively large number of target examples, which is impractical for HSIs. In addition, the target HSIs suffer from class imbalance, which limits the adaptation performance. To address the above issues, this article proposes a prototype-guided class-balanced ADA (PCADA) method for HSI classification. PCADA alternately aligns the distributions between domains through prototype guidance and selects the most valuable target examples for annotation. Specifically, a prototype-guided domain alignment (PGDA) module is introduced, which generates target prototypes based on highly confident pseudolabels and aligns the distributions of two domains. The inconsistency-aware example selection (IES) module identifies target-specific examples and selects the most valuable ones for annotation. Furthermore, we propose a class-balanced self-training (CBST) module that generates pseudolabels with balanced class distribution to solve the class imbalance issue in the target domain. The experimental results conducted on multiple benchmark HSI datasets demonstrate the superior performance of our proposed method. The code will be available at: https://github.com/Leap-luohaiyang/PCADA-2025**

***Index Terms*— Active domain adaptation (ADA), class imbalance, hyperspectral image (HSI) classification.**

I. Introduction

HYPERSPECTRAL images (HSIs) capture reflectance across hundreds of narrow and contiguous spectral bands. The rich spatial–spectral information have led to the widespread application in various fields, including environmental monitoring [1], mineral exploration [2], agricultural remote sensing [3], [4], and military surveillance [5], [6]. HSI classification (HSIC) is a crucial task, which aims to assign a specific category to each pixel in the HSI. Early HSIC techniques primarily relied on traditional machine learning methods, such as k-nearest neighbors [7] and support vector machines [8]. These approaches utilized only the spectral information of HSIs while neglecting their spatial context. To overcome this limitation, spatial–spectral joint features-based methods [9], [10] significantly improved the classification accuracy by incorporating the spatial information of HSIs. Due to the ability to effectively extract spatial–spectral features from HSIs and handle fine image classification tasks, deep learning-based methods [11], [12], [13], [14] have gradually become mainstream in the field of HSIC. However, deep learning-based methods usually rely on a large amount of labeled data for supervision, which is costly. Due to the complexity and high dimensionality of hyperspectral data, annotating pixels is both time-consuming and labor-intensive. Domain adaptation (DA) [15], [16] offers a promising solution by transferring knowledge from a labeled dataset (i.e., source domain) to an unlabeled or rarely labeled but related dataset (i.e., target domain). DA enhances the generalization ability of the model across varying environments and sensor conditions. Most studies focus on unsupervised DA (UDA), where no labels in the target domain are available [17], [18].

The main idea of UDA is to reduce the domain gap between source and target domains, enabling models trained on the source domain to perform well on the target domain. In recent years, many deep UDA methods have achieved promising results. However, their performance still significantly falls short of supervised learning methods. In practical scenarios, it is often feasible to annotate a small number of target examples to enhance the model performance. However, the effectiveness of performance improvement largely depends on which target examples are annotated. Thus, selecting the most valuable examples from the target domain within a limited annotation budget has become an increasingly critical issue.

This consideration naturally leads to the exploration of active learning (AL) [19], [20] techniques. AL improves the model performance by strategically selecting the most informative examples for annotation. The core idea is to reduce the need for extensive labeled data by focusing on examples that offer the greatest contribution to the learning process. However, traditional AL methods typically assume that labeled and unlabeled data share the same distribution. The sampling strategies are not applicable for UDA scenarios with domain shifts. Merely using AL and fine-tuning under domain shift

Received 20 March 2025; revised 28 April 2025; accepted 23 May 2025. Date of publication 4 June 2025; date of current version 16 June 2025. This work was supported in part by the National Natural Science Foundation (NSF) of China under Grant 62101261, Grant 62336003, and Grant 12371510; and in part by the Fundamental Research Funds for the Central Universities under Grant 30923011027. *(Corresponding authors: Shengwei Zhong; Chen Gong.)*

The authors are with the PCA Laboratory, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, Ministry of Education, Jiangsu Key Laboratory of Image and Video Understanding for Social Security, and the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: haiyang.luo@njust.edu.cn; zhongsw_91@foxmail.com; chen.gong@njust.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3576654

results in suboptimal performance [21]. In this article, a new paradigm known as active DA (ADA) [22] has emerged, which actively selects target examples for annotation in the presence of domain shift to assist in DA.

The key aspects of ADA involve developing efficient example selection strategies to select the most valuable target examples. The selected target examples should maximize the transfer performance of the model after annotation. In addition, it emphasizes effectively utilizing the limited annotated data to enhance adaptation. However, current ADA methods typically require annotations for approximately 10% of target examples [23], [24], [25], and this requirement greatly exceeds the annotation budget for HSIs. Furthermore, most existing ADA methods overlook the class distribution discrepancies between domains. The class distribution of target data is usually imbalanced in HSIs, where the ratio of majority class to minority class can reach thousands [26]. Models trained on such imbalanced data tend to favor the majority class, resulting in the suboptimal classification performance.

To address the aforementioned issues, we propose a prototype-guided class-balanced ADA (PCADA) framework for HSIC. As described in [27], target examples can generally be divided into two categories: source-like examples, whose feature distributions are similar to source data, and target-specific examples, which exhibit unique characteristics of the target domain. Allocating the annotation budget to target-specific examples is more beneficial for the model generalizing to the target domain than annotating source-like examples. With the consideration above, we design a pseudolabel selection strategy for source-like examples. This enables the generation of target prototypes and facilitates initial domain alignment, guided by the interaction between source and target prototypes. Afterward, target-specific examples are manually annotated.

In the example selection stage, we identify target-specific examples as those where the classifier predicted labels are inconsistent with the labels of the nearest target prototypes. In addition, the selected examples which maximize uncertainty, diversity, and representativeness are encouraged. Diversity is measured using the combination of the classifier predicted labels and the labels of the nearest target prototypes to ensure the variety of "label pairs." The uncertainty of each unlabeled target example under the current model is quantified by the predicted probabilities of the classifier. Given the limited annotation budget, we maintain a candidate set comprising the most frequent label combinations and select examples from each combination that are both uncertain and representative for annotation. Considering that a small number of annotated examples often fail to capture the overall distribution of target data, and the imbalanced class distribution in the target domain, inspired by [28] and [29], we propose the class-balanced self-training (CBST) module. It jointly estimates the class distribution in the target domain using the labels of annotated examples and the predictions of the classifier, and samples pseudolabeled examples following a class-balanced rule for self-training.

The main contributions of this article are summarized as follows.

1) We propose a novel ADA method for HSIC, named PCADA. To the best of the authors' knowledge, this work stands as one of the few works that applies ADA to HSIC while attempting to address the class imbalance issue. The training pipeline of PCADA follows a progressive three-stage framework. The first stage performs initial domain alignment using source-like examples and learns good representations of target prototypes. In the second stage, the model is fine-tuned with target-specific examples annotated by the oracle, improving both DA and pseudolabel accuracy for the next stage. The third stage further optimizes the model through a large amount of high-quality pseudolabeled examples.
2) We introduce an innovative example selection strategy that efficiently identifies the most valuable examples under domain shift. This method is suitable for HSIs, where annotation budgets are strictly constrained.
3) We propose a CBST module to tackle the prevalent yet often overlooked issue of class imbalance in the target domain. This module mitigates the model bias toward majority classes while leveraging pseudolabeled target examples to improve the overall adaptation performance.
4) Comprehensive experiments demonstrate that our proposed method achieves state-of-the-art performance on multiple public benchmark datasets.

II. Related Works

A. Unsupervised Domain Adaptation

Existing UDA methods can be broadly categorized into three types: instance-based, classifier-based, and feature-based. Instance-based methods [30], [31] focus on reweighting or selecting source instances to better align with target distribution. These methods often involve techniques such as importance sampling or instance reweighting to reduce domain discrepancy. Classifier-based methods [32], [33] aim to adapt the decision boundary of the classifier, allowing models trained on source data to generalize to the target domain. This can be achieved by adjusting classifier parameters or employing ensemble techniques to enhance robustness across domains.

Compared with the above two methods, feature-based deep UDA methods have received more attention. These methods focus on aligning the distributions of source and target domains in the feature space. Among these, adversarial learning-based UDA methods introduce an adversarial objective to reduce the distribution discrepancy between source and target domains. Specifically, the feature extractor and the domain discriminator engage in an adversarial game and this iterative process ensures that the feature extractor ultimately learns domain-invariant feature representations. Notable representative methods include domain adversarial training of neural networks (DANNs) [34], adversarial discriminative DA (ADDA) [35], and conditional domain adversarial network (CDAN) [36]. Some other methods employ adversarial training between two task-specific classifiers to obtain domain decision boundaries to drive the alignment of class-level distributions between domains, such as confident learning-based DA [37], two-branch attention adversarial DA [38], and unsupervised

joint adversarial DA [39]. These methods have achieved excellent performance in the cross-domain HSIC.

Another popular branch in the field is the metric discrepancy-based UDA methods, which focus on directly optimizing discrepancy metrics to reduce the domain gap. The essence of these methods lies in designing various metric paradigms. Common metric paradigms include maximum mean discrepancy (MMD) [40], correlation alignment (CORAL) [41], and central moment discrepancy (CMD) [42]. Pioneering works, such as deep adaptation network (DAN) [43] and joint adaptation network (JAN) [44], perform domain matching by minimizing variants of MMD, namely, multikernel MMD and joint MMD. Yan et al. [45] introduced class-specific auxiliary weights into the original MMD, proposing a weighted MMD model to address the issue of DA performance degradation caused by class weight bias across domains. Sun and Saenko [46] proposed deep CORAL, which achieves feature alignment by minimizing the covariance difference between source and target features. Some studies have expanded upon these metric discrepancy-based frameworks and applied them to HSIC tasks, such as topological structure and semantic information transfer network [47] and discriminative cooperative alignment [48].

While the above methods primarily rely on single-granularity alignment, our work reduces multigranularity domain discrepancies through two alignment mechanisms. The feature-level domain alignment aims to minimize domain discrepancies in the overall distribution of examples within the same category, whereas task-level domain alignment is designed to optimize domain discrepancies at the individual example level.

B. Active Learning

AL aims to improve the model performance by selecting a small number of informative examples for annotation. Researchers have explored a variety of criteria for developing example selection strategies. For example, uncertainty-based AL methods quantify the uncertainty of model predictions using metrics, such as prediction confidence [49], entropy [50], or margin [51], and select those examples for annotation where the model is most uncertain. The core objective is to obtain a more precise decision boundary. Representative-based AL methods aim to select examples that represent the distribution of unlabeled data, based on the intuition that once the selected representative examples are annotated, they can serve as effective substitutes for unlabeled data. Typical methods include clustering [52] and CoreSet [53]. Diversity-based AL methods tend to select examples that are different from the former-selected ones, ensuring that the selected examples cover different areas of the feature space and capture a comprehensive representation of the data. Current AL methods typically employ combination strategies based on multiple criteria, which aims to achieve a tradeoff between various criteria when selecting examples [54], [55].

However, the above methods are not well-suited for DA scenarios, as they fail to consider the domain shift between labeled and unlabeled data. In our work, we introduce a novel AL strategy to identify valuable examples under domain shift and present a simple yet effective label combination mechanism for selecting diverse and representative examples.

C. Active Domain Adaptation

Unlike semi-supervised DA [56], [57], which passively relies on randomly annotated target instances, ADA actively selects the most informative examples using AL strategies to maximize annotation resource efficiency. ADA was first introduced by Rai et al. [22] for sentiment analysis in text data. Its first application in visual tasks, active adversarial DA (AADA) [21], combines the predictive entropy of the classifier with the output of the domain discriminator to select uncertain target examples that are distanced from the source domain. Submodular subset selection for virtual adversarial active DA (S^3VAADA) [23] introduces a novel submodule criterion for selecting the most informative example subset rather than individual example, thus avoiding redundancy. Clustering uncertainty-weighted embeddings (CLUEs) [58] performs entropy-weighted k-means clustering to select target examples that are both uncertain and diverse. Unlike methods that primarily focus on example uncertainty and diversity, select-by-distinctive-margin (SDM) [59] explores target examples similar to potential hard source examples via a maximum margin loss function. Energy-based ADA (EADA) [60] demonstrates the feasibility of energy models in ADA and selects examples based on energy values. Divide-and-adapt (DiaNA) [27] divides target data into subsets with different levels of transferable attributes and applies customized learning strategies to each subset. Easy-to-hard DA with human interaction (IEH-DA) [61] incorporates the idea of curriculum learning and is the first work to apply AL to cross-domain HSIC task. Nevertheless, the general ADA methods are not applicable to HSIC due to the need to annotate a large number of target examples, which motivates our method. Furthermore, current ADA methods often overlook the issue of class imbalance in the target domain, which is particularly prevalent in HSIs. Our proposed PCADA aims to address these challenges.

III. Proposed Method

In ADA, we have access to a labeled source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labeled examples, where $\mathbf{x}_i^s$ represents the ith source domain training example and $y_i^s \in \{1, 2, \ldots, C\}$ denotes the associated labels, with C being the number of class types, and an unlabeled target domain $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$ with n_t unlabeled examples. To facilitate the AL process, we define two sets within the target domain: $\mathcal{D}_{lt} = \{(\mathbf{x}_i^{lt}, y_i^{lt})\}_{i=1}^{n_{lt}}$ stores the target examples labeled during the example selection process, while $\mathcal{D}_{ut} = \{\mathbf{x}_i^{ut}\}_{i=1}^{n_{ut}}$ contains the remaining unlabeled target examples. Initially, $\mathcal{D}_{lt}$ is an empty set $\emptyset$, and $\mathcal{D}_{ut}$ is equal to $\mathcal{D}_t$. Given a fixed annotation budget B, the example selection is conducted over R rounds. In each round, $b = B/R$ examples are selected from $\mathcal{D}_{ut}$ and labeled by an oracle. These examples are then removed from $\mathcal{D}_{ut}$ and added to $\mathcal{D}_{lt}$. Once updated, $\mathcal{D}_s \cup \mathcal{D}_{lt}$ is utilized in training. The updated

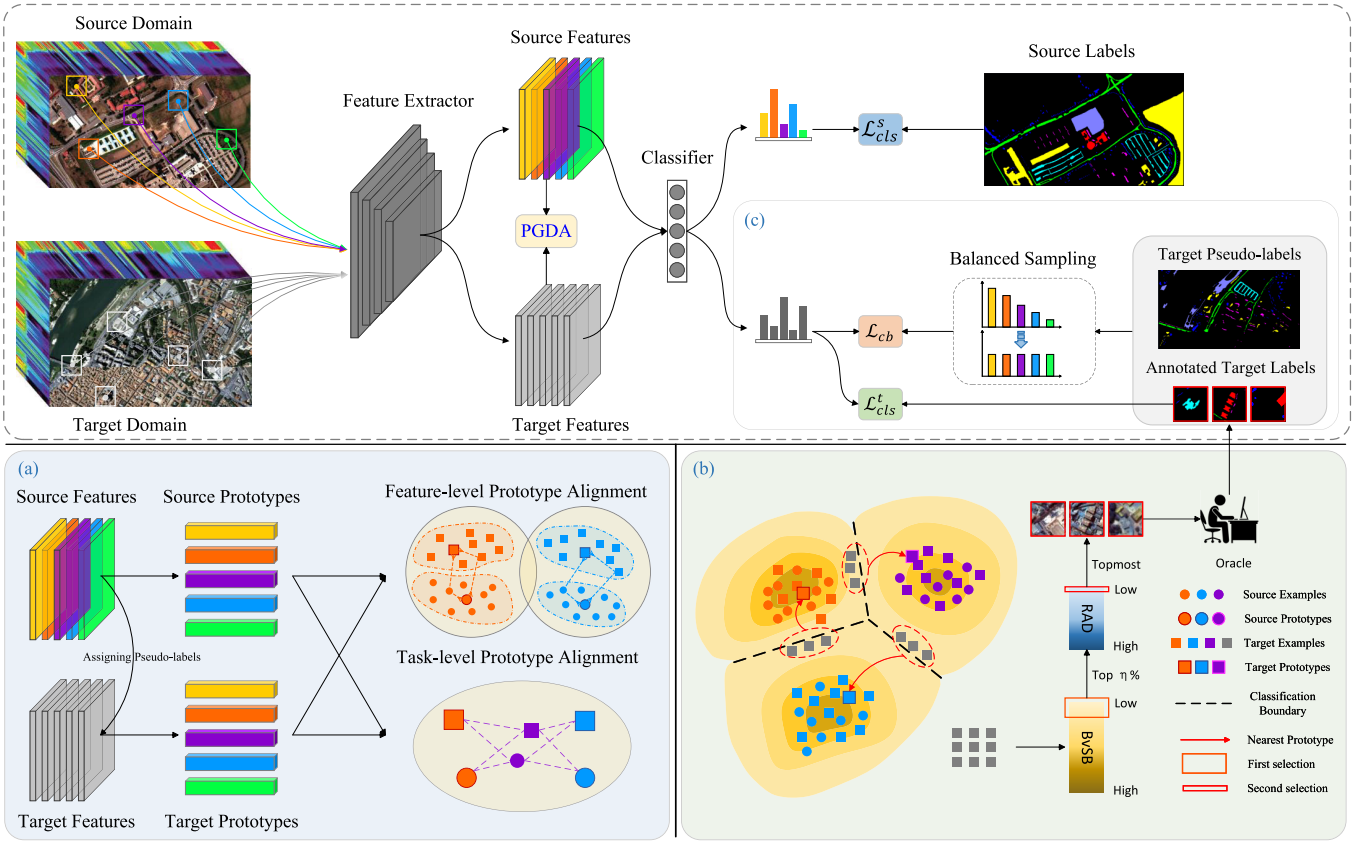

Fig. 1. Framework of the proposed PCADA method. (a) PGDA: prototype alignment across domains is performed at both the feature level and the task level to reduce multigranular domain discrepancies. (b) IES: utilizing the disparity between the nearest target prototype and the classifier prediction, the most representative examples are chosen for annotation from the diverse and uncertain target examples. (c) CBST: by iteratively retraining with the generated pseudolabels using class-balanced sampling, the class imbalance problem in the target domain is effectively alleviated.

model is subsequently employed in the next round to identify new target examples for annotation, iteratively improving the model performance.

As illustrated in Fig. 1, our proposed PCADA method consists of three parts: prototype-guided domain alignment (PGDA) module, inconsistency-aware example selection (IES) module, and CBST module. The PGDA module employs two alignment mechanisms to mitigate multigranular domain discrepancies, ensuring robust distribution alignment between domains. The IES module selects uncertain and diverse target examples based on the inconsistency between the prototype-based labels and the classifier predicted labels. The CBST module generates pseudolabels with balanced class distribution by sampling varying proportions of examples from each class, which addresses class imbalance within the target domain. Finally, we summarize the entire algorithm process.

A. Backbone Network

The backbone network consists of a feature extractor $\mathcal{G}$ and a classifier $\mathcal{F}$. The feature extractor $\mathcal{G}$ processes HSI patches from both source and target domains, encoding them into a common feature space. The common feature space is designed to ensure the features are both domain invariant and class discriminative. The classifier $\mathcal{F}$ projects the high-dimensional features generated by $\mathcal{G}$ into the probability space corresponding to each class. It serves for the supervised classification of source examples and for generating pseudolabels for unlabeled target examples.

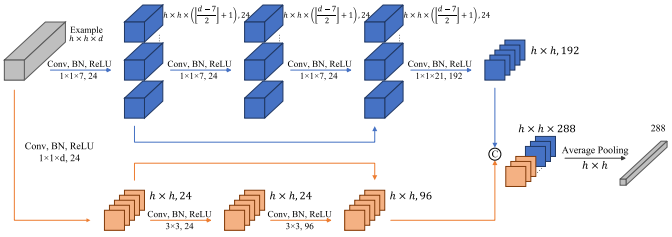

Fig. 2. Network architecture of the feature extractor. The upper network (spectral branch) is used to extract the spectral features of the input HSI patch, and the lower network (spatial branch) is used to extract the spatial features of the input HSI patch. Finally, the spatial and spectral features are concatenated to obtain the spatial–spectral features.

The detailed architecture of the feature extractor $\mathcal{G}$ is shown in Fig. 2. It is built upon the dual-channel residual network (DCRN) architecture [62], which has demonstrated promising performance in a previous cross-domain HSIC task [63]. The entire network mainly consists of three components. The spatial feature extraction branch is designed to extract spatial features from the input data, while the spectral feature extraction branch is dedicated to extract spectral features. The spatial–spectral feature fusion module then concatenates the

outputs of both branches to form the final comprehensive spatial–spectral feature representation. An example utilizing the input data with shape $h \times h \times d$ is employed to illustrate the network architecture, where $h \times h$ represents the patch size and d is the number of bands in HSI.

The classifier $\mathcal{F}$ is a fully connected layer. The backbone network is first trained in a fully supervised manner to accurately classify source examples

$$\mathcal{L}_{\text{cls}}^{s} = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathcal{L}_{\text{ce}}\left(\mathcal{F}\left(\mathcal{G}\left(\mathbf{x}_i^s\right)\right), y_i^s\right) \tag{1}$$

where $\mathcal{L}_{\text{ce}}(\cdot, \cdot)$ is the cross-entropy loss.

B. Prototype-Guided Domain Alignment

The purpose of the PGDA module is to initially align the distributions between source and target domains, guided by the prototypes from both domains. In addition, it aims to learn effective feature representations of target prototypes to facilitate subsequent example selection.

1) Acquisition of Prototypes: Using the labels of source data, the prototype of the cth class in the source domain can be calculated as follows:

$$\boldsymbol{\mu}_c^s = \frac{1}{n_c^s} \sum_{i=1}^{n_c^s} \mathbf{v}_{c_i}^s \tag{2}$$

where n_c^s is the number of source examples belonging to class c and $\mathbf{v}_{c_i}^s$ denotes the l_2-normalized feature of the ith example among all the source examples labeled c.

In order to effectively identify source-like examples and assign high-confidence pseudolabels to them, we design a pseudolabel selection strategy that integrates information from individual examples, prototypes, and classifier predictions. Specifically, the first pseudolabel of each target example can be obtained from the output of the classifier, which can be expressed as follows:

$$\hat{y}_{i,1}^t = \arg \max_{c \in [1,C]} \mathcal{F}\left(\mathcal{G}\left(\mathbf{x}_i^t\right)\right). \tag{3}$$

Subsequently, we allocate $1 + K$ additional possible pseudolabels to each target example. The first pseudolabel is derived from the nearest source prototype, while the remaining K pseudolabels are determined by the K nearest source examples. The consistency of these pseudolabels is utilized to evaluate the confidence of each target example, resulting in

$$\chi_i = \begin{cases} 1, & \hat{y}_{i,1}^t = \hat{y}_{i,j}^t, \quad j = 2, 3, \ldots, (2+K) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\chi_i = 1$ indicates that target example $\mathbf{x}_i^t$ is a source-like target example and its pseudolabel is highly confident. By retaining such examples, we can obtain a set of target examples with high-confidence pseudolabels, denoted by $\hat{\mathcal{D}}_t = \{(\mathbf{x}_i^t, \hat{y}_i^t)\}_{i=1}^{\hat{n}_t}$. Then, the prototype of the cth class in the target domain can be calculated as follows:

$$\boldsymbol{\mu}_c^t = \frac{1}{n_c^t} \sum_{i=1}^{n_c^t} \mathbf{v}_{c_i}^t \tag{5}$$

where n_c^t is the number of target examples with a pseudolabel of c and $\mathbf{v}_{c_i}^t$ denotes the l_2-normalized feature of the ith example among all the target examples pseudolabeled as c.

2) Feature-Level Prototype Alignment: In the feature space, we aim for examples and prototypes of the same class to be closely aligned, regardless of their domain labels. This ensures that the learned features are both domain invariant and class discriminative.

To learn compact feature representations within each class, inspired by [64], we minimize the total distance between each example and the prototypes of its corresponding class in both source and target domains. Formally, the feature-level prototype alignment loss is defined as follows:

$$\mathcal{L}_{\text{fpa}} = \frac{1}{n_s}\left(\sum_{i=1}^{n_s} \left\|\mathbf{v}_i^s - \boldsymbol{\mu}_{y_i^s}^s\right\|_2^2 + \sum_{i=1}^{n_s} \left\|\mathbf{v}_i^s - \boldsymbol{\mu}_{y_i^s}^t\right\|_2^2\right) + \frac{1}{\hat{n}_t}\left(\sum_{j=1}^{\hat{n}_t} \left\|\mathbf{v}_j^t - \boldsymbol{\mu}_{\hat{y}_j^t}^s\right\|_2^2 + \sum_{j=1}^{\hat{n}_t} \left\|\mathbf{v}_j^t - \boldsymbol{\mu}_{\hat{y}_j^t}^t\right\|_2^2\right) \tag{6}$$

where $\mathbf{v}_i^s$ denotes the l_2-normalized feature of the ith source example and $\mathbf{v}_j^t$ denotes the l_2-normalized feature of the jth example among all the pseudolabeled target examples.

3) Task-Level Prototype Alignment: While the feature-level prototype alignment only emphasizes the similarity of feature distributions across domains, it does not consider the relationship between prototypes and example classifications. Even without relying on a parametric classifier, the class of an example can be inferred based on its distance to the prototypes. Specifically, for a given source/target example $\mathbf{x}_i$, the class probability distribution $\mathbf{P}_i^t$ predicted by target prototypes can be obtained by applying the distances between $\mathbf{x}_i$ and the target prototype of each class to the softmax function. The cth element of $\mathbf{P}_i^t$ represents the probability of $\mathbf{x}_i$ belonging to class c

$$\mathbf{P}_{ic}^t = p(y_i = c|\mathbf{x}_i) = \frac{e^{-d\left(\mathbf{v}_i, \boldsymbol{\mu}_c^t\right)}}{\sum_{c'=1}^{C} e^{-d\left(\mathbf{v}_i, \boldsymbol{\mu}_{c'}^t\right)}} \tag{7}$$

where $\mathbf{v}_i$ denotes the l_2-normalized feature of $\mathbf{x}_i$ and $d(\cdot, \cdot)$ represents the Euclidean distance metric. The predicted probability distribution $\mathbf{P}_i^s$ of source prototypes for each example can be obtained in a similar manner.

In addition, we incorporate the novel alignment mechanism proposed in [65], namely, task-level prototype alignment, to minimize the domain discrepancies. This is achieved by constraining the predicted probability distributions of cross-domain prototypes for each example are consistent. The core idea is that when the distributions of two domains are well aligned, the predicted probability distributions of cross-domain prototypes for the same example should be similar.

KL divergence is used to measure the difference between the probability distributions predicted by prototypes from different domains for the same example. The task-level prototype alignment loss is defined as follows:

$$\mathcal{L}_{\text{tpa}} = \frac{1}{n_s} \sum_{\mathbf{x}_i \in \mathcal{D}_s} D_{\text{KL}}\left(\mathbf{P}_i^s, \mathbf{P}_i^t\right) + \frac{1}{n_t} \sum_{\mathbf{x}_j \in \mathcal{D}_t} D_{\text{KL}}\left(\mathbf{P}_j^s, \mathbf{P}_j^t\right) \tag{8}$$

where $D_{\text{KL}}(\cdot, \cdot)$ denotes the symmetric pairwise KL divergence.

C. Inconsistency-Aware Example Selection

During the example selection stage, we aim to select the most informative target examples for annotation. To achieve this, we first assign a prototype-based label to each unlabeled target example according to target prototypes. Subsequently, the target-specific examples in the target domain are identified based on the inconsistency between the classifier predicted labels and the prototype-based labels. Finally, we propose to comprehensively consider the uncertainty, diversity, and representativeness to choose the most valuable examples for annotation.

1) Prototype-Based Label: As introduced in the task-level prototype alignment section of the PGDA module, for an example $\mathbf{x}_i \in \mathcal{D}_{ut}$, its prototype-based classification probability $\mathbf{P}_i^t$ can be obtained through (7). Then, its prototype-based label $\ddot{y}(\mathbf{x}_i)$ can be formulated as follows:

$$\ddot{y}(\mathbf{x}_i) = \arg\max_{c\in[1,C]} \mathbf{P}_i^t. \tag{9}$$

2) Target-Specific Example Identification: We propose to exploit the inconsistency between the classifier predicted labels and the prototype-based labels to identify target-specific examples, the effectiveness of which has been demonstrated in [27]. However, different from using source prototypes in [27], we suggest utilizing the target prototypes obtained in the PGDA module. Since target prototypes contain more target domain-specific information, they can more effectively identify target-specific examples than source prototypes. Formally, the classifier predicted label $\dot{y}(\mathbf{x}_i)$ for an example $\mathbf{x}_i$ can be obtained as described in (3). The set of target-specific examples can be expressed as follows:

$$\mathcal{D}_{\text{ts}} = \left\{\mathbf{x}_i \in \mathcal{D}_{ut} \mid \dot{y}(\mathbf{x}_i) \neq \ddot{y}(\mathbf{x}_i)\right\}. \tag{10}$$

3) Example Selection: For a query example $\mathbf{x} \in \mathcal{D}_{\text{ts}}$, we define the combination of its classifier predicted label and prototype-based label as the label pair for that example, denoted by $(\dot{y}(\mathbf{x}), \ddot{y}(\mathbf{x}))$. Intuitively, selecting examples with different label pairs can better ensure the diversity of selected examples. Furthermore, the more examples a label pair contains, the greater the improvement in the transfer performance of the model attributed to annotating examples belonging to that label pair. Considering the annotation budget in each round, we designate the top b label pairs with the highest example counts as the candidate label pair set $\mathcal{P}_{\text{cd}}$. From $\mathcal{P}_{\text{cd}}$, one example from each label pair is selected for annotation. Specifically, for each label pair in $\mathcal{P}_{\text{cd}}$, we first identify the subset of examples with the highest uncertainty and then select the example best that represents the distribution of that subset for annotation.

We use BvSB [66] to measure the uncertainty of examples, quantifying the difference between the top two predicted class probabilities. Specifically, this criterion is defined as follows:

$$\text{BvsB}(\mathbf{x}) = P_B(\mathbf{x}) - P_{\text{SB}}(\mathbf{x}) \tag{11}$$

where $P_B(\mathbf{x})$ denotes the highest class prediction probability of example $\mathbf{x}$ and $P_{\text{SB}}(\mathbf{x})$ denotes the second-highest class prediction probability of example $\mathbf{x}$. We select the top $\eta\%$ of examples with the lowest BvSB values in each label pair to constitute the uncertain subset for that label pair.

Let $\mathcal{X}_p^u$ be the uncertain subset of examples with label pair p. The example that best represents the data distribution of $\mathcal{X}_p^u$ should minimize the statistical distance between itself and $\mathcal{X}_p^u$. Inspired by [24], we employ the squared MMD [67] as a measure of the statistical distance between the example $\mathbf{x}$ and $\mathcal{X}_p^u$ in the feature space, which is formally given by

$$\begin{aligned}\text{MMD}^2\left(\mathbf{x}, \mathcal{X}_p^u\right) = {} & k(\mathcal{G}(\mathbf{x}), \mathcal{G}(\mathbf{x})) \\ & - \frac{2}{|\mathcal{X}_p^u|}\sum_{\mathbf{x}_i\in\mathcal{X}_p^u} k(\mathcal{G}(\mathbf{x}), \mathcal{G}(\mathbf{x}_i)) \\ & + \frac{1}{|\mathcal{X}_p^u|^2}\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}_p^u} k\left(\mathcal{G}(\mathbf{x}_i), \mathcal{G}\left(\mathbf{x}_j\right)\right)\end{aligned} \tag{12}$$

where $k(\mathbf{x}, \mathbf{x}') = \exp(-(\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2))$ denotes the radial basis function (RBF) kernel. It should be noted that, except for the first round, subsequent rounds may select label pairs that have been selected in previous rounds. To ensure diversity across rounds, the statistical distance between the currently selected example and the set of previously selected examples with the same label pair is maximized. Thus, the representativeness and diversity of the example can be comprehensively measured as follows:

$$\begin{aligned}\text{RAD}\left(\mathbf{x}, \mathcal{X}_p^u, \mathcal{X}_p'\right) &= \text{MMD}^2\left(\mathbf{x}, \mathcal{X}_p^u\right) - \text{MMD}^2\left(\mathbf{x}, \mathcal{X}_p'\right) \\ &= \frac{2}{|\mathcal{X}_p'|}\sum_{\mathbf{x}_i\in\mathcal{X}_p'} k(\mathcal{G}(\mathbf{x}), \mathcal{G}(\mathbf{x}_i)) \\ &\quad - \frac{2}{|\mathcal{X}_p^u|}\sum_{\mathbf{x}_i\in\mathcal{X}_p^u} k(\mathcal{G}(\mathbf{x}), \mathcal{G}(\mathbf{x}_i)) \\ &\quad + \frac{1}{|\mathcal{X}_p^u|^2}\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}_p^u} k\left(\mathcal{G}(\mathbf{x}_i), \mathcal{G}\left(\mathbf{x}_j\right)\right) \\ &\quad - \frac{1}{|\mathcal{X}_p'|^2}\sum_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}_p'} k\left(\mathcal{G}(\mathbf{x}_i), \mathcal{G}\left(\mathbf{x}_j\right)\right)\end{aligned} \tag{13}$$

where $\mathcal{X}_p'$ denotes the set of all examples with label pair p that were selected in previous rounds. When $\mathcal{X}_p' = \emptyset$, we retain only the term $\text{MMD}^2(\mathbf{x}, \mathcal{X}_p^u)$. The overall process of the IES module is described by the pseudocode in Algorithm 1.

D. Class-Balanced Self-Training

Annotating a small number of valuable target examples can effectively assist DA. However, these examples often fail to fully reflect the overall target distribution. In addition, the imbalanced class distribution in the target domain can lead to preferential adaptation toward the majority class. Therefore, in this module, we first estimate the frequency of each class in the target domain. Then, based on the class frequencies and the pseudolabels predicted by the classifier, we perform CBST to fully utilize target data while addressing the class imbalance issue in the target domain.

1) Class Frequency Estimation: We estimate the frequency of each class in the target domain through $\mathcal{D}_{lt}$ and $\mathcal{D}_{ut}$.

Algorithm 1 IES

Input: $\mathcal{D}_{lt}$, $\mathcal{D}_{ut}$, the annotation budget for each round b, the proportion of the most uncertain examples selected from each label pair η, the set of label pairs have been selected $\mathcal{P}$

Output: $\mathcal{D}_{lt}$, $\mathcal{D}_{ut}$, the target examples selected in each round $\mathcal{X}_r$

Calculate the classifier predicted labels and the prototype-based labels of each example in $\mathcal{D}_{ut}$ according to Eq. (3) and Eq. (9)
$\mathcal{D}_{ts} \leftarrow$ select the set of target-specific examples from $\mathcal{D}_{ut}$ according to Eq. (10)
$\mathcal{P}_{ts} \leftarrow$ obtain the set of label pairs for the examples in $\mathcal{D}_{ts}$
$\mathcal{P}_{cd} \leftarrow$ select the top b label pairs from $\mathcal{P}_{ts}$ with the highest example counts
for *label pair* p *in* $\mathcal{P}_{cd}$ **do**
  $\mathcal{X}_p \leftarrow$ select the examples with label pair p from $\mathcal{D}_{ts}$
  $\mathcal{X}'_p \leftarrow$ select the examples with label pair p that were selected in previous rounds from $\mathcal{D}_{lt}$
  $\forall \mathbf{x} \in \mathcal{X}_p$, compute $BvSB(\mathbf{x})$ (Eq. (11)) to serve as measure of uncertainty
  $\mathcal{X}^u_p \leftarrow$ select $\eta\%$ of $BvSB$ with the lowest values
  $\forall \mathbf{x} \in \mathcal{X}^u_p$, compute $RAD(\mathbf{x}, \mathcal{X}^u_p, \mathcal{X}'_p)$ (Eq. (13)) to serve as measure of representativeness and diversity
  $\mathbf{x}^r_p \leftarrow$ select RAD with the lowest value
  $\mathcal{X}_r \leftarrow \mathcal{X}_r \cup \mathbf{x}^r_p$
end
$\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{P}_{cd}$
$\mathcal{D}_{lt} \leftarrow \mathcal{D}_{lt} \cup \mathcal{X}_r$
$\mathcal{D}_{ut} \leftarrow \mathcal{D}_{ut} \setminus \mathcal{X}_r$

The number of examples in $\mathcal{D}_{lt}$ belonging to class c can be directly calculated

$$n_{lt,c} = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{lt}} \mathbb{I}(y_i = c) \tag{14}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. For $\mathcal{D}_{ut}$, we compute the number of examples per class in a weighted manner based on the predicted probability

$$\hat{n}_{ut,c} = \sum_{(\mathbf{x}_i, \hat{y}_i) \in \mathcal{D}_{ut}} \mathbb{I}(\hat{y}_i = c) \delta_c(\mathcal{F}(\mathcal{G}(\mathbf{x}_i))) \tag{15}$$

where $\delta_c(\mathbf{a}) = (e^{\mathbf{a}_c} / \sum_i e^{\mathbf{a}_i})$ denotes the cth element in the softmax output of a C-dimensional vector $\mathbf{a}$. Then, the frequency of each class in the target domain can be estimated as follows:

$$f_c = \frac{n_{lt,c} + \hat{n}_{ut,c}}{n_{lt} + \sum_c \hat{n}_{ut,c}}. \tag{16}$$

2) Self-Training With Class-Balanced Sampling: We select a subset $\mathcal{D}_{\text{cb}} \subset \mathcal{D}_t$ with balanced class distribution for self-training. In choosing $\mathcal{D}_{\text{cb}}$, we follow a class-balanced rule: the less frequent a class c is, the higher the sampling proportion for the examples predicted as class c. Specifically, we sample ratio ρ of the examples from the least frequent class to include into $\mathcal{D}_{\text{cb}}$. The sampling ratios for other classes are defined as follows:

$$\rho_c = \left(\frac{\min_c(f_c)}{f_c} \right)^{\lambda} \rho \tag{17}$$

where $\min_c(f_c)$ denotes the frequency of the least frequent class. For each class, we select the most confident examples. Using $\mathcal{D}_{\text{cb}}$, we can conduct CBST by optimizing the following loss function:

$$\mathcal{L}_{\text{cb}} = \frac{1}{|\mathcal{D}_{\text{cb}}|} \sum_{(\mathbf{x}_i, \hat{y}_i) \in \mathcal{D}_{\text{cb}}} \mathcal{L}_{\text{ce}}\big(\mathcal{F}(\mathcal{G}(\mathbf{x}_i)), \hat{y}_i\big). \tag{18}$$

E. Training Pipeline

The overall training pipeline of our proposed PCADA method consists of three stages, which are illustrated with pseudocode in Algorithm 2. The first stage is UDA, during which the feature extractor and the classifier are trained through the classification loss of the source domain and the prototype alignment loss

$$\min_{\theta_{\mathcal{G}}, \theta_{\mathcal{F}}} \mathcal{L}^s_{\text{cls}} + \alpha\big(\mathcal{L}_{\text{fpa}} + \mathcal{L}_{\text{tpa}}\big) \tag{19}$$

where α is the tradeoff parameter and $\theta_{\mathcal{G}}$ and $\theta_{\mathcal{F}}$ are the parameters of $\mathcal{G}$ and $\mathcal{F}$, respectively.

After $\mathcal{M}$ epochs, we enter the second stage, i.e., ADA. During this stage, we select a portion of target examples for annotation by an oracle, as described in the IES module. The annotated target examples can enhance the performance of the classifier through the classification loss and further align the feature distributions between domains using the feature-level prototype alignment loss

$$\mathcal{L}^t_{\text{cls}} = \frac{1}{n_{lt}} \sum_{i=1}^{n_{lt}} \mathcal{L}_{\text{ce}}\big(\mathcal{F}\big(\mathcal{G}\big(\mathbf{x}^{lt}_i\big)\big), y^{lt}_i\big) \tag{20}$$

$$\mathcal{L}^{lt}_{\text{fpa}} = \frac{1}{n_{lt}} \left(\sum_{i=1}^{n_{lt}} \|\mathbf{v}^{lt}_i - \boldsymbol{\mu}^s_{y^{lt}_i}\|^2_2 + \sum_{i=1}^{n_{lt}} \|\mathbf{v}^{lt}_i - \boldsymbol{\mu}^t_{y^{lt}_i}\|^2_2 \right). \tag{21}$$

Therefore, at this stage, we additionally minimize the two losses mentioned above

$$\min_{\theta_{\mathcal{G}}, \theta_{\mathcal{F}}} \mathcal{L}^t_{\text{cls}} + \alpha \mathcal{L}^{lt}_{\text{fpa}}. \tag{22}$$

After $\mathcal{H}$ epochs come to the third stage. As outlined in the CBST module, during this stage, we conduct CBST using a subset $\mathcal{D}_{\text{cb}}$ selected from $\mathcal{D}_t$, with the optimization objective being

$$\min_{\theta_{\mathcal{G}}, \theta_{\mathcal{F}}} \gamma \mathcal{L}_{\text{cb}}. \tag{23}$$

IV. Experiments and Analysis

In this section, we first present the experimental results on six HSI datasets to evaluate the effectiveness of the proposed PCADA method, using overall accuracy (OA), average accuracy (AA), and the kappa coefficient as performance metrics. Next, we provide a sensitivity analysis of the hyperparameters. Finally, ablation studies are conducted to demonstrate the effectiveness of each module.

Algorithm 2 Training Pipeline of PCADA

Input: $\mathcal{D}_s$, $\mathcal{D}_t$, $\mathcal{D}_{lt} = \emptyset$, $\mathcal{D}_{ut} = \mathcal{D}_t$, the total annotation budget B, the total annotation round R, the total epoch number $\mathcal{E}$, the epoch number to start example selection $\mathcal{M}$, the epoch number to start CBST $\mathcal{H}$, hyperparameters α, η, λ, ρ, γ

Output: The parameters $\theta_\mathcal{G}$ of the feature extractor $\mathcal{G}$ and $\theta_\mathcal{F}$ of the classifier $\mathcal{F}$

Randomly initialize the parameters $\theta_\mathcal{G}$ of the feature extractor $\mathcal{G}$ and $\theta_\mathcal{F}$ of the classifier $\mathcal{F}$
for *epoch=1 to* $\mathcal{E}$ **do**
  Update $\theta_\mathcal{G}$, $\theta_\mathcal{F}$ with $\mathcal{D}_s$, $\mathcal{D}_t$ via Eq. (19)
  if *epoch* $\geq \mathcal{M}$ **then**
    if *example selection is needed* **then**
      Update $\mathcal{D}_{lt}$, $\mathcal{D}_{ut}$ via Algorithm 1
    Update $\theta_\mathcal{G}$, $\theta_\mathcal{F}$ with $\mathcal{D}_{lt}$ via Eq. (22)
  if *epoch* $\geq \mathcal{H}$ **then**
    Update $\theta_\mathcal{G}$, $\theta_\mathcal{F}$ with $\mathcal{D}_t$ via Eq. (23)
end

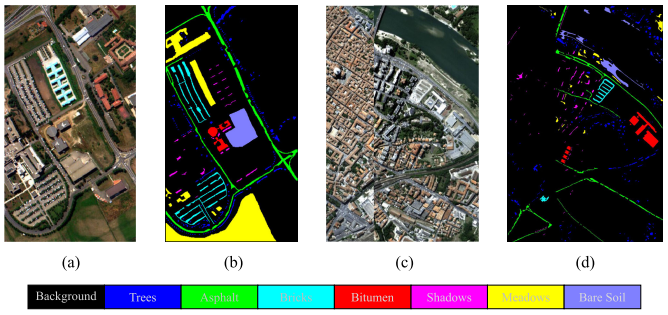

Fig. 3. False-color images and ground-truth maps of Pavia University and Pavia Center datasets. (a) False-color image of Pavia University. (b) Ground-truth map of Pavia University. (c) False-color image of Pavia Center. (d) Ground-truth map of Pavia Center.

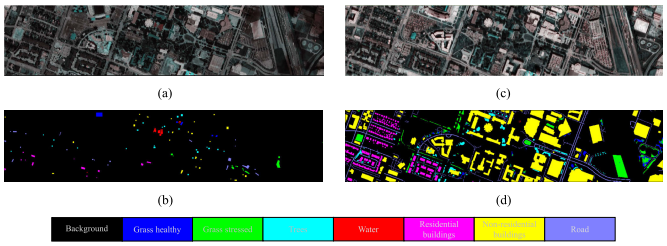

Fig. 4. False-color images and ground-truth maps of Houston2013 and Houston2018 datasets. (a) False-color images of Houston2013. (b) Ground-truth map Houston2013. (c) False-color images of Houston2018. (d) Ground-truth maps of Houston2018.

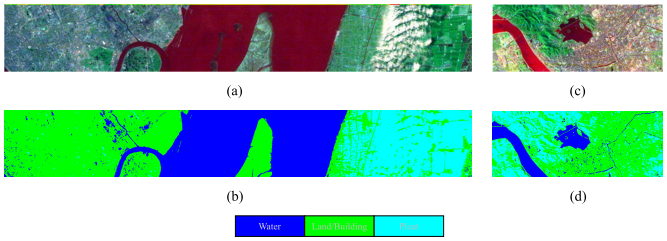

Fig. 5. False-color images and ground-truth maps of Shanghai and Hangzhou datasets. (a) False-color image of Shanghai. (b) Ground-truth map of Shanghai. (c) False-color image of Hangzhou. (d) Ground-truth map of Hangzhou.

TABLE I
NUMBER OF EXAMPLES IN EACH CLASS FOR PAVIA UNIVERSITY AND PAVIA CENTER DATASETS

Class		Number of Examples	
No.	Name	Pavia University (Source)	Pavia Center (Target)
1	Trees	3064	7598
2	Asphalt	6631	9248
3	Bricks	3682	2685
4	Bitumen	1330	7287
5	Shadows	947	2863
6	Meadows	18649	3090
7	Bare Soil	5029	6584
Total		39332	39355

A. Dataset Description

1) Pavia University and Pavia Center: Both datasets were captured by the reflective optics system imaging spectrometer (ROSIS) sensor during a flight campaign over Pavia in northern Italy. Due to differences in regional locations, there exists a domain shift between the two datasets. The Pavia University dataset covers an area of 610 × 610 pixels and contains 103 spectral bands, while the Pavia Center dataset covers an area of 1096 × 1096 pixels and contains 102 spectral bands. After removing some examples with no information, Pavia University consists of 610 × 315 pixels, and Pavia Center consists of 1096 × 715 pixels. In addition, to ensure the two datasets that have the same number of spectral bands, the last band of Pavia University was discarded. The false-color representations and ground-truth maps of both datasets are shown in Fig. 3.

2) Houston2013 and Houston2018: The two datasets cover the same place on the University of Houston campus. However, due to differences in acquisition times and sensors, there are domain discrepancies between them. The wavelength range of both datasets spans from 380 to 1050 nm. The Houston2013 dataset has 144 spectral bands; the Houston2018 dataset has 48 spectral bands. To match the number of spectral bands and spatial locations of the Houston2018 dataset, we averaged every three bands of the Houston2013 dataset and chose the overlapping areas of the two datasets. The false-color representations and ground-truth maps of both datasets are shown in Fig. 4.

3) Shanghai and Hangzhou: The Shanghai and Hangzhou datasets were captured by the EO-1 Hyperion hyperspectral sensor, which retains 198 bands after removing the bad bands. The image size of the Shanghai dataset is 1600 × 230 pixels, and the image size of the Hangzhou dataset is 590 × 230 pixels. The three shared land cover classes between the two datasets are water, ground/buildings, and plants. The false-color representations and ground-truth maps of both datasets are shown in Fig. 5.

Taking the six HSI datasets, we formulated three DA tasks, which are delineated as follows.

4) Pavia Task: The Pavia University dataset is considered as the source domain, and the Pavia Center dataset is considered as the target domain. Seven shared classes are selected for the task, and the number of examples in each class is listed in Table I.

5) Houston Task: The Houston2013 dataset serves as the source domain, and the Houston2018 dataset serves as the

TABLE II
NUMBER OF EXAMPLES IN EACH CLASS FOR HOUSTON2013 AND HOUSTON2018 DATASETS

Class		Number of Examples	
No.	Name	Houston2013 (Source)	Houston2018 (Target)
1	Grass healthy	345	1353
2	Grass stressed	365	4888
3	Trees	365	2766
4	Water	285	22
5	Residential buildings	319	5347
6	Non-residential buildings	408	32459
7	Road	443	6365
Total		2530	53200

TABLE III
NUMBER OF EXAMPLES IN EACH CLASS FOR SHANGHAI AND HANGZHOU DATASETS

Class		Number of Examples	
No.	Name	Shanghai (Source)	Hangzhou (Target)
1	Water	123123	18043
2	Land/Building	161689	77450
3	Plant	83188	40207
Total		368000	135700

target domain. Seven common classes are chosen from two datasets to conduct this task, which are presented in Table II.

6) Shanghai–Hangzhou Task: The Shanghai dataset is used as the source domain, and the Hangzhou dataset is used as the target domain. Three common classes in both datasets are considered in this task, which are detailed in Table III.

B. Experimental Setting

To evaluate the effectiveness of the proposed PCADA method, we compare it with eight other methods, including four ADA methods, AADA [21], SDM [59], label distribution matching through density-aware active sampling (LAMDA) [24], local context-aware ADA (LADA) [25], three UDA methods proposed for HSIC, contrastive learning based on category matching for DA (CLCM) [68], classwise prototype-guided alignment network (CPGAN) [69], masked self-distillation DA (MSDA) [70], and a semi-supervised DA method integrated with AL for HSIC and IEH-DA [61].

In the experiments, we selected 180 examples from each class in the source domain and used all the target examples for training. In addition, to eliminate the impact of random sampling, all experiments were repeated ten times, and the average results were taken as the final outcome.

For the four compared ADA methods, the annotation budget B for the Pavia task is set to 30, while for the Houston and Shanghai–Hangzhou tasks, B is set to 40. In addition, following the settings from the original papers, the number of annotation rounds R is fixed at 5 for all tasks. Since these methods were originally designed for natural images, we adjusted their hyperparameters to optimize their performance on HSI data. For the three UDA methods, to ensure fairness, we randomly selected 30 labeled target examples for the Pavia task and 40 labeled target examples for the Houston and Shanghai–Hangzhou tasks, using them in training. For IEH-DA, we ensure that the total number of examples ultimately labeled in each task remains consistent with other methods. As these UDA methods were originally proposed for HSI, we did not modify their parameters for datasets shared with their original studies. For tasks involving datasets not used in their works, we tuned their hyperparameters accordingly.

For our proposed method, we ensure that the annotation budgets for each task are consistent with those of the corresponding tasks in the compared ADA methods, and the annotation budget b for each round is fixed at 5. We set the total number of epochs as 100. During the training process, we used the stochastic gradient descent optimizer for backpropagation, with a weight decay of 0.0005 and a momentum of 0.9. For the Pavia task, the learning rate is set to 0.001, while for the Houston and Shanghai–Hangzhou tasks, the learning rates are set to 0.01 and 0.0003, respectively. α is set to 1 in all three tasks. Following [63], the patch size is set as 9×9, 7×7, and 1×1 in three tasks. The parameters in the CBST module are configured as follows: for the Houston task, $\rho = 1$, $\lambda = 0.3$, and $\gamma = 0.01$; for the Pavia task, $\rho = 0.5$, $\lambda = 0$, and $\gamma = 0.03$; and for the Shanghai–Hangzhou task, $\rho = 0.5$, $\lambda = 0$, and $\gamma = 0.005$. Also, batch-level-balanced sampling is further integrated into the CBST module to approximate a balanced class distribution within each mini-batch. Consistent with [71], we set K in (4) to 3. $2\sigma^2$ of the RBF kernel in (12) is set equal to the feature dimension, which is 288 for our model. Moreover, $\mathcal{M}$ and $\mathcal{H}$ are established as 40 and 75, respectively. Following [72], [73], and [74], the hyperparameters were determined through the validation set.

C. Experimental Results

Table IV shows the experimental results of different methods on the Pavia task. It can be observed that among the compared ADA methods, AADA achieves an accuracy of around 85%. This method relies solely on a hybrid informativeness criterion that combines the predictive entropy of the classifier with the output of the domain discriminator for example selection. In contrast, SDM and LADA implement more effective example selection strategies. SDM focuses on hard examples, while LADA leverages local context to guide its example selection. Consequently, SDM and LADA achieve accuracies of 89.49% and 91.25%, respectively. In addition, among the compared DA methods proposed for HSIC, IEH-DA demonstrates the most outstanding performance, with its accuracy exceeding 93%. In IEH-DA, the target examples are used in a way from easy to hard, which forms a curriculum sequence for orderly model training, and the "hard" examples with high informativity are annotated to provide supervision information for adaptation. Notably, our PCADA method achieves optimal results on all three evaluation metrics and has the highest classification accuracy for the third category. Compared with IEH-DA, PCADA improves OA, AA, and kappa coefficient by 1.56%, 1.99%, and 1.89%, respectively. Furthermore, PCADA is the only method that exceeds 90% classification accuracy for each category. Fig. 6 shows the classification maps obtained from all the methods.

Due to different acquisition years and sensors of the datasets, there is a significant domain gap between the source and target domains in the Houston task. Furthermore, the class

TABLE IV
CLASSIFICATION RESULTS (MEAN ± STANDARD DEVIATION) ON THE PAVIA TASK

Class No.	Methods								
	AADA [21]	SDM [59]	LAMDA [24]	LADA [25]	CLCM [68]	CPGAN [69]	MSDA [70]	IEH-DA [61]	PCADA
1	84.42	86.65	**95.97**	86.64	91.88	91.92	93.44	95.19	94.12
2	94.16	98.20	98.61	93.70	98.24	97.76	98.49	**99.34**	98.05
3	83.79	68.51	66.59	63.00	77.94	88.54	79.92	87.75	**91.99**
4	70.98	81.98	86.13	**92.56**	76.50	83.70	83.87	87.86	90.69
5	98.81	95.43	99.77	99.75	97.51	99.83	99.84	**100.00**	98.49
6	89.66	92.31	87.40	**97.11**	92.79	94.93	86.47	90.20	95.06
7	78.62	93.48	96.57	96.75	95.25	90.38	**96.95**	90.65	96.51
OA	84.67±2.92	89.49±1.68	92.47±1.50	91.25±1.81	90.62±4.09	92.09±0.96	92.44±0.98	93.50±0.52	**95.06±0.57**
AA	85.78±2.01	88.08±2.49	90.15±3.36	89.93±2.53	90.01±3.16	92.44±1.06	91.28±1.85	93.00±0.83	**94.99±0.94**
Kappa	81.74±3.39	87.36±2.01	90.90±1.83	89.48±2.17	88.69±4.97	90.52±1.15	90.90±1.19	92.17±0.63	**94.06±0.68**

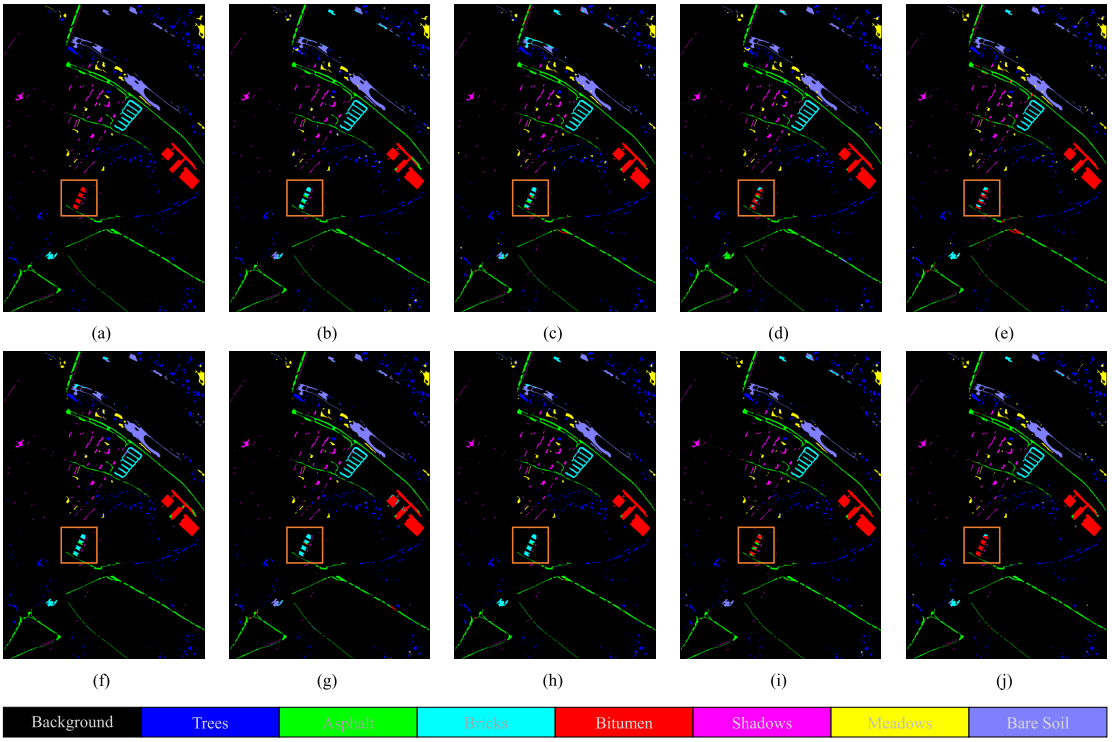

Fig. 6. Classification maps for the Pavia task. (a) Ground truth. (b) AADA. (c) SDM. (d) LAMDA. (e) LADA. (f) CLCM. (g) CPGAN. (h) MSDA. (i) IEH-DA. (j) PCADA.

TABLE V
CLASSIFICATION RESULTS (MEAN ± STANDARD DEVIATION) ON THE HOUSTON TASK

Class No.	Methods								
	AADA [21]	SDM [59]	LAMDA [24]	LADA [25]	CLCM [68]	CPGAN [69]	MSDA [70]	IEH-DA [61]	PCADA
1	71.97	43.81	43.44	43.91	52.95	61.86	59.65	**80.24**	63.41
2	78.15	88.27	84.52	87.48	86.63	83.42	**88.97**	70.66	86.35
3	69.72	70.31	75.34	**79.14**	67.26	67.15	67.61	74.56	74.97
4	75.91	73.64	55.45	74.09	71.36	17.73	80.91	68.69	**84.55**
5	70.21	78.78	86.57	**93.59**	85.96	85.61	92.53	78.27	83.91
6	51.78	86.48	84.30	70.16	84.03	80.54	81.69	86.99	**88.30**
7	69.29	75.52	77.95	**85.90**	70.17	53.72	72.27	69.81	79.05
OA	59.61±7.13	82.63±1.42	82.27±2.17	75.79±3.02	81.13±1.58	76.91±2.19	81.03±1.59	81.74±3.66	**85.24±2.01**
AA	69.58±4.78	73.83±5.18	72.51±5.50	76.32±2.68	74.05±2.74	64.29±3.38	77.66±2.01	75.60±3.88	**80.08±1.72**
Kappa	45.95±6.32	71.64±2.21	71.83±3.06	64.44±3.68	69.84±2.14	63.26±3.34	70.25±2.00	70.03±5.33	**76.03±2.67**

distribution in the target domain is extremely imbalanced. These factors make the Houston task particularly challenging. Nevertheless, our method still produces superior performance compared with other methods. The experimental results of various methods on the Houston task are presented in Table V. ADA methods that can more effectively select

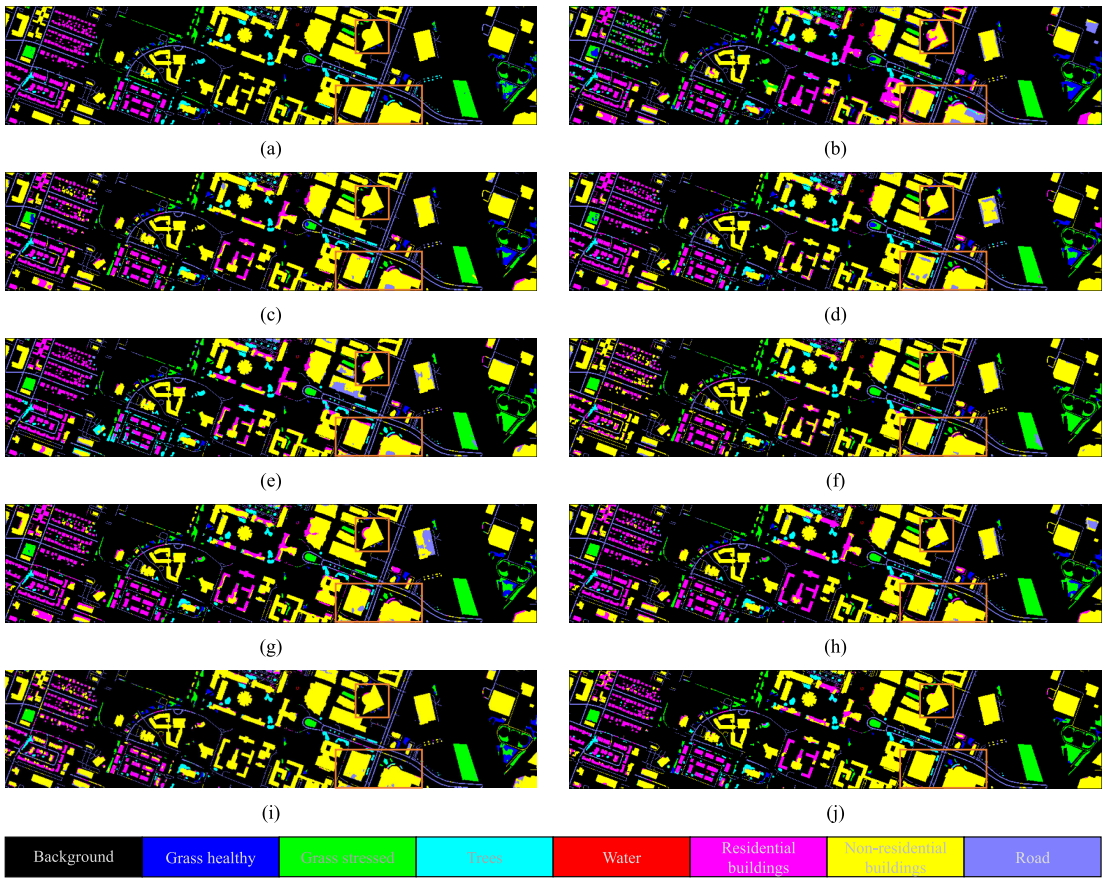

Fig. 7. Classification maps for the Houston task. (a) Ground truth. (b) AADA. (c) SDM. (d) LAMDA. (e) LADA. (f) CLCM. (g) CPGAN. (h) MSDA. (i) IEH-DA. (j) PCADA.

TABLE VI
CLASSIFICATION RESULTS (MEAN ± STANDARD DEVIATION) ON THE SHANGHAI–HANGZHOU TASK

Class No.	Methods								
	AADA [21]	SDM [59]	LAMDA [24]	LADA [25]	CLCM [68]	CPGAN [69]	MSDA [70]	IEH-DA [61]	PCADA
1	98.68	97.95	97.56	99.13	96.73	**99.77**	96.57	92.76	96.06
2	60.54	91.65	97.32	90.23	95.79	84.37	95.04	89.48	**98.64**
3	85.54	91.41	84.62	88.08	76.29	89.73	86.69	**91.88**	91.22
OA	73.02±5.26	92.41±1.05	93.59±0.74	90.77±2.88	90.14±1.55	88.01±2.98	92.77±1.66	90.63±1.70	**96.10±1.33**
AA	81.58±3.99	93.67±0.76	93.17±1.03	92.48±2.54	89.60±1.62	91.29±2.30	92.76±1.46	91.38±0.85	**95.30±1.75**
Kappa	58.49±7.21	86.87±1.71	88.57±1.38	84.24±4.78	82.26±2.83	79.95±4.79	87.26±2.93	83.76±2.68	**93.03±2.46**

informative examples, such as SDM, LAMDA, and LADA, still outperform AADA, which designs example selection strategies based on simple criteria. Both CLCM and MSDA are impressive, with their accuracies reaching approximately 81%. CLCM aligns the discriminative features between domains through cross-domain contrastive learning. MSDA reduces domain differences through adversarial training between the two classifiers and integrates masked self-distillation into DA to enhance feature discriminability. In contrast, our method achieves the best classification performance on the fourth and sixth categories and obtains the best performance with an accuracy of 85.24%. The classification maps of different methods are presented in Fig. 7.

As shown in Table VI, for the Shanghai–Hangzhou task, LAMDA achieves the best performance among the compared ADA methods. In particular, LAMDA seeks target examples that best approximate the entire target distribution as well as being representative, diverse, and uncertain. The selected examples are utilized not only for supervised learning but also for matching the label distribution between domains. Our method exhibits a superiority of 8.09% compared with CPGAN, which also employs a prototype learning approach. In addition, our method exhibits a higher AA and kappa coefficient, achieving an impressive AA of 95.30% and an outstanding kappa coefficient of 93.03%. Fig. 8 displays the classification maps on the Shanghai–Hangzhou task.

However, we observe that PCADA does not achieve optimal performance in some categories compared with the baseline methods. For the ADA methods in the baseline, this phenomenon can be largely attributed to the distribution of the

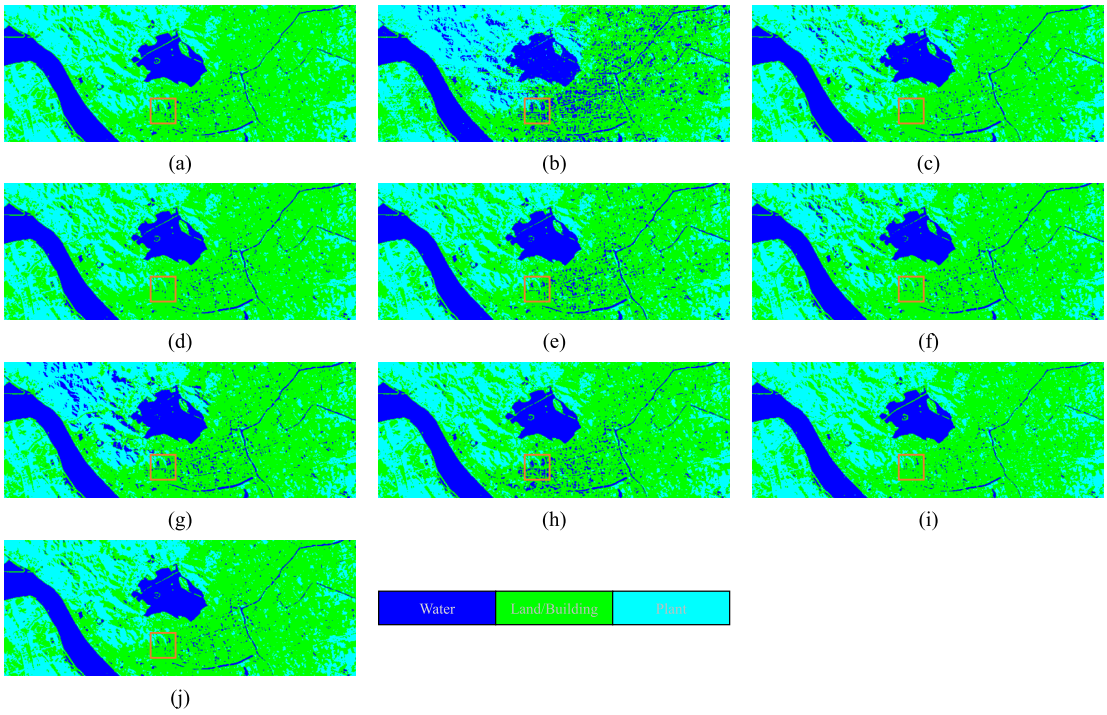

Fig. 8. Classification maps for the Shanghai–Hangzhou task. (a) Ground truth. (b) AADA. (c) SDM. (d) LAMDA. (e) LADA. (f) CLCM. (g) CPGAN. (h) MSDA. (i) IEH-DA. (j) PCADA.

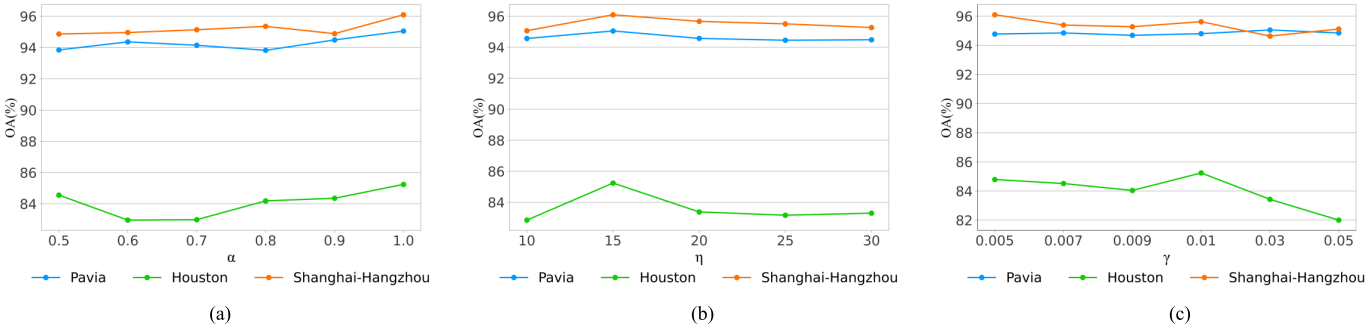

Fig. 9. Sensitivity analysis of the hyperparameter. (a) α. (b) η. (c) γ.

annotated examples. Specifically, the prediction accuracy of individual categories shows a strong positive correlation with the number of annotated examples in those categories. For instance, on the Pavia task, PCADA underperforms LADA in the fourth and sixth categories. This can be explained by the fact that LADA assigned relatively more annotations to target examples to these two categories, leading to better performance in these categories. Furthermore, the baseline methods typically apply uniform training strategies to both majority and minority classes. Consequently, the model inevitably develops bias toward majority classes, resulting in the misclassification of examples that originally belonged to minority classes, particularly in distinguishing spectrally similar categories. In contrast, our method prioritizes balanced predictive accuracy across all classes. While this design improves overall fairness in classification, it may introduce marginal performance degradation for some majority classes as a deliberate tradeoff.

D. Parameter Sensitivity Analysis

In the proposed method, there are five critical hyperparameters that need to be tuned, namely, α, η, ρ, λ, and γ. α controls the weights of the feature-level prototype alignment loss and the task-level prototype alignment loss in the PGDA module, while η adjusts the size of the uncertain subset for each label pair in the IES module. ρ, λ, and γ pertain to the CBST module, where ρ and λ control the sampling ratio from each category, and γ regulates the weight of the cross-entropy loss. We adjusted each hyperparameter individually while keeping all other hyperparameters constant.

We tested six different values for α, specifically 0.5, 0.6, 0.7, 0.8, 0.9, and 1. Fig. 9(a) illustrates the trend of OA for each task as α varies. It can be observed that all three tasks reach their maximum OA when α is set to 1, which is consistent with the parameter setting in [65]. For η, we selected five values within the range of 10–30 for testing. As shown in Fig. 9(b), a smaller η appears to yield more promising results.

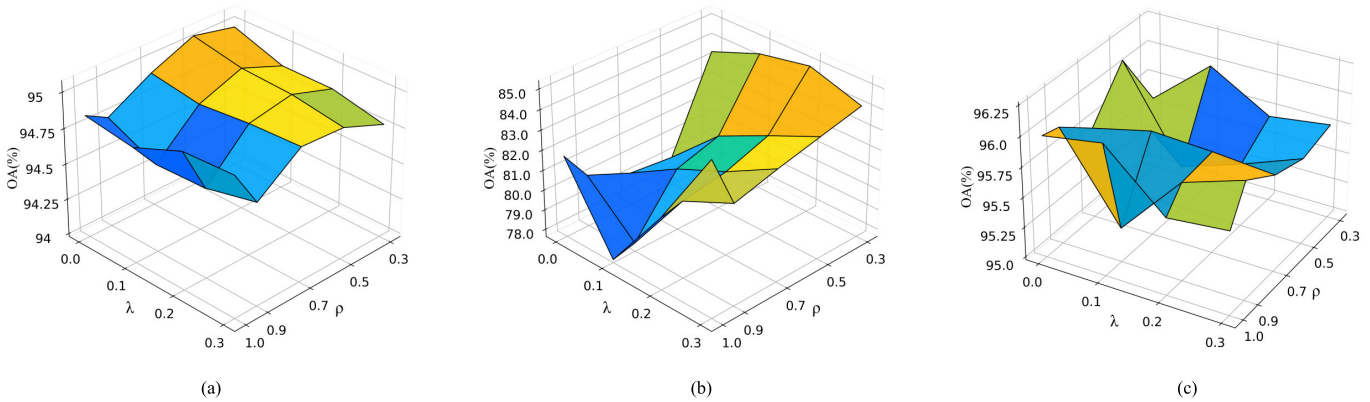

Fig. 10. Sensitivity analysis of the hyperparameters ρ and λ on different tasks. (a) Pavia. (b) Houston. (c) Shanghai–Hangzhou.

TABLE VII
ABLATION RESULTS OF THE PROPOSED METHOD WITH DIFFERENT MODULES

Method	PGDA	IES	CBST	Pavia Task			Houston Task			Shanghai-Hangzhou Task		
				OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
PCADA w/o PGDA		✓	✓	93.19	93.25	91.83	81.03	79.18	70.32	93.85	93.01	89.00
PCADA w/o CBST	✓	✓		94.49	94.33	93.37	83.52	76.88	73.51	95.95	95.27	92.78
PCADA w/o IES w/o CBST	✓			80.47	80.43	76.49	70.20	71.32	56.95	88.15	89.70	79.42
PCADA w/o IES w/ CBST	✓		✓	86.30	87.36	83.59	70.55	71.34	57.29	88.46	89.31	79.67
PCADA	✓	✓	✓	**95.06**	**94.99**	**94.06**	**85.24**	**80.08**	**76.03**	**96.10**	**95.30**	**93.03**

Consequently, we choose a relatively small η for all three tasks.

We tested six different values for γ, specifically 0.005, 0.007, 0.009, 0.01, 0.03, and 0.05. Fig. 9(c) shows how OA varies as γ is modified. The experiments indicate that γ has a minor impact on the classification results for the Pavia task, with optimal performance observed at γ set to 0.03. In the Houston task, OA reaches its maximum at γ valued at 0.01, after which it begins to decline. In the Shanghai–Hangzhou task, OA decreases with an increase of γ. Therefore, we set γ to 0.005 for this task.

We varied ρ in the range of 0.3–1 and λ in the range of 0–0.3, and then observed the fluctuations in classification performance caused by different combinations on the three tasks. As shown in Fig. 10(a), the performance on the Pavia task fluctuates smoothly across different combinations, with the best performance observed at $\rho = 0.5$ and $\lambda = 0$. In Fig. 10(b), we find that the performance on the Houston task varies between 78% and 85% across most combinations, with the best performance occurring when $\rho = 1$ and $\lambda = 0.3$. For the Shanghai–Hangzhou task, the performance exhibits a more complex variation with changes in the combinations of ρ and λ, as observed from Fig. 10(c). Ultimately, the best classification result is achieved at $\rho = 0.5$ and $\lambda = 0$.

E. Ablation Study

To further verify the effectiveness of each module of PCADA, we conducted a series of ablation studies. The corresponding results are recorded in Table VII, where the check mark (✓) indicates that the module is included in the experiment.

1) PCADA Without PGDA: In this case, the framework fails to utilize prototypes to perform multigranular alignment of distributions between domains and instead relies solely on the supervised loss for source examples and the selected target examples for optimization. It can be seen that the evaluation metrics for the three tasks decrease to varying degrees.

2) PCADA Without IES: Under the experimental condition where the active selection and annotation of target examples are removed, the evaluation metrics across all three tasks show significant degradation, with a drop exceeding that observed when removing the other two modules. This clearly demonstrates the crucial importance of the IES module. The IES module is essential because it provides target domain-specific information: when the selected examples are misclassified by the classifier, the annotation information effectively corrects the errors, allowing the model to generate more discriminative features, particularly between confusing categories. In addition, since the examples selected by IES typically exhibit high uncertainty, the annotation information help the model establish a more refined decision boundary. In conclusion, the IES module is a critical component for enhancing the generalization capability of the model in the target domain.

3) PCADA Without CBST: This condition removes the CBST part during the later stages of training. It can be observed that in the Houston task, which has an extremely imbalanced example size of each category in its target domain, OA decreases by 1.72%, kappa decreases by 2.52%, and AA shows a more pronounced decrease of 3.2%. The evaluation metrics of the other two tasks also decline.

In addition, we also compared the performance difference between using and not using the CBST module in the absence of any annotated target examples to explore the effectiveness

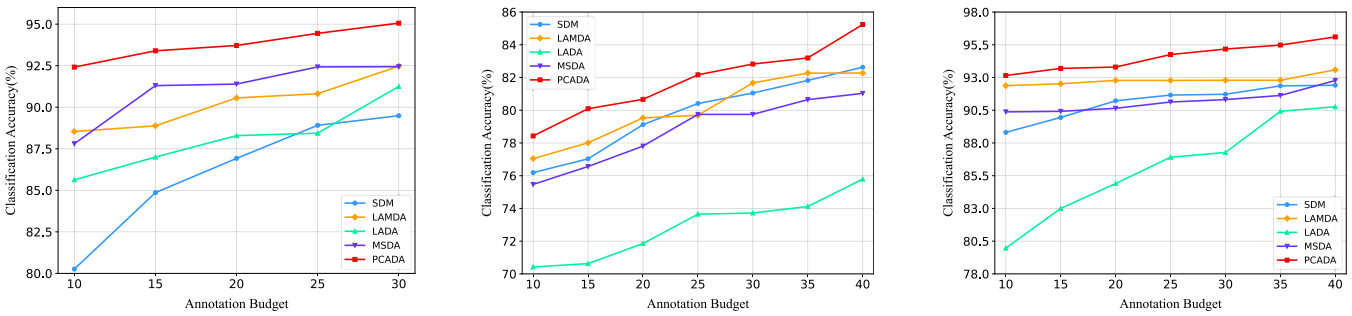

Fig. 11. Performance comparison under different annotation budgets. (a) Pavia. (b) Houston. (c) Shanghai–Hangzhou.

of the CBST module in UDA. The hyperparameters in the CBST module were tuned accordingly to achieve the best performance in this case. As shown in Table VII, the CBST module remains effective when the target domain is unsupervised. However, on the Houston task and the Shanghai–Hangzhou task, the improvement from the CBST module is relatively limited. This is mainly due to the lack of supervision information of the target domain, which leads to a decrease in the accuracy of pseudolabels in the later training stages, thus limiting the effect of self-training. For this reason, we perform CBST after example selection. With this design, the annotated target examples can improve the reliability of pseudolabels, enabling the CBST module to have a greater impact.

F. Varying Annotation Budget

We selected several well-performing baseline methods, including SDM, LAMDA, LADA, and MSDA, and compared their performance with PCADA under different annotation budgets. As shown in Fig. 11, as the budget increases, PCADA performs better on each task and consistently outperforms the compared methods, which means that our method is applicable to different budgets. More importantly, even with a small annotation budget, PCADA still shows relatively impressive performance, which fully demonstrates its ability to effectively and continuously select the most valuable examples.

V. Conclusion

In this article, we propose an ADA framework named PCADA for HSIC. In PCADA, the PGDA module assigns high-confidence pseudolabels to source-like examples and generates target prototypes. Guided by the interaction between source and target prototypes, the distributions of two domains are initially aligned. In the subsequent training phase, the IES module evaluates uncertainty and representativeness to annotate diverse target-specific examples, ensuring that the selected target examples are the most valuable despite domain shift. In addition, a CBST module is introduced, which samples highly confident pseudolabeled target examples in a class-balanced manner to address the issue of imbalanced class distribution in the target domain. The effectiveness of our method is validated through experiments conducted on multiple benchmark HSI datasets.

References

[1] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[2] P. Ghamisi et al., "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[3] J. Liu et al., "Estimating the forage neutral detergent fiber content of alpine grassland in the Tibetan Plateau using hyperspectral data and machine learning algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4405017.

[4] B. Luo, C. Yang, J. Chanussot, and L. Zhang, "Crop yield estimation based on unsupervised linear unmixing of multidate hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 162–173, Jan. 2013.

[5] C.-I. Chang, C.-Y. Lin, and P. F. Hu, "Band sampling of hyperspectral anomaly detection in effective anomaly space," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5502729.

[6] X. Fu, S. Jia, L. Zhuang, M. Xu, J. Zhou, and Q. Li, "Hyperspectral anomaly detection via deep plug-and-play denoising CNN regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9553–9568, Nov. 2021.

[7] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k -nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.

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

[9] L. Sun, Z. Wu, J. Liu, L. Xiao, and Z. Wei, "Supervised spectral–spatial hyperspectral image classification with weighted Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1490–1503, Mar. 2015.

[10] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

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

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

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

[14] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

[15] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," in *Proc. Adv. Data Sci. Inf. Eng.*, Jan. 2021, pp. 877–894.

[16] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.

[17] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.

[18] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–46, Oct. 2020.

[19] P. Ren et al., "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, Oct. 2021.

[20] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, May 1994.

[21] J.-C. Su, Y.-H. Tsai, K. Sohn, B. Liu, S. Maji, and M. Chandraker, "Active adversarial domain adaptation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 739–748.

[22] P. Rai, A. Saha, H. Daumé, and S. Venkatasubramanian, "Domain adaptation meets active learning," in *Proc. NAACL HLT Workshop Act. Learn. Nat. Lang. Process.*, 2010, pp. 27–32.

[23] H. Rangwani, A. Jain, S. K. Aithal, and R. V. Babu, "S3VAADA: Submodular subset selection for virtual adversarial active domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 7516–7525.

[24] S. Hwang, S. Lee, S. Kim, J. Ok, and S. Kwak, "Combating label distribution shift for active domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2022, pp. 549–566.

[25] T. Sun, C. Lu, and H. Ling, "Local context-aware active domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18588–18597.

[26] J. Feng et al., "Class-aligned and class-balancing generative domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5509617.

[27] D. Huang, J. Li, W. Chen, J. Huang, Z. Chai, and G. Li, "Divide and adapt: Active domain adaptation via customized learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7651–7660.

[28] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 289–305.

[29] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "CReST: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10857–10866.

[30] Z. Deng, K. Zhou, D. Li, J. He, Y.-Z. Song, and T. Xiang, "Dynamic instance domain adaptation," *IEEE Trans. Image Process.*, vol. 31, pp. 4585–4597, 2022.

[31] H. Li, J. Li, Y. Zhao, M. Gong, Y. Zhang, and T. Liu, "Cost-sensitive self-paced learning with adaptive regularization for classification of image time series," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11713–11727, 2021.

[32] G. Matasci, D. Tuia, and M. Kanevski, "SVM-based boosting of active learning strategies for efficient domain adaptation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 5, pp. 1335–1343, Oct. 2012.

[33] J. Wang, Y. Chen, H. Yu, M. Huang, and Q. Yang, "Easy transfer learning by exploiting intra-domain structures," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2019, pp. 1210–1215.

[34] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–35, Jan. 2016.

[35] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[36] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Jan. 2017, pp. 1645–1655.

[37] Z. Fang et al., "Confident learning-based domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5527116.

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

[39] X. Tang, C. Li, and Y. Peng, "Unsupervised joint adversarial domain adaptation for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536415.

[40] A. Gretton, K. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. Conf. Neural Inf. Process. Syst.*, Sep. 2007, pp. 513–520.

[41] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2016, vol. 30, no. 1, pp. 2058–2065.

[42] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2017.

[43] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 97–105.

[44] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[45] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2272–2281.

[46] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 443–450.

[47] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 6, pp. 2817–2830, Sep. 2021.

[48] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, and Z. Cai, "Cross-scene hyperspectral image classification with discriminative cooperative alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9646–9660, Nov. 2021.

[49] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[50] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 112–119.

[51] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *Proc. Eur. Conf. Mach. Learn.* Cham, Switzerland: Springer, 2006, pp. 413–424.

[52] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proc. 21st Int. Conf. Mach. Learn.-ICML*, 2004, p. 79.

[53] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017, *arXiv:1708.00489*.

[54] C. Yin et al., "Deep similarity-based batch mode active learning with exploration-exploitation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 575–584.

[55] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," 2019, *arXiv:1906.03671*.

[56] M. Ye, C. Wang, Z. Meng, F. Xiong, and Y. Qian, "Domain-invariant attention network for transfer learning between cross-scene hyperspectral images," *IET Comput. Vis.*, vol. 17, no. 7, pp. 739–749, Oct. 2023.

[57] M. Ye, J. Chen, F. Xiong, and Y. Qian, "Adaptive graph modeling with self-training for heterogeneous cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5503815.

[58] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active domain adaptation via clustering uncertainty-weighted embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 8505–8514.

[59] M. Xie et al., "Learning distinctive margin toward active domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 7993–8002.

[60] B. Xie, L. Yuan, S. Li, C. H. Liu, X. Cheng, and G. Wang, "Active learning for domain adaptation: An energy-based approach," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 8708–8716.

[61] C. Zhang, S. Zhong, S. Wan, and C. Gong, "Easy-to-hard domain adaptation with human interaction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5506813.

[62] Y. Xu et al., "Dual-channel residual network for hyperspectral image classification with noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502511.

[63] Z. Li et al., "Supervised contrastive learning-based unsupervised domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524017.

[64] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vision. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 499–515.

[65] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2239–2247.

[66] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2372–2379.

[67] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, Jan. 2012.

[68] Y. Ning, J. Peng, Q. Liu, Y. Huang, W. Sun, and Q. Du, "Contrastive learning based on category matching for domain adaptation in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5301814.

[69] Z. Xie, P. Duan, X. Kang, W. Liu, and S. Li, "Classwise prototype-guided alignment network for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5529211.

[70] Z. Fang, W. He, Z. Li, Q. Du, and Q. Chen, "Masked self-distillation domain adaptation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5525720.

[71] L. Li, J. Yang, Y. Ma, and X. Kong, "Pseudo-labeling integrating centers and samples with consistent selection mechanism for unsupervised domain adaptation," *Inf. Sci.*, vol. 628, pp. 50–69, May 2023.

[72] X. Li, Y. Gu, and A. Pižurica, "A unified multiview spectral feature learning framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540614.

[73] X. Li, M. Ding, and A. Pižurica, "Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2615–2629, Apr. 2020.

[74] X. Li, M. Ding, and A. Pižurica, "Spectral feature fusion networks with dual attention for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5508614.

Haiyang Luo received the B.E. degree from Hefei University of Technology, Hefei, China, in 2023. He is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

His research interests include computer vision, machine learning, and hyperspectral image processing.

Shengwei Zhong (Member, IEEE) received the B.E. degree in information countermeasure technology and the M.S. and Ph.D. degrees in electronics and communication engineering from Harbin Institute of Technology, Harbin, China, in 2013, 2015, and 2020, respectively.

She was an Exchange Ph.D. Student visiting the Remote Sensing Signal and Image Processing Laboratory (RSSIPL), University of Maryland, Baltimore County (UMBC), Baltimore, MD, USA, as a Faculty Research Assistant. She is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. Her research interests include hyperspectral image processing, remote sensing image fusion, and applications.

Chen Gong (Senior Member, IEEE) received the dual Ph.D. degree from Shanghai Jiao Tong University (SJTU), Shanghai, China, and the University of Technology Sydney (UTS), Ultimo, NSW, Australia, in 2016 and 2017, respectively.

He is currently a Full Professor with Nanjing University of Science and Technology, Nanjing, China. He has published more than 130 technical papers at prominent journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), *Journal of Machine Learning Research* (JMLR), *International Journal of Computer Vision* (IJCV), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), International Conference on Machine Learning (ICML), Conference on Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR), IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), AAAI Conference on Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), and IEEE International Conference on Data Mining (ICDM). His research interests mainly include machine learning, data mining, and learning-based vision problems.

Dr. Gong won the ICDM Best Student Paper Runner-Up Award, the Second Prize of Natural Science Award of Chinese Institute of Electronics, "Excellent Doctorial Dissertation Award" of Chinese Association for Artificial Intelligence, "Wu Wen-Jun AI Excellent Youth Scholar Award," and the Scientific Fund for Distinguished Young Scholars of Jiangsu Province. He was also selected as the "Global Top Chinese Young Scholars in AI" released by Baidu, and "World's Top 2% Scientists" released by Stanford University. He serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *Neural Networks*, and *Neural Processing Letters* (NePL), and also serves as the Area Chair or a Senior PC Member for several top-tier conferences, such as AAAI, IJCAI, ICML, ICLR, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), International Conference on Artificial Intelligence and Statistics (AISTATS), ICDM, and ACM Multimedia (ACM MM).