Mind the Gap: Multi-Level Unsupervised Domain Adaptation for Cross-scene Hyperspectral Image Classification

Mingshuo Cai, Bobo Xi, *Member, IEEE,* Jiaojiao Li, *Senior Member, IEEE,* Shou Feng, *Member, IEEE,* Yunsong Li, *Member, IEEE,* Zan Li, *Senior Member, IEEE,* and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—Recently, cross-scene hyperspectral image classification (HSIC) has attracted increasing attention, alleviating the dilemma of no labeled samples in the target domain. Although collaborative source and target training has dominated this field, training effective feature extractors and overcoming intractable domain gaps remains challenging. To cope with this issue, we propose a multi-level unsupervised domain adaptation (MLUDA) framework, which comprises image-, feature-, and logic-level alignment between domains to fully investigate the comprehensive spectral-spatial information. Specifically, at the image level, we propose an innovative domain adaptation method named GuidedPGC based on classic image matching techniques and the guided filter. The adaptation results are physically explainable with intuitive visual observations. Regarding the feature level, we design a multi-branch cross attention structure (MBCA) specifically for HSIC, which enhances the interaction between the features from the source and target domains through dot-product attention. Finally, at the logic level, we adopt a supervised contrastive learning (SCL) approach that incorporates a pseudo-label strategy and local maximum mean discrepancy loss, increasing inter-class distance across diverse domains and further improving the classification performance. Experimental results on three benchmark cross-scene datasets demonstrate that our proposed method consistently outperforms the compared approaches. The source code is available at https://github.com/cfcys/MLUDA.**

***Index Terms*—Cross-scene, domain adaptation, guided filter, cross attention, supervised contrastive learning**

I. Introduction

With the rapid development of sensor and image processing technologies, hyperspectral image (HSI) has attracted extensive attention due to its abundant land cover spatial distribution information and detailed spectral reflection data [1]–[3]. HSI classification (HSIC) is currently a hot research topic with the objective of assigning specific land cover categories to each pixel [4]–[8]. Additionally, HSIC holds considerable practical significance and has been extensively used in various fields, including urban monitoring, ecological science, and deep space exploration [9], [10].

In the last two decades, significant progress has been achieved in HSIC tasks, profiting from the advancements in deep learning (DL) and increased computational power [11]–[14]. Compared to the traditional methods, the DL-based approaches utilize deep neural networks to extract more sophisticated features from the raw data. Chen et al. [15] [16] are pioneers in integrating HSIC with stacked autoencoders (SAE) and deep belief networks (DBN), demonstrating the suitability of DL frameworks. Subsequently, numerous scholars have developed methods specifically for HSIC based on various architectures, such as the recurrent neural networks (RNN) [17], the classical convolutional neural networks (CNN) [18], 2-D CNN [19], 3-D CNN [20], a hybrid spectrum CNN incorporating both 3-D CNN and 2-D CNN [21], and the generative adversarial networks (GAN) [22], which yielded significant results.

Nowadays, the vision Transformer (ViT) based on self attention mechanism has demonstrated promising performance in computer vision. Numerous studies indicate that ViT can effectively capture global semantic information in the HSI. For instance, Hong et al. [23] introduced a backbone network called SpectralFormer, which is built upon the foundational ViT structure and reconsiders the HSIC from a sequential perspective. To fully exploit the spectral-spatial information, Farooque et al. [24] proposed a multi-scale three-dimensional atrous convolution approach based on the swin transformer (SwinT), which involves parallel branches to fuse spectral-spatial features in various scales, and then linearly embeds them into the lightweight SwinT. Huang et al. [25] integrated 3-D SwinT with contrastive learning to accommodate the 3-D properties of HSI, proving the effectiveness of ViT and contrastive learning in HSIC. Additionally, Peng et al. [26] introduced cross attention mechanism that investigates the mutual information between spectral and spatial aspects, leading to superior results than the conventional ViT structure.

This work was supported by the China Postdoctoral Science Foundation under Grant (2023M732742), the Natural Science Basic Research Plan in Shaanxi Province of China under Grant (2024JC-YBQN-0641), the Postdoctoral Science Foundation of Shaanxi Province (2023BSHYDZZ97), the Youth Open Project of National Space Science Data Center (NSSDC2302004), the Fundamental Research Funds for the Central Universities (XJSJ23086), Guangdong Basic and Applied Basic Research Foundation under Grant (2023A1515110504), and the Youth Innovation Team of Shaanxi Universities. *(Corresponding author: Bobo Xi, Zan Li.)*

Mingshuo Cai is with the State Key Laboratory of Integrated Service Networks, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: caimingshoo@gmail.com).

Bobo Xi is with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China, and also with the National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: xibobo@xidian.edu.cn).

Jiaojiao Li, Yunsong Li, and Zan li are with the State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: jjli@xidian.edu.cn; ysli@mail.xidian.edu.cn; zanli@xidian.edu.cn).

Shou Feng is with the School of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China (e-mail: fengshou@hrbeu.edu.cn).

Jocelyn Chanussot is with the Université Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France (e-mail: jocelyn@hi.is).

In practice, due to the increasingly obtained HSIs and the dilemma in acquiring annotated samples, it is pretty hard to predict class labels of new HSI scenes [27]. The direct transfer of classification models trained on the source domain (SD) to the target domain (TD) is full of difficulties. The problem arises from the distinct changes in the conditional distribution and spectral features of the same object in different scenarios, creating a significant domain gap between SD and TD and sharply reducing the classification performance. Therefore, overcoming this obstacle and achieving better cross-scene HSIC without annotated samples in TD has high practical significance, which is identified as the unsupervised domain adaptation (UDA) problem [28].

Existing UDA methods are generally categorized into three major classes, i.e., instance-based, classifier-based, and feature-based. Among them, the instance-based methods involve reweighting samples or extracting crucial instances to mitigate the domain shift. The classifier-based methods aim to train a robust classifier that can generalize to the target domain during the training process. The feature-based methods, which are the most prevalent, aim to extract domain-invariant features to minimize distribution differences between two domains, [29]. Early seminal UDA methods, such as space alignment [30] and correlation alignment [31], both focus on acquiring features that exhibit effective domain invariance. However, constrained by their inherent limitations, these methods can only extract manually crafted shallow features.

Afterward, plenty of UDA methods based on DL have been developed benefited from automatically extracting features, showcasing advantages over the above traditional methods. For instance, Long et al. [32] proposed a deep adaptation network (DAN) that embeds representations into a common space and reduces domain shift by minimizing the maximum mean discrepancy (MMD) across multiple kernels. Using the distinctive adversarial strategy [33], Ganin et al. [34] integrated GAN into domain adaptation and proposed a domain adversarial neural network (DANN), which extracts domain-invariant features through an adversarial learning mechanism.

With the above progress, the collaborative source and target training scheme to extract domain-invariant information has been widely utilized in cross-scene HSIC. For instance, Zhu et al. [35] proposed the deep subdomain adaptive network (DSAN) inspired by the DAN, introducing a local MMD (LMMD) to address the alignment between the same category from different domains. Subsequently, numerous methods emerged to achieve domain alignment by optimizing the variance of statistical distributions. For example, Liu et al. [36] proposed a class-wise distribution adaptation (CDA) network, combining class-wise adversarial adaptation with a probability-prediction-based MMD for effective adaptation. Zhang et al. [37] presented a discriminative co-alignment (DCA) method, which addresses geometric and statistical shifts by aligning subspaces and distributions. Then, Huang et al. [38] proposed a two-branch attention adversarial network (TAADA), where a dual-branch feature extraction network is designed as a generator to extract attention-based domain-invariant spectral-spatial features. In addition, Fang et al. [39] utilized confident learning for domain adaptation (CLDA), which employs high-confidence samples to enhance the discriminative capability of the network and achieves impressive classification performance.

Despite the significant success of the above methods, we notice that the previous attention-based approaches mainly focus on extracting and interacting within the spatial and spectral information in the HSI, while overlooking the attentive interaction between the source and target domains. Most importantly, although massive efforts have focused on reducing the gaps between the source and target domains, little emphasis has been placed on accommodating the domain shift in various levels of the networks. Thus, the improvements in the classification performance are limited.

To address these issues, we propose a multi-level unsupervised domain adaptation (MLUDA) method for cross-scene HSIC, which considers the image level, feature level, and logic level to accommodate the domain gaps. At the image level, we utilize gamma correction and color histogram to transfer the style of the principal components of source and target domains and then transform the original HSI through a guide filter [40]. This module, abbreviated as GuidedPGC, produces visually compelling results with physical implications. At the feature level, we introduce cross attention mechanism into cross-scene HSIC and propose a multi-branch cross attention (MBCA) module, which allows sufficient interactions between features of the source and target domains by the classical dot-product attention. At the logic level, we adopt supervised contrastive learning (SCL) in the source and target scenes with a pseudo-label strategy conducted in the target domain. Besides, we incorporate the LMMD loss function into the optimization, enhancing the intra-class compactness in the aligned features and consequently improving the classification accuracy.

In summary, the contributions of the study are as follows:

1) We propose the GuidedPGC by incorporating the guided filter and classical style transfer techniques, achieving image-level domain adaptation for the specific three-dimensional HSI. This approach is characterized by its explainable physical meaning, allowing visual verification of the effective domain adaptation process.

2) We first introduce the cross attention mechanism into the cross-scene HSIC and devise the MBCA module, which facilitates comprehensive interactions between source and target domain information, fully using the rich high-level spectral-spatial features in the HSI through a multi-head attention layer structure.

3) We involve the SCL in the source and the target domains with a pseudo-label strategy. And the LMMD is integrated into the network to impose the inter-class features separable and intra-class compact. Extensive experiments on three public cross-scene HSIC datasets validate the superiority of the proposed method.

The paper is divided into five sections. Section II provides an overview of the related works. Section III details our proposed method. Section IV presents our experimental results and analysis. Section V summarizes this study.

II. Related works

A. Image-level Domain Adaptation

For HSI applications, the classification task exhibits certain similarities with semantic segmentation for normal remote sensing images [41], [42], which provides valuable enlightenment for our research. For example, Li et al. [43] drew inspiration from the concept of white balance and proposed an unsupervised color mapping unification module to normalize the color space of source and target domain datasets. It mitigates the covariate shifts caused by varying capturing conditions for the very high-resolution (VHR) images and achieves promising performance. Additionally, Ma et al. [44] addressed the issue of image-level domain shift by employing gamma correction on the luminance channel to globally align the source domain images with reference images from the target domain. This method has strong physical significance and the effectiveness of the domain adaptation can be visualized. Furthermore, Tasar et al. [45] proposed a novel ColorMapGAN to generate synthetic training images to achieve domain transfer. The model learns to transform the color of the training data to match the test data by performing only one element-wise matrix multiplication and one matrix addition. The produced images are semantically identical to the training images and they have similar spectral distributions to the test images, demonstrating good visual fidelity. Based on our investigation, the image-level style transfer is effective for RS semantic segmentation tasks. However, holistic image-level domain adaptation has not been widely developed for cross-scene HSIC tasks.

B. Supervised Contrastive Learning

The contrastive learning gained prominence with the milestone study of momentum contrast (MoCo) [46]. The InfoNCE loss function introduced in MoCo brings positive samples close to each other in the feature space while simultaneously pushing them apart from negative samples. Subsequent approaches such as SimCLR [47] and MoCoV2 [48] further harnessed the significant potential of contrastive learning. Afterwards, SCL [49] is proposed by extending the concepts of contrastive learning to fully supervised learning. The framework first augments the data and projects them into the embedding space, then selects samples of the same category as positive samples and aligns them in the embedding space, while increasing the distance between samples from different classes. This way, the extracted features are more discriminative than the general fully supervised learning.

In HSIC, Liu et al. [50] designed a deep contrastive learning network (DCLN) to address the small-sample classification problems by constructing contrastive groups that enable contrastive learning to train the network to extract highly domain-invariant features. Guan et al. [51] considered the inherent characteristics of HSI and proposed a spatial-spectral contrastive learning (SSCL) method, which extracts spectral and spatial information to create a pair of samples representing two perspectives in contrastive learning. Additionally, Ref. [52] proposed a fine-grained prototype contrastive learning network, which combines SCL and prototype learning to mitigate the domain shift in few-shot learning and proves the significant potential of the SCL in HSIC tasks.

III. Proposed MLUDA for HSIC

The flowchart of the proposed MLUDA for cross-scene HSIC is depicted in Fig. 1. It can be observed that the framework mainly consists of three levels of domain adaptation: image level, feature level, and logic level. In the framework, data from the source and target domains initially undergo image-level domain adaptation to align the source domain data guided by the target domain in overall features such as color and brightness. Subsequently, we apply two types of data augmentation to both the source and target domain data. The augmented data and the preliminary data are then input into the spectral-spatial dual-branch attention feature extraction network. The obtained features undergo feature-level domain adaptation based on cross attention, leading to the extraction of domain-invariant features. Finally, at the logic level, we adopt an SCL loss based on pseudo-labels and the LMMD loss, which is done to widen the gap between the source and target domains of different classes to enhance the classification performance.

A. GuidedPGC for Image-level Domain Adaptation

GuidedPGC is our newly proposed method based on the guided filter, where P represents principal component analysis (PCA), G denotes gamma correction, and C stands for color histogram correction. This approach aims to reduce the gap in attributes such as color and brightness between the source and target domains at the image level, and it is completed by the following steps.

1) Guided filtering: Guided filter [40] is an edge-preserving filter widely used in image processing tasks, such as denoising, detail enhancement, and HDR imaging. The core concept involves integrating a guidance image into the filtering process to ensure the resulting filter retains clarity at the edges. An essential feature of the guided filter is its computational efficiency, which is not affected by the window size and depends only on the pixel count of the image.

An important assumption of the guided filter is that the output image denoted as *out* and the guidance image denoted as *Ref* have a local linear relationship within the filtering window w_k, it can be modeled as

$$out_i = a_k Ref_i + b_k, \forall i \in w_k \tag{1}$$

where a_k and b_k are the parameters need to be determined. At the same time, within the window w_k, out is obtained by subtracting the noise n (the part that needs to be filtered out) from the input image Ori.

$$out_i = Ori_i - n_i, \forall i \in w_k \tag{2}$$

Suppose there is an edge in *out*, it can be preserved while making the filtered result *out* similar to the input image *Ori*, minimizing the information loss caused by filtering. The obtained closed-form solution after imposing the constraint is:

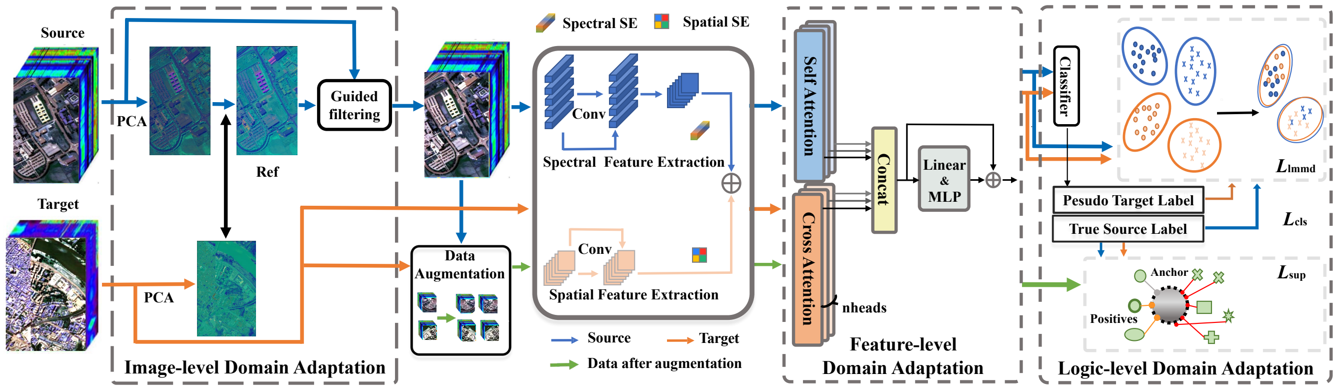


Fig. 1. The flowchart of the MLUDA, illustrating the domain adaptation strategies at the image-, feature-, and logic-level.

$$out_i = \frac{1}{|\omega|}\sum_{i\in\omega_k}(a_k Ref_i + b_k) \tag{3}$$

$$a_k = \frac{\frac{1}{|\omega|}\sum_{i\in\omega_k} Ref_i \cdot Ori_k - \mu_k \bar{Ori}_k}{\sigma_k^2+\epsilon} \tag{4}$$

$$b_k = \bar{Ori}_k - a_k\mu_k \tag{5}$$

where $\bar{Ori}_k = \frac{1}{|\omega|}\sum_{i\in\omega_k} Ori_i$. μ_k and σ_k represent the mean and variance pixel values of the *Ref* within w_k, respectively.

2) Brightness gamma correction: Brightness gamma correction is a common technique to conveniently adjust overall brightness and contrast of an image. The fundamental principle involves applying a nonlinear transformation to each pixel in the image, typically represented as a power function expressed as $I_{out} = I_{in}^{\gamma}$, where I_{in} and I_{out} denote the pixel intensity of the input and output images, respectively. γ is a predetermined parameter, and when γ is smaller than 1, gamma correction increases the brightness and contrast of the image, making dark details more noticeable. Conversely, when γ is greater than 1, gamma correction reduces the brightness and contrast, emphasizing bright details. Due to variations in color representation among different display and printing devices, brightness gamma correction ensures consistent image display across various devices. In our task, HSIs from the source and target domains may have different brightness and contrast due to acquisition time and equipment variations. Thus, gamma correction is employed to align these fundamental properties.

3) Color histogram matching: The image histogram is a statistical representation of the distribution of pixels within an image. Histogram matching aligns the histogram of an image or a specific region with another image to ensure that the tonal features are consistent between the two images. Specifically, the procedure involves the following steps: converting the original and target image to the HSV color space, calculating their histograms and normalizing them, and computing the cumulative distribution functions for them. For each pixel in the original image, the algorithm identifies the corresponding pixel in the target image with the closest match, and then assigns its color value to the original image pixel. The modified image is then converted back to the RGB color space, ensuring the color spaces between the two images are harmonized.

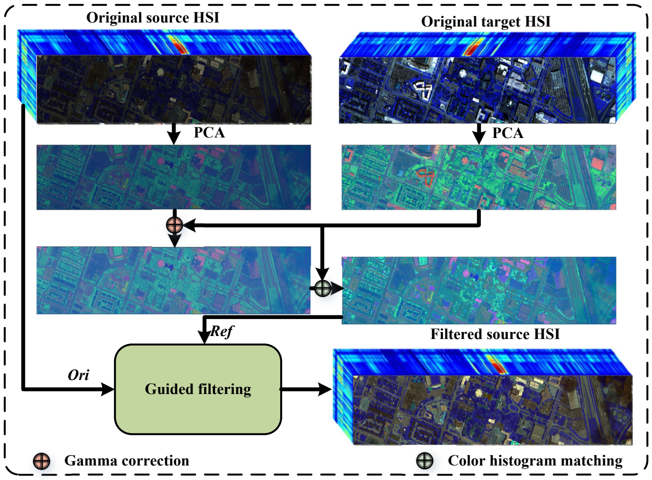


Fig. 2. Illustration of the proposed GuidedPGC.

4) Image-level domain adaptation: Fig. 2 illustrates the proposed GuidedPGC for image-level adaptation. We utilize up to np principal components obtained through PCA as the reference for brightness gamma correction and color histogram matching. Taking the Houston dataset as an example, the input images of the source and target domains have the dimensions of $w \times h \times 48$. After performing PCA, the dimensions become $w \times h \times np$. Subsequently, we employ the data from the target domain as a reference to apply gamma correction and color histogram matching to the source domain data, minimizing the gaps in brightness and color between the source and target domains. Finally, the transformed image is utilized as the reference (*Ref*) in the guided filtering, while the original image serves as the input (*Ori*). After applying guided filtering band-by-band, the resulting *out* maintains the same dimensions as the original data but exhibits a distribution of brightness and color features more similar to the target domain.

B. MBCA for Feature-level Domain Adaptation

MBCA aims to address the domain shift by leveraging the advantages of the cross attention and self attention mechanisms. The features from the source and target domains are effectively integrated through dot-product attention, enhancing

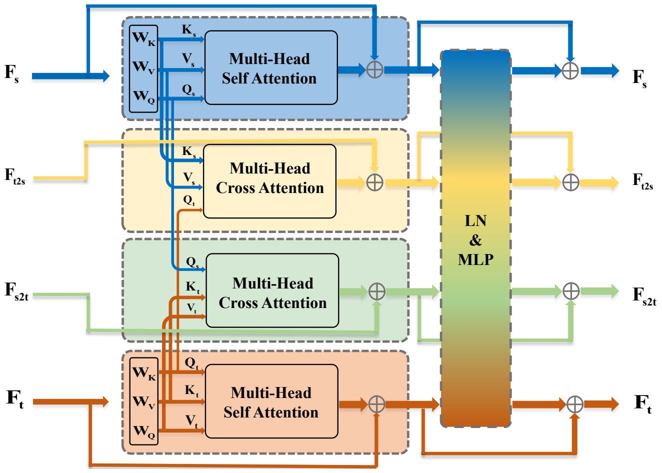

F_s
F_t2s
F_s2t
F_t
W_K
W_V
W_Q
K_s
V_s
Q_s
Multi-Head Self Attention
K_s
V_s
Q_t
Multi-Head Cross Attention
Q_s
K_t
V_t
Multi-Head Cross Attention
Q_t
K_t
V_t
Multi-Head Self Attention
LN & MLP
F_s
F_t2s
F_s2t
F_t

Fig. 3. The schematic diagram of the MBCA module.

their complementary characteristics and facilitating the extraction of domain-invariant features.

Fig. 3 illustrates the proposed MBCA, where $\mathbf{F}_{\mathrm{t2s}}$ and $\mathbf{F}_{\mathrm{s2t}}$ represent the cross-domain features extracted by using the cross attention mechanism. In our experiments, $\mathbf{F}_{\mathrm{t2s}}$ and $\mathbf{F}_{\mathrm{s2t}}$ are initialized with values equal to $\mathbf{F}_{\mathrm{s}}$ and $\mathbf{F}_{\mathrm{t}}$, respectively. $\mathbf{F}_{\mathrm{t2s}}$ represents features derived from the source domain with the target domain as the auxiliary information. Similarly, $\mathbf{F}_{\mathrm{s2t}}$ represents features derived from the target domain with the source domain as auxiliary. Taking advantage from the multi-head attention, cross tokens are split into 2^n nonoverlapping copies denoted as $\mathbf{F}_{\mathrm{s}_i} \in \mathbb{R}^{Z\times(\frac{D}{2^n})}$ and $\mathbf{F}_{\mathrm{t}_i} \in \mathbb{R}^{Z\times(\frac{D}{2^n})}$, where i = 1,2,...,n and D represents the dimension of the input data. Each head of the MBCA module consists of three groups of linear matrices $\mathbf{W}_{\mathrm{Q}_i}, \mathbf{W}_{\mathrm{K}_i}$ and $\mathbf{W}_{\mathrm{V}_i} \in \mathbb{R}^{\frac{D}{2^n}\times\frac{D}{2^n}}$. $\mathbf{F}_{\mathrm{s}}$ and $\mathbf{F}_{\mathrm{t}}$ are the inputs to one head of the MBCA module because each head is identical in actual operations. Here, we present the computation of two sets of matrices $\{\mathbf{Q}_{\mathrm{s}}, \mathbf{K}_{\mathrm{s}}, \mathbf{V}_{\mathrm{s}}\}$ and $\{\mathbf{Q}_{\mathrm{t}}, \mathbf{K}_{\mathrm{t}}, \mathbf{V}_{\mathrm{t}}\}$ for each attention head in the structure of dot-product attention as follows:

$$\mathbf{Q}_{\mathrm{s}} = \mathbf{F}_{\mathrm{s}}\mathbf{W}_{\mathrm{Q}}, \mathbf{K}_{\mathrm{s}} = \mathbf{F}_{\mathrm{s}}\mathbf{W}_{\mathrm{K}}, \mathbf{V}_{\mathrm{s}} = \mathbf{F}_{\mathrm{s}}\mathbf{W}_{\mathrm{V}} \tag{6}$$

$$\mathbf{Q}_{\mathrm{t}} = \mathbf{F}_{\mathrm{t}}\mathbf{W}_{\mathrm{Q}}, \mathbf{K}_{\mathrm{t}} = \mathbf{F}_{\mathrm{t}}\mathbf{W}_{\mathrm{K}}, \mathbf{V}_{\mathrm{t}} = \mathbf{F}_{\mathrm{t}}\mathbf{W}_{\mathrm{V}} \tag{7}$$

We further derive two self-similarity matrices $\mathbf{S}_{\mathrm{s}}$ and $\mathbf{S}_{\mathrm{t}}$, and two cross-similarity matrices $\mathbf{S}_{\mathrm{t2s}}$ and $\mathbf{S}_{\mathrm{s2t}}$ after performing the matrix calculations. Specifically, $\mathbf{S}_{\mathrm{s}}$ and $\mathbf{S}_{\mathrm{t}}$ retrieve information for each element from their respective domain feature maps. $\mathbf{S}_{\mathrm{t2s}}$ uses the target query $\mathbf{Q}_{\mathrm{t}}$ to explore each element $\mathbf{K}_{\mathrm{s}}$ in the source domain, while $\mathbf{S}_{\mathrm{s2t}}$ uses the source query $\mathbf{Q}_{\mathrm{s}}$ to explore each element $\mathbf{K}_{\mathrm{t}}$ in the target domain. The self attention matrix $\{\mathbf{A}_{\mathrm{s}},\mathbf{A}_{\mathrm{t}}\}$ and the cross attention matrix $\{\mathbf{A}_{\mathrm{s2t}}, \mathbf{A}_{\mathrm{t2s}}\}$ can be calculated by utilizing the self-similarity matrices and cross-similarity matrices as follows:

$$\mathbf{A}_{\mathrm{s}} = \mathbf{S}_{\mathrm{s}}\mathbf{V}_{\mathrm{s}}^{\mathrm{T}} = \varphi\left(\frac{\mathbf{Q}_{\mathrm{s}}^{\mathrm{T}}\mathbf{K}_{\mathrm{s}}}{\sqrt{d}}\right)\mathbf{V}_{\mathrm{s}}^{\mathrm{T}} \tag{8}$$

$$\mathbf{A}_{\mathrm{t}} = \mathbf{S}_{\mathrm{t}}\mathbf{V}_{\mathrm{t}}^{\mathrm{T}} = \varphi\left(\frac{\mathbf{Q}_{\mathrm{t}}^{\mathrm{T}}\mathbf{K}_{\mathrm{t}}}{\sqrt{d}}\right)\mathrm{V}_{\mathrm{t}}^{\mathrm{T}} \tag{9}$$

$$\mathbf{A}_{\mathrm{s2t}} = \mathbf{S}_{\mathrm{s2t}}\mathbf{V}_{\mathrm{t}}^{\mathrm{T}} = \varphi\left(\frac{\mathbf{Q}_{\mathrm{s}}^{\mathrm{T}}\mathbf{K}_{\mathrm{t}}}{\sqrt{d}}\right)\mathrm{V}_{\mathrm{t}}^{\mathrm{T}} \tag{10}$$

$$\mathbf{A}_{\mathrm{t2s}} = \mathbf{S}_{\mathrm{t2s}}\mathbf{V}_{\mathrm{s}}^{\mathrm{T}} = \varphi\left(\frac{\mathbf{Q}_{\mathrm{t}}^{\mathrm{T}}\mathbf{K}_{\mathrm{s}}}{\sqrt{d}}\right)\mathbf{V}_{\mathrm{s}}^{\mathrm{T}} \tag{11}$$

where d is the channel size per head. $\varphi()$ and $()^{\mathrm{T}}$ represent the softmax function and the matrix transpose, respectively.

By utilizing the pairs $\{\mathbf{A}_{\mathrm{s2t}}, \mathbf{F}_{\mathrm{s2t}}\}$ and $\{\mathbf{A}_{\mathrm{t2s}}, \mathbf{F}_{\mathrm{t2s}}\}$, the MBCA can integrate information from different domains and concurrently acquire more transferable features from one domain to another. The fused features are then input to a multi-layer perceptron, which is a two-layer fully connected neural network designed to recover perceptual information across all channels. In the MBCA module, information from two different domains is integrated by combining the information from each point on both feature maps across all channel dimensions. Furthermore, the entire MBCA incorporates normalization layers and dropout layers after the attention modules. For different datasets, the MBCA module can produce different effects with multiple cascaded levels. Additionally, different numbers of heads can also effect the classification results. We will further discuss the parameter settings in the Section IV-E.

C. SCL for Logic-level Domain Adaptation

The SCL [53] is a prevalent approach in the field of self-supervised learning, which aims to maximize the consistency across different augmented views of samples from the same class. This, in turn, reduces the distance between samples belonging to the same category, while simultaneously increasing the separation between samples from different classes.

1) Data augmentation: Data augmentation is a critical element in SCL. For each sample x in the source and target domains, we conduct two data augmentations to generate related views of the same, $\widetilde{x} = Aug(x)$. The data augmentation techniques used in this study are random Gaussian noise and random flipping.

- **Random Gaussian noise:** The method involves Gaussian distribution random noise to the original data.
- **Random flipping:** The scheme horizontally or vertically flips the samples to increase the diversity of the data.

2) SCL loss: We randomly select N sample pairs $\{x_k^s, y_k^s\}_{k=1,\ldots,N}$, then through two data augmentation modules, we can obtain $2N$ samples pairs denoted as $\{\tilde{x}_i^s, \tilde{y}_i^s\}_{i=1,\ldots,2N}$. The loss function for SCL can be formulated as follows:

$$L_{\mathrm{scl}} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i^s \cdot \boldsymbol{z}_p^s/\tau\right)}{\sum_{a\in A(i)} \exp\left(\boldsymbol{z}_i^s \cdot \boldsymbol{z}_a^s/\tau\right)} \tag{12}$$

where $i \in I = \{1, \ldots, 2N\}$ represents the index of each augmented sample, referred to as "anchor". $\boldsymbol{z}_i^s = Pro(E(\tilde{x}_i^s)) \in \mathbb{R}^{D_P}$ denotes the vector projected from the extracted features and D_p is set to 128 in our experiments. $A(i) = I \setminus \{i\}$ represents the set of all data excluding the anchor itself. $P(i) = \{p \in A(i) : \tilde{\boldsymbol{y}}_p^s = \tilde{\boldsymbol{y}}_i^s\}$ represents the set of all "positive" samples where the samples in the set share the same label as the anchor. $\tau \in \mathbb{R}^+$ is a temperature coefficient. In the framework, the SCL is applied to both the source and target domains. For the target domain, we utilize the classifier from the current iteration to generate pseudo-labels based on the already obtained features.

D. Details of the Feature Extractor and Model Implementation

1) Feature extractor: In the proposed method, we adopt the spectral-spatial dual-branch attention feature extraction network proposed in TAADA [38] as the feature extractor, which is illustrated in Fig. 4.

As shown in the right branch, the data undergo spectral feature extraction by utilizing 24 convolutional kernels with dimensions of $1 \times 1 \times 7$. After traversing three convolutional layers and a single residual operation, the data is transformed into the Spectral squeeze-excite (SE) module. Compared to the conventional SE model [54], Spectral SE considers the significance of each frequency band and enhances various bands. We define $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_C]$ as the 2-D spatial patches, the Spectral SE operation is formulated as

$$\mathbf{z}_k = F_{sq}(\boldsymbol{v}_k) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \boldsymbol{v}_k(i,j), \quad k = 1, \ldots, C \tag{13}$$

where $\boldsymbol{v}_k \in \mathbb{R}^{H \times W}$ and $H = W = 7$, $C = 128$, the operation $F_{sq}(\cdot)$ is known as the squeeze operator.

The spatial feature extractor is shown in the left branch, which consists of two convolutional layers, a residual block, and a Spatial SE structure. This design enables the extractor to compress spectral information and emphasize the spatial features. In detail, 24 convolutional kernels with size $1 \times 1 \times 48$ are initially employed to process the data, with an emphasis on extracting spectral information. After passing through two convolutional layers and undergoing a residual operation, the information is input to the Spatial SE structure to capture the inherent spatial information effectively. We define the 49 vectors as $\mathbf{V} = [\boldsymbol{v}^{1,1}, \boldsymbol{v}^{1,2}, \ldots, \boldsymbol{v}^{i,j}, \ldots, \boldsymbol{v}^{W,H}]$ which has a dimension of $1 \times 1 \times 24$ and can be converted to 24 patches with the size of 7×7, where $\boldsymbol{v}^{i,j} \in \mathbb{R}^{1 \times 1 \times C}$, $H = W = 7$, and $C = 24$. Finally, the Spatial SE operation is

$$\mathbf{q} = \mathbf{F}_{ex}\big(\mathbf{F}_{sq}(\mathbf{V})\big) = \sigma(\mathbf{W} \otimes \mathbf{V}) \tag{14}$$

where $\mathbf{q} \in \mathbb{R}^{W \times H}$ and $\mathbf{W} \in \mathbb{R}^{1 \times 1 \times C}$. Then, we concatenate the features obtained from the two branches to obtain the final extracted features.

2) Model implementation: By combining GuidedPGC, MBCA, and SCL for domain adaptation, we achieve additional performance improvement by accommodating the domain gaps at multiple levels. A basic classifier is devised, projecting features into the category dimensionality, followed by a softmax operation to produce the final predicted category with the highest probability. The normal cross-entropy loss is employed in the optimization process. In addition, we adopt the LMMD loss proposed in [35] to effectively align each category between the source and target domains. In summary, the overall loss function is as follows:

$$Loss = L_{\text{cls}} + L_{\text{scl}} + \lambda L_{\text{lmmd}} \tag{15}$$

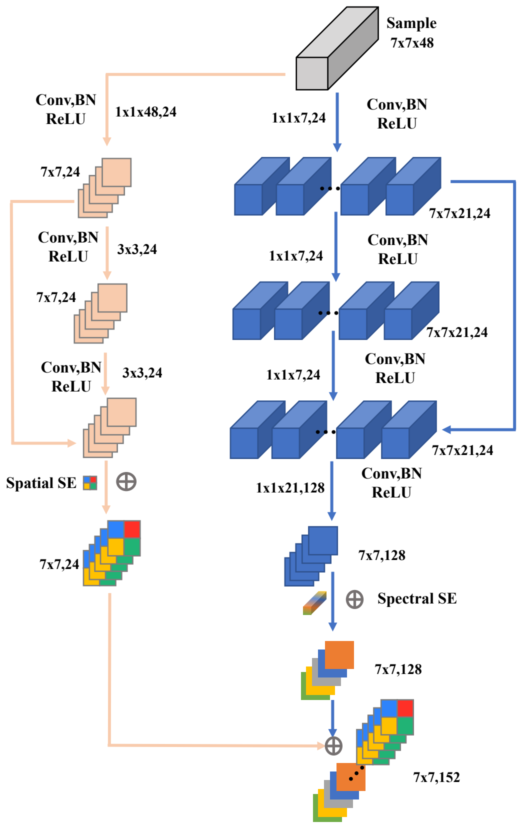


Fig. 4. The spectral-spatial dual-branch attention feature extraction network.

where λ is a balancing parameter. We empirically set $\lambda = 0.01 * 2/(1 + e^{(-10*(epoch)/epochs)}) - 1$ which produces better results than ordinary constant parameter [53].

IV. Experimental Results and Analysis

A. Datasets

To demonstrate the effectiveness of the proposed methods, we conducted experiments on six HSI datasets: Houston2013, Houston2018, Pavia University, Pavia Center, Shanghai, and Hangzhou.

Houston2013 and Houston2018: The Houston2013 dataset has spatial dimension of 349×1905, with a 2.5-meter resolution, and comprises 144 spectral bands for analysis across 15 categories. In contrast, the Houston2018 dataset has dimensions of 209×955 pixels, with a 1-meter resolution, and includes 48 spectral bands covering 20 categories. The datasets are acquired in 2013 and 2018 utilizing disparate sensors, encompassing the University of Houston campus and its environs. This highlights the discernible gap between the source and target domains due to temporal disparities and variations in sensor characteristics. Fig. 5 presents false-color images and ground truth maps for the datasets. Besides, both datasets span a wavelength range from 0.38 to 1.05 μm.

Pavia University and Pavia Center: The Pavia University contains 610×610 pixels, and initially includes 115 spectral bands. After removing 12 noisy bands, the dataset consists of 103 spectral bands. Furthermore, the size is reduced to 610 ×

TABLE I: NUMBER OF SAMPLES IN THE HOUSTON DATASET

No.	Class	Houston13 (Source)	Houston18 (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

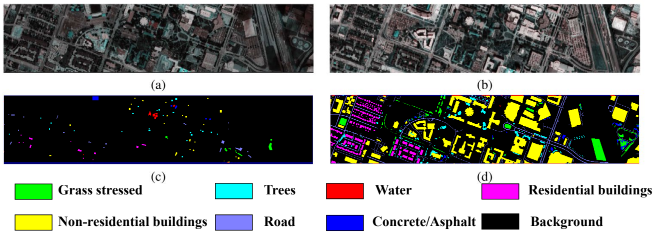


Fig. 5. HOUSTON dataset visualization.

340 pixels due to the presence of invalid data in the datasets. The Pavia Center dataset comprises 1096 × 1096 pixels and 102 spectral bands. Additionally, we adjust its size to 1906 × 715 pixels due to the presence of invalid data. The two datasets are acquired using ROSIS sensors in Pavia, a city in the northern region of Italy. Thus, the two datasets effectively demonstrate the gap between distinct locations within the same urban area. Besides, the last spectral band of the Pavia University dataset is removed for consistency. Fig. 6 shows the false-color images and ground truth maps for both datasets.

Shanghai and Hangzhou: The SH2HZ dataset accurately captures the gap that exists between remotely sensed images of different urban laneways taken by the same sensor. The datasets are collected using the EO-1 Hyperion hyperspectral sensor. The Shanghai dataset consists of 1600 × 230 pixels, while the Hangzhou dataset comprises 590 × 230 pixels. Both datasets contain three common categories: water, land/buildings, and plants. After removing the defective bands, we obtain a total of 198 spectral channels. Fig. 7 shows false-color images and ground truth maps for both datasets.

Taking the six HSI datasets, we design three UDA tasks, each approaching the gap between the source and target domains from the perspectives of distinct collection dates, scenes, and cities, which are delineated as follows:

1) Houston Task: The source and target domain are represented by the Houston2013 and Houston2018 datasets, respectively. To achieve this, seven shared categories are considered in the datasets, which is detailed in Table I. The primary challenge of this task lies in the considerable temporal gap and the utilization of disparate sensors during data collection between the source and target domains.

2) Pavia Task: The Pavia University dataset defines the source domain, while the Pavia Center dataset represents the target domain. We consider seven shared categories as illustrated in Table II. This task effectively illustrates how our model addresses the HSIC task across diverse scenes within the same city.

TABLE II: NUMBER OF SAMPLES IN THE PAVIA DATASET

No.	Class	PaviaU (Source)	PaviaC (Target)
1	Tree	3064	7598
2	Asphalt	6631	9248
3	Brick	3682	2685
4	Bitumen	1330	7287
5	Shadow	947	2863
6	Meadows	18649	3090
7	Bare soil	5029	6584
	Total	39332	39355

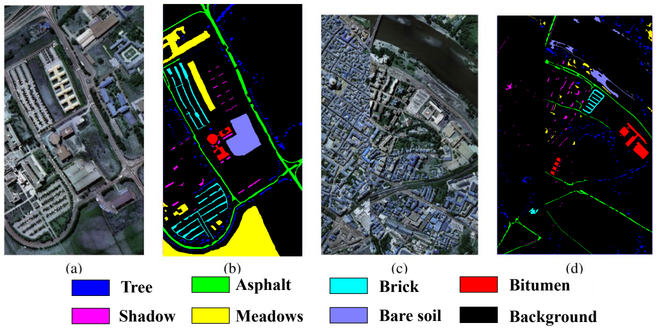


Fig. 6. PAVIA dataset visualization.

TABLE III: NUMBER OF SAMPLES IN THE SH2HZ DATASET

No.	Class	ShangHai (Source)	HangZhou (Target)
1	Water	123123	18043
2	Land/Building	161689	77450
3	Plant	83188	40307
	Total	368000	135700

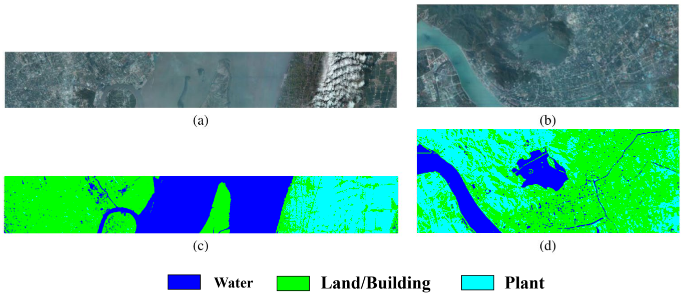


Fig. 7. SH2HZ dataset visualization.

3) SH2HZ Task: The source domain is the Shanghai dataset, while the target domain involves the Hangzhou dataset. This task focuses on three common categories in the datasets as listed in Table III. The primary challenge of this task is to address the gap between the source and target domains caused by similar scenes across different cities.

B. Experimental Setup

To evaluate the proposed approach, we conduct comparisons with seven other domain adaptation methods, including DAN [32], DANN [34], DSAN [35], CDA [36] TAADA [38], and the recent state-of-the-art CLDA [39] method. All experiments are conducted ten times to mitigate the impact of random sampling, and the average value is utilized as the final classification result. Given the current distribution shift between the source and target domains, we conduct the necessary normalization on the input data. For the experiments, we select 180 samples for each class in the source domain and use all of the target samples for training.

In DAN, we incorporate three adaptive layers, each utilizing multiple kernel variants of the MMD metric. The number of Gaussian kernels is fixed to five. In addition, we set the epoch to 200 and the batch size to 32 for training. To maximize its effectiveness, we use a specific configuration for the Pavia dataset with a batch size of 64 and 300 epochs. For DSAN, the spectral-spatial dual-branch attention convolutional network mentioned earlier is utilized as the feature extractor. In each task, we incorporate five transfer layers and set the weighted parameter of the LMMD loss function to 0.1. Additionally, we adopt the optimal parameters as identified in their original papers for the methods DANN, CDA, TAADA, and CLDA, which are mainly designed for HSIC.

For our proposed method, we set the number of epochs to 100 for the three tasks. The initial learning rate η_0 is set to 0.01, 0.001, and 0.0003 for three tasks during training. For the Houston task, the learning rate is adjusted dynamically using the formula $\eta_\theta = \eta_0/(1 + a\theta)^b$, where $a = 10$, $b = 0.75$, and θ is a variable that changes linearly from 0 to 1 in the training process. The patch size is set to 7 × 7, 11 × 11, and 1 × 1 for the three tasks. The number of the principal components is experimentally set to 2, the number of the head in the cross attention is set to 2, and τ is set to 0.1. Additionally, we utilize stochastic gradient descent as the optimizer and the batch size is set to 32, 64, and 32 for three tasks.

C. Experiments and Results Analysis

We utilize overall accuracy (OA), average accuracy (AA), and the Kappa coefficient (Kappa) to evaluate the classification performance of different methods. From Table IV to Table VI, we present the experimental results with the best outcomes in bold. For visual comparison, we present the best classification maps of various methods from Fig. 8 to Fig. 10. The ground truth is also shown for convenient comparison.

1) Experiments on Houston task: Table IV displays the experimental results of various comparative algorithms on the Houston dataset. The CDA, TAADA, and CLDA built specially for HSI analysis demonstrate superior performance compared to DAN and DSAN. The CLDA and TAADA perform relatively well, but their metrics exhibit higher variability across multiple independent experiments. The DANN is designed explicitly for cross-scene HSIC, but as an early attempt that only focuses on aligning feature distributions, its performance lags behind other methods except for the DAN. The proposed method, which conducts domain adaptation at multiple levels, is superior to other comparisons, demonstrating stable classification performance. For example, the MLUDA achieves an impressive peak accuracy of 76.64% in OA, surpassing the second-ranked CDA by a notable improvement exceeding 3%. Concerning AA and Kappa, MLUDA significantly outperforms the alternative methods.

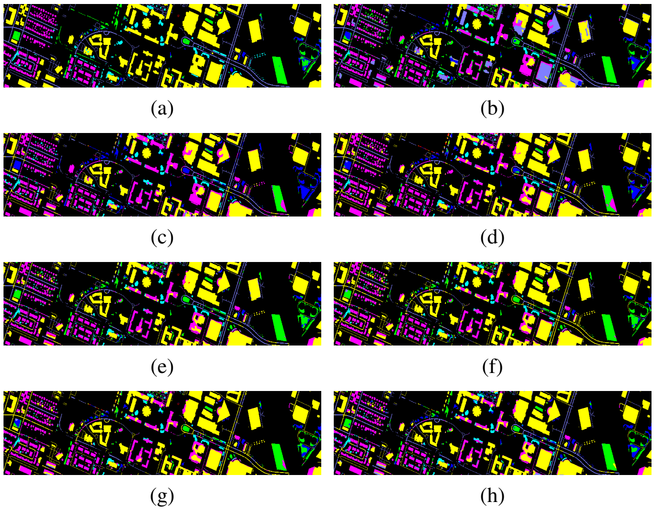


Fig. 8. Classification maps obtained via various algorithms for HOUSTON DATASET. (a) Ground truth, (b) DAN, (c) DANN, (d) DSAN, (e) CDA, (f) TAADA, (g) CLDA, (h) MLUDA.

In the visual classification results, the challenge lies in distinguishing the “Non-residual building” and “Residual building” classes. The accuracy improvement is primarily associated with this category. Methods such as DAN, DANN, and DSAN, which lack specialized designs for hyperspectral data, struggle to differentiate these categories and perform poorly in the “Grass stressed” class. Due to the limited number of samples, we attribute these challenges to the failure of these methods to extract representative hyperspectral features from the source domain. In contrast, our method excels at classifying these complex categories in Houston.

Let’s recall Fig. 2, which shows the visual domain adaptation process of the Houston dataset in the GuidedPGC module. It can be observed that the Houston18 image has a strong color contrast and brightness compared to Houston13. There are discernible gaps between the source and target domain due to temporal disparities and variations in sensor characteristics. After the GuidedPGC module, significant improvements can be found in the color contrast and brightness of the Houston13 image. This enhancement in GuidedPGC reduces the apparent gap between the source and target domain, thereby elucidating the gains in classification accuracy.

2) Experiments on Pavia task: Table V displays the experimental results on the Pavia dataset. It can be seen that all methods achieve relatively excellent performance in domain adaptation classification, which may be attributed to the abundant sample quantity and balanced sample distribution in this dataset. The DAN and DANN exhibit lower performance. This may be due to the fact that DAN only uses MMD loss, and DANN relies solely on basic domain obfuscation techniques.

TABLE IV: CLASSIFICATION PERFORMANCE (%) OF VARIOUS METHODS ON THE HOUSTON DATASET

No.	Class	DAN [32]	DANN [34]	DSAN [35]	CDA [36]	TAADA [38]	CLDA [39]	MLUDA
1	Grass healthy	78.74	74.52	**93.71**	93.21	83.79	73.73	83.26
2	Grass stressed	76.23	77.31	46.55	53.17	51.10	**92.43**	75.69
3	Trees	70.37	69.32	59.67	64.28	61.32	**70.68**	54.13
4	Water	92.43	93.99	59.09	90.21	91.76	70.91	**94.55**
5	Residential buildings	83.31	77.62	76.61	74.31	71.22	82.64	**84.28**
6	Non-residential buildings	45.32	56.31	72.07	**83.21**	72.15	54.74	79.33
7	Road	59.65	75.12	56.23	31.34	73.98	**80.83**	65.74
OA	-	55.43 ±1.54	60.37 ±1.86	68.19 ±2.46	72.95 ±1.41	72.88 ±3.16	69.63 ±2.82	**76.64** **±1.18**
AA	-	71.32 ±1.94	71.38 ±2.31	66.28 ±6.59	70.21 ±1.67	73.21 ±2.64	71.99 ±2.76	**76.71** **±1.99**
Kappa	-	45.33 ±2.24	46.31 ±2.21	52.31 ±3.33	54.21 ±2.01	60.33 ±2.41	56.84 ±3.59	**63.62** **±1.47**

TABLE V: CLASSIFICATION PERFORMANCE (%) OF VARIOUS METHODS ON THE PAVIA DATASET

No.	Class	DAN [32]	DANN [34]	DSAN [35]	CDA [36]	TAADA [38]	CLDA [39]	MLUDA
1	Tree	72.63	75.03	82.42	92.13	92.13	**95.51**	93.02
2	Asphalt	77.39	82.63	85.61	77.21	91.11	**98.31**	96.36
3	Brick	82.65	77.11	88.89	84.97	76.31	74.14	**96.81**
4	Bitumen	67.22	72.91	72.54	82.39	78.53	81.66	**85.42**
5	Shadow	98.15	95.32	88.77	99.26	87.15	91.24	**99.91**
6	Meadows	73.01	87.11	86.5	63.19	86.64	85.32	**97.45**
7	Bare soil	75.16	72.57	78.57	73.94	86.78	**90.19**	79.97
OA	-	76.43 ±1.46	79.14 ±3.23	81.92 ±2.75	84.56 ±2.08	89.17 ±1.94	91.02 ±2.10	**91.26** **±0.53**
AA	-	78.18 ±1.34	80.33 ±2.32	83.33 ±2.02	83.15 ±1.31	88.21 ±1.23	90.02 ±1.78	**92.70** **±0.80**
Kappa	-	71.23 ±1.71	72.35 ±3.19	78.37 ±2.14	80.38 ±2.33	87.88 ±2.01	89.23 ±2.21	**89.63** **±0.63**

TABLE VI: CLASSIFICATION PERFORMANCE (%) OF VARIOUS METHODS ON THE SH2HZ DATASET

No.	Class	DAN [32]	DANN [34]	DSAN [35]	CDA [36]	TAADA [38]	CLDA [39]	MLUDA
1	Water	99.81	88.37	99.47	98.91	93.23	95.88	**99.94**
2	Land/building	77.21	86.31	73.66	81.23	84.83	89.12	**89.90**
3	Plant	87.12	81.24	87.12	88.17	95.13	87.23	**93.05**
OA	-	83.12 ±2.12	85.31 ±3.12	81.08 ±3.88	87.92 ±1.10	82.09 ±1.77	89.68 ±0.98	**92.15** **±0.94**
AA	-	87.91 ±1.31	84.19 ±2.89	86.75 ±2.42	91.36 ±1.77	84.91 ±0.92	90.32 ±1.49	**94.30** **±0.63**
Kappa	-	73.14 ±2.91	72.89 ±3.42	69.58 ±5.58	79.31 ±1.72	67.37 ±2.41	82.16 ±2.20	**92.15** **±1.49**

In contrast, DSAN employs an additional LMMD loss, while CDA is equipped with multiple discriminators, and both of them achieve improvements. TAADA employs adversarial domain adaptation with two classifiers which exhibit strong capabilities in extracting domain-invariant features and gain improvements. Besides, CLDA with the confident learning strategy performs relatively better.

Notably, our approach achieves higher metrics compared to other methods, with an OA of 91.26%, an AA of 92.70%, and a Kappa of 89.63%. Nevertheless, considering that TAADA and CLDA effectively address the gaps of various scenarios, the enhancements achieved by our method are limited. However, our approach shows smaller deviations across multiple experiments, highlighting the robustness of the model.

In the visual classification results, we can observe that the DAN, DANN, and DSAN algorithms incorrectly predicted the "Meadows" as the "Tree", instead of the adjacent "Bare soil". In the CDA, TAADA, and CLDA, the categories "Bitumen", "Shadow" and "Meadows" show significant differences. This variation may be attributed to the insufficient domain adaptation capability of these methods. Our approach achieves the highest classification accuracy for these three categories, which is consistent with the results presented in Table V.

3) Experiments on SH2HZ task: Table VI displays the experimental results on the SH2HZ dataset. Unlike the previous two datasets, DSAN and TAADA did not outperform DAN and DANN. Meanwhile, the adversarially structured CDA and CLDA exhibit relatively stable performance in this task, achieving commendable metrics across the three indicators. Similar to the other two tasks, our method achieves higher OA, AA, and Kappa with better stability. Therefore, we can conclude that our multi-level structure can handle the task of UDA in various complex situations.

In the visual classification results, the challenge lies in distinguishing the dispersed "Water" within the "Land" due to the fewer categories. Methods such as CDA, CLDA, and TAADA misclassify numerous instances of "Land" as "Water". On the other hand, general methods such as DAN often

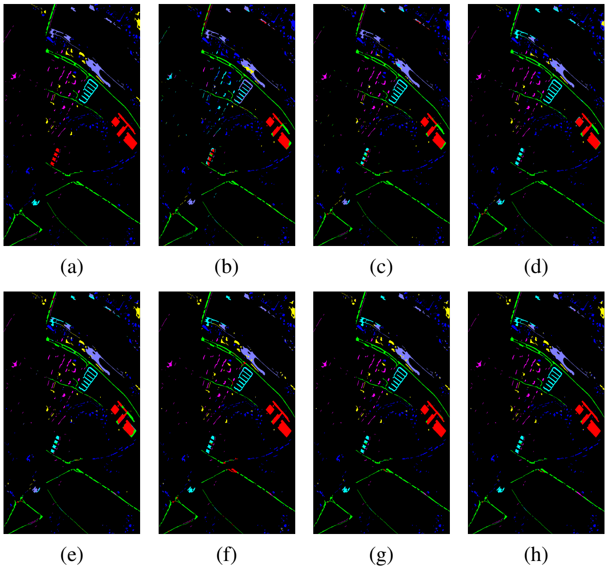


Fig. 9. Classification maps obtained via various algorithms for PAVIA DATASET. (a) Ground truth, (b) DAN, (c) DANN, (d) DSAN, (e) CDA, (f) TAADA, (g) CLDA, (h) MLUDA.

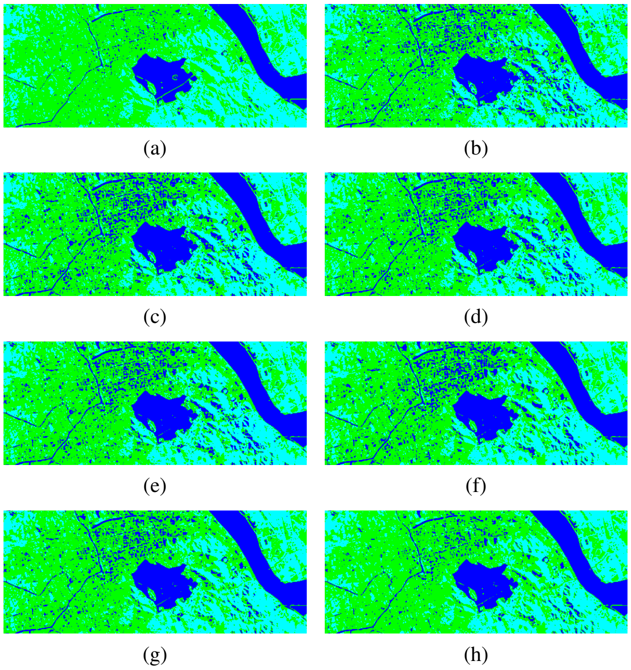


Fig. 10. Classification maps obtained via various algorithms for SH2HZ DATASET. (a) Ground truth, (b) DAN, (c) DANN, (d) DSAN, (e) CDA, (f) TAADA, (g) CLDA, (h) MLUDA.

misclassify many instances of “Plant” as “Water”. In contrast, MLUDA performs best by using multi-level domain adaptation to reduce the misclassifications in the mentioned scenarios.

D. Ablation Experiments

To further illustrate the effectiveness of each level in our proposed multi-level framework, referred to as image-level domain adaptation (A), feature-level domain adaptation (B), and logic-level domain adaptation (C), we conduct ablation experiments. The complete structure is denoted as ABC, where AB represents the structure without the C module, i.e., the logic-level domain adaptation is removed. Similarly for AC and BC. The ablation results for various datasets are presented in Tables VIII, VII, and IX. The ablation experiments clearly indicate that removing any portions from the multi-level structure (ABC) results in a decrease in OA, AA, Kappa, and even the stability of the model. This supports the reasonableness and effectiveness of each level within the proposed structure.

TABLE VII: ABLATION EXPERIMENTS ON HOUSTON DATESET

Class	AB	AC	BC	ABC
1	66.85	81.03	74.52	**83.26**
2	69.83	70.14	75.17	**75.69**
3	52.08	**61.87**	57.15	54.13
4	83.64	82.73	**95.00**	94.55
5	85.94	**86.08**	84.99	84.28
6	79.99	70.57	**80.09**	79.33
7	44.39	**71.29**	61.06	65.74
OA	73.59 ±2.08	71.99 ±3.25	76.18 ±0.88	**76.64** **±1.18**
AA	68.96 ±5.08	74.82 ±3.64	75.85 ±1.96	**76.71** **±1.99**
Kappa	57.69 ±3.87	58.68 ±3.52	63.13 ±1.45	**63.62** **±1.47**

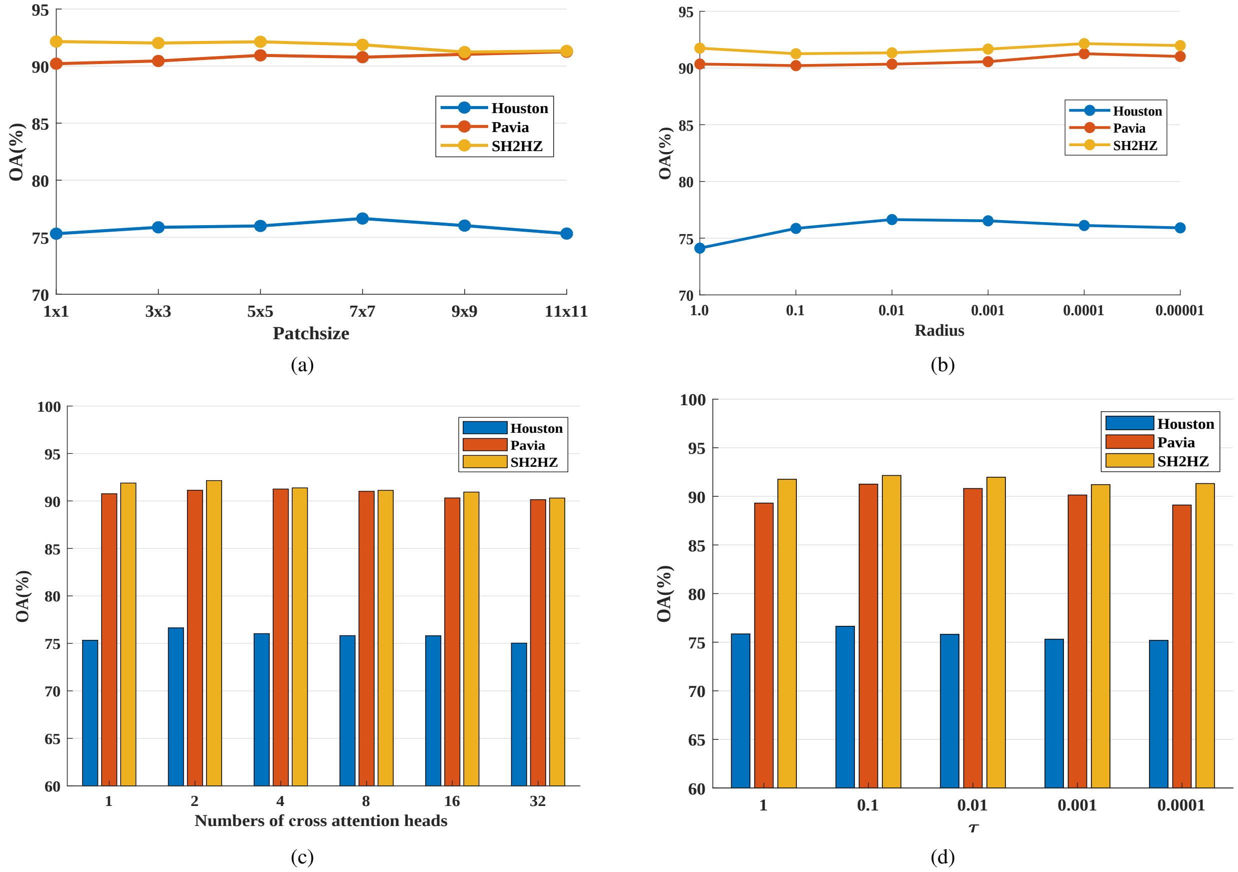
E. Parameter and Sensitivity Analysis

We conduct a sensitivity analysis on critical parameters at various levels within the proposed framework, including the patch size of the input data, the filtering radius in the GuidedPGC module, the number of heads in the cross attention of MBCA, and the temperature parameter (τ) in the SCL. In the following, we adjust these parameters individually while keeping other parameters at their optimal values to observe their influence on OA.

We test seven different patch sizes as 1×1, 3×3, 5×5, 7×7, 9×9, 11×11, and 13×13. The variation in classification accuracy is illustrated in Fig. 11(a). For the Pavia task, the OA gradually increased with growing patch size, reaching its peak at the size of 11 × 11. In the Houston task, the OA initially increases with increasing patch size, reaching its maximum at 7×7, and then decreases. For the SH2HZ task, increasing the patch size results in a decrease in classification accuracy. Thus, for the SH2HZ task, the patch size is set to 1×1.

Regarding GuidedPGC for the image-level domain adaptation, we test six different filtering radius parameters including 1.0, 0.1, 0.01, 0.001, 0.0001, and 0.00001. As depicted in Fig. 11(b), for the Pavia and SH2HZ, the best performance occurs when the radius equals 0.0001, and the highest performance is achieved at 0.01 for the Houston. Furthermore, the influence of the radius on the Houston dataset is more pronounced than on the Pavia and SH2HZ datasets, possibly because of the relatively sparse nature of its source domain.

In the proposed MBCA, multiple heads can calculate diverse attention to enhance the feature representation. We test six different numbers of cross attention heads as 1, 2, 4, 8, 16,



(a)

(b)

(c)

(d)

Fig. 11. Sensitivity analysis of parameters. (a) Patchsize, (b) Radius, (c) Numbers of cross attention heads, (d) τ.

TABLE VIII: ABLATION EXPERIMENTS ON PAVIA DATESET

Class	AB	AC	BC	ABC
1	**93.99**	90.92	92.28	93.02
2	90.66	95.97	96.22	**96.36**
3	**99.95**	99.37	97.43	96.81
4	84.48	83.36	85.29	**85.42**
5	99.76	99.62	**99.94**	99.91
6	96.53	97.54	**97.87**	97.45
7	70.97	77.18	78.08	**79.97**
OA	88.56	90.13	90.82	**91.26**
	±1.23	±0.85	±0.63	**±0.53**
AA	90.9	91.99	92.44	**92.70**
	±0.92	±0.64	±0.74	**±0.80**
Kappa	86.45	88.24	89.12	**89.63**
	±1.45	±1.00	±0.75	**±0.63**

and 32. Fig. 11(c) illustrates the change in OA as the number of heads increases. All three datasets show optimal accuracy when the number of heads is set to 2. The overall effect of the number of heads on performance is relatively small, indicating low sensitivity to this parameter.

TABLE IX: ABLATION EXPERIMENTS ON SH2HZ DATESET

Class	AB	AC	BC	ABC
1	99.69	99.68	**99.97**	99.94
2	88.26	85.72	89.12	**89.90**
3	89.83	91.78	93.85	**93.05**
OA	90.23	89.37	91.95	**92.15**
	±1.07	±2.58	±1.06	**±0.94**
AA	92.59	92.39	94.11	**94.46**
	±0.82	±1.94	±0.79	**±0.63**
Kappa	83.37	82.19	86.35	**86.65**
	±1.68	±4.17	±1.71	**±1.49**

The temperature coefficient, denoted as τ, is a critical parameter in the SCL module. We test five different temperature coefficients including 1, 0.1, 0.01, 0.001, and 0.0001. As shown in Fig. 11(d), our experiments reveal that all three tasks performed best when the value of τ is set to 0.1. As the value of τ decreases, the performance for each task consistently declines, indicating the importance of the temperature coefficient.

F. Analysis of the Computational Complexity

In this section, we analyze the computational complexity of the proposed MLUDA. Specifically, we compare the training efficiency of MLUDA with related cross-domain HSIC methods, e.g., CLDA, CDA, and TAADA. All methods are performed using the NVIDIA GeForce GTX 4090 GPU and the Intel Xeon Silver 4210R CPU. And the experiments utilize the open-source deep learning framework PyTorch, Version 2.11. Table X shows the training time of different algorithms

TABLE X: EXECUTION TIME (IN SECONDS) OF ONE EPOCH TRAINING OF DIFFERENT METHODS

Datasets	DAN	DANN	DSAN	CDA	TAADA	CLDA	MLUDA
HOUSTON	18.91	9.11	14.15	14.37	14.23	13.19	16.66
PAVIA	28.13	6.43	21.21	22.23	21.01	19.46	23.98
SH2HZ	17.35	9.48	12.14	15.14	12.31	9.33	13.44

for one epoch with identical parameter settings. It can be observed that the proposed MLUDA incurs a relatively longer training time compared to other algorithms. As an early UDA method, DANN has a much shorter training time than DAN due to the adversarial structure, while DAN has the longest training time.

V. Conclusion

In this paper, we propose a new MLUDA framework for cross-scene HSIC. First, we design a GuidedPGC module based on classic image matching techniques and guided filter to achieve image-level domain adaptation, which enhances the classification accuracy and improves the robustness of the model. Second, at the feature level, we introduce a cross attention structure called MBCA for HSIC, enhancing the interaction of features between the source and target domain. Third, at the logic level, we implement the SCL strategy based on pseudo-labels and LMMD loss to further increase the inter-domain class distance, thereby making the domain adaptation of previous levels more achievable. The proposed MLUDA features a clear and effective multi-level structure to reduce the gap between the source and target domains, and the experiments show its superiority over other related methods. In the future, we will consider migrating the multi-level UDA structure into different tasks such as semantic segmentation and change detection.

References

[1] J. Peng, W. Sun, H.-C. Li, W. Li, X. Meng, C. Ge, and Q. Du, "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 10–43, 2022.

[2] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7726–7738, 2020.

[3] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[4] Y. Tang, S. Feng, C. Zhao, Y. Fan, Q. Shi, W. Li, and R. Tao, "An object fine-grained change detection method based on frequency decoupling interaction for high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[5] C. Zhao, B. Qin, S. Feng, W. Zhu, L. Zhang, and J. Ren, "An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[6] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, X. Jia, A. Plaza *et al.*, "Spectralgpt: Spectral foundation model," *arXiv preprint arXiv:2311.07113*, 2023.

[7] B. Xi, J. Li, Y. Li, R. Song, D. Hong, and J. Chanussot, "Few-shot learning with class-covariance metric for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 5079–5092, 2022.

[8] B. Xi, J. Li, Y. Li, R. Song, Y. Xiao, Q. Du, and J. Chanussot, "Semisupervised cross-scale graph prototypical network for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 9337–9351, 2023.

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

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

[11] L. Fang, D. Zhu, J. Yue, B. Zhang, and M. He, "Geometric-spectral reconstruction learning for multi-source open-set classification with hyperspectral and lidar data," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 10, pp. 1892–1895, 2022.

[12] J. Yan, C. Deng, H. Huang, and W. Liu, "Causality-invariant interactive mining for cross-modal similarity learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[13] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 2023.

[14] J. Yan, L. Luo, C. Deng, and H. Huang, "Adaptive hierarchical similarity metric learning with noisy labels," *IEEE Transactions on Image Processing*, vol. 32, pp. 1245–1256, 2023.

[15] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[16] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2381–2392, 2015.

[17] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, 2017.

[18] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, pp. 1–12, 2015.

[19] S. M. Jun Yue, Wenzhi Zhao and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sensing Letters*, vol. 6, no. 6, pp. 468–477, 2015.

[20] M. He, B. Li, and H. Chen, "Multi-scale 3d deep convolutional neural network for hyperspectral image classification," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3904–3908.

[21] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277–281, 2019.

[22] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5046–5063, 2018.

[23] Y. Peng, Y. Zhang, B. Tu, Q. Li, and W. Li, "Spatial–spectral transformer with cross-attention for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[24] G. Farooque, Q. Liu, A. B. Sargano, and L. Xiao, "Swin transformer with multiscale 3d atrous convolution for hyperspectral image classification," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107070, 2023.

[25] X. Huang, M. Dong, J. Li, and X. Guo, "A 3-d-swin transformer-based hierarchical contrastive learning method for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[26] Y. Peng, Y. Zhang, B. Tu, Q. Li, and W. Li, "Spatial–spectral transformer

with cross-attention for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[27] P. Duan, X. Kang, P. Ghamisi, and S. Li, "Hyperspectral remote sensing benchmark database for oil spill detection with an isolation forest-guided unsupervised detector," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[28] X. Tang, C. Li, and Y. Peng, "Unsupervised joint adversarial domain adaptation for cross-scene hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[29] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE geoscience and remote sensing magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[30] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2960–2967.

[31] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[32] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," 2016.

[35] Y. Zhu, F. Zhuang, J. Wang, G. Ke, J. Chen, J. Bian, H. Xiong, and Q. He, "Deep subdomain adaptation network for image classification," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 4, pp. 1713–1722, 2020.

[36] Z. Liu, L. Ma, and Q. Du, "Class-wise distribution adaptation for unsupervised classification of hyperspectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 508–521, 2021.

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

[38] Y. Huang, J. Peng, W. Sun, N. Chen, Q. Du, Y. Ning, and H. Su, "Two-branch attention adversarial domain adaptation network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[39] Z. Fang, Y. Yang, Z. Li, W. Li, Y. Chen, L. Ma, and Q. Du, "Confident learning-based domain adaptation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[40] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[41] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sensing of Environment*, vol. 299, p. 113856, 2023.

[42] Q. Wang, C. Yin, H. Song, T. Shen, and Y. Gu, "Utfnet: Uncertainty-guided trustworthy fusion network for rgb-thermal semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[43] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, and Q. Du, "A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[44] H. Ma, X. Lin, and Y. Yu, "I2f: A unified image-to-feature approach for domain adaptive semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[45] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7178–7193, 2020.

[46] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[47] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[48] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[49] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[50] Q. Liu, J. Peng, G. Zhang, W. Sun, and Q. Du, "Deep contrastive learning network for small-sample hyperspectral image classification," *Journal of Remote Sensing*, vol. 3, p. 0025, 2023.

[51] P. Guan and E. Y. Lam, "Spatial-spectral contrastive learning for hyperspectral image classification," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 1372–1375.

[52] Q. Liu, J. Peng, Y. Ning, N. Chen, W. Sun, Q. Du, and Y. Zhou, "Refined prototypical contrastive learning for few-shot hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[53] Z. Li, Q. Xu, L. Ma, Z. Fang, Y. Wang, W. He, and Q. Du, "Supervised contrastive learning-based unsupervised domain adaptation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.