Domain Adaptation for Multilabel Remote Sensing Image Annotation With Contrastive Pseudo-Label Generation

Rui Huang, *Member, IEEE*, Mingyang Ma, and Wei Huang

***Abstract*—Deep-learning-based multilabel remote sensing image annotation (MLRSIA) is receiving increasing attention in recent years. MLRSIA needs a large volume of labeled samples for effective training of the deep models. However, the scarcity of labeled samples is a common challenge in this field. Domain adaptation (DA), aiming to transfer knowledge from label-rich datasets (source domains) to label-scarce datasets (target domains), has become an effective means to address this problem of limited labeled samples. But most of the existing DA models are primarily designed for single-label annotation tasks, leaving the application of DA to multilabel annotation tasks as an open issue. In this article, a DA method for MLRSIA, named contrastive pseudo-label generation (CPLG), is proposed. CPLG mainly consists of two parts: generating and selecting pseudo-labels for the samples in the target domain, and enhancing the cross-domain feature consistency through contrastive learning. Specifically, the soft predictions (or posterior probabilities) and the corresponding pseudo-labels of the target samples are first generated using neighborhood aggregation. Then, a positive and negative pseudo-label selection strategy is designed to refine these pseudo-label. Finally, a contrastive loss is introduced to align the similar sample features between the source and target domains to avoid the pseudo-labels of the target samples being overly biased toward the source domain, further improving the precision of these pseudo-labels. The MLRSIA experiments, conducted across four different DA scenarios on three benchmark datasets, demonstrate the advantages of the proposed CPLG compared to other state-of-the-art methods.**

***Index Terms*—Contrastive learning, domain adaptation (DA), multilabel image annotation, neighborhood aggregation (NA), remote sensing images.**

I. Introduction

With the rapid advances in earth observation technology, large quantities of high-resolution and content-rich remote sensing images have become available for land cover monitoring, urban planning, disaster prediction, and intelligent transportation [1], [2], [3], [4]. To fully utilize the ever-increasing remote sensing images, it is urgent to classify and interpret these images and obtain the associated semantic information from them. It is common for a single remote sensing image to contain multiple land covers, so the single-label annotation [5], [6], [7], which assigns a label to the image, is insufficient to fully understand it. Therefore, multilabel remote sensing image annotation (MLRSIA) [8], [9], [10] has received increasing attention and become an active research area. Fig. 1 specifically illustrates the difference between single-label and multilabel annotations. Single-label annotation typically assigns a relatively general single label to describe the entire image, such as “river,” “bridge,” “grassland,” etc. However, multilabel annotation describes the images with more specific object-level labels, aiming to identify fine-grained information.

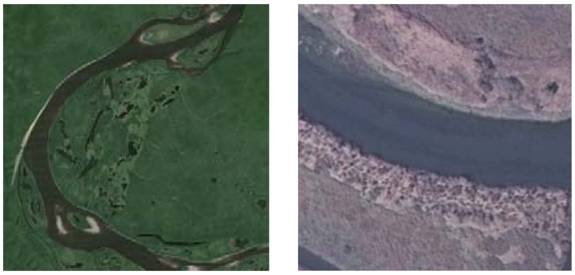
(a) (b)

Fig. 1. Example images from two remote sensing datasets. Scene label: river. Object label: bare soil, grass, water. (a) Image from AID dataset. (b) Image from UCM dataset.

Despite significant progress in MLRSIA tasks in recent years, these supervised methods often struggle with the issue of limited labeled training samples. Labeling is a labor-intensive and time-consuming task, and the single-label annotated remote sensing images available could not be directly used for multilabel learning [12]. Therefore, a significant degradation in performance can be observed in the deep neural network where more training samples are needed to achieve a full-trained model with better generalization capability. To address the issue, the idea of domain adaptation (DA) [13], [14], [15], [16], [17], [18], [19], [20], [21], which aims to transfer knowledge from a label-rich dataset (source domain) to a label-scarce dataset (target domain), has recently attracted increasing attentions. However, there exists a significant distribution discrepancy between the domains of source and target, which is called the domain shift. As illustrated in Fig. 1, while the two datasets share similarities at the object level, there are distinct stylistic differences between them. Effectively learning the common features from image samples that belong to the source domain and target domain is key to successful domain knowledge transfer.

Received 5 September 2024; revised 11 October 2024; accepted 23 October 2024. Date of publication 1 November 2024; date of current version 19 November 2024. *(Corresponding author: Rui Huang.)*

The authors are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: huangr@shu.edu.cn; mamy@shu.edu.cn; lyxhw@shu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3490596

Many DA models have been proposed so far, including adversarial DA methods (such as DANN [15], MCD [16], JAN [17], CDAN [19], etc.) and pseudo-label generation DA methods (such as ATDOC [20], MixMatch [21], etc.). Adversarial DA methods first align source and target features by confusing a domain discriminator and then obtain prediction results through a shared classifier which is not specifically designed for the samples of the target domain. Therefore, these methods may fail in the fine-grained multilabel classification tasks. Pseudo-label generation DA methods, which are inspired by semisupervised learning (SSL) [22], [23], primarily generate pseudo-labels for unlabeled target samples to guide the model training for target domain data. Compared to the adversarial DA methods, these methods directly focus on classifier training to enhance DA performance without addressing the issue of feature alignment. Therefore, the methods inevitably lean toward source data during the training process and result in low-quality pseudo-labels for samples in the target domain. In addition, most existing DA methods are primarily tailored for single-label classification tasks. Single-label learning, where each instance is only assigned with a single label, can be deemed as a simplified version of multilabel learning. Consequently, multilabel learning poses greater challenges, and the same is true for DA in multilabel learning scenarios. In MLRSIA tasks, an image may belong to multiple classes, and thus the boundaries between these classes are less distinct. Compared to the single-label DA, the label space in multilabel DA is significantly larger because each sample can be associated with a subset of labels, leading to a more complex label-level alignment or adaptation.

In this article, we propose an unsupervised DA method for multilabel learning, which is named contrastive pseudo-label generation (CPLG). In the proposed framework, the neighborhood aggregation (NA) idea is first used to generate the soft predictions (or posterior probabilities) for the target domain samples. Specifically, for a target sample, its prediction is obtained by averaging the posterior probabilities of its multiple nearest neighbors. Then, a positive–negative pseudo-label confidence selection strategy is designed to refine the target sample's pseudo-label based on its soft prediction. Next, we introduce a contrastive loss [24], [25], [26] to align features of similar samples between source and target domains, aiming to address the issue that the generated pseudo-labels of target samples are overly biased toward source domain and to further improve the precision of these pseudo-labels.

The main contributions of this article can be summarized as follows.

1) Presenting an unsupervised DA method for MLRSIA, which involves NA-based pseudo-label generation for target samples and contrastive-learning-based consistency enhancement for feature alignment across domains.
2) Designing a strategy for multilabel positive-negative pseudo-label confidence selection to obtain credible pseudo-labels of target samples for network training.
3) Introducing contrastive learning to enhance feature consistency between samples with similar labels across domains, thereby aligning feature differences.
4) Conducting extensive experiments in four DA scenarios to demonstrate the advantages of the proposed method CPLG.

II. Related Work

A. Multilabel Remote Sensing Image Annotation

Early multilabel image annotation is composed of two separate steps, namely feature extraction and classification. The handcrafted features are extracted and used as the inputs of traditional machine learning methods, such as support vector machines [27], and ML-KNN [28]. These methods usually fail due to the less representational capacities of the handcrafted features and the lack of interaction between the two steps.

In recent years, deep-learning-based methods have been widely applied to remote sensing image classification [29]. These methods present end-to-end frameworks with powerful feature learning abilities and have shown promising performances. In the deep learning models, the attention mechanism [30] is introduced for better feature representation. For example, Tong et al. [7] proposed a Channel-Attention DenseNet network for scene classification, Li et al. [31] presented a CNN multi-augmentation scheme based on the attention mechanism, and Yu et al. [32] used hierarchical attention and bilinear pooling for feature fusion in remote sensing image classification. In addition, dependencies among different labels, namely label correlations, are helpful for performance improvement for multilabel learning. The CNN-RNN framework proposed by Wang et al. [33] allow the RNN model to learn joint image-label embeddings from CNN features and use the memory mechanism of RNN to predict labels in an orderly prediction path. Hua et al. [34] fed the image features into a bidirectional LSTM for classification, implicitly learning label co-occurrence information. In ML-GCN [11], label correlations are modeled as a graph and integrated into feature classification through a graph convolutional network. Liu et al. [29] added a global channel attention mechanism and label correlation fusion to the shallow feature extraction network. However, these methods need sufficient labeled samples for model training. As is known, manual annotation is labor-intensive, and thus the number of training samples is insufficient. The situation is particularly severe in MLRSIA, leading to performance degradation of deep learning models.

B. Unsupervised DA

Unsupervised DA (UDA), which aims to transfer the knowledge learned from the labeled source domain to the unlabeled target domain, presents a promising solution to the issue of lack of training samples in MLRSIA.

In UDA, there exists the domain shift or distribution difference between source and target data. To minimize domain shift, various UDA methods have been proposed. Alignment-based methods, such as maximum mean discrepancy [35] and HΔH-distance [36], reduce domain shift by minimizing the differences in statistical distribution measures. Inspired by the generative adversarial networks [37], adversarial-based methods introduce domain discriminators and use the idea of adversarial learning for domain confusion [16], [17], [18], [19]. In the field of remote sensing, Du et al. [38] utilized middle-layer feature extraction to address feature heterogeneity in remote sensing images. Lin et al. [12] combined the idea of an adversarial domain discriminator with GCN for multilabel classification. DAA-MLIC [39] leverages the classifier's predicted probability distribution as the basis for domain adversarial alignment, effectively eliminating the need for an additional discriminator. However, the discriminator only aligns global domain statistical information and ignores key semantic information on each category, making it relatively coarse-grained for DA. Sample-matching-based methods can effectively address this issue by matching samples with similar categories through the analysis of their features. For example, ATDOC [20] generates pseudo-labels for target samples by comparing feature similarities. Clustering-based methods [40], [41] apply unsupervised clustering for the unlabeled samples to reduce domain shift. PCLUDA [42] leverages pseudo-labels and consistency regularization to enhance retrieval performance while minimizing class confusion without relying on adversarial learning. However, these fine-grained DA methods are mostly designed for single-label classification.

C. Contrastive Learning

Contrastive learning [24], [25], aiming to align categories by comparing distances between samples, has been widely studied in self-supervised learning [43], SSL [44], and supervised learning [45]. In these works, the basic idea is to encourage similar sample pairs to have more similar feature representations in the feature space, while also separating the features of dissimilar sample pairs. It is noted that sample pairs all come from the same domain. In UDA, contrastive learning is used for sample pairs crossing domains. SpCL [46] proposes a self-paced contrastive learning framework, jointly distinguishing source domain classes, target domain clusters, and noncluster instances. In CPGA [47], representative features for each source class are generated from the source model rather than source samples, and each pseudo-labeled target data are aligned to source representative features via contrastive learning. PCS [41] aggregates unlabeled samples through intradomain and cross-domain contrastive loss. CMFT [25] solves the class imbalance problem in DA through a centroid memory-based directed memory transfer mechanism and fine-grained neighborhood prototypes. MemSAC [48] proposes a variant of contrastive loss to improve DA classification performance in the case of a large number of categories. SRKT [49] derives domain-invariant as well as discriminative representations by adversarial pattern and contrastive learning. However, these studies also mainly focus on single-label classification.

III. Methodology

In the article, the effectiveness of UDA in MLRSIA tasks is explored and a DA method CPLG using ideas of sample matching and contrastive learning is proposed. In this section, we will discuss the details of the proposed CPLG framework.

A. Problem Definition

In the UDA task for MLRSIA, two domains, namely source domain and target domain with different distributions, are involved in the model's training and test procedures. During the training phase, all the labeled source data and part of the unlabeled target data are input into the model. During the test phase, the test samples of the target data are fed to the trained model for performance evaluation.

The source domain dataset is represented as $\boldsymbol{D}_S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)|1 \leq i \leq n_s\}$, where n_s is the number of source samples and $y_i = \{y_i^1, y_i^2, \ldots, y_i^C\} \in \{0,1\}^C$ is the C-dimensional label vector of the sample $\boldsymbol{x}_i \in R^D$. For the sample $\boldsymbol{x}_i$, $y_i^c = 1$ indicates that category c is associated with it, while $y_i^c = 0$ indicates that the category does not belong to it. The target domain dataset is represented as $\boldsymbol{D}_T = \{\boldsymbol{x}_j|1 \leq j \leq n_t\}$, where n_t is the number of target samples. In this article, we mainly study closed-set DA, that is, all categories in the source domain and target domain are known and exactly the same. Our goal is to better transfer the knowledge learned in the source domain to the target domain in the same label space.

As shown in Fig. 2, the framework of CPLG includes a feature extractor G and a classifier F, with network parameters θ_G and θ_F, respectively. In addition, two nonparametric memory banks [50], [51] named source memory bank and target memory bank are used to store the features and corresponding labels for the source samples or posterior probabilities for the target samples. In CPLG, two stages of transfer learning are involved: 1) Generating confident pseudo-labels for the target domain through NA and pseudo-label confidence selection to overcome domain classifier bias, and seamlessly integrating output-level DA with feature-level DA. 2) Introducing contrastive DA to enhance the feature similarity of similar samples across domains to achieve fine-grained interdomain consistency enhancement.

B. Generation of Pseudo-Labels for Target Domain Data

The NA classifier [20], inspired by the idea of passing message via neighbors, is an effective classifier that characterizes local data structures. In DA, this classifier is used to describe the data structure of target domain through the neighborhood centers of samples with local similarity. It needs to build a global memory bank to store features and their posterior probabilities of all target samples. Based on the classifier for single-label learning, we further design a multilabel NA classifier by introducing a novel positive–negative pseudo-label confidence selection strategy to obtain more reliable pseudo-labels for the training samples in the target domain to guide the training process.

1) Multilabel Pseudo-Label Generation via NA: Memory bank initialization and update. The memory bank is essentially a key-value store. In the target memory bank, the key is the feature representation of each target sample, and the value corresponds to its posterior probability predicted by the classifier. During training, an iterative update is applied to the memory bank with a mini-batch of samples for consistency between the features and predictions. The initial posterior probabilities in the target memory bank are set to 0.5. For each target sample in the mini-batch, the feature vector and probability vector are updated

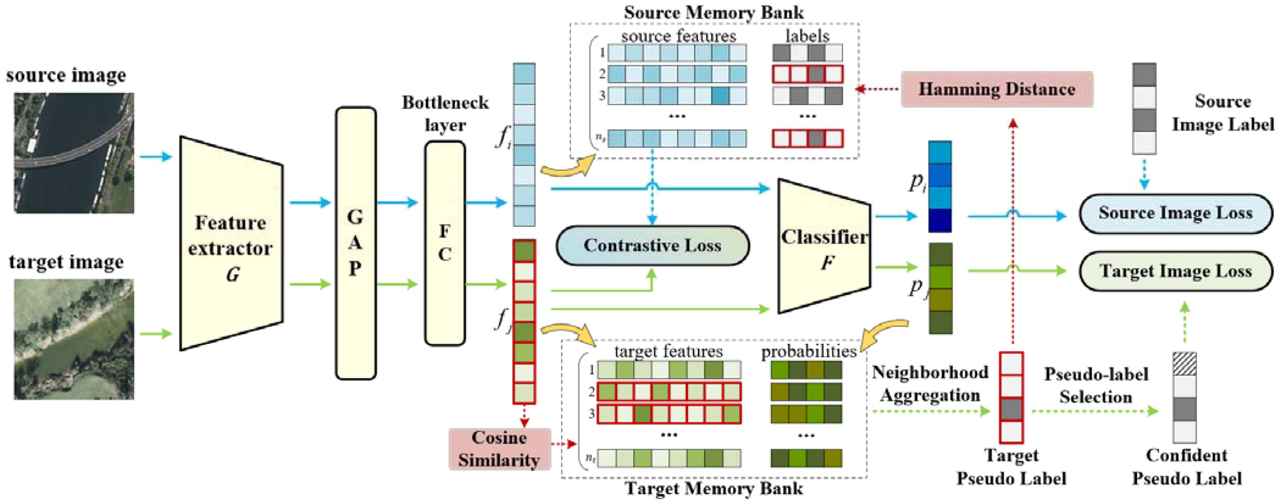

Fig. 2. Framework of our proposed method CPLG, which consists of a shared feature extractor and classifier, as well as the source and target memory banks. Source images and target images are input into the network, and their forward propagation processes are indicated in blue and green arrows, respectively. The role of the bottleneck layer is to reduce the dimensionality of image feature obtained from the feature extractor and GAP operation. Yellow arrows indicate the update procedures of the two memory banks. The target features with red borders have similar data distributions. In the confident pseudo-labels obtained from pseudo-label selection, dark color represents negative labels, light color represents positive labels, and shaded areas represent less confident labels. There are three kinds of losses, namely source image loss, target image loss, and contrastive loss. The parts enclosed by the dashed lines do not participate in the gradient computation.

according to the sample's global index which records the storage location of the sample in the bank.

Pseudo-label generation based on NA. The multilabel pseudo-label generation adopts a nonparametric NA strategy which greatly reduces the training time. First, the feature of the target sample $\boldsymbol{x}_j$ in the current mini-batch is obtained through $\boldsymbol{f}_j = G(\boldsymbol{x}_j)$, and the corresponding initial posterior probability is obtained through $\boldsymbol{p}_j = F(\boldsymbol{f}_j)$. Then, the cosine similarity metrics between the feature and all other features in the target memory bank are calculated. Based on the metrics, the top *m* nearest neighbors of sample $\boldsymbol{x}_j$ are retrieved. The clustering center of these *m* nearest neighbors' posterior probabilities is defined as the final posterior probability of sample $\boldsymbol{x}_j$ as follows:

$$\tilde{\boldsymbol{p}}_j = \frac{1}{m} \sum_{k \in N_j,\, k \neq j} \boldsymbol{p}_k \tag{1}$$

where N_j is the set of *m* nearest neighbor indices of the target sample $\boldsymbol{x}_j$ in the memory bank, $\boldsymbol{p}_k \in R^C$ $(k \in N_j)$ is the posterior probability of the sample in the nearest neighbors of $\boldsymbol{x}_j$. The pseudo-label $\tilde{\boldsymbol{y}}_j = \{\tilde{y}_i^1, \tilde{y}_i^2, \ldots, \tilde{y}_i^C\} \in \{0,1\}^C$ of the target sample $\boldsymbol{x}_j$ can be generated by thresholding the probability with 0.5.

Based on the true labels of the source samples and the pseudo-labels of the target samples, the multilabel binary cross-entropy loss is used as the objective function for the source sample $\boldsymbol{x}_i$ and target sample $\boldsymbol{x}_j$, respectively

$$L_s = -\frac{1}{C} \sum_{c=1}^{C} \left[y_i^c \log\left(p_i^c\right) + \left(1 - y_i^c\right) \log\left(1 - p_i^c\right) \right] \tag{2}$$

$$L_t = -\frac{1}{C} \sum_{c=1}^{C} \left[\tilde{y}_j^c \log\left(p_j^c\right) + \left(1 - \tilde{y}_j^c\right) \log\left(1 - p_j^c\right) \right] \tag{3}$$

where p_i^c and p_j^c denote the probabilities of source sample $\boldsymbol{x}_i$ and target sample $\boldsymbol{x}_j$ belonging to class *c*, respectively.

The loss function L_s aims to increase the classification accuracy of the model based on the true labels of the source domain data, while the loss function L_t tries improve the generalization performance to the target domain with the help of the pseudo-labels of the target domain data.

2) Positive–Negative Pseudo-Label Confidence Selection: In UDA, it is necessary to obtain more reliable pseudo-labels of target samples to guide the training. Therefore, we further propose a positive–negative pseudo-label confidence selection strategy. Let $\boldsymbol{g}_j = \{g_j^1, g_j^2, \ldots, g_j^c\} \in \{0,1\}^c$ be a binary vector of the confidence level of the pseudo-label vector for the target sample $\boldsymbol{x}_j$, where $g_j^c = 1$ when $\tilde{y}_j^c$ is considered reliable, and $g_j^c = 0$ when $\tilde{y}_j^c$ is considered unreliable. g_j^c is defined as follows:

$$g_j^c = \begin{cases} 1, & \tilde{p}_j^c \geq \gamma_p \text{ or } \tilde{p}_j^c \leq \gamma_n \\ 0, & \gamma_n < \tilde{p}_j^c < \gamma_p \end{cases} \tag{4}$$

where γ_p and γ_n are the confidence thresholds for positive and negative labels. Specifically, γ_p is set to 0.6 and γ_n is set to 0.1. The probability $\tilde{p}_j^c$ exceeding γ_p or below γ_n can be determined that class *c* is a credible positive or negative label for $\boldsymbol{x}_j$. Equation (3) can be rewritten as

$$L_{t,ps} = -\frac{1}{s_j} \sum_{c=1}^{C} g_j^c \cdot \left[\tilde{y}_j^c \log\left(p_j^c\right) + \left(1 - \tilde{y}_j^c\right) \log\left(1 - p_j^c\right) \right] \tag{5}$$

where $s_j = \sum_c g_j^c$ is the number of pseudo-labels with high confidence. Strategically selecting high-confidence pseudo-labels effectively mitigates the impact of noise on the training process, thereby enhancing the correctness of pseudo-label generation. The selection strategy plays an important role in the learning

TABLE I
NUMBERS OF IMAGES IN UCM, AID, DFC15 DATASETS FOR DIFFERENT CLASSES

Classes	Dataset	Images	Training	Testing
17	UCM	2100	1680	420
	AID	3000	2400	600
8	DFC15	3342	2674	668
6	UCM	1726	1378	348
	AID	2740	2190	550
	DFC15	2190	1752	438

process, in which it curtails the propagation of errors from low-confidence predictions. Its effectiveness will be demonstrated through subsequent ablation studies.

C. Consistency Enhancement Based on Contrastive Loss

In this section, the consistency enhancement based on contrastive learning for multilabel cross-domain alignment is presented. The fine-grained sample-level information is explored to enhance the consistency of similar samples between domains. The similar sample-pairs between domains are identified, and the features associated with these pairs are expected to exhibit similarity. For each target sample in the target mini-batch $\mathcal{B}_t$, the hamming distance d_{ham} is employed to measure the similarity between its pseudo-label and the ground-truth labels of the source samples. The samples are considered similar if the hamming distance is below a predefined threshold d_{th}. Based on this criterion, a subset of source samples similar to the target sample $\boldsymbol{x}_j$ can be obtained, which is denoted as $\mathcal{B}_{s+}^{\boldsymbol{x}_j} = \{\boldsymbol{x}_i \in \mathcal{B}_s | d_{\text{ham}}(\tilde{\boldsymbol{y}}_j, \boldsymbol{y}_i) < d_{\text{th}}\}$, where $\mathcal{B}_s$ represents the mini-batch of source samples. A sample consistency loss function based on hamming distance and contrastive loss is defined as follows:

$$L_{\text{sc},\mathcal{B}} = -\frac{1}{|\mathcal{B}_t|\,|\mathcal{B}_{s+}|}\sum_{j\in\mathcal{B}_t}\log\left(\sum_{i\in\mathcal{B}_{s+}}\frac{\exp\left(\frac{\varphi_{ij}}{\tau}\right)}{\sum_{i\in\mathcal{B}_s}\exp\left(\frac{\varphi_{ij}}{\tau}\right)}\right) \tag{6}$$

where φ_{ij} is the cosine similarity between two feature vectors $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ with $\varphi_{ij} = \varphi\left(\boldsymbol{f}_i,\ \boldsymbol{f}_j\right) = \frac{\boldsymbol{f}_i^T \cdot \boldsymbol{f}_j}{||\boldsymbol{f}_i||\cdot||\boldsymbol{f}_j||}$, τ is the temperature parameter used to measure the contribution of positive and negative pairs [48]. This loss function enhances the feature similarities of similar sample-pairs between the source and target mini-batches. However, due to the complexity of label distributions in multilabel classification tasks, it is possible that there are no similar sample-pairs in a local mini-batch. An alternative solution is to expand the mini-batch size. However, it will lead to a considerable memory augmentation, impacting the scalability of the model. Inspired by the previously discussed concept of a global target memory bank, a global source memory bank is established to address the issue where no similar samples are present in the local mini-batches. In the global source memory bank, more source samples similar to the target sample can be found, and the global similar sample pairs are more in line with our expectations for domain-level style feature alignment. Based on the global sample-level similarity, the consistency loss L_{sc} evolves as follows:

$$L_{\text{sc}} = -\frac{1}{|\mathcal{B}_t|\,|\mathcal{M}_{s+}|}\sum_{j\in\mathcal{B}_t}\log\left(\sum_{i\in\mathcal{M}_{s+}}\frac{\exp\left(\frac{\varphi_{ij}}{\tau}\right)}{\sum_{i\in\mathcal{M}_s}\exp\left(\frac{\varphi_{ij}}{\tau}\right)}\right) \tag{7}$$

where $\mathcal{M}_s$ is the source memory bank, and $\mathcal{M}_{s+} = \{\boldsymbol{x}_i \in \mathcal{M}_s | d_{\text{ham}}(\tilde{\boldsymbol{y}}_j, \boldsymbol{y}_i) < d_{\text{th}}\}$ is the subset of source samples similar to the target sample $\boldsymbol{x}_j$. To ensure the source memory bank $\mathcal{M}_s$ contains reliable and representative features, we first train the feature extractor with L_s and $L_{t,ps}$ losses for several epochs, and then incorporate the consistency loss L_{sc} to align the features of similar samples across domains. The source memory bank $\mathcal{M}_s$ is updated in the same way as the targe memory bank $\mathcal{M}_t$, in which the update happens at the end of each training iteration in mini-batch sizes.

D. Training Procedure

The loss function of CPLG consists of three components: the source domain loss L_s, the target domain loss $L_{t,ps}$, and the contrastive loss L_{sc}. It is can be formulated as

$$L_{\text{CPG}} = L_s\ + \lambda_t L_{t,ps} + \lambda_{\text{sc}} L_{\text{sc}} \tag{8}$$

where λ_t and λ_{sc} are the hyperparameters to balance the loss. The model is trained in an end-to-end way. The source domain loss L_s is designed to boost the network's ability to accurately classify images from the source domain. The target domain loss $L_{t,ps}$ is intended to promote better generalization of the model from the source to the target domain. Meanwhile, the contrastive loss L_{sc} serves to minimize the feature space discrepancies between similar samples across domains, thereby facilitating sample-level domain feature alignment.

In the early stage of training, the classification performance of the model is poor, and thus the quality of pseudo-labels generated by NA is less satisfactory. Therefore, a dynamically increasing weight parameter $\lambda_t \in (0, 1)$ is applied to L_t. The parameter ensures that the influence of the target domain loss on the network training is small at the beginning of the training but will continuously strengthen with the ongoing training. This indicates that the training of the model initially focuses more on improving the classification accuracy of the source domain data, and gradually shifts the emphasis to inter-domain generalization and feature alignment.

For the same reason, the contrastive loss weight λ_{sc} is set to 0 during the first 10 training epochs in which the source and target memory banks are updated. After 10 epochs of training, λ_{sc} is set to 0.01 to refine the feature alignment across domains.

IV. EXPERIMENT

In this section, we conduct DA experiments on three multilabel remote sensing image datasets to validate the performance of our method CPLG. The experiments are primarily divided into two parts. First, CPLG is compared with other state-of-the-art multilabel DA methods. Second, the roles of different components of our proposed method are evaluated.

TABLE II
COMPARISONS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS FROM UCM TO AID (MEAN%±STD%)

Method	UCM→AID								
	OP	OR	OF1	OF2	CP	CR	CF1	CF2	mAP
SIGNA	70.56±0.54	46.36±1.09	55.96±0.81	49.77±1.03	60.58±1.32	47.27±1.03	53.10±0.82	49.44±1.16	53.13±0.97
CDAN	80.79±0.72	61.06±0.60	69.55±0.47	64.20±0.64	65.31±1.21	58.92±0.72	61.95±0.67	60.10±1.87	70.01±1.44
ATDOC	**82.65±0.41**	52.73±0.29	64.38±0.25	56.84±0.31	63.52±0.87	51.84±1.09	57.09±0.74	53.82±1.07	73.19±0.43
DA-MAIC	72.68±1.01	47.13±1.87	57.18±1.41	50.69±1.77	57.84±1.45	45.92±1.28	51.20±0.98	47.89±1.37	74.43±1.16
CPLG	82.07±0.45	66.23±0.61	73.30±0.42	68.89±0.59	68.18±0.28	63.71±0.39	65.87±0.25	64.56±0.38	75.57±0.65
w/ CDAN	81.54±1.00	**68.95±0.77**	**74.72±0.62**	**71.15±0.87**	**69.71±0.35**	**66.14±0.83**	**67.88±0.47**	**66.82±0.73**	**76.17±0.59**

TABLE III
COMPARISONS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS UCM TO DFC15 (MEAN%±STD%)

Method	UCM→DFC15								
	OP	OR	OF1	OF2	CP	CR	CF1	CF2	mAP
SIGNA	42.37±1.84	42.86±1.56	42.61±1.21	42.76±1.93	43.91±1.45	37.13±1.19	40.24±0.93	38.13±1.36	50.06±1.24
CDAN	46.37±1.74	50.06±1.37	48.14±1.13	49.28±1.91	51.75±1.47	40.28±0.98	45.30±0.84	42.15±1.17	55.41±0.65
ATDOC	44.74±1.31	46.45±0.41	45.58±0.71	46.10±1.17	49.59±1.02	35.85±0.39	41.62±0.44	37.95±0.59	53.72±0.74
DA-MAIC	30.76±1.29	46.21±0.57	36.93±0.96	41.99±2.02	45.25±1.54	26.58±1.22	33.49±1.06	28.97±1.26	61.05±0.93
CPLG	45.41±0.53	51.75±0.62	48.37±0.40	50.34±0.70	51.62±0.40	41.15±0.81	45.79±0.52	42.89±0.74	57.46±0.71
w/ CDAN	**51.41±0.59**	**67.90±1.01**	**58.52±0.53**	**63.81±1.01**	**54.75±1.14**	**50.73±0.93**	**52.66±0.73**	**51.49±1.03**	**61.63±1.19**

TABLE IV
COMPARISONS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS FROM AID TO UCM (MEAN%±STD%)

Method	AID→UCM								
	OP	OR	OF1	OF2	CP	CR	CF1	CF2	mAP
SIGNA	52.37±0.88	69.61±0.97	59.77±0.68	65.31±1.29	51.22±1.76	59.69±1.46	55.31±1.19	57.78±2.08	57.45±1.22
CDAN	60.31±1.54	76.81±0.99	67.57±1.04	72.83±1.92	55.82±0.99	66.27±1.98	60.60±1.02	63.88±1.79	68.62±1.49
ATDOC	56.42±0.14	73.16±0.81	63.71±0.33	69.06±0.61	52.51±1.64	63.78±1.43	57.60±1.14	61.15±2.06	69.68±1.87
DA-MAIC	52.25±1.59	73.17±1.87	60.97±1.23	67.75±2.28	49.52±1.09	60.30±1.00	54.38±0.77	57.78±1.39	66.88±1.06
CPLG	60.63±1.33	**80.36±1.09**	69.11±0.96	75.45±1.83	54.09±0.98	67.31±1.59	59.98±0.88	64.17±1.58	71.83±1.09
w/ CDAN	**64.03±0.80**	77.03±1.20	**69.93±0.68**	**74.02±1.31**	**57.41±1.24**	**67.73±1.32**	**62.14±0.92**	**65.38±1.66**	**73.86±0.87**

TABLE V
COMPARISONS OF THE PROPOSED METHOD WITH STATE-OF-THE-ART METHODS AID TO DFC15 (MEAN%±STD%)

Method	AID→DFC15								
	OP	OR	OF1	OF2	CP	CR	CF1	CF2	mAP
SIGNA	40.81±1.97	59.62±2.32	48.45±1.59	54.59±3.25	42.76±1.43	30.12±1.49	35.34±1.13	32.01±1.49	48.48±2.01
CDAN	42.81±1.78	73.46±0.79	54.10±1.45	64.26±3.01	44.73±2.04	35.38±1.37	39.51±1.17	36.92±1.55	50.17±1.95
ATDOC	38.17±0.71	62.95±0.95	47.52±0.62	55.71±1.42	43.56±1.49	32.35±0.95	37.13±0.83	34.11±1.12	50.00±0.39
DA-MAIC	42.58±2.24	80.70±1.99	55.75±1.88	68.44±3.28	45.73±1.54	34.59±1.53	39.39±1.15	36.36±1.56	54.18±1.95
CPLG	45.49±0.79	82.59±0.89	58.67±0.69	71.01±1.83	46.38±1.51	39.66±1.32	42.76±1.03	40.84±1.39	56.76±1.04
w/ CDAN	**49.43±1.24**	**85.59±1.03**	**62.67±1.06**	**74.67±2.50**	**48.96±0.97**	**43.06±1.34**	**45.82±0.95**	**44.12±1.28**	**58.43±1.43**

A. Experiment Details

1) Datasets: We conduct classification experiments on three multilabel remote sensing image datasets, namely UCM [52], AID [2], and DFC15 [53]. The details of these datasets are as follows.

a) UCM multilabel dataset: The UCM multilabel dataset is reannotated based on the UCM dataset [52], which originates from aerial images provided by the US Geological Survey National Map. It contains 2100 images, with a size of 256 × 256 × 3 and a spatial resolution of 0.3 m. There are 17 object-level labels, including airplane, bare-soil, buildings, cars, chaparral, court, dock, field, grass, mobile-home, pavement, sand, sea, ship, tanks, trees, and water.

b) AID multilabel dataset: The AID multilabel dataset is reannotated from the AID dataset [2] released by Wuhan University in 2017. It contains 3000 multilabel images with an image size of 600 × 600 × 3 and a spatial resolution of 0.5–8 m. There are 17 labels which are the same as those of the UCM dataset.

c) DFC15 multilabel dataset: The DFC15 multilabel dataset [53] was first used in the 2015 IEEE GRSS Data Fusion Contest. It contains 3342 images with a higher spatial resolution (5 cm) than the UCM dataset. The label set includes eight objects: building, boat, car, clutter, impervious, water, vegetation, and trees.

There exist six classes, namely water, grass, building, trees, boat, and car, in all three datasets. For the closed-set UDA

TABLE VI
COMPONENT EFFECTIVENESS EVALUATION OF CPLG (MEAN%±STD%)

Method	OP	OR	OF1	OF2	CP	CR	CF1	CF2	mAP
					UCM→AID				
CPLG	82.07±0.45	**66.23±0.61**	**73.30±0.42**	**68.89±0.59**	**68.18±0.28**	**63.71±0.39**	**65.87±0.25**	**64.56±0.38**	**75.57±0.65**
w/o PS	**84.34±0.64**	55.38±1.11	66.86±0.84	59.46±1.06	64.75±0.94	54.28±1.07	59.05±0.74	56.09±1.08	73.48±0.69
w/o CE	80.16±1.01	64.17±0.39	71.28±0.47	66.84±0.66	66.26±0.39	61.33±0.45	63.70±0.30	62.26±0.46	74.33±0.51
					UCM→DFC15				
CPLG	**45.41±0.53**	**51.75±0.62**	**48.37±0.40**	**50.34±0.70**	**51.62±0.40**	**41.15±0.81**	**45.79±0.52**	**42.89±0.74**	**57.46±0.71**
w/o PS	44.75±0.50	48.79±0.82	46.68±0.46	47.92±0.78	50.01±0.81	36.01±1.35	41.87±0.96	38.15±1.27	54.64±2.01
w/o CE	43.85±0.93	49.34±1.08	46.43±0.71	48.13±1.19	47.24±1.69	37.85±1.45	42.03±1.12	39.42±1.56	55.69±1.10
					AID→UCM				
CPLG	**60.63±1.33**	**80.36±1.09**	**69.11±0.96**	**75.45±1.83**	**54.38±0.98**	**67.31±1.59**	**60.16±0.88**	**64.25±1.58**	**71.83±1.09**
w/o PS	56.43±1.49	73.68±1.99	63.91±1.21	69.43±2.29	54.09±0.75	64.97±1.45	59.03±0.75	62.46±1.33	70.13±1.33
w/o CE	54.89±2.15	78.87±1.48	64.73±1.58	72.53±3.19	52.83±1.38	66.42±1.31	58.85±1.00	63.17±1.86	69.88±1.25
					AID→DFC15				
CPLG	**45.49±0.79**	**82.59±0.89**	**58.67±0.69**	**71.01±1.83**	**46.38±1.51**	**39.66±1.32**	**42.76±1.03**	**40.84±1.39**	**56.76±1.04**
w/o PS	41.25±1.90	67.73±2.01	51.27±1.09	60.02±1.79	44.79±1.81	34.25±2.19	38.82±1.55	35.94±2.14	53.11±1.01
w/o CE	43.54±0.91	73.94±1.42	54.81±0.83	64.88±1.92	45.97±1.35	37.64±1.44	41.39±1.02	39.06±1.47	54.46±1.33

TABLE VII
CLASSIFICATION RESULTS OF THREE IMAGES FROM UCM TO AID

Samples from the AID multi-label dataset			
Ground Truth	buildings, cars, grass, pavement, trees	bare-soil, buildings, cars, grass, pavement, trees	buildings, cars, grass, pavement, trees, water
SIGNA	buildings, cars, grass, pavement, trees	bare-soil, buildings, grass, cars, pavement, trees	buildings, cars, grass, pavement, trees, water
CDAN	buildings, cars, grass, pavement, trees	bare-soil, buildings, cars, grass, pavement, trees	buildings, cars, grass, pavement, trees, water
ATDOC	buildings, cars, grass, pavement, trees, mobile-home	bare-soil, buildings, cars, grass, pavement, trees	buildings, cars, grass, pavement, trees, water
DA-MAIC	buildings, cars, grass, pavement, trees	bare-soil, buildings, cars, grass, pavement, trees	buildings, cars, grass, pavement, trees, water
CPLG	buildings, cars, grass, pavement, trees	bare-soil, buildings, cars, grass, pavement, trees	buildings, cars, grass, pavement, trees, water

task, only the six class labels are considered in the transfer experiments from UCM or AID to DFC15. After excluding images with none of the six labels from the three datasets, there are 1726, 2740, and 2190 images in them, respectively. Table I lists the total number of images for each dataset as well as the number of images when only six classes are considered. In the experiments, we randomly select 80% images of each dataset for training and the rest images for test.

2) Evaluation Metrics: The mean average precision (mAP) is adopted as the evaluation metric to assess the overall performance of the model. The precision, recall, and F-score based on samples and labels are also calculated, respectively.

For sample-based metrics, the Average Precision (OP), Recall (OR), and F-score (OF1 and OF2)

$$\mathrm{OP} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}, \ \mathrm{OR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{9}$$

$$\mathrm{OF}\beta = \left(1+\beta^2\right)\frac{\mathrm{OP}\cdot\mathrm{OR}}{\beta^2\mathrm{OP}+\mathrm{OR}}, \beta = 1, 2 \tag{10}$$

where TP_i represents the number of correctly predicted positive labels for sample $\boldsymbol{x}_i$, FP_i represents the number of positive labels that are not recognized for sample $\boldsymbol{x}_i$, FN_i represents

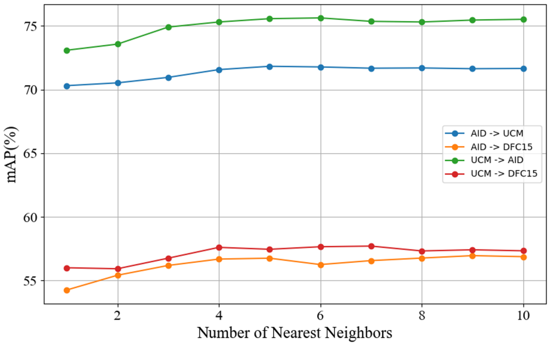

Fig. 3. mAP values in four DA scenarios under different numbers of nearest neighbors.

the number of negative labels that are incorrectly predicted for sample $\boldsymbol{x}_i$, and n represents the number of images.

For label-based metrics, the Average Precision (CP), Recall (CR) and F-score (CF1 and CF2) are formulated as

$$\mathrm{CP} = \frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c}, \ \mathrm{CR} = \frac{1}{C}\sum_{c=1}^{C}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FN}_c} \tag{11}$$

$$\mathrm{CF}\beta = \left(1+\beta^2\right)\frac{\mathrm{CP}\cdot\mathrm{CR}}{\beta^2\mathrm{CP}+\mathrm{CR}}, \ \beta = 1,2 \tag{12}$$

where TP_c represents the number of correctly predicted positive labels for class c, FP_c represents the number of positive labels that are not recognized for class c, FN_c represents the number of incorrectly predicted negative labels for class c, and C represents the number of categories in the dataset.

3) Implementation Details: We employ SGD as the network optimizer to fine-tune the feature encoder based on the ImageNet pre-trained ResNet-50 model. The learning rate is set to 0.001. The 2048-dimensional image features are then fed into the bottleneck layer and classifier. The bottleneck layer and classifier are trained from scratch using a learning rate of 0.01. The bottleneck layer's dimensionality is set to 256, and the output features are stored in the memory banks. We adopt the learning rate scheduler from ATDOC [20], setting the momentum to 0.9, weight decay to 0.001, and batch size to 16. In CPLG, the number of nearest neighbors m is set to 5, and the consistency loss temperature parameter τ is 0.07. The network is trained for a total of 40 epochs. Before conducting the classification experiments, all dataset images are processed by resizing to 256 × 256 pixels and cropping to 224 × 224 pixels.

All experiments are based on an NVIDIA GeForce GTX 1080 Ti GPU, and implemented in the PyTorch framework.

B. Quantitative Analysis

In the subsection, performance comparison of CPLG and other state-of-the-art methods and ablation analysis of CPLG are presented.

1) Comparisons With State-of-the-Art Methods: We compare our proposed method CPLG with the following state-of-the-art methods, including one MLRSIA method, two DA methods, and one DA for MLRSIA method.

SIGNA [29]: An MLRSIA method that integrates the label correlations into shallow features of the image.

CDAN [19]: An adversarial DA method that employs a ResNet-50 architecture as well as a conditional domain discriminator.

ATDOC [20]: A domain adaptation method inherently designed for single-label classification and modified for multilabel classification.

DA-MAIC [12]: A domain adaption method for MLRSIC which integrates MLGCN and DANN.

The experiments are carried out in four scenarios. The first two are with the UCM dataset as the source domain and the AID and DFC15 datasets as the target domains, respectively; the last two are with the AID dataset as the source domain and the UCM and DFC15 datasets as the target domains, respectively. The experimental results are reported as the mean and standard deviation over 10 runs.

Tables II–V show the experimental results of six methods across different DA scenarios. In the tables, CPLG+CDAN represents the proposed CPLG method combined with the CDAN domain discriminator. The best value is in bold and the second-best value is underlined.

From the tables, it can be seen that CPLG and CPLG+CDAN achieve very competitive results on all metrics. Since we have 4 scenarios and 9 evaluation metrics, there are a total of 36 comparing conditions. Among them, CPLG ranks in the top two 31 times and CPLG+CDAN ranks first 33 times. Among the five methods, SIGNA performs worse than the other DA methods, indicating that the DA methods can effectively align interdomain features. In UCM→AID, CPLG and CPLG+CDAN perform better than the other methods in terms of all metrics except for OP where ATDOC outperforms them. In UCM→DFC15, CPLG+CDAN achieves the best results on all metrics, while CPLG is inferior to CDAN on OP and CP and to DA-MAIC on mAP. In AID→UCM and AID→DFC15, both CPLG+CDAN and CPLG almost attained the best or second-best metric values. Compared to the other DA methods, CPLG introduces the classifier with a pseudo-label confidence selection strategy to better generalize toward the target domain by aligning the feature representations, and consistency enhancement based on contrastive loss to minimize feature discrepancies between analogous samples across different domains. Therefore, it shows notable performance improvements.

Comparing Tables III and V with the DFC15 dataset as the target domain, it can be observed that when all methods except DA-MAIC use the UCM dataset as the source domain, they perform better than when the AID dataset is used as the source domain. This may suggest that the spatial resolutions of the source and target domains can affect the DA performance. Specifically, the spatial resolutions of the UCM, AID, and DFC15 datasets are 0.3, 0.5-8, and 0.05 m, respectively. When the images in the AID and DFC15 datasets are resized to 256 × 256 pixels, their spatial resolutions will be reduced. The spatial resolution of the DFC15 dataset becomes closer to that of the UCM dataset and remains higher than that of the AID dataset. Better performance can be achieved when the source and target domains exhibit similar spatial resolution.

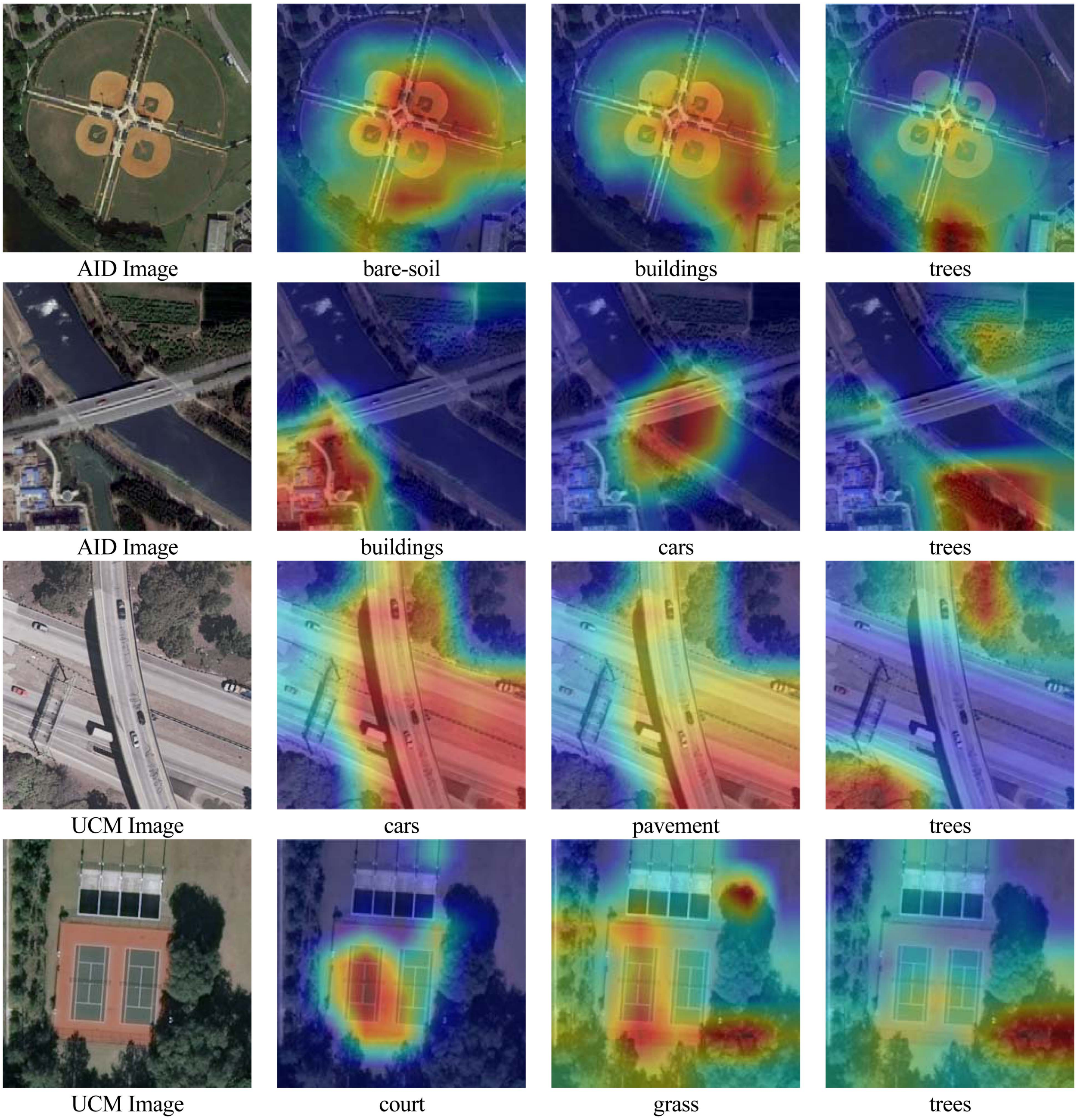

Fig. 4. Heat maps of CPLG. The top two images are from the AID dataset in the scenario of UCM→AID, while the bottom two images are from the UCM dataset in the scenario of AID→UCM.

The performance is suboptimal when images from the source domain have a lower spatial resolution compared to the target domain images. However, it should be noted that CPLG and CPLG+CDAN significantly enhance performance even when employing the AID dataset with lower spatial resolution as the source domain.

2) Ablation Study: The proposed method CPLG is comprised of two key components: the positive–negative pseudo-label confidence selection (PS) and the consistency enhancement based on contrastive loss (CE). To verify the effectiveness of each module, the following ablation experiments are carried out, and the results are shown in Table VI. The best value is in bold.

From the table, we can see that the PS component provides a notable enhancement in F-scores, recall, and mAP. Specifically, in the scenarios of UCM→AID, UCM→DFC15, AID→UCM, and AID→DFC15, the average increases across various metrics are 6.08%, 2.88%, 3.26%, and 6.33%, respectively. At the same time, the CE component substantially boosts the F-scores and mAP. The average increases across various metrics are 2.01%, 2.76%, 2.37%, and 3.16%, respectively.

We also investigate the impact of the number of nearest neighbors in the NA on the performance of CPLG. As shown in Fig. 3, performance improvement can be observed with the increase in the number of nearest neighbors. It is because the model becomes more susceptible to local noise when the number of nearest neighbors is smaller. But the enhancement in performance is not significant when the number is larger than 6. Considering both local sensitivity and global robustness, the

TABLE VIII
VALUES OF THE TWO EFFICIENCY METRICS

Method	FLOPs(G)	params(M)
Source-only	4.1099	24.034
SIGNA	4.1916	26.438
CDAN	4.1209	29.536
ATDOC	4.1119	24.034
DA-MAIC	4.1916	26.438
CPLG	4.1166	24.034

number of nearest neighbors can be selected within the range of 4–6.

3) Efficiency Analysis: A comparison of computational complexity between CPLG and other methods is listed in Table VII. The number of floating-point operations (FLOPs) and the number of model parameters are used as the efficiency metrics. "Source-only" means directly applying a ResNet-50 model trained on the source domain to the target domain for classification. From the table, we can see that CPLG has small numbers of FLOPs and parameters. CPLG is only composed of a feature extractor and a classifier like source-only and ATDOC, so they have the same minimum number of parameters. On the other hand, FLOPs of CPLG are slightly higher than those of source-only and ATDOC, as the pseudo-label generation and consistency enhancement increase a small amount of computational load. It demonstrates the efficiency of CPLG.

In addition, we assess the inference speed of CPLG on an NVIDIA GeForce GTX 1080 Ti GPU. With batch sizes of 1, 4, 8, and 16, the average inference times are 0.0075, 0.0121, 0.0172, and 0.0292 s, respectively. In other words, the inference rates can reach 133, 83, 58, and 34 fps, respectively. This shows that CPLG can meet the real-time performance requirements.

C. Qualitative Analysis

For a more intuitive assessment of CPLG, we present a qualitative analysis with UCM as the source domain and AID as the target domain. Table VIII illustrates the annotation results of three representative images from the AID test set using our method and four alternative approaches. In the table, the ground-truth labels and the labels predicted by different methods are listed. The correct predictions are highlighted in green, the false positive predictions are in red, and the false negative predictions are in blue. It is observed that CPLG distinguishes itself by being the only one capable of correctly predicting the labels for all three images.

To further demonstrate the effectiveness of CPLG in detecting objects within images, we conduct a visualization analysis on four images from the AID and UCM datasets under the scenarios of UCM→AID and AID→UCM, respectively. The heat maps, as shown in Fig. 4, indicate that the proposed model can effectively focus on the regions where objects are located.

V. CONCLUSION

In this article, we propose an unsupervised DA method CPLG for the MLRSIA task. In CPLG, pseudo-labels for samples in the target domain are generated through NA and refined by the positive–negative pseudo-label confidence selection. Additionally, a contrastive loss is introduced into the loss function to minimize feature discrepancies between analogous samples across different domains. The DA experimental results on the UCM, AID and DFC15 multilabel datasets show that our proposed CPLG method can effectively improve the classification performance for the target domain. Moreover, when integrated with other adversarial DA methods (e.g., CDAN), CPLG can further improve the classification performance.

CPLG does not exploit the dependencies among labels, which are considered helpful for performance improvement in multilabel classification. In future work, we will explore how to incorporate label correlations to further enhance the precision of the pseudo-labels for samples in the target domain.

REFERENCES

[1] T. Gao, Z. Li, Y. Wen, T. Chen, Q. Niu, and Z. Liu, "Attention-free global multiscale fusion network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5603214, doi: 10.1109/TGRS.2023.3346041.

[2] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.

[3] S. Wen, W. Zhao, F. Ji, R. Peng, L. Zhang, and Q. Wang, "Recognizing unknown disaster scenes with knowledge graph-based zero-shot learning (KG-ZSL) model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5621315, doi: 10.1109/TGRS.2024.3394653.

[4] Y. Wen, T. Gao, J. Zhang, Z. Li, and T. Chen, "Encoder-free multiaxis physics-aware fusion network for remote sensing image dehazing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4705915, doi: 10.1109/TGRS.2023.3325927.

[5] C. Shi, L. Fang, Z. Lv, and H. Shen, "Improved generative adversarial networks for VHR remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Feb. 2022, Art. no. 8001805, doi: 10.1109/LGRS.2020.3025099.

[6] Q. Xu, Y. Shi, X. Yuan, and X. X. Zhu, "Universal domain adaptation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4700515, doi: 10.1109/TGRS.2023.3235988.

[7] W. Tong, W. Chen, W. Han, X. Li, and L. Wang, "Channel-attention based DenseNet network for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4121–4132, 2020, doi: 10.1109/JSTARS.2020.3009352.

[8] B. Ma et al., "Label-driven graph convolutional network for multilabel remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2245–2255, 2024, doi: 10.1109/JSTARS.2023.3344106.

[9] X. Tan, Z. Xiao, J. Zhu, Q. Wan, K. Wang, and D. Li, "Transformer-driven semantic relation inference for multilabel classification of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1884–1901, 2022, doi: 10.1109/JSTARS.2022.3145042.

[10] R. Huang, F. Zheng, and W. Huang, "Multilabel remote sensing image annotation with multiscale attention and label correlation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6951–6961, 2021, doi: 10.1109/JSTARS.2021.3091134.

[11] Z. M. Chen, X. S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5172–5181.

[12] D. Lin, J. Lin, L. Zhao, Z. J. Wang, and Z. Chen, "Multilabel aerial image classification with unsupervised domain adaptation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609613, doi: 10.1109/TGRS.2021.3115484.

[13] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2962–2971.

[14] C. Zhang and G. H. Lee, "GeT: Generative target structure debiasing for domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 23520–23531.

[15] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 1–35, May 2016.

[16] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

[17] S. Li and J. Yu, "Deep transfer network with adaptive joint distribution adaptation: A new process fault diagnosis model," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 3507813, doi: 10.1109/TIM.2022.3157007.

[18] X. Yu et al., "Conditional adversarial domain adaptation with discrimination embedding for locomotive fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 3503812, doi: 10.1109/TIM.2020.3031198.

[19] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 1647–1657.

[20] J. Liang, D. Hu, and J. Feng, ""Domain adaptation with auxiliary target domain-oriented classifier," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16627–16637.

[21] D. Berthelot et al., "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2019, pp. 5049–5059.

[22] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8934–8954, Sep. 2023.

[23] W. Ma, O. Karakuş, and P. L. Rosin, "Confidence guided semi-supervised learning in land cover classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 5487–5490.

[24] T. Xiao, S. Liu, S. De Mello, Z. Yu, J. Kautz, and M.-H. Yang, "Learning contrastive representation for semantic correspondence," *Int. J. Comput. Vis.*, vol. 130, pp. 1293–1309, May 2022.

[25] G. Xiao, S. Peng, W. Xiang, H. Chen, J. Guo, and Z. Gong, "CMFT: Contrastive memory feature transfer for nonshared-and-imbalanced unsupervised domain adaption," *IEEE Trans. Ind. Inf.*, vol. 19, no. 8, pp. 9227–9238, Aug. 2023.

[26] M. Bi, M. Wang, Z. Li, and D. Hong, "Vision transformer with contrastive learning for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 738–749, 2023, doi: 10.1109/JSTARS.2022.3230835.

[27] S. Koda, A. Zeggada, F. Melgani, and R. Nishii, "Spatial and structured SVM for multilabel image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5948–5960, Oct. 2018.

[28] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit*, vol. 40, pp. 2038–2048, Jul. 2007.

[29] Y. Liu, K. Ni, Y. Zhang, L. Zhou, and K. Zhao, "Semantic interleaving global channel attention for multilabel remote sensing image classification," *Int. J. Remote Sens.*, vol. 45, no. 2, pp. 393–419, 2024.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[31] F. Li, R. Feng, W. Han, and L. Wang, "An augmentation attention mechanism for high-spatial-resolution remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3862–3878, 2020, doi: 10.1109/JSTARS.2020.3006241.

[32] D. Yu, H. Guo, Q. Xu, J. Lu, C. Zhao, and Y. Lin, "Hierarchical attention and bilinear fusion for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6372–6383, 2020, doi: 10.1109/JSTARS.2020.3030257.

[33] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2285–2294.

[34] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multilabel aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, 2019.

[35] B. Schölkopf, J. Platt, and T. Hofmann, "A kernel method for the two-sample-problem," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 513–520.

[36] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 12, pp. 151–175, May 2010.

[37] M. Krichen, "Generative adversarial networks," in *Proc. Int. Conf. Comput. Commun. Netw. Technol.*, 2023, pp. 1–7.

[38] R. Du, G. Wang, N. Zhang, L. Chen, and W. Liu, "Domain adaptive remote sensing scene classification with middle-layer feature extraction and nuclear norm maximization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2448–2460, 2024, doi: 10.1109/JSTARS.2023.3339336.

[39] I. P. Singh, E. Ghorbel, A. Kacem, A. Rathinam, and D. Aouada, "Discriminator-free unsupervised domain adaptation for multi-label image classification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 3924–3933.

[40] H. Tang, K. Chen, and K. Jia, "Unsupervised domain adaptation via structurally regularized deep clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8725–8735.

[41] X. Yue et al., "Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13829–13839.

[42] D. Hou, S. Wang, X. Tian, and H. Xing, "PCLUDA: A Pseudo-label consistency learning- based unsupervised domain adaptation method for Cross-domain optical remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600314, doi: 10.1109/TGRS.2022.3233133.

[43] L. Zhao, W. Luo, Q. Liao, S. Chen, and J. Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6008205, doi: 10.1109/LGRS.2022.3159549.

[44] I. Hernandez-Sequeira, R. Fernandez-Beltran, and F. Pla, "Semi- and self-supervised metric learning for remote sensing applications," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6006305, doi: 10.1109/LGRS.2024.3381228.

[45] Z. Li, K. Bi, Y. Wang, Z. Fang, and J. Zhang, "Supervised contrastive learning for open-set hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5509805, doi: 10.1109/LGRS.2023.3319403.

[46] Y. Ge et al., "Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 11309–11321.

[47] Z. Qiu et al., "Source-free domain adaptation via avatar prototype generation and adaptation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1–10.

[48] T. Kalluri, A. Sharma, and M. Chandraker, "MemSAC: Memory augmented sample consistency for large scale domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 550–568.

[49] Y. Zhao, S. Li, C. H. Liu, Y. Han, H. Shi, and W. Li, "Domain adaptive remote sensing scene recognition via semantic relationship knowledge transfer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2001013, doi: 10.1109/TGRS.2023.3267149.

[50] Y. Chen, X. Zhu, and S. Gong, "Semi-supervised deep learning with memory," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 268–283.

[51] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2015, vol. 2, pp. 2440–2448.

[52] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semi-supervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[53] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.

Rui Huang (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in signal and information processing from the Northwestern Polytechnical University, Xi'an, China, in 1999, 2002, and 2006, respectively.

She is currently an Associate Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. Her current research interests include remote sensing information processing, pattern recognition, and machine learning.

Mingyang Ma received the B.S. degree in communication engineering from the School of Communication and Information Engineering, Shanghai University, Shanghai, China, in 2022. He is currently working toward the M.S. degree in signal and information processing with the School of Communication and Information Engineering, Shanghai University.

His research interests include remote sensing information processing and multilabel learning.

Wei Huang received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from the Wuhan University, Wuhan, China, in 2002 and 2008, respectively.

She is currently a Lecturer with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. Her research interests include remote sensing, images denoising and analysis, and image enhancement and restoration for display processing.