Class-Aware Domain Adaptation for Semantic Segmentation of Remote Sensing Images

Qingsong Xu, Xin Yuan, *Senior Member, IEEE*, and Chaojun Ouyang

***Abstract*— Unsupervised domain adaptation (UDA) for the semantic segmentation of remote sensing images is challenging since the same class of objects may have different spectra while the different class of objects may have the same spectrum. To address this issue, we propose a class-aware generative adversarial network (CaGAN) for UDA semantic segmentation of multisource remote sensing images, which explicitly models the discrepancies of intraclass and the interclass between the source domain images with labels and the target domain images without labels. Specifically, first, to enhance the global domain alignment (GDA), we propose a transferable attention alignment (TAA) procedure to add more fine-grained features into the adversarial learning framework. Then, we propose a novel class-aware domain alignment (CDA) approach in semantic segmentation. CDA mainly includes two parts: the first one is adaptive category selection, which is to alleviate the class imbalance and select the reliable per-category centers in the source and target domains; the second one is adaptive category alignment, which is to model the intraclass compactness and interclass separability from source-only, target-only, and joint source and target images. Finally, the CDA plays as a penalty of GDA to train GaGAN in an alternating and iterative manner. Experiments on domain adaptation of space to space, spectrum to spectrum, both space-to-space and spectrum-to-spectrum data sets demonstrate that CaGAN outperforms the current state-of-the-art methods, which may serve as a starting point and baseline for the comprehensive applications of semantic segmentation in cross-space and cross-spectrum remote sensing images.**

***Index Terms*— Class-aware domain alignment (CDA), class-aware generative adversarial network (CaGAN), cross-scene and cross-spectrum remote sensing images, global domain alignment (GDA), unsupervised domain adaptation (UDA) semantic segmentation.**

Manuscript received April 3, 2020; revised August 17, 2020; accepted October 10, 2020. Date of publication November 17, 2020; date of current version December 2, 2021. This work was supported in part by the Strategic Priority Research Program of CAS under Grant XDA23090303; in part by the National Key Research and Development Program of China under Grant 2017YFC1501000; in part by NSFC under Grant 42022054; and in part by the CAS Youth Innovation Promotion Association. *(Qingsong Xu and Xin Yuan contributed equally to this article.) (Corresponding author: Chaojun Ouyang.)*

Qingsong Xu is with the Key Laboratory of Mountain Hazards and Surface Process, Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Chengdu 610041, China, and also with the School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China.

Xin Yuan is with Bell Labs, Murray Hill, NJ 07974 USA.

Chaojun Ouyang is with the Key Laboratory of Mountain Hazards and Surface Process, Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Chengdu 610041, China, also with the School of Engineering Science, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Tibetan Plateau Earth Sciences, Chinese Academy of Sciences (CAS), Beijing 100101, China (e-mail: cjouyang@imde.ac.cn).

Digital Object Identifier 10.1109/TGRS.2020.3031926

I. Introduction

SEMANTIC segmentation is a pixel-wise class classification task that assigns a label to every pixel in an input image, which describes the class of its enclosing region [1]. The semantic segmentation is extremely useful in various applications of remote sensings, such as traffic analysis and management [2], urban area monitoring and planning [3], and hazard detection and avoidance [4], [5]. Semantic segmentation is typically modeled as a supervised learning task, which requires a large amount of labeled training samples to train the model. Semantic segmentation on remote sensing applications have achieved high performance with active research studies [6]–[9]. Recently, several different types of sensors are jointly used to capture the images of the earth's surface and atmosphere with the development of modern remote sensing technology. These sensors capture numerous images at varying electromagnetic wavebands and different resolutions from different perspectives. However, labeling those images is quite challenging due to the huge amount of images. Thus, adapting the segmentation models trained only with labeled images from one remote sensing data to inference semantic segments of the unlabeled images from other remote sensing data has been attracting much attention.

Toward this end, the concept of domain shift [10] has emerged, and the unsupervised domain adaptation (UDA) methods have been proposed to address the domain shift problem among different data sets, without the need of labeling new image data set. UDA aims to learn the invariant representations between the source domain with labels and target domain without labels. Among the recent works on UDA semantic segmentation, adversarial learning frameworks have attracted significant interest because of the improved quality of alignment between different distributions by adapting representations of different domains, e.g., the source and target's pixel, and feature space [11]–[14].

Despite the success of these adversarial learning frameworks by global domain alignment (GDA) in the pixel, feature, and output spaces, most methods neglected the *category information* and the *marginal distributions across domains* cannot be optimally aligned (Fig. 1). Furthermore, the traditional generative adversarial models do not work well for UDA semantic segmentation of cross-scene and cross-spectrum data sets. The main reason is that the fine structural features of objects in remote sensing images are difficult to be obtained due to high intraclass variance and low interclass variance. Consequently, it is necessary to select the reliable semantic

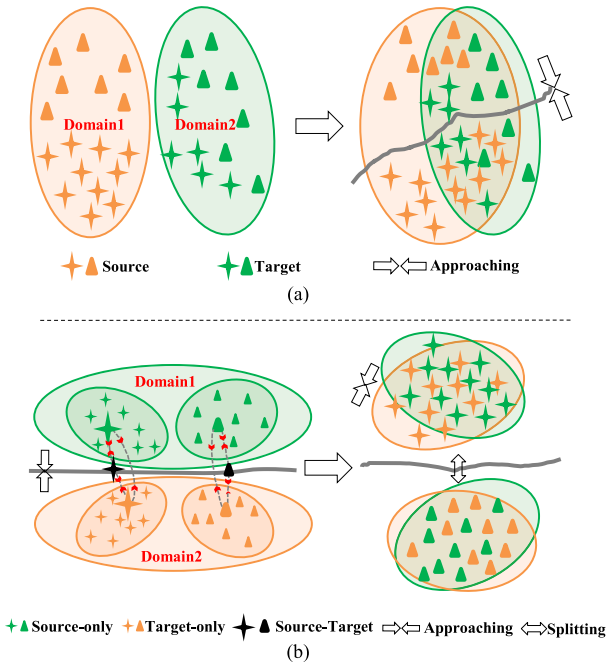

Fig. 1. Comparison between traditional GAN and CaGAN. In Traditional GAN (a) the marginal distributions of the source domain and target domain cannot be optimally aligned since it neglects the category information of the samples. (b) Proposed CaGAN enhances the GDA by adding more fine-grained features, and models the intra-class and inter-class discrepancies across source-only, target-only and source–target domains. Consequently, the model can seek a suitable match between source and target domain, which may deal with the challenge for the UDA semantic segmentation of remote sensing data sets.

feature maps of each class and align these feature maps for the semantic segmentation of remote sensing images.

Bearing these concerns in mind, in this article, we propose the *class-aware generative adversarial network (CaGAN)* to optimize the adversarial learning framework with GDA and class-aware domain alignment (CDA), which models the intraclass and the interclass discrepancies across domains. First, a transferable attention alignment (TAA) in the GDA module is proposed to add more fine-grained features into the adversarial learning framework. Second, we propose a CDA module which includes two building blocks: adaptive category selection and adaptive category alignment. Adaptive category selection alleviates the class imbalance and selects reliable per-category features. Specifically, we found that classes with higher accuracy always have more pixels on the label map through extensive experiments. Therefore, we propose the class-balanced weighting label factor to tackle the issue. Then, since it is challenging to evaluate the correctness of the "pseudo" label of each target sample, the adaptive reliable-category selection is proposed. We take a batch of labeled source and unlabeled target examples and get the "pseudo" labels of the source and target by the segmentation model, and then the per-category feature center of the source is used as a guide to adaptively select the reliable per-category feature of the target. The second building block, adaptive category alignment, weighs the intraclass compactness and the interclass separability. Particularly, the distribution differences of spectrum and space are reduced by minimizing the intraclass domain discrepancy, from source-only, source-target, and target-only dimensions, whereas the source–target representations of each other become farther away from the decision boundary by maximizing the interclass domain discrepancy (Fig. 1).

Considering the stability of GAN training and GDA to drive CDA work, we put the CDA as a penalty of GDA to seek a suitable match across domains. Inspired by Chen *et al.* [15], we employ the asymptotic training scheme to avoid the insufficiency of categorical information in each mini-batch. Particularly, an increasing amount of samples of source and target are taken into account during training. Also, if the segmentation model is over-saturated on the source data, we can retard the convergence speed of the source segmentation by adding a parameter in the softmax function to control the convergence.

A. Deep Adversarial Domain Adaptation

A popular procedure for UDA is to reduce the gap between the source and target domains by learning invariant feature representation to domain shift through a variety of statistical matching approaches. For example, transfer component analysis (TCA) [16] and joint distribution adaptation (JDA) [17] used traditional methods as feature extraction, and then employed margin adaptation or joint adaptation, respectively, to represent the differences between different distributions. Maximum mean discrepancy (MMD) [16]–[19] mapped the data distribution of different domains to reproducing kernel Hilbert space (RKHS), in which the distance between two distributions is measured to reduce the domain shift.

Another direction of research studies addressed the domain adaptation problem by leveraging the adversarial learning behaviors of GANs to perform distribution alignment in the pixel, feature, and output spaces [11], [20]–[23]. For instance, GANs are commonly used at feature spaces generated from CNNs where a pixel-level or object-level domain discriminator is trained to correctly distinguish the domain of each input feature [24], [25]. Some advanced GAN-based methods [15], [26] utilized attention mechanisms and self-supervision mechanisms in feature spaces for domain adaptation problems. Benjdira *et al.* [27] designed a GAN-based algorithm to perform pixel space translation from the source domain to the target domain. Tsai *et al.* [12] presented a domain adaptation scheme in output space by an adversarial network. Although these methods have made outstanding contributions to GAN-based domain adaptation, they usually considered the marginal distribution adaptation for the global domain, but lacking category information. Later this technique was applied to the decision boundary, which utilized task-specific classifiers to align distributions [28], [29]. These approaches focused on directly reshaping the target data regions instead of aligning manifold in feature space under the heuristic assumptions. Recently, several GAN models were proposed for UDA tasks: Wasserstein GAN [14], [30], Siamese-based GAN [20], and ColorMapGAN [31]. In addition, self-supervised model was used by the generator [29], [32]. Most recently, deep adversarial attention alignment methods were proposed for a better adaptation of the source network to the target one

by aligning the *attention maps* of the source network and target network [33]–[35].

B. Class-Aware Domain Alignment

It is necessary to align the class-aware domain in remote sensing semantic segmentation. Kang *et al.* [36] proposed a contrastive adaptation network to explicitly model the intraclass compactness and the interclass separability on adaptation across domains. Chen *et al.* [15] utilized the intraclass variance of the target domain and cross-domain class consistency to address UDA problems. Luo *et al.* [37] proposed the category-level adversarial network (CLAN) aimed to address the problem of semantic inconsistency. Transferable prototypical networks (TPNs) [38] were proposed to construct an embedding space of each class. Category anchor-guided UDA model (CAG-UDA) [39] was presented to adapt the segmentation model by aligning category-wise features guided by category anchors. However, most of these methods were focused on class-aware issues of natural image classification and segmentation. Limited research studies have been devoted to a cross-scene and cross-spectrum segmentation with the remote sensing images.

C. Contributions

This article proposes CaGAN adopting the two powerful techniques: task-specific distribution alignment and CDA. GDA serves as the task-specific segmentation, which can strengthen the GDA by adding more fine-grained features. Furthermore, learning similar prototypes of each class in different domains reduces class-aware domain differences to achieve intraclass compactness and interclass separability. Our specific contributions are listed as follows.

1) A novel per-category selection scheme is proposed to alleviate the class imbalance and select the reliable per-category centers for UDA semantic segmentation.
2) A new metric is proposed to reduce the class-level domain discrepancy to weigh the intraclass compactness and the interclass separability from source-only, target-only, and source–target data.
3) A TAA approach is developed to strengthen the GDA, by adding more fine-grained features (such as the details contained in the lower layers) into the adversarial learning framework.
4) A novel and practical paradigm, *CaGAN* is proposed to optimize the GAN with GDA and CDA by the end-to-end asymptotic training scheme.
5) Our proposed method is verified on four remote sensing data sets, including two high-resolution data sets Postdam and Vaihingen, a synthetic Panchromatic data set, and a hyperspectral data set PaviaU. The results on the four data sets prove the performance of the domain adaption of space to space, spectrum to spectrum, both space to space and spectrum to spectrum, respectively. The experimental results demonstrate that the proposed method outperforms the current state-of-the-art methods. Also, the ablation study is presented to verify the effectiveness of CDA and GDA.

TABLE I
SUMMARY OF ABBREVIATIONS

Abbreviations	Meaning
UDA	Unsupervised Domain Adaptation
CaGAN	Class-aware Generative Adversarial Network
GDA	Global Domain Alignment
CDA	Class-aware Domain Alignment
TAA	Transferable Attention Alignment
DAT	Domain Adversarial Training
OA	Overall pixel Accuracy
mIoU	mean Intersection over Union

D. Organization of This Article

The rest of this article is organized as follows. Section II describes the proposed model with details of components. Section III presents experimental results with the ablation study. Section IV concludes this article with some future works. The abbreviations used in this article are summarized in Table I.

II. PROPOSED CLASS-AWARE GAN

A. Problem Statement

This article focuses on the problem of UDA in the semantic segmentation of remote sensing images. We consider having a source domain $\mathcal{S}$, with both images $\mathbf{X}^{\mathcal{S}}$ and pixel-level labels $\mathbf{Y}^{\mathcal{S}}$. Meanwhile, we have a target domain $\mathcal{T}$, with images $\mathbf{X}^{\mathcal{T}}$ but no annotations. The UDA semantic segmentation problem is defined as finding a model G predicting pixel-wise class labels of the target domain. The traditional GAN trains a prediction network G to learn domain-invariant features by confusing a domain discriminator network D which is trying to distinguish domains. This is achieved by minimizing the segmentation loss L_{seg} and minimaxing the adversarial loss L_{adv}, which are defined as follows:

$$L_{\text{seg}}(G) = \mathrm{E}[\ell(G(\mathbf{X}^{\mathcal{S}}), \mathbf{Y}^{\mathcal{S}})] \tag{1}$$

$$L_{\text{adv}}(G, D) = -\mathrm{E}[\log(D(G(\mathbf{X}^{\mathcal{S}})))] - \mathrm{E}[\log(1 - D(G(\mathbf{X}^{\mathcal{T}})))] \tag{2}$$

where E[ ] denotes the statistical expectation operator and $\ell(\,)$ is an appropriate loss function, such as multiclass cross-entropy.

However, significant limitations exist in using traditional GAN for the semantic segmentation of remote sensing images. First, the marginal distributions of the source domain and target domain cannot be optimally aligned because most adversarial learning methods align only high-level representations, such as activations in the fully connected (FC) layers. In this manner, the details contained in the lower layers cannot be aligned well. Second, there might be a negative transfer that leads to different mapping in feature space of the same class of objects from different domains because objects in remote sensing images usually have high intraclass variance and low interclass variance. More importantly, as the adversarial training going on, the interclass errors will be accumulated. Thus, these models are incapable of preserving cross-domain category consistency. This phenomenon is called "deterioration

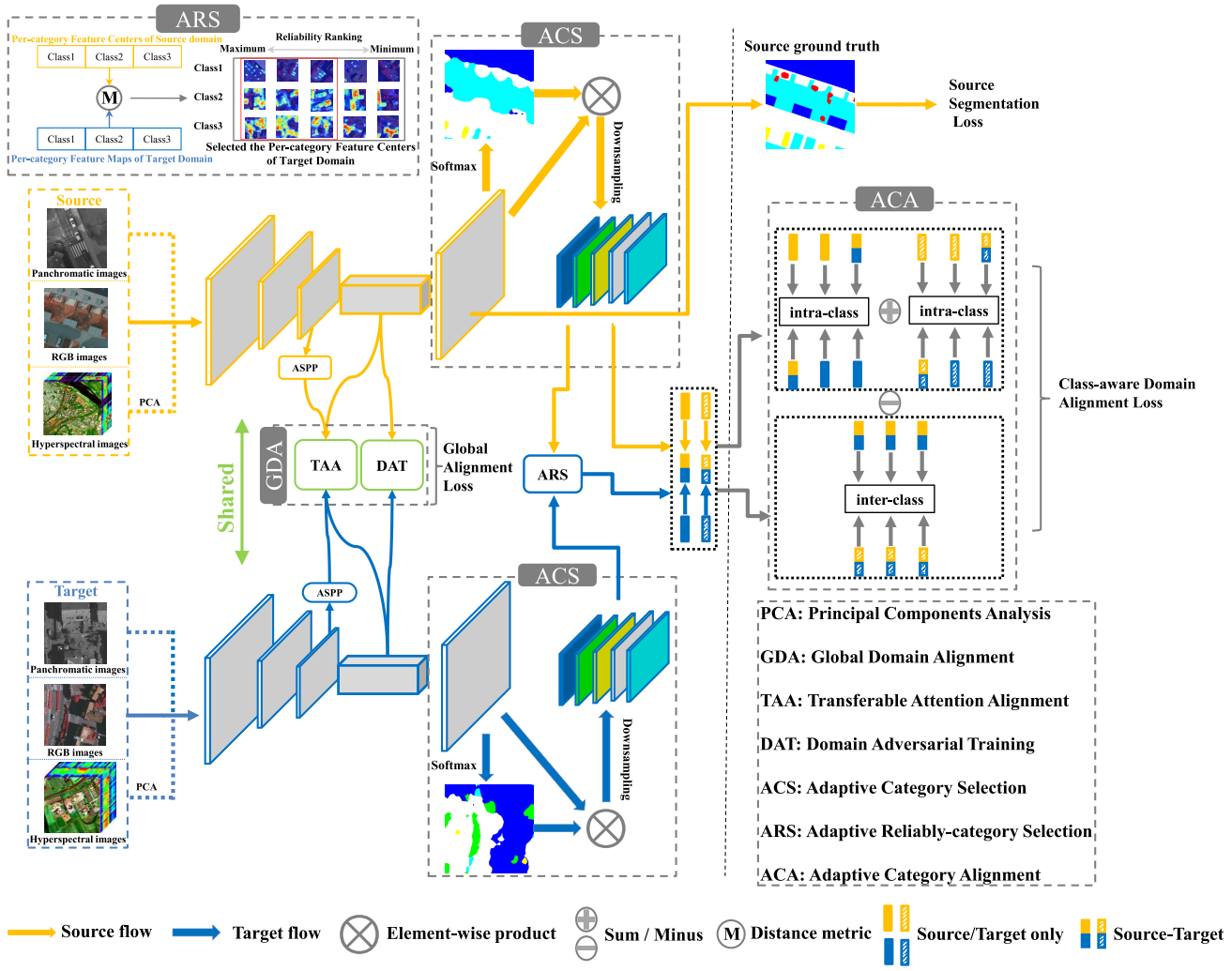

Fig. 2. Overview of the proposed CaGAN, the input of the model includes RGB/Panchromatic/Hyperspectral remote sensing images. The model consists of a semantic segmentation model G and a discriminator D. In the source flow, a segmentation loss is first computed based on the source ground truth, and we calculate the per-category aggregation centers of source feature maps, encoded by model G. In the target flow, we select the reliable per-category feature centers of target samples by the model G and adaptive category selection. To push the distributions of the two domains as close as possible, first, we minimize the domain shift between representations of the source and target data by GDA-based domain adversarial training, and we add more fine-grained features into the adversarial learning framework by TAA; second, CDA is proposed as a penalty of GDA to further model the intraclass compactness and the interclass separability. Detailed descriptions can be found in Section II-B. The back-propagating with minimax loss [(29)] is used to align the global and class-aware domain.

of interclass performance" by us. Therefore, it is necessary to extract semantic feature maps of each class, and guarantee correct pseudo-labels in the target domain. Inspired by this, we propose our CaGAN model depicted in Fig. 2.

As in any UDA model, we first make the necessary assumption that the source and target domains share the same label space and the source model achieves higher performance than that on the target domain. A source domain $\mathcal{D}_S$ is given as

$$\mathcal{D}_S = \left\{\left(x_i^{\mathcal{S}}, y_i^{\mathcal{S}}\right)\right\}_{i=1}^{n_s}, \quad x_i^{\mathcal{S}} \in \mathbf{X}^{\mathcal{S}}, \quad y_i^{\mathcal{S}} \in \mathbf{Y}^{\mathcal{S}} \tag{3}$$

with n_s labeled samples and a target domain $D_{\mathcal{T}}$ is given as

$$D_{\mathcal{T}} = \left\{x_j^{\mathcal{T}}\right\}_{j=1}^{n_t}, \quad x_j^{\mathcal{T}} \in X^{\mathcal{T}} \tag{4}$$

with n_t unlabeled samples. In addition, we consider C classes in the CaGAN.

B. Class-Aware Domain Alignment

CDA mainly includes two parts, namely adaptive category selection and adaptive category alignment. Adaptive category selection aims to extract the semantic features in each class, and adaptive category alignment aims to adaptively match the features from the source domain to the target domain. In the following, we describe these two parts in detail.

1) Adaptive Category Selection: Inspired by self-supervised learning, we directly utilize a semantic segmentation model G (i.e. the feature extractor in Fig. 2) learned on labeled source data by minimizing the segmentation loss, and then assign the source/target sample a "pseudo" pixel label. In this way, the source and target samples are defined with pseudo pixel labels as follows:

$$\hat{\mathcal{S}}^s = \left\{\left(x_i^s, \hat{y}_i^s\right)\right\}_{i=1}^{n_s}, \quad x_i^s \in \mathbb{R}^{\boldsymbol{w}\times h\times c}, \quad \hat{y}_i^s \in \mathbb{R}^{\boldsymbol{w}\times h} \tag{5}$$

$$\hat{\mathcal{S}}^t = \left\{\left(x_i^t, \hat{y}_i^t\right)\right\}_{i=1}^{n_t}, \quad x_i^t \in \mathbb{R}^{\boldsymbol{w}\times h\times c}, \quad \hat{y}_i^t \in \mathbb{R}^{\boldsymbol{w}\times h} \tag{6}$$

where $\hat{y}_i^s$ and $\hat{y}_i^t$ denote "pseudo" pixel labels of the source sample x_i^s and target sample x_i^t, respectively. n_s and n_t are the number of the source and target samples. w and h are the width and height of the sample, and c is the number of channels of the sample. After obtaining the pseudo pixel labels of the source–target data, per-category feature maps can be selected by class-balanced weighting label factor and adaptive reliable-category selection.

a) Class-balanced weighting label factor: Through sufficient experiments, we found that classes with higher accuracy always have more pixels on the label maps, which leads to an imbalance in quantity. To address this issue, we count the class frequency of the entire target and source data sets, respectively. Given an image $x \in \mathbb{R}^{w\times h\times c}$ and a pseudo label map $\hat{y} \in \mathbb{R}^{w\times h}$

$$w_{i,j}^k|_{k=1}^C = \begin{cases} 1, & \hat{y}_{i,j} = k \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$N^k|_{k=1}^C = \sum_j \sum_i w_{i,j}^k|_{k=1}^C \tag{8}$$

$$\hat{\mathbf{F}}^k|_{k=1}^C = \left(1 + \frac{1}{N_n^k|_{k=1}^C}\right) w^k|_{k=1}^C \odot G(x) \tag{9}$$

where $\odot$ denotes the element-wise product, (i, j) are the indices of the pixel location, and $k = 1, \ldots, C$ represents the class number. $w \in \mathbb{R}^{C\times w\times h}$ is the category mask of each image. Furthermore, we balance the influence of category pixels by the number of classes $N \in \mathbb{R}^C$. Ultimately, we acquire per-category feature maps $\hat{\mathbf{F}} \in \mathbb{R}^{C\times w\times h\times \hat{c}}$, where $\hat{c}$ is the number of channels of feature maps.

b) Adaptive reliable-category selection: In order to decrease the sparsity of the per-category feature maps meanwhile to increase the compactness of the per-category feature maps, we impose downsampling to the feature maps. For the source domain samples, we calculate the aggregation centers for each class. Then a set of per-category feature centers of source samples $c_{\mathcal{S}}^k|_{k=1}^C$ is obtained. This strategy can be formulated as

$$c_{\mathcal{S},n}^k\big|_{k=1}^C = \text{Downsampling}\big(\mathbf{F}_{\mathcal{S},n}^k\big|_{k=1}^C\big) \tag{10}$$

$$c_{\mathcal{T},n}^k\big|_{k=1}^C = \text{Downsampling}\big(\mathbf{F}_{\mathcal{T},n}^k\big|_{k=1}^C\big) \tag{11}$$

$$c_{\mathcal{S}}^k|_{k=1}^C = \frac{1}{N_{\mathcal{S}}^k} \sum_n c_{\mathcal{S},n}^k|_{k=1}^C \tag{12}$$

where $N_{\mathcal{S}}^k$ denotes the number of the set of samples including class k in the source domain.

Inspired by the easy-to-hard transfer strategy [15], we find reliable pseudo labels of target samples are closer with the feature centers of source samples. Thus, similar to [15] and [40], we use a similarity measurement ψ to select the reliable per-category feature centers of target samples. To control the similarity process, we set an increasing threshold T during training. The per-category feature maps are determined by the relationship between the similarity and this threshold. This strategy can be formulated as

$$\psi\big(c_{\mathcal{T},n}^k\big) = \rho\big(c_{\mathcal{S}}^k|_{k=1}^C, c_{\mathcal{T},n}^k|_{k=1}^C\big) \tag{13}$$

$$T = \frac{1}{1.5 + e^{-\mu(m+1)}} \tag{14}$$

$$R_n = \begin{cases} 1, & \text{if } \psi\big(c_{\mathcal{T},n}^k\big) \geq T \\ 0, & \text{if } \psi\big(c_{\mathcal{T},n}^k\big) < T \end{cases} \tag{15}$$

$$\hat{c}_{\mathcal{T}}^k|_{k=1}^C = \frac{1}{\hat{N}_{\mathcal{T}}^k} \sum_n c_{\mathcal{T},n}^k|_{k=1}^C \tag{16}$$

where $\rho(x_i, x_j) = ((x_i \cdot x_j)/(\|x_i\| \times \|x_j\|))$ denotes the cosine similarity function. μ is a hyperparameter, and m denotes the training step. In (15), $R_n = 1$ indicates $c_{\mathcal{T},n}^k$ to be selected; otherwise, $R_n = 0$ indicates $c_{\mathcal{T},n}^k$ not to be selected. $\hat{N}_{\mathcal{T}}^k$ denotes the number of the set of target samples selected by the adaptive reliable-category selection. Finally, a set of per-category feature centers of target samples $\hat{c}_{\mathcal{T}}^k|_{k=1}^C$ is obtained. It is worth noting that if the kth class $\hat{c}_{\mathcal{T}}^k = 0$, the feature map $c_{\mathcal{T}}^k$ corresponding to the maximum value of $\psi(c_{\mathcal{T},n}^k)$ in all the kth samples will be selected as the kth category feature centers of target samples, in order to further ensure the reliability of $\hat{c}_{\mathcal{T}}^k|_{k=1}^C$ and the stability of the training process.

2) Adaptive Category Alignment: Adaptive category alignment explicitly models the intraclass and the interclass discrepancies across domains. The domain discrepancy of intraclass is minimized to reduce the spectrum and space distribution differences from three dimensions: source-only, source-target, and target-only, whereas the domain discrepancy of interclass is maximized to put the source–target representations further away from the decision boundary. As shown in Fig. 2, we define the class-level discrepancy loss as follows:

$$\begin{aligned} L_{\text{intra}}\big(\hat{c}_{\mathcal{T}}^k, c_{\mathcal{S}}^k, c_{\mathcal{S}-\mathcal{T}}^k\big) = &\frac{1}{C}\sum_{k=1}^{C}\big\|\hat{c}_{\mathcal{T}}^k - c_{\mathcal{S}}^k\big\| \\ &+ \frac{1}{C}\sum_{k=1}^{C}\big\|\hat{c}_{\mathcal{T}}^k - c_{\mathcal{S}-\mathcal{T}}^k\big\| \\ &+ \frac{1}{C}\sum_{k=1}^{C}\big\|c_{\mathcal{S}}^k - c_{\mathcal{S}-\mathcal{T}}^k\big\| \end{aligned} \tag{17}$$

$$L_{\text{inter}}\big(c_{\mathcal{S}-\mathcal{T}}^k, c_{\mathcal{S}-\mathcal{T}}^{k'}\big) = \frac{1}{C'}\sum_{k=1}^{C-1}\sum_{k'=k+1}^{C}\big\|c_{\mathcal{S}-\mathcal{T}}^k - c_{\mathcal{S}-\mathcal{T}}^{k'}\big\| \tag{18}$$

$$c_{\mathcal{S}-\mathcal{T}}^k\big|_{k=1}^C = c_{\mathcal{S}-\mathcal{T}}^{k'}\big|_{k'=1}^C = \frac{1}{2}\big(\hat{c}_{\mathcal{T}}^k\big|_{k=1}^C + c_{\mathcal{S}}^k\big|_{k=1}^C\big) \tag{19}$$

where $C' = ((C(C-1))/2)$. ℓ_1-norm is used as the distance measure. Different from [15], [36] and [38], we experimentally found that Euclidean distance cannot lead to the high performance of our model, and MMD is limited to the mini-batch size, which does not fit the semantic segmentation. More importantly, because the category information in each mini-batch is insufficient and false pseudo pixel labels of source and target images may cause huge model bias in semantic segmentation, the adaptive category alignment is employed here to solve the problem. The adaptive category alignment

first computes the initial per-category feature maps of two domains based on the Adaptive Category Selection. Then, in each iteration, we compute a set of local category features $c_t^k|_{k=1}^C$ using samples of mini-batch. The accumulated features are computed by averaging all previous local features in each iteration. Finally, we update $\hat{c}_{\mathcal{T}}^{k(I)}$ based on the similarity of $\hat{c}_{\mathcal{T}}^{k(I-1)}$ and $\bar{c}_t^{k(I)}$. Specifically

$$\hat{c}_{\mathcal{T}}^{k(0)} = \hat{c}_{\mathcal{T}}^k|_{k=1}^C \tag{20}$$

$$\bar{c}_t^{k(I)} = \frac{1}{I}\sum_{i=1}^{I} c_t^{k(i)} \tag{21}$$

$$\hat{c}_{\mathcal{T}}^{k(I)} \leftarrow \rho_t^2 \bar{c}_t^{k(I)} + \left(1-\rho_t^2\right)\hat{c}_{\mathcal{T}}^{k(I-1)} \tag{22}$$

$$\rho_t = \rho\left(\bar{c}_t^{k(I)}, \hat{c}_{\mathcal{T}}^{k(I-1)}\right) \tag{23}$$

where I denotes the iteration time in the current training step. ρ_t is a balancing parameter. $c_{\mathcal{S}}^{k(I)}$ is similarly updated for the source domain. Ultimately, the CDA loss $\mathcal{L}_{cda}$ is defined as follows:

$$\mathcal{L}_{cda}(\theta_g) = L_{\text{intra}}\left(\hat{c}_{\mathcal{T}}^{k(I)}, c_{\mathcal{S}}^{k(I)}, c_{\mathcal{S}-\mathcal{T}}^{k(I)}\right) - L_{\text{inter}}\left(c_{\mathcal{S}-\mathcal{T}}^{k(I)}, c_{\mathcal{S}-\mathcal{T}}^{k'(I)}\right). \tag{24}$$

C. Global Domain Alignment

GDA is proposed to minimize the domain shift between representations of the source and target data. Similar to the traditional adversarial loss, we use the features in the last layer of the generator G of both source and target domains, which can be described as

$$\mathcal{L}_{\text{adv}}(\theta_g, \theta_d) = -\text{E}[\log(D(G(X^{\mathcal{S}})))] - \text{E}[\log(1 - D(G(X^{\mathcal{T}})))]. \tag{25}$$

a) TAA: To further eliminate the domain shift and using the GDA during training to align the class-aware domains, it is necessary to add more fine-grained features into the adversarial learning framework. We propose TAA here to reduce the discrepancies across domains. Formally, given an input image $\mathbf{X}$, the attention map $\mathbf{A}_l(\mathbf{X})$ is obtained by feeding $\mathbf{X}$ to G. The corresponding feature maps for the layer l is represented by $\mathbf{G}_l(\mathbf{X})$. Inspired by the image-to-image translation [41], the attention map $\mathbf{A}_l(\mathbf{X})$ is defined as follows:

$$\mathbf{A}_l(\mathbf{X}) = \sum_c |\mathbf{G}_{l,c}(\mathbf{X})|^2 \tag{26}$$

where $\mathbf{G}_{l,c}(\mathbf{X})$ denotes the cth channel of the feature maps. The operations are element-wise across the channel dimension. Thus, we penalize the distance between the last layer and the penultimate layer attention maps of the source and the target network models to minimize their discrepancies. The transferable attention maps of different layers in a segmentation network focus on different features. For instance, the lower layer attention maps have higher activations on local regions of transferable targets, while the attention maps of the higher layers focus on global semantic information of targets. Thus, the TAA loss is defined as

$$\mathcal{L}_{taa}(\theta_g) = \sum_l \sum_i \left\| \frac{\mathbf{A}_l^{\mathcal{S}}\left(\mathbf{X}_i^{\mathcal{S}}\right)}{\left\|\mathbf{A}_l^{\mathcal{S}}\left(\mathbf{X}_i^{\mathcal{S}}\right)\right\|^2} - \frac{\mathbf{A}_l^{\mathcal{T}}\left(\mathbf{X}_i^{\mathcal{S}}\right)}{\left\|\mathbf{A}_l^{\mathcal{T}}\left(\mathbf{X}_i^{\mathcal{S}}\right)\right\|^2} \right\|^2. \tag{27}$$

Algorithm 1 Optimization of CaGAN

Require: source data: $D_s = \left\{\left(x_i^s, y_i^s\right)\right\}_{i=1}^{n_s}$,
target data: $D_t = \left\{x_j^t\right\}_{j=1}^{n_t}$ $epoch = 0$

1: **Initialize**: G_0 and D_0
2: **while** not converge or $epoch < maxepoch$ **do**
3: $epoch = epoch + 1$;
4: Run Adaptive Category Selection based on $G_{epoch-1}$
5: Calculate the per-category feature center $c_S^{k(epoch-1)}|_{k=1}^C$ and $\hat{c}_T^{k(epoch-1)}|_{k=1}^C$
6: **for** $I = 1, I \le I_{\max}; I++$ **do**
7: $I = I_{\max}(epoch-1) + I$;
8: Derive B_s and B_t sampled from D_s and D_t
9: **Train** G_I on labeled source data:
10: Train G_I by $\min_{\theta_g} \mathcal{L}_{seg}$
11: **Train** G_I to fool D_I
12: Calculate local per-category feature center $c_T^{k(I)}$ and $\hat{c}_T^{k(I)}$
13: Update: $c_S^{k(I)}|_{k=1}^C$, $\hat{c}_T^{k(I)}|_{k=1}^C$ bu using Eq. (22) and Eq. (23)
14: Train G_I on unlabel target data by $\min_{\theta_g} \beta\mathcal{L}_{adv}(\theta_g, \theta_d) + \gamma\,\mathcal{L}_{cda}\left(\theta_g\right) + \alpha\mathcal{L}_{taa}(\theta_g)$
15: **Train** D_I on source and target data by $\max_{\theta_d} \beta\mathcal{L}_{adv}(\theta_g, \theta_d)$
16: **end for**
17: **end while**

D. Optimization

Given a source domain $D_{\mathcal{S}} = \{(x_i^{\mathcal{S}}, y_i^{\mathcal{S}})\}$, ($x_i^{\mathcal{S}} \in \mathbf{X}^{\mathcal{S}}$, $y_i^{\mathcal{S}} \in \mathbf{Y}^{\mathcal{S}}$), the semantic segmentation model G is trained by minimizing the multiclass cross-entropy loss

$$\mathcal{L}_{\text{seg}}(\theta_g) = \sum_{i=1}^{h\times w}\sum_{c=1}^{C} -y_{ic}^{\mathcal{S}} \log p_{ic}^{\mathcal{S}} \tag{28}$$

where $p_{ic}^{\mathcal{S}}$ and $y_{ic}^{\mathcal{S}}$ denote the predicted probability and the ground truth of class c on the pixel i, respectively. Formally, the goal of training is to optimize the following minimax objective:

$$\min_{\theta_g}\max_{\theta_d}\ \mathcal{L}_{seg}(\theta_g) + \beta\mathcal{L}_{adv}(\theta_g, \theta_d) + \gamma\,\mathcal{L}_{cda}(\theta_g) + \alpha\mathcal{L}_{taa}(\theta_g) \tag{29}$$

where β, γ, and α are weights that control the adversarial loss, the CDA loss, and the TAA loss, respectively. Algorithm 1 shows the flow of our CaGAN procedure. Specifically, per-category feature centers of the source domain and target domain are initialized in every epoch (Steps 4 and 5). For each iteration, G_I is first trained on labeled source data (Step 10), G_I is then trained to fool D_I (Steps 12–14), and D_I is trained on source and target data (Step 15). Note that the first step is pre-trained for an epoch, and then all the three steps are iteratively trained in the training procedure. In addition, if the segmentation model G is oversaturated on the source data, we can retard the convergence speed of the source segmentation by

adding a speed parameter $S(S > 1)$ in the softmax function:

$$Z_i = \frac{\exp(q_i/S)}{\sum_j \exp(q_j/S)} \tag{30}$$

where q_i is the final output feature maps by source segmentation model, Z_i denotes the class probabilities of pixels for a source sample.

E. Theoretical Analysis

We utilize the theory of domain adaptation [10] to conduct the theoretical analysis of our model. Let $\mathcal{H}$ be the hypothesis class. Given two domains $\mathcal{S}$ and $\mathcal{T}$

$$\forall h \in \mathcal{H}, \quad R_{\mathcal{T}}(h) \le R_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda. \tag{31}$$

The $R_{\mathcal{T}}(h)$ is thus limited by three terms.

1) The expected error on the source domain $R_{\mathcal{S}}(h)$, which is expected to be small and prone to be optimized by the segmentation model G based on the source labels.
2) $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ measures the domain discrepancy in the hypothesis space $\mathcal{H}$. According to the prior work [10], [42], it can be minimized by the global domain adversarial training, proved by the XOR-function [42]

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) &= 2 \sup_{(h,h') \in \mathcal{H}^2} \Big| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}} \mathrm{P}[h(\mathbf{x}) \neq h'(\mathbf{x})] \\ &\qquad - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}} \mathrm{P}[h(\mathbf{x}) \neq h'(\mathbf{x})] \Big| \\ &\le 2 \sup_{h \in \mathcal{H}} \Big| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}} \mathrm{P}[h(\mathbf{x}) = 1] \\ &\qquad - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}} \mathrm{P}[h(\mathbf{x}) = 1] \Big| \\ &= 2 \sup_{h \in \mathcal{H}} \Big| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{S}} \mathrm{P}[h(\mathbf{x}) = 0] \\ &\qquad + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{T}} \mathrm{P}[h(\mathbf{x}) = 1] - 1 \Big|. \end{aligned} \tag{32}$$

3) λ is related to $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$. However, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ does not guarantee the alignment of λ on cross-domain category distributions. Thus, it is necessary to further analyze λ. Let $f_{\hat{\mathcal{T}}} \in \hat{\mathcal{D}}_t$ and $f_{\hat{\mathcal{S}}} \in \hat{\mathcal{D}}_s$ be the pseudo-labeling functions of target and source domains, respectively. And $f_{\mathcal{T}}$ and $f_{\mathcal{S}}$ are the corresponding true labels. Thus, λ is approximately evaluated by the true labels of source domain and the pseudo-labels of target domain. In addition, this proof relies on the triangle inequality [43]

$$\begin{aligned} \lambda &= \min_{\forall h \in \mathcal{H}} R_{\mathcal{S}}(h, f_{\mathcal{S}}) + R_{\hat{\mathcal{T}}}(h, f_{\mathcal{T}}) \\ &\le \min_{\forall h \in \mathcal{H}} R_{\mathcal{S}}\left(h, \frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}\right) + R_{\mathcal{S}}\left(\frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}, f_{\hat{\mathcal{S}}}\right) \\ &\quad + R_{\mathcal{S}}(f_{\hat{\mathcal{S}}}, f_{\mathcal{S}}) + R_{\hat{\mathcal{T}}}\left(h, \frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}\right) \\ &\quad + R_{\hat{\mathcal{T}}}\left(\frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}, f_{\hat{\mathcal{T}}}\right) + R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{T}}}, f_{\mathcal{T}}) \\ &\le \min_{\forall h \in \mathcal{H}} R_{\mathcal{S}}(h, f_{\hat{\mathcal{S}}}) + 2R_{\mathcal{S}}\left(\frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}, f_{\hat{\mathcal{S}}}\right) \\ &\quad + R_{\mathcal{S}}(f_{\hat{\mathcal{S}}}, f_{\mathcal{S}}) + R_{\hat{\mathcal{T}}}(h, f_{\hat{\mathcal{T}}}) \\ &\quad + 2R_{\hat{\mathcal{T}}}\left(\frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}, f_{\hat{\mathcal{T}}}\right) + R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{T}}}, f_{\mathcal{T}}) \\ &\le \min_{\forall h \in \mathcal{H}} \left\{ R_{\mathcal{S}}(h, f_{\hat{\mathcal{S}}}) + R_{\mathcal{S}}(f_{\hat{\mathcal{S}}}, f_{\mathcal{S}}) + R_{\hat{\mathcal{T}}}(h, f_{\hat{\mathcal{T}}}) \right. \\ &\quad \left. + R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{T}}}, f_{\mathcal{T}}) \right\} \\ &\quad + \left\{ 2R_{\mathcal{S}}\left(\frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}, f_{\hat{\mathcal{S}}}\right) + 2R_{\hat{\mathcal{T}}}\left(\frac{f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}}}{2}, f_{\hat{\mathcal{T}}}\right) \right. \\ &\quad \left. + 2R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{S}}}, f_{\hat{\mathcal{T}}}) \right\}. \end{aligned} \tag{33}$$

We thus have the following conclusions.

Remark 1: Minimizing $\{R_{\mathcal{S}}(h, f_{\hat{\mathcal{S}}}) + R_{\mathcal{S}}(f_{\hat{\mathcal{S}}}, f_{\mathcal{S}}) + R_{\hat{\mathcal{T}}}(h, f_{\hat{\mathcal{T}}}) + R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{T}}}, f_{\mathcal{T}})\}$. This term is minimized by the proposed adaptive category selection and the segmentation model G. The proposed adaptive category selection is to alleviate the class imbalance and select the reliable per-category centers in the source and target domains, which minimizes $R_{\mathcal{S}}(f_{\hat{\mathcal{S}}}, f_{\mathcal{S}})$ and $R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{T}}}, f_{\mathcal{T}})$. Furthermore, the softmax function in the segmentation model G aims to get a better target performance, i.e., minimizing $R_{\mathcal{S}}(h, f_{\hat{\mathcal{S}}}) + R_{\hat{\mathcal{T}}}(h, f_{\hat{\mathcal{T}}})$. Ultimately, the gaps between pseudo-labels and true labels are gradually narrowed.

Remark 2: Minimizing $\{R_{\mathcal{S}}(((f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}})/2), f_{\hat{\mathcal{S}}}) + R_{\hat{\mathcal{T}}}(((f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}})/2), f_{\hat{\mathcal{T}}}) + R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{S}}}, f_{\hat{\mathcal{T}}})\}$. The proposed adaptive category alignment is to reduce the spectrum and space distribution differences within the class across domains from three dimensions: source-only, source-target and target-only, which minimizes $R_{\mathcal{S}}(((f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}})/2), f_{\hat{\mathcal{S}}})$, $R_{\hat{\mathcal{T}}}(((f_{\hat{\mathcal{S}}} + f_{\hat{\mathcal{T}}})/2), f_{\hat{\mathcal{T}}})$ and $R_{\hat{\mathcal{T}}}(f_{\hat{\mathcal{S}}}, f_{\hat{\mathcal{T}}})$, respectively. Finally, adaptive category alignment can align features in category-level, i.e., when the categories are aligned, $f_{\hat{\mathcal{S}}} = f_{\hat{\mathcal{T}}}$, then λ is expected to be minimized.

III. EXPERIMENTS

A. Setups

1) Data Sets: Four remote sensing data sets are employed in our experiments, including two high-resolution data sets `Postdam`[1] and `Vaihingen`,[2] a synthetic `Panchromatic` data set, and a hyperspectral data set `PaviaU`.[3] These data sets are used to validate the domain adaption ability of the proposed method for space to space, spectrum to spectrum, and both space to space and spectrum to spectrum scenes.

The Postdam data set consists of 3-band IRRG[4] and 3-band RGB[5] image data. The data set includes 38 annotated images where a spatial resolution is 5 cm. We randomly split the data set into 28 images of the training set and ten images of the testing set. Images are labeled with six classes: building, low vegetation, impervious surface, car, tree, and clutter. We randomly sample 512×512 patches from the original images and generate 3024 patches for training while 500 patches for testing.

The Vaihingen data set consists of 3-band IRRG image data with 33 images. Twenty-three images and 10 images are

[1] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[2] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

[3] http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

[4] IRRG: Infra-red, Red, and Green.

[5] RGB: Red, Green, and Blue.

selected for training and testing, respectively. We randomly sample 512×512 patches from the original images and generate 2484 patches for training while 500 patches for testing.

The synthetic Panchromatic data set is created from the Vaihingen data set, from which we select 3-band IRRG images and normalize the pixel values into [0, 255]. There are 3000 panchromatic gray images of 512×512 pixels, and we randomly sample 2400 images as the training set with the rest as the testing set.

The PaviaU data set is one of the public hyperspectral image data sets, which has a single HSI with a size of $489 \times 388 \times 103$. We select the first 20 bands as the source domain and the bands from 21 to 40 as the target domain. We randomly sample $128 \times 128 \times 20$ patches from the first 20 band images. 3000 patches are generated for training with 1000 patches for testing. Similarly, we randomly sample $128 \times 128 \times 20$ patches from the second 20 band images, 2000 patches are generated for training and 1000 patches for testing.

2) Implementation Details: We use Deeplab-V2 [44] with ResNet-101 [45] pretrained on ImageNet [46] as the basic semantic segmentation model G. To better align the global feature distribution, Atrous spatial pyramid pooling (ASPP) is applied on the last and penultimate layer's feature outputs, and TAA is used on multilevel outputs coming from the last layer and the penultimate layer, as illustrated in Fig. 2. We modify the sampling rates of ASPP as $\{6, 12, 24, 36\}$ and the dilation rates of the last layers so that they produce dense feature maps with larger field-of-views. For discriminator network D, we adopt the same architecture as the one used in DCGAN [47], which is composed of five convolution layers with the kernel size of 4×4, channel numbers are $\{64, 128, 256, 512, 1\}$, and a stride of 2. Leaky-ReLU with the slope of 0.2 follows each convolution layer except the last layer. All the network weights are shared for the source domain and the target domain.

The stochastic gradient descent (SGD) is used as the optimizer for G with momentum $= 0.9$, weight decay $= 5\ e-4$, while using Adam to optimize D with $\beta_1 = 0.9, \beta_2 = 0.99$. For SGD, initial learning rate $= 2.5\ e-4$, while initial learning rate $= 1\ e-4$ for Adam. A poly learning rate policy is used for two optimizers. max_iteration $= 7000$ in every epoch. We adopt the overall pixel accuracy (OA) and mean intersection over union (mIoU) as evaluation criterion [48]. We use Pytorch for implementation on a high-performance computing cluster, with four Tesla K80 12 GB GPUs. The code will be made publicly available.[6]

B. Model Analysis

1) Hyperparameter Analysis: To study the effect of hyperparameter settings on our proposed method, we perform sensitivity analysis on the weights of training loss (β, γ, and α), the threshold parameter (μ), and the speed parameter (S).

a) Weights of training loss: β, α, and γ correspond to the adversarial loss, the TAA loss, and the CDA loss, respectively. A series of cross-validation experiments are conducted to investigate the impact of these three weights on CaGAN. For instance, a space-to-space adaptation scenery IRRG Vaihingen $\rightarrow$ IRRG Postdam with $\mu = 0$, $S = 1$ is illustrated in Table II. First, our adversarial loss performs the best with $\beta = 0.001$, hence the same value of β is utilized in the following experiments. Second, by evaluating the impact of adding a TAA loss in GDA, we find that the weight of $\alpha = 0.0002$ achieves the best result. A larger power of TAA (e.g., $\alpha = 0.001$) results in a decreased performance. In the case of fixing the weight of $\beta = 0.001$ and $\alpha = 0.0002$, we further analyze the impact of CDA for our CaGAN. It can be seen that higher performance is achieved when γ is set to 0.0002. In addition, the GAN training process is more stable when β and γ are set to 0.001 and 0.0002, respectively. This is because the global distribution alignment should guide and drive the alignment of class-aware distributions. Thus, we use the same $\beta = 0.001$, $\gamma = \alpha = 0.0002$ for all our experiments.

TABLE II

HYPERPARAMETER ANALYSIS OF THE TRAINING LOSS WEIGHTS ON TRANSFER TASK IRRG VAIHINGEN $\rightarrow$ IRRG POSTDAM. THE VALUES IN **BOLD** ARE THE BEST

Parameters	β	α	γ	μ	S	OA	mIoU
β	0.0005	0	0	0	1	65.60	45.83
	0.001	0	0	0	1	**67.89**	**47.67**
	0.002	0	0	0	1	67.41	47.42
	0.004	0	0	0	1	67.58	47.45
α	0.001	0.0001	0	0	1	68.48	48.44
	0.001	0.0002	0	0	1	**68.73**	**48.99**
	0.001	0.0005	0	0	1	68.05	48.57
	0.001	0.001	0	0	1	66.71	47.10
γ	0.001	0.0002	0.0001	0	1	68.98	49.07
	0.001	0.0002	0.0002	0	1	**69.70**	**49.26**
	0.001	0.0002	0.0005	0	1	69.33	49.24
	0.001	0.0002	0.001	0	1	68.14	47.93

b) Threshold parameter: To investigate the impact of the reliable pseudo label selection parameter μ of target samples on different domain adaptation scenarios, we conduct cross-space domain (different objects with the same spectra) and cross-spectrum domain (the same object with different spectra) experiments with μ in {0,0.4,0.8,1.2,1.6,2} and $S = 1$. As depicted in Fig. 3(a) and (b), it has been seen that the best result is obtained when μ is set as 0.8 in terms of cross-space experiments (IRRG Postdam $\rightarrow$ IRRG Vaihingen). In terms of cross-spectrum domains, it is shown that better performance is achieved when $\mu = 0.8$ on cross-spectrum scenarios with a small number of bands (IRRG Postdam $\rightarrow$ RGB Postdam), while the peak performance is reached when $\mu = 1.6$ on cross-spectrum scenarios with a large number of bands (First 20-band PaviaU $\rightarrow$ Second 20-band PaviaU). Digging further, μ between 0.8 and 1.6 leads to favorable performances for all adaptation scenarios. However, the performance is relatively poor when μ is either too small or too large. When μ is small (the threshold is low), reliable pseudo labels of target samples cannot be selected. When μ is large (the threshold is high), the generated pseudo label samples cannot provide effective per-category feature centers for subsequent training.

c) Speed parameter: We perform a sensitivity analysis of the speed parameter S on cross-space and cross-spectrum

[6]https://github.com/xupine/CaGAN

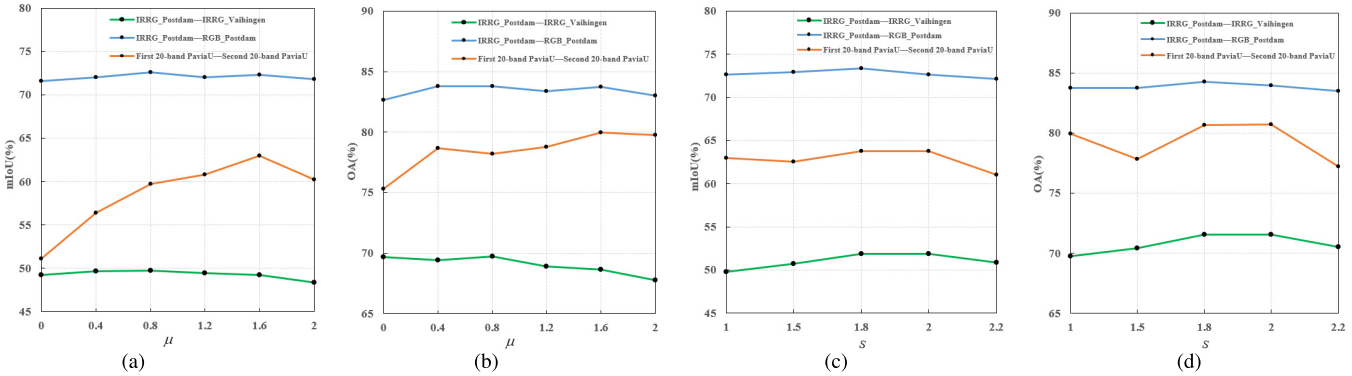

Fig. 3. Sensitivity analysis for the parameters μ and S. (a) Threshold parameter μ on mIoU. (b) Threshold parameter μ on OA. (c) Speed parameter S on mIoU. (d) Speed parameter S on OA.

domain adaptive experiments. Fig. 3(c) and (d), respectively, shows the mIoU and OA results by changing S in {1,1.5,1.8,2,2.2}. The performance first increases and then decreases as S varies. The peak performances are obtained when S is set as 1.8 for cross-space experiments (IRRG Postdam → IRRG Vaihingen) and cross-spectrum scenarios with a small number of bands (IRRG Postdam → RGB Postdam). When S is set as 2.0, the CaGAN model performs better for cross-spectrum scenarios with a large number of bands (First 20-band PaviaU → Second 20-band PaviaU). Furthermore, the results implicitly demonstrate that a good UDA model needs a non-saturated source segmentation network.

2) Ablation Study for Each Module in CaGAN: The proposed model has two main components: CDA and GDA. To analyze the effects of those two components, the ablation study was conducted in the UDA semantic segmentation of remote sensing images. Fig. 4 presents the contrastive analysis of the segmentation results and visualized features obtained from source-only, CaGAN with CDA and CaGAN without CDA. In addition, to analyze the enhancement of TAA to the GDA, we visualize the feature map of the models with/without TAA on every category. We randomly sample some target images from the transfer task IRRG Postdam → IRRG Vaihingen. The category heatmaps of the segmentation network before softmax operation are then overlaid with a target image. The visualization analysis results are shown in Fig. 5.

First, it has been shown that GDA can be used to minimize the domain shift according to (f) and (g) in Fig. 4. The segmentation model G-based source-only can capture the features, which are originally aligned between domains. However, the fine structural features of the objects with the same label possibly have huge differences in different spatial or spectral domains because of the same objects with different spectra and different objects with the same spectra. Thus, GDA can encourage G to capture the domain-invariant features of pixels, as a result of reducing the intra-class dispersion. As shown in Fig. 4(b), the result is poor without GDA. Furthermore, it has been demonstrated that TAA in GDA can focus more transferable regions (especially in the objects in different domains), by adding more fine-grained features into the adversarial learning framework according to Fig. 5.

Second, the results in Fig. 4 show that CDA can further model the intraclass compactness and the interclass

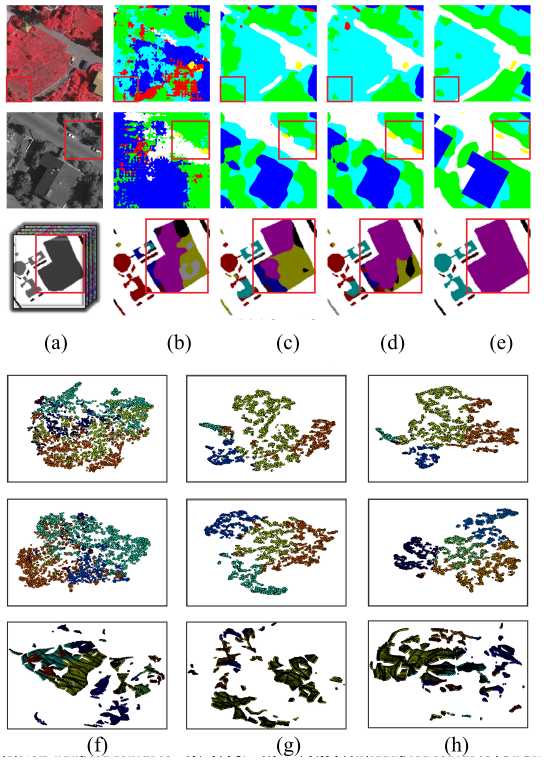

Fig. 4. Contrastive analysis of proposed CaGAN without CDA and with CDA. (a) Target images of IRRG Vaihingen, synthetic panchromatic Vaihingen, and Second 20-band PaviaU. Red boxes are the focused regions. (b) Nonadapted segmentation results based on source-only. (c) Adapted (CaGAN without CDA) results of three experiments, in which a decent segmentation map is produced. However, the inter-class discrepancy is low and this leads to "deterioration of inter-class performance." (d) Adapted results from CaGAN. CaGAN reduces the distribution differences within the class and improves appropriately the inter-class domain discrepancies. (e) Ground truth of target images. The h-dimensional features of (b)–(d) are mapped to a 2-D space with t-SNE [49] in (f)–(h). Each color represents a class. The comparison of feature distributions proves that CaGAN can model the intra-class compactness and the inter-class separability.

separability, especially in the space domain and spectrum domain adaptive experiments. Different from other category-level alignment models [37], [39], [50], CDA visually builds the class-aware model. Furthermore, according to Table II, $\gamma \mathcal{L}_{\text{cda}}(\theta_g)$ is a penalty for global adaptation. When γ is large, CDA is dominant, but it cannot capture correct pseudo-labels for target samples, and thus G cannot

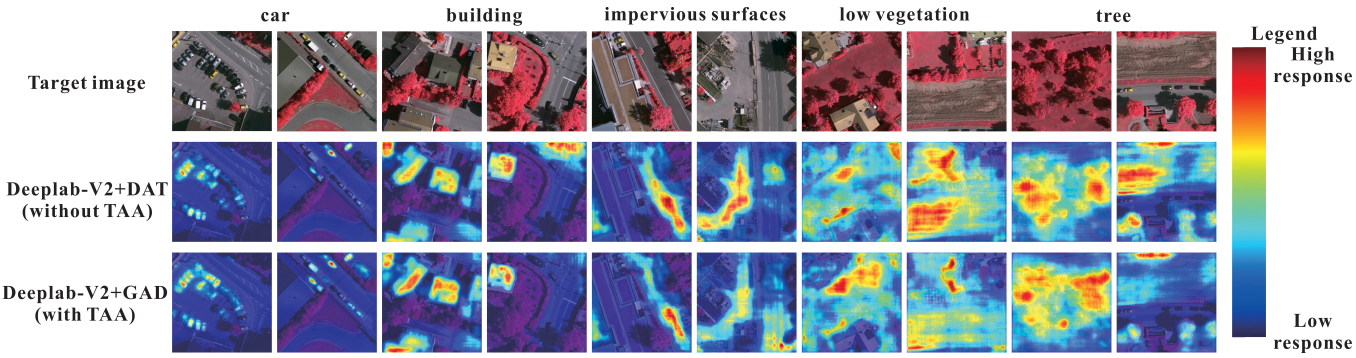

Fig. 5. Visualization analysis results of category features with/without TAA module in target images. (From Left to Right) First two columns represent car category, the second two columns represent building category, the third two columns represent impervious surfaces category, the fourth two columns represent low vegetation category, the fifth two columns represent tree category.

TABLE III

ADAPTATION FROM IRRG POSTDAM TO IRRG VAIHINGEN. PER-CLASS IOU: 5–9TH COLUMNS, OA: OVERALL PIXEL ACCURACY, AND MIOU: MEAN IOU

IRRG Postdam → IRRG Vaihingen										
Method	Backbone	Model parameters	Approach	Imp. Surf.	Build.	Low veg.	Tree	Car	OA	mIoU
Source-only	ResNet-101	6.223×10^7	Seg.	24.36	58.86	16.85	52.56	18.30	49.77	34.19
MCD	ResNet-101	17.831×10^7	Adv-C.	53.55	64.43	40.17	56.25	16.77	68.68	46.23
AdasegNet(multi-level)	ResNet-101	4.870×10^7	Adv.	61.77	**74.44**	**46.67**	50.13	18.00	**71.60**	50.20
AdveNet(multi-level)	ResNet-101	4.870×10^7	Adv.	53.97	72.02	38.86	**56.65**	28.69	69.67	50.03
CLAN	ResNet-101	4.615×10^7	Adv.	52.79	71.31	38.29	56.20	31.10	69.54	49.94
Siamese-based GAN	ResNet-101	–	Adv-R*.	65.6	70.8	29.6	53.3	38.0	–	51.5
Deeplab-V2+DAT(without TAA)	ResNet-101	6.500×10^7	Adv.	57.59	65.91	35.20	52.16	33.34	66.98	48.84
Deeplab-V2+GDA(without CDA)	ResNet-101	6.500×10^7	Adv.	60.17	63.37	39.02	50.86	33.43	68.50	49.37
Deeplab-V2+GDA+CDA(CaGAN)	ResNet-101	6.500×10^7	Adv.	**63.71**	72.81	33.97	54.79	**33.91**	71.57	**51.84**
Target-only	ResNet-101	6.223×10^7	Seg.	74.41	78.93	60.92	72.32	39.14	82.83	65.14
Source-only	VGG-16	2.030×10^7	Seg.	22.81	51.87	8.57	33.29	27.81	39.76	28.87
CaGAN	VGG-16	2.307×10^7	Adv.	50.70	62.12	23.38	53.50	31.98	64.39	44.34
Target-only	VGG-16	2.030×10^7	Seg.	70.77	73.90	55.19	70.14	37.81	80.13	61.56
Source-only	ResNet-50	4.324×10^7	Seg.	21.28	53.09	8.57	50.31	25.61	45.27	31.77
CaGAN	ResNet-50	4.601×10^7	Adv.	54.72	64.08	25.88	57.62	30.17	65.14	46.49
Target-only	ResNet-50	4.324×10^7	Seg.	73.66	78.35	57.35	71.89	38.42	82.15	63.93

TABLE IV

ADAPTATION FROM IRRG VAIHINGEN TO IRRG POSTDAM. PER-CLASS IOU: 4–8TH COLUMNS, OA: OVERALL PIXEL ACCURACY, AND MIOU: MEAN IOU

IRRG Vaihingen → IRRG Postdam									
Method	Backbone	Approach	Imp. Surf.	Build.	Low veg.	Tree	Car	OA	mIoU
Source-only	ResNet-101	Seg.	29.74	39.47	27.26	23.88	11.20	47.00	26.31
MCD	ResNet-101	Adv-C.	54.66	41.05	18.65	29.18	26.53	55.67	34.01
AdasegNet(multi-level)	ResNet-101	Adv.	55.54	30.65	**44.67**	**42.12**	37.34	62.03	42.06
AdveNet(multi-level)	ResNet-101	Adv.	55.56	25.99	42.70	33.20	**52.00**	59.97	41.89
CLAN	ResNet-101	Adv.	**58.21**	44.55	29.45	28.01	46.72	60.91	41.39
Deeplab-V2+DAT(without TAA)	ResNet-101	Adv.	52.12	48.84	35.46	37.56	28.63	61.53	40.52
Deeplab-V2+GDA(without CDA)	ResNet-101	Adv.	52.54	45.75	43.60	37.54	28.34	62.05	41.55
Deeplab-V2+GDA+CDA(CaGAN)	ResNet-101	Adv.	56.38	**50.76**	37.15	37.13	32.81	**63.90**	**42.85**
Target-only	ResNet-101	Seg.	75.18	75.25	58.88	58.38	61.51	80.75	65.84

fool D on the global domain. This causes diverse model training. When γ is small, G tends to have similar joint distributions and similar class-aware marginal distributions between two domains. D is encouraged to be trained to distill more knowledge from images suffering from class inconsistency rather than well-aligned classes.

We further evaluate the performance of TAA, GDA, and CDA in different scenes of remote sensing images. The results of space-to-space scenes are shown in Tables III and IV. The results of spectrum to spectrum scenes are provided in Tables V and VI. In addition, The more complex scenes (both space to space and spectrum to spectrum) are considered in Tables VII and VIII. The tables show that introducing the TAA in GDA improves the mIoU performance by an average 12.57% (between 7.99% and 15.24%) and an average 0.88% (between 0.37% and 2.15%) compared with the source-only segmentation model and the Deeplab-V2 + DAT model. Digging further, it has been shown that the TAA module

TABLE V

SMALL DOMAIN ADAPTATION FROM IRRG POSTDAM TO RGB POSTDAM. PER-CLASS IOU: 4–8TH COLUMNS, OA: OVERALL PIXEL ACCURACY, AND MIOU: MEAN IOU

IRRG Postdam → RGB Postdam									
Method	Backbone	Approach	Imp. Surf.	Build.	Low veg.	Tree	Car	OA	mIoU
Source-only	ResNet-101	Seg.	70.60	78.64	53.36	53.92	68.51	78.65	65.00
MCD	ResNet-101	Adv-C.	74.67	83.53	57.61	62.76	76.34	82.33	70.98
AdasegNet(multi-level)	ResNet-101	Adv.	75.72	85.37	61.36	**64.40**	78.35	83.75	73.04
AdveNet(multi-level)	ResNet-101	Adv.	77.54	85.29	58.53	63.18	**80.25**	82.96	72.96
CLAN	ResNet-101	Adv.	**78.11**	85.89	59.51	61.11	79.99	83.42	72.92
Deeplab-V2+DAT(without TAA)	ResNet-101	Adv.	74.86	84.34	60.42	63.06	77.15	82.80	71.97
Deeplab-V2+GDA(without CDA)	ResNet-101	Adv.	76.35	85.79	**61.73**	63.69	77.42	84.04	72.99
Deeplab-V2+GDA+CDA(CaGAN)	ResNet-101	Adv.	77.76	**86.72**	61.44	62.83	77.95	**84.28**	**73.34**
Target-only	ResNet-101	Seg.	75.18	75.25	58.88	58.38	61.51	80.75	65.84

TABLE VI

LARGE DOMAIN ADAPTATION FROM FIRST 20-BAND PAVIAU TO SECOND 20-BAND PAVIAU. PER-CLASS IOU: 4–12TH COLUMNS, OA: OVERALL PIXEL ACCURACY, AND MIOU: MEAN IOU

First 20-band PaviaU → Second 20-band PaviaU													
Method	Backbone	Approach	Asphalt	Meadows	Gravel	Trees	metal sheets	Bare Soil	Bitumen	Self-Blocking Bricks	Shadows	OA	mIoU
Source-only	ResNet-101	Seg.	67.15	73.86	44.14	32.24	69.83	43.79	17.60	46.09	34.33	69.51	47.67
AdasegNet(multi-level)	ResNet-101	Adv.	74.06	**84.00**	38.12	39.57	79.06	91.16	**59.87**	67.20	30.64	76.06	62.63
AdveNet(multi-level)	ResNet-101	Adv.	**78.55**	82.28	16.97	**48.30**	72.86	**97.90**	39.62	**74.18**	28.10	**84.51**	59.86
Deeplab-V2+DAT(without TAA)	ResNet-101	Adv.	74.18	82.29	56.95	33.43	81.52	75.21	35.29	50.73	39.47	78.20	58.79
Deeplab-V2+GDA(without CDA)	ResNet-101	Adv.	74.55	74.98	62.92	38.04	82.51	65.30	23.72	66.50	43.90	77.84	59.16
Deeplab-V2+GDA+CDA(CaGAN)	ResNet-101	Adv.	75.64	76.37	**65.21**	42.32	**87.58**	72.32	39.28	66.75	**48.52**	80.70	**63.78**
Target-only	ResNet-101	Seg.	92.05	98.29	91.19	76.61	87.99	98.03	91.84	91.32	51.71	96.05	86.56

TABLE VII

ADAPTATION FROM RGB POSTDAM TO IRRG VAIHINGEN. PER-CLASS IOU: 4–8TH COLUMNS, OA: OVERALL PIXEL ACCURACY, AND MIOU: MEAN IOU

RGB Postdam → IRRG Vaihingen									
Method	Backbone	Approach	Imp. Surf.	Build.	Low veg.	Tree	Car	OA	mIoU
Source-only	ResNet-101	Seg.	41.64	52.73	11.82	26.71	23.93	44.47	31.37
AdasegNet(multi-level)	ResNet-101	Adv.	56.79	71.43	**44.10**	47.78	8.29	**68.87**	45.68
AdveNet(multi-level)	ResNet-101	Adv.	51.57	65.15	19.82	48.20	21.82	62.91	41.31
CLAN	ResNet-101	Adv.	61.38	**77.50**	40.16	22.00	24.10	66.60	45.03
Deeplab-V2+DAT(without TAA)	ResNet-101	Adv.	56.25	62.73	23.17	53.56	29.08	64.02	44.96
Deeplab-V2+GDA(without CDA)	ResNet-101	Adv.	57.09	66.49	16.40	56.85	28.90	65.68	45.14
Deeplab-V2+GDA+CDA(CaGAN)	ResNet-101	Adv.	**58.38**	68.26	21.70	**57.39**	**31.86**	66.96	**47.52**
Target-only	ResNet-101	Seg.	74.41	78.93	60.92	72.31	39.14	82.83	65.14

TABLE VIII

ADAPTATION FROM IRRG POSTDAM TO SYNTHETIC PANCHROMATIC VAIHINGEN. PER-CLASS IOU: 4–8TH COLUMNS, OA: OVERALL PIXEL ACCURACY, AND MIOU: MEAN IOU

IRRG Postdam → synthetic panchromatic Vaihingen									
Method	Backbone	Approach	Imp. Surf.	Build.	Low veg.	Tree	Car	OA	mIoU
Source-only	ResNet-101	Seg.	20.55	46.05	4.62	35.26	20.90	41.60	25.48
AdasegNet(multi-level)	ResNet-101	Adv.	43.96	**68.65**	18.65	36.76	24.04	58.33	38.41
Deeplab-V2+DAT(without TAA)	ResNet-101	Adv.	**46.10**	68.30	25.14	18.84	17.03	55.66	35.08
Deeplab-V2+GDA(without CDA)	ResNet-101	Adv.	39.95	60.86	**28.65**	30.73	25.96	56.89	37.23
Deeplab-V2+GDA+CDA(CaGAN)	ResNet-101	Adv.	41.11	61.50	20.69	**49.64**	**29.56**	**60.33**	**40.50**
Target-only	ResNet-101	Seg.	68.17	81.27	53.03	69.29	35.20	80.03	61.39

has more advantages in space-to-space scenes than spectrum-to-spectrum scenes, which is because the cross-space scenes have a more considerable domain shift. Overall, the results in different scenes of remote sensing images show that TAA can be used as a novel pipeline in the UDA semantic segmentation to enhance the global distribution alignment. Furthermore, the use of the CDA module improves the mIoU performance by more than 8.34% (between 8.34% and 17.65%) compared with the source-only model. Importantly, CDA can bring an average 2.40% (between 0.35% and 4.62%) performance compared with the Deeplab-V2 + GDA (without CDA), which performs better than introducing the TAA module in GDA. Thus, CDA can be used as a practical module in different scenes of remote sensing images.

3) Model Parameter Analysis: Table III shows the number of parameters under different methods and different backbones (including ResNet-101, VGG-16 and ResNet-50). It is shown that the introduction of TAA and CDA modules did not bring additional parameters. Thus, TAA and CDA in this article can be widely used in other related UDA semantic segmentation models, because these modules improve model performance without increasing the number of model parameters. In terms of different backbones, although the parameter amount of CaGAN with ResNet-101 is 2.7 times larger than that of CaGAN with VGG-16, the mIoU performance of CaGAN with ResNet-101 is improved by 7.5%. Thus, it has been demonstrated that the backbone has a stronger ability to express multiscale features in remote sensing images, and CaGAN will achieve better domain adaptation effects. Furthermore, CaGAN improves the mIoU performance by an average of 16% compared with the source-only model in different backbones, which shows the effectiveness of CaGAN in different models.

C. Comparing Methods

For a comprehensive evaluation, we compare our model with existing state-of-the-art methods, including MCD [28], AdasegNet [12], and Advenet [51] with multilevel and CLAN [37].

1) Source-only is a baseline model which only utilizes source domain data for training and tests on target data.
2) Target-only is a traditional supervised semantic segmentation model which means training target domain with annotations. It is necessary to verify the effectiveness and reliability of unsupervised adaptive methods in remote sensing image application scenarios, by the comparison with the performance of the supervised model (target-only).
3) MCD [28] is proposed to align distributions of source and target by utilizing the task-specific decision boundaries.
4) AdasegNet [12] with multilevel is an effective adversarial learning method for UDA in semantic segmentation, which performs output space domain adaptation at different feature levels by a multilevel adversarial network. Compared with adaptation in the feature space, output space adaptation with multilevel can contain rich information and low computational cost for semantic segmentation.
5) Advenet [51] with multilevel is presented to address domain shift in the UDA semantic segmentation based on the entropy-based adversarial training approach. An entropy-based loss is proposed to directly penalize low-confident predictions on the target domain, which does not add significant overhead to segmentation frameworks. In addition, the entropy-based adaptation scheme is performed on multilevel outputs coming from different scale features.
6) CLAN [37] is to align category-level joint distributions by adaptively weighting the adversarial loss for different features. CLAN is a cooperative adversarial training approach by increasing the weight of adversarial loss for category-level features poorly aligned.
7) The existing adversarial learning methods proposed for the UDA semantic segmentation of the same remote sensing data sets (such as Postdam and Vaihingen) are considered, including Siamese-based GAN [20], traditional GAN [27] and conditional GAN [25].
8) Deeplab-V2 + DAT is the traditional domain adversarial training method without applying the TAA and CDA modules; Deeplab-V2 + GDA is our proposed GDA (DAT + TAA) adversarial training method without applying the CDA module; Deeplab-V2 + GDA + CDA is our full method that employs the CDA module to Deeplab-V2 + GDA.

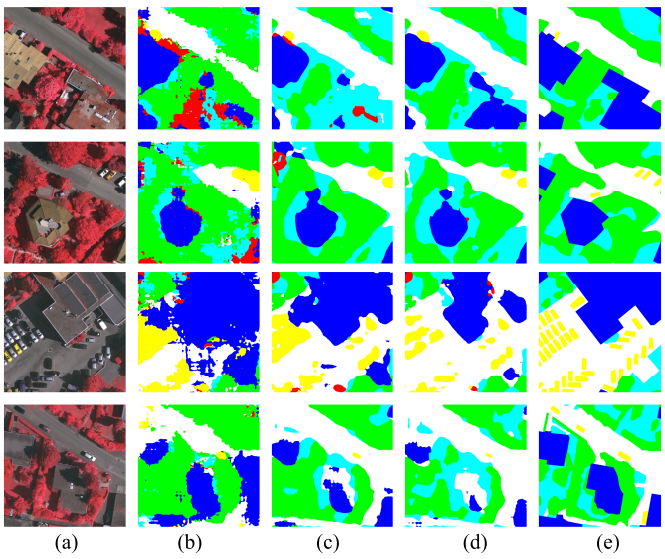

Fig. 6. UDA segmentation for the space-to-space scene (IRRG Postdam → IRRG Vaihingen). Legend–blue: buildings, white: impervious surfaces, green: trees, yellow: cars, cyan: low vegetation. (a) Target image. (b) Before DA (Source only). (c) After DA (Adapt with multilevel). (d) After DA (CaGAN). (e) Ground truth.

D. UDA Results in Different Scenarios

In all the tables, the values in **bold** are the best and the values underlined are the second best. "Seg," "Adv," and "Adv-C" represent the segmentation-based source, adversarial learning, and task-specific classifiers adversarial learning-based DA, respectively. "Adv-R*" denotes the most excellent adversarial learning method proposed for remote sensing images, where * represents the results directly selected from the corresponding article.

1) UDA Cases of Space to Space Scenes: The adaptation results on tasks IRRG Postdam → IRRG Vaihingen and IRRG Vaihingen → IRRG Postdam are given in Tables III and IV, respectively. These tasks focused on the cases of space to space scenes in remote sensing data. The baseline method is a source-only model achieving an mIoU of 34.19% for IRRG Postdam to IRRG Vaihingen, and an mIoU of 26.31% for IRRG Vaihingen to IRRG Postdam. CaGAN significantly outperforms the two source-only segmentation methods by 17.65% and 16.54%, respectively. Also, CaGAN outperforms

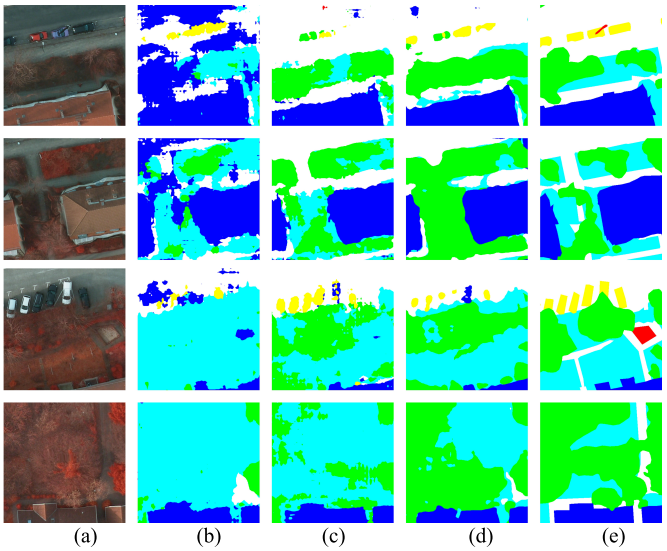

Fig. 7. UDA segmentation for the space-to space scene (IRRG Vaihingen → IRRG Postdam). Legend–blue: buildings, white: impervious surfaces, green: trees, yellow: cars, cyan: low vegetation. (a) Target image. (b) Before DA (Source only). (c) After DA (Adapt with multilevel). (d) After DA (CaGAN). (e) Ground truth.

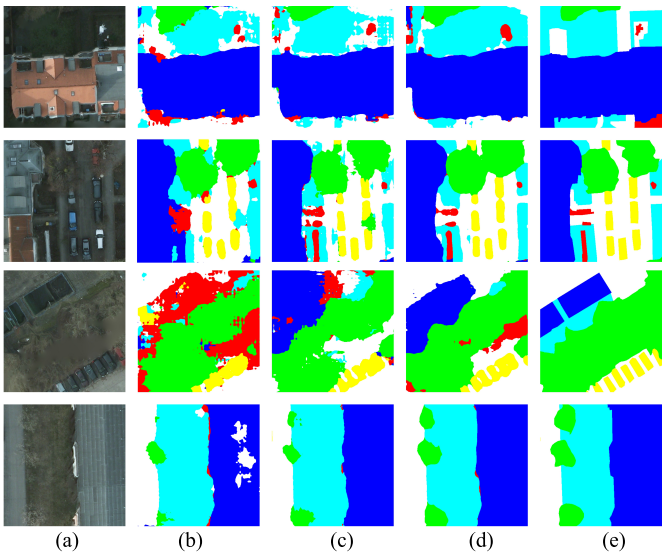

Fig. 8. UDA segmentation for the spectrum to spectrum scene (IRRG Postdam → RGB Postdam). Legend–blue: buildings, white: impervious surfaces, green: trees, yellow: cars, cyan: low vegetation. (a) Target images. (b) Before DA (Source only). (c) After DA (Adapt with multilevel). (d) After DA (CaGAN). (e) Ground truth.

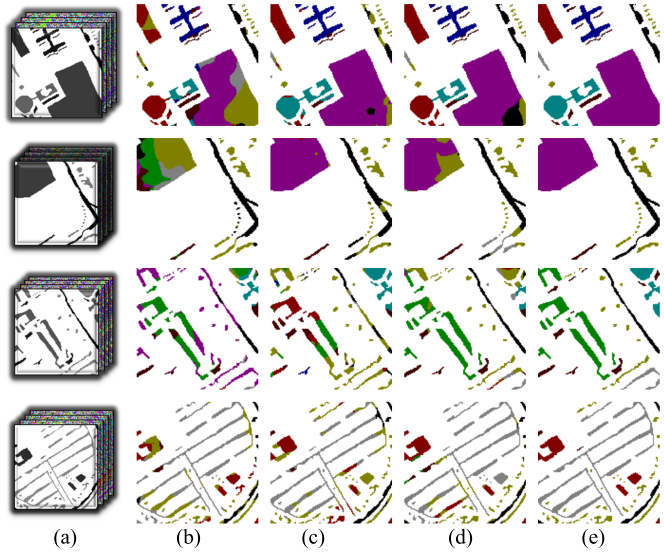

Fig. 9. UDA segmentation for the spectrum to spectrum scene (First 20-band PaviaU → Second 20-band PaviaU). Legend–black:Asphalt, maroon: Meadows, green: Gravel, olive: Trees, navy: Metal sheets, purple: Bare Soil, teal: Bitumen, gray: Self-blocking bricks, red: Shadows.

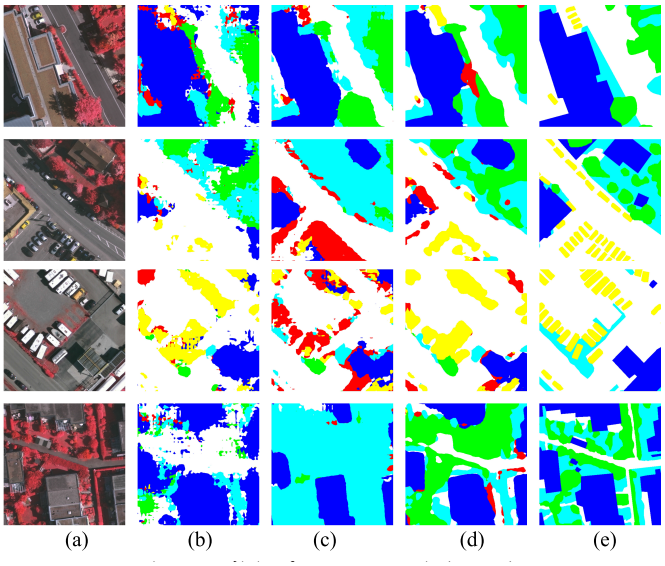

Fig. 10. UDA segmentation for the space to space and spectrum to spectrum scene (RGB Postdam → IRRG Vaihingen). Legend–blue: buildings, white: impervious surfaces, green: trees, yellow: cars, cyan: low vegetation. (a) Target images. (b) Before DA (Source only). (c) After DA (Adapt with multilevel). (d) After DA (CaGAN). (e) Ground truth.

the AdasegNet [12] with a multilevel adaptation model by 1.64% for IRRG Postdam to IRRG Vaihingen and 0.79% for IRRG Vaihingen to IRRG Postdam, which is the second-best among the benchmark methods. Importantly, although the transfer tasks on IRRG Postdam to IRRG Vaihingen and IRRG Vaihingen to IRRG Postdam have the same domain difference, it has been shown that the different source domain (different transfer order) can lead to a considerable difference in performance by further comparing the two cross-space domain adaptation tasks. Specifically, the number of samples (the spatial features) in the source domain is more sufficient, and the effect of domain alignment is better. Furthermore, per-class mIoU is computed to assess the segmentation performance for different objects. As shown in the quantitative results, CaGAN produces better segmentation results. In terms of IRRG Postdam to IRRG Vaihingen, cars, impervious surfaces, and building are better; impervious surfaces and building are better for IRRG Vaihingen to IRRG Postdam.

We also compare the CaGAN with the adversarial learning methods proposed for remote sensing images, for example, in terms of the segmentation adaptation from IRRG Potsdam to IRRG Vaihingen, Siamese-based GAN [20] achieves an mIoU of 51.5%, Benjdira *et al.* [27] achieves a mIoU of 30% based on traditional GAN, Liu *et al.* [25] reaches 38.70% from IRRG Vaihingen to IRRG Postdam based on conditional GAN. Overall, our CaGAN achieves higher results than these methods used in space to space remote sensing images.

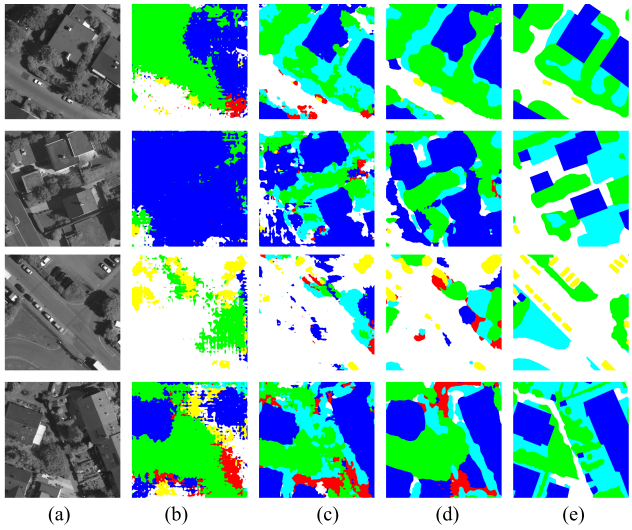

Fig. 11. UDA segmentation for the space to space and spectrum to spectrum scene (IRRG Postdam → synthetic panchromatic Vaihingen). Legend–blue: buildings, white: impervious surfaces, green: trees, yellow: cars, cyan: low vegetation. (a) Target images. (b) Before DA (Source only). (c) After DA (Adapt with multilevel). (d) After DA (CaGAN). (e) Ground truth.

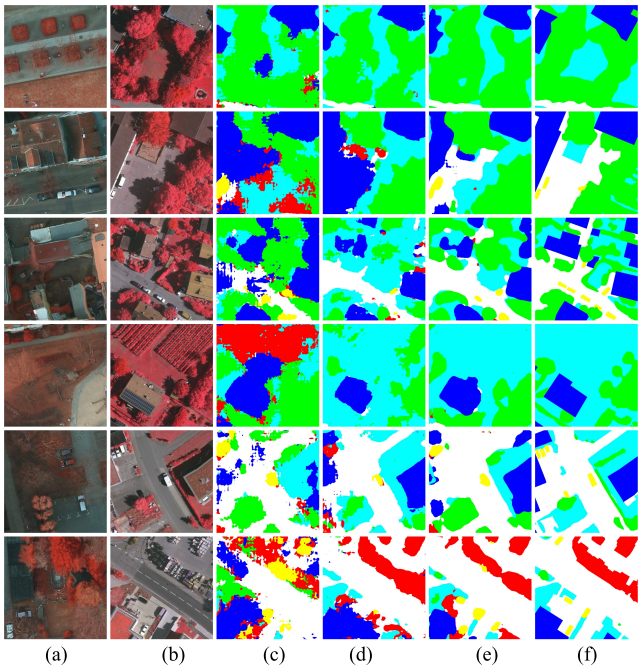

Fig. 12. More results on UDA segmentation for the space to space scene (IRRG Postdam → IRRG Vaihingen). Legend–white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars. (a) Source images. (b) Target images. (c) Before DA (Source only). (d) After DA (Adapt with multilevel). (e) After DA (CaGAN). (f) Ground truth.

2) UDA Cases of Spectrum to Spectrum Scenes: The same objects with different spectra in different remote sensing images can also lead to huge domain differences. To verify the effectiveness of our proposed CaGAN in cross-spectrum scenes, we conduct a 3-band domain adaptation task from IRRG Postdam to RGB Postdam (small domain shift) and a 20-band domain adaptation task from first 20-band PaviaU to second 20-band PaviaU (large domain shift).

a) Small domain shift from IRRG Postdam to RGB Postdam: Table V presents the comparison of our proposed CaGAN and other state-of-the-art methods from IRRG Postdam to RGB Postdam, which focused on the cross-spectrum scenes with a small number of bands. As shown in Table V, the mIoU and OA for our CaGAN are higher than those competitive methods. In terms of mIoU, CaGAN outperforms the baseline model (source-only) by 8.3%, and state-of-the-art method (AdasegNet [12] with a multilevel adaptation model)

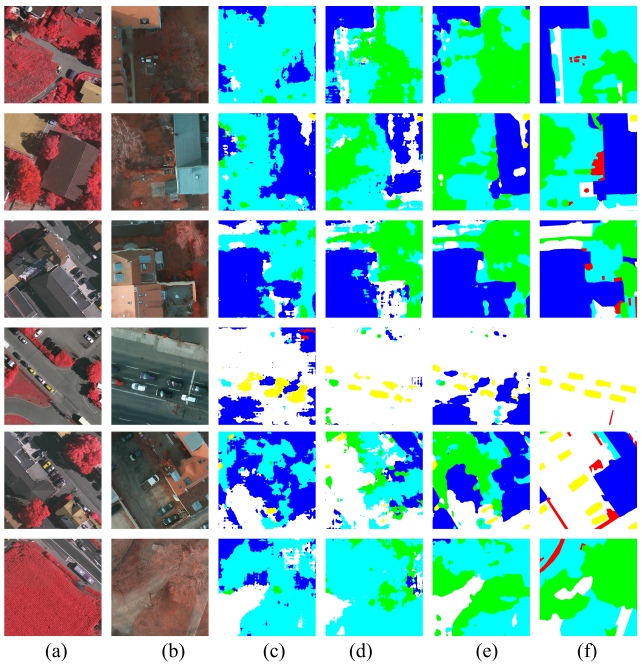

Fig. 13. More results on UDA segmentation for the space to space scene (IRRG Vaihingen → IRRG Postdam). Legend–white: impervious surfaces, blue: buildings,cyan: low vegetation, green: trees, yellow: cars. (a) Source images. (b) Target images. (c) Before DA (Source only). (d) After DA (Adapt with multilevel). (e) After DA (CaGAN). (f) Ground truth.

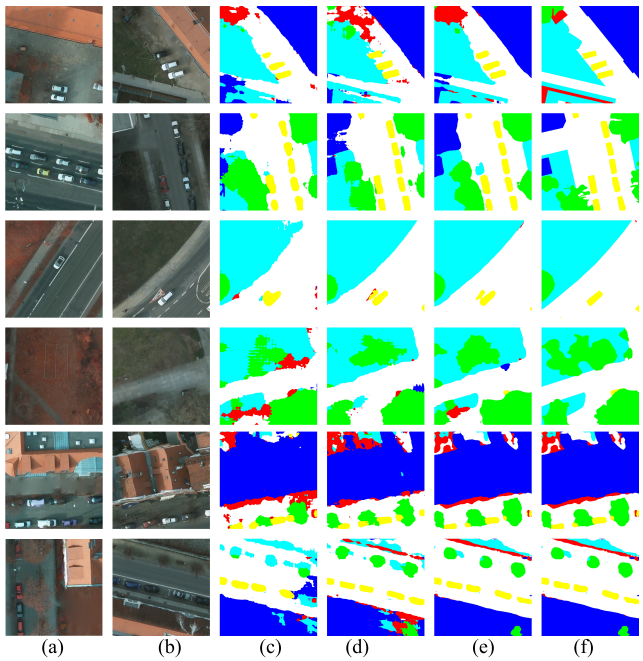

Fig. 14. More results on UDA segmentation for the spectrum to spectrum scene (IRRG Postdam → RGB Postdam). Legend—white: impervious surfaces, blue: buildings,cyan: low vegetation, green: trees, yellow: cars. (a) Source images. (b) Target images. (c) Before DA (Source only). (d) After DA (Adapt with multilevel). (e) After DA (CaGAN). (f) Ground truth.

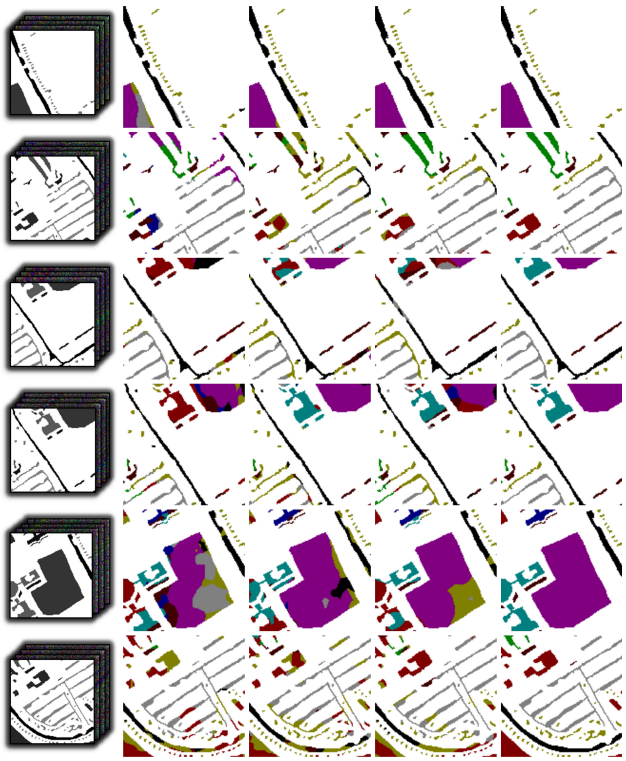

Fig. 15. More results on UDA segmentation for the spectrum to spectrum scene (First 20-band PaviaU → Second 20-band PaviaU). Legend—black: Asphalt, maroon: Meadows, green: Gravel, olive: Trees, navy: Metal sheets, purple: Bare Soil, teal: Bitumen, gray: Self-Blocking Bricks, red: Gravel. (a) Target images. (b) Before DA (Source only). (c) After DA (Adapt with multilevel). (d) After DA (CaGAN). (e) Ground truth.

by 0.3%. Specifically, it has been shown that CaGAN improves the IoU performance in each category by more than 7% (between 7% and 10%) compared with the source-only model and it has more advantages in building and low vegetation compared with other methods. Furthermore, it is worth noting that the performance of the methods after applying domain adaptation is better than that of a fully supervised model (target-only) for the same space scenes with small spectral domain differences, especially for our proposed CaGAN. This phenomenon indicates that the domain adaptive methods can increase the distinguishability of features by adversarial training. In addition, domain adaptive methods can also learn more complete context features in remote sensing images by adding the adversarial loss, which is similar to the inpainting tasks from context information [52].

b) Large domain shift from first 20-band PaviaU to second 20-band PaviaU: Table VI provides the comparative results on the Hyperspectral data sets, first 20-band PaviaU → second 20-band PaviaU, which focused on the cross-spectrum scenes with a large number of bands. To facilitate the direct use of Deeplab-V2, HSIs are converted into three-channel images by principal components analysis (PCA) [53]. As Table VI shows, equipped with ResNet-101 backbone, CaGAN outperforms the source-only segmentation model by 16.11%. Besides, CaGAN also outperforms state-of-the-art methods by over 1% compared with AdasegNet [12] and AdveNet [51] with multilevel adaptation model. However,

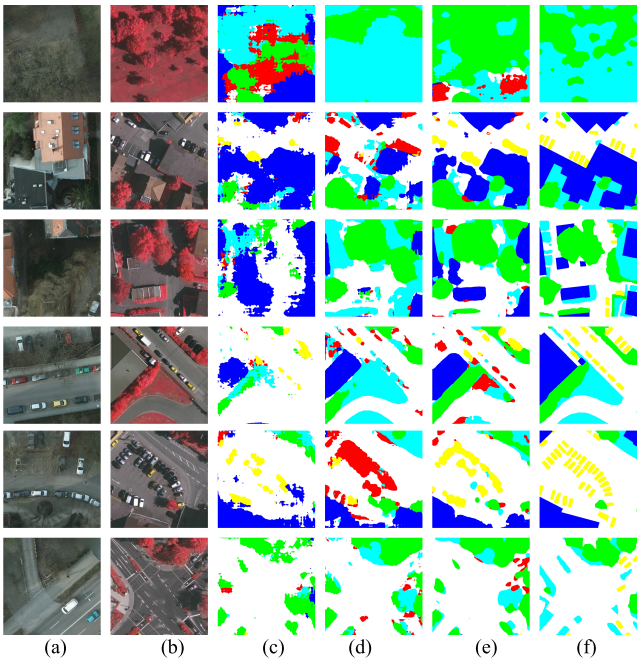

Fig. 16. More results on UDA segmentation for the space to space and spectrum to spectrum scene (RGB Postdam → IRRG Vaihingen). Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars. (a) Source images. (b) Target images. (c) Before DA (Source only). (d) After DA (Adapt with multilevel). (e) After DA (CaGAN). (f) Ground truth.

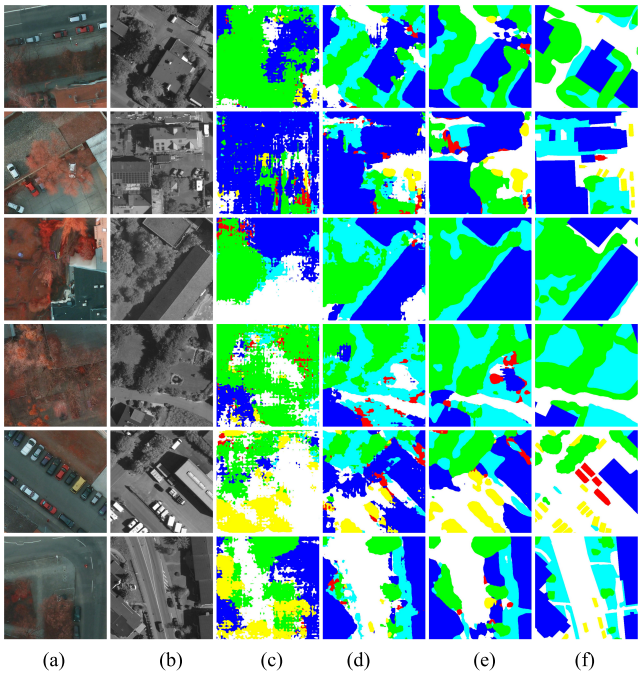

Fig. 17. More results on UDA segmentation for the space to space and spectrum to spectrum scene (IRRG Postdam → synthetic panchromatic Vaihingen). Legend—white: impervious surfaces, blue: buildings, cyan: low vegetation, green: trees, yellow: cars. (a) Source images. (b) Target images. (c) Before DA (Source only). (d) After DA (Adapt with multilevel). (e) After DA (CaGAN). (f) Ground truth.

compared to the stronger baseline, we observe a significant drop in class "Bare Soil" and "Bitumen." Note that it is due to the large layout gaps in the hyperspectral data sets, where the hyperspectral objects have fewer pixels and more background pixels, and the sample pixel distributions are

extremely unbalanced. Thus, early stopping has been taken during the model training to prevent target data over-fitting.

3) UDA Cases of Space to Space and Spectrum to Spectrum Scenes: To further verify the effectiveness of CaGAN in the remote sensing data sets of more complex and larger domain shift (both spatial domain and spectral domain differences are considered), we construct two space to space and spectrum to spectrum domain adaptation tasks, including RGB Postdam to IRGB Vaihingen and IRRG Postdam to synthetic panchromatic Vaihingen.

Table VII presents comparative results of CaGAN and three leading models on task RGB Postdam → IRRG Vaihingen. It has been demonstrated that CaGAN is competitive in the domain adaptation semantic segmentation of cross-space and cross-spectrum remote sensing images by comparing with these results. Specifically, our proposed CaGAN has increased mIoU from 31.37% of the baseline (source-only) to 47.52%. CaGAN outperforms the strong model (AdasegNet [12] with a multilevel adaptation model) by 1.84%. Notably, the IoU of a car is only 8.29% in AdasegNet, whose performance degrades compared to source-only. This phenomenon is caused by the limitation of global distribution alignment, which leads to "deterioration of inter-class performance." In contrast, CaGAN can avoid this deterioration, achieving a car IoU of 31.86%.

Table VIII shows the results on task IRRG Postdam → synthetic panchromatic Vaihingen. We compare our model with the best baseline AdasegNet with a multilevel adaptation model [12] in the first task. Without domain adaptation, the model trained only on the source domain achieves a mIoU of 25.48%. Our model achieves 15.02% and 2.09% improvements compared to source-only and AdasegNet, respectively.

E. Qualitative Results

Qualitative UDA segmentation results of space to space, spectrum to spectrum, and both space to space and spectrum to spectrum are presented in Figs. 6–11, respectively. The second column demonstrates the results before DA based source only. The segmentation results of different classes in cross-space domains (including space to space scenes, both space to space and spectrum to spectrum scenes) are poor while relatively better in spectrum to spectrum scenes. We argue that it is due to the fact that spectrum to spectrum scenes perform less domain shift and more spatial features are originally aligned between the same targets with different bands in remote sensing data. The third and fourth columns demonstrate the results after DA based AdasegNet with a multilevel adaption model [12] and CaGAN model, respectively. The comparison between the third and fourth columns show that our model provides relatively accurate predictions on target images (especially in both space to space and spectrum to spectrum scenes). Furthermore, CaGAN can effectively capture the small objects in the remote sensing images and obtain more precise boundaries of objects for UDA in the multisource remote sensing images. To make the experimental results more solid and convincing, more qualitative results are given in Figs. 12–17.

IV. Conclusion

In this article, we propose CaGAN to strengthen the GDA and explicitly model the intra-class and the inter-class discrepancies in the UDA semantic segmentation. CaGAN is trained in an end-to-end asymptotic training manner. Experimental results on domain adaptation of space to space, spectrum to spectrum, both space to space and spectrum to spectrum data sets demonstrate that CaGAN outperforms the state-of-the-art remote sensing UDA semantic segmentation methods. Ablation studies show that GDA encourages the generator to capture the domain-invariant aspect of pixels, as a result of reducing the intra-class dispersion, while CDA further models the intraclass compactness and the interclass separability. However, the proposed CaGAN framework lacks prior knowledge of remote sensing images (such as the spectral curve features, geospatial information, and the spatial distribution of objects in remote sensing images). In future works, this will be investigated.

References

[1] R. Mottaghi *et al.*, "The role of context for object detection and semantic segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 891–898.

[2] A. Youssef, D. Albani, D. Nardi, and D. D. Bloisi, "Fast traffic sign recognition using color segmentation and deep convolutional networks," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.*, vol. 10016, 2016, pp. 205–216.

[3] R. Cao *et al.*, "Deep learning-based remote and social sensing data fusion for urban region function recognition," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 82–97, May 2020.

[4] C. Ouyang *et al.*, "Early identification and dynamic processes of ridge-top rockslides: Implications from the Su village landslide in Suichang county, Zhejiang Province, China," *Landslides*, vol. 16, no. 4, pp. 799–813, Apr. 2019.

[5] K. Wei *et al.*, "Reflections on the catastrophic 2020 Yangtze river basin flooding in Southern China," *Innovation*, vol. 1, no. 2, Aug. 2020, Art. no. 100038.

[6] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, Mar. 2018.

[7] Z. Niu, W. Liu, J. Zhao, and G. Jiang, "DeepLab-based spatial feature extraction for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 251–255, Feb. 2019.

[8] Q. Xu, C. Ouyang, T. Jiang, X. Fan, and D. Cheng, "DFPENet-geology: A deep learning framework for high precision recognition and segmentation of co-seismic landslides," 2019, *arXiv:1908.10907*. [Online]. Available: http://arxiv.org/abs/1908.10907

[9] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.

[10] 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, nos. 1–2, pp. 151–175, May 2010.

[11] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv:1612.02649*. [Online]. Available: http://arxiv.org/abs/1612.02649

[12] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[13] S. Lee, D. Kim, N. Kim, and S.-G. Jeong, "Drop to adapt: Learning discriminative features for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 91–100.

[14] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced Wasserstein discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10285–10295.

[15] C. Chen *et al.*, "Progressive feature alignment for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 627–636.

[16] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[17] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 2208–2217.
[18] A. Gretton *et al.*, "Optimal kernel choice for large-scale two-sample tests," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1205–1213.
[19] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015, *arXiv:1502.02791*. [Online]. Available: http://arxiv.org/abs/1502.02791
[20] L. Yan, B. Fan, S. Xiang, and C. Pan, "Adversarial domain adaptation with a domain similarity discriminator for semantic segmentation of urban areas," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1583–1587.
[21] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.
[22] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.
[23] B. Benjdira, A. Ammar, A. Koubaa, and K. Ouni, "Data-efficient domain adaptation for semantic segmentation of aerial imagery using generative adversarial networks," *Appl. Sci.*, vol. 10, no. 3, p. 1092, Feb. 2020.
[24] Q. Wang, J. Gao, and X. Li, "Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4376–4386, Sep. 2019.
[25] W. Liu and F. Su, "Unsupervised adversarial domain adaptation network for semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1978–1982, Nov. 2020.
[26] Y. Yu, X. Li, and F. Liu, "Attention GANs: Unsupervised deep feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 519–531, Jan. 2020.
[27] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, p. 1369, Jun. 2019.
[28] 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.
[29] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.
[30] P. Xu, P. Gurram, G. Whipps, and R. Chellappa, "Wasserstein distance based domain adaptation for object detection," 2019, *arXiv:1909.08675*. [Online]. Available: http://arxiv.org/abs/1909.08675
[31] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," 2019, *arXiv:1907.12859*. [Online]. Available: http://arxiv.org/abs/1907.12859
[32] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," 2019, *arXiv:1909.11825*. [Online]. Available: http://arxiv.org/abs/1909.11825
[33] X. Chen, C. Xu, X. Yang, and D. Tao, "Attention-GAN for object transfiguration in wild images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 164–180.
[34] G. Kang, L. Zheng, Y. Yan, and Y. Yang, "Deep adversarial attention alignment for unsupervised domain adaptation: The benefit of target expectation maximization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 401–416.
[35] X. Wang, L. Li, W. Ye, M. Long, and J. Wang, "Transferable attention for domain adaptation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 5345–5352.
[36] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4893–4902.
[37] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.
[38] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2239–2247.
[39] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," 2019, *arXiv:1910.13049*. [Online]. Available: http://arxiv.org/abs/1910.13049
[40] Q. Feng, G. Kang, H. Fan, and Y. Yang, "Attract or distract: Exploit the margin of open set," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7990–7999.
[41] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for Image-to-Image translation," 2019, *arXiv:1908.06616*. [Online]. Available: http://arxiv.org/abs/1908.06616
[42] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2014, *arXiv:1409.7495*. [Online]. Available: http://arxiv.org/abs/1409.7495
[43] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *J. Mach. Learn. Res.*, vol. 9, no. 8, pp. 1757–1774, 2008.
[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[47] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[48] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.
[49] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[50] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2090–2099.
[51] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2517–2526.
[52] R. Uittenbogaard, C. Sebastian, J. Viiverberg, B. Boom, and P. H. N. de With, "Conditional transfer with dense residual attention: Synthesizing traffic signs from street-view imagery," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 553–559.
[53] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.

Qingsong Xu received the B.S. degree in geotechnical engineering from the Chengdu University of Technology, Chengdu, China, in 2018. He is pursuing the master's degree with the Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Beijing, China.

His research interests include unsupervised/self-supervised semantic segmentation, remote sensing image perception and cognition, and combination of machine learning (deep reinforcement learning, graph convolution) and dynamic pysical model.

Xin Yuan (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from Xidian University, Xi'an, China, in 2007 and 2009, respectively, and the Ph.D. degree from the Hong Kong Polytechnic University, Hong Kong, in 2012.

He had been a Post-Doctoral Associate with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, from 2012 to 2015. He is a Video Analysis and Coding Lead Researcher with Bell Labs, Murray Hill, NJ, USA. He is the author of 2 book chapters and more than 100 journal and conference articles. He holds dozens of international patents. His research interests are in signal processing, computational imaging and machine learning.

Chaojun Ouyang received the B.S. and Ph.D. degrees from the Department of Mechanics, Huazhong University of Science and Technology, Huazhong, China, in 2005 and 2009, respectively.

He is a Researcher with the Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Beijing, China. His research interest lies in computational fluid dynamics, machine learning, remote sensing, and construction of dynamic model of landslide and debris flow.