A Fine-Grained Unsupervised Domain Adaptation Framework for Semantic Segmentation of Remote Sensing Images

Luhan Wang, Pengfeng Xiao, *Senior Member, IEEE*, Xueliang Zhang, *Member, IEEE*, and Xinyang Chen

***Abstract*—Unsupervised domain adaptation (UDA) aims at adapting a model from the source domain to the target domain by tackling the issue of domain shift. Cross-domain segmentation of remote sensing images (RSIs) remains a big challenge due to the unique properties of RSIs. On the one hand, the divergence of data distribution in different local regions leads to negative transfer by directly applying the global alignment method in RSIs. On the other hand, the underlying category-level structure in the target domain is often ignored, which confuses the decision of semantic boundaries on the dispersed category features caused by large intraclass variance and small interclass variance in RSIs. In this study, we propose a novel fine-grained adaptation framework combining two stages of global-local alignment and category-level alignment to solve the above-mentioned problems. In the first stage of global-local adaptation, an attention map is derived from an intermediate discriminator and focuses on hard-to-align regions to mitigate negative transfer due to global adversarial learning. In the second stage of category-level adaptation, the category feature compact module is utilized to address the issue of dispersed features in the target domain attained by the cross-domain network, which will facilitate the fine-grained alignment of categories. Experiments under various scenarios, including geographic location variation and spectral band composition variation, demonstrate that the local adaptation and category-level adaptation of RSIs are complementary in the cross-domain segmentation, and the integrated framework helps achieve outstanding performance for UDA semantic segmentation of RSIs.**

***Index Terms*—Adversarial learning, category-level alignment, global-local alignment, remote sensing images (RSIs), semantic segmentation, unsupervised domain adaptation (UDA).**

I. Introduction

With the rapid development of remote sensing technologies, massive high-resolution remote sensing images (RSIs) are becoming widely available for Earth observation. Semantic segmentation of RSIs is a pixel-wise classification task that assigns a category label to every pixel in an input image [1], which plays an increasingly significant role in many applications, such as land management [2], traffic analysis [3], and disaster prevention [4].

Deep learning methods [5], [6], [7], [8] show promising results in semantic segmentation of RSIs [9], [10], [11]. However, semantic segmentation model training requires huge amounts of pixel-level ground truth data obtained by expensive human labor [12], which makes it more difficult for RSIs that cover large areas and have various objects. In addition, large domain shift in RSIs, which refers to data discrepancy caused by different illuminations, geographic locations, and wavelength bands, usually degrades model accuracy in practical applications. Therefore, it is necessary to reduce the labeling cost and improve the generalization ability of segmentation models when faced with domain shift.

Unsupervised domain adaptation (UDA) transfers the model from the source domain to the target domain in an unsupervised way by tackling the issue of domain shift. Generally, the samples in the source domain are labeled but not in the target domain, and the data distribution of the source domain is inconsistent with the target domain. To reduce the upper limit of error in the target domain, the core concept is to minimize the difference between domains [13]. Originally, several distance metrics were utilized to fulfill this goal [13], [14], [15].

Adversarial learning employing generative adversarial network (GAN) structure [16] is widely used to learn domain-invariant features by generators and discriminators. Adversarial UDA methods can be divided into image level [17], [18], [19], feature level [20], [21], [22], and output level [23], [24], [25]. The image-level adversarial methods translate images from the source domain to the target domain by cycle-consistency, and train the network by the translated images, in which the performance of UDA is affected by the quality of image-to-image translation. The approaches of feature-level and output-level introduce a discriminator alongside the segmentation network to align cross-domain features and semantic structures. Some recent studies [26], [27], [28], [29] combined these UDA methods to better alleviate domain shift, which have been applied in cross-domain segmentation of RSIs [30], [31], [32], [33], [34], [35], [36], [37]. For geographic location variation and spectral band composition variation in multisource RSIs, an image-level method [34] through DualGAN [38] was introduced to generate

Manuscript received 15 February 2023; revised 27 March 2023; accepted 21 April 2023. Date of publication 25 April 2023; date of current version 3 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297 and Grant 41871235, in part by the Fundamental Research Funds for the Central Universities under Grant 020914380095, and in part by the High-Level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China. *(Corresponding author: Pengfeng Xiao.)*

The authors are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: mg21270086@smail.nju.edu.cn; xiaopf@nju.edu.cn; zxl@nju.edu.cn; mg21270060@smail.nju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3270302

images with similar styles to the target domain, and then multiple weak supervision constraints were added to overcome the difficulties of complex image structure. Deng et al. [31] designed an adaptation network by output-level adversarial learning to extract cross-city roads and cross-country buildings from large-scale RSIs. In addition, a full-space adaptation strategy was proposed to perform cross-domain land cover classification of RSIs [33]. Recently, MBATrans combined Transformer and GAN-based network to align the high-level features [36]. MemoryAdaptNet embed an invariant feature memory module to preserve domain-level information in adversarial learning, which overcomes the deficiency of pseudo invariant features [37].

Although the above-mentioned methods have been applied in UDA semantic segmentation, most of them aligned the source domain to the target domain from a global perspective, which ignores the implicit local alignment and the explicit category-level alignment, resulting in negative transfer accordingly. To achieve local alignment, many methods [39], [40], [41], [42], [43], [44] focus on regions with large domain shift. For example, Luo et al. [39] and [40] conducted cotraining of two classifiers to obtain a local alignment score map, where the high score represents the difficult alignment regions, and then weighted the score map into the adversarial loss. An entropy-guided adaptation algorithm was proposed in [44] to decrease the wrong mapping of well-aligned features in UDA of aerial images, in which the target entropy maps were used to measure interdomain discrepancy. Boundaries of object are usually clear in entropy maps while the interiors tend to be confused, which means entropy-guided local alignment fails to achieve satisfactory performance within objects. To achieve category-level alignment, several studies considered learning category structure explicitly [45], [46], [47], [48], [49]. The methods in [45] and [46] implemented fine-grained alignment by category-level adversarial learning while preserving the internal structure of cross-domain semantics. Such an idea was also exploited in [47] and [48] to model the category discrepancies between different domains for adaptive segmentation in RSIs. However, the structure of category-level discriminator is complex, resulting in more difficult training process in adversarial learning for UDA semantic segmentation.

The above-mentioned studies have promoted the development of cross-domain adaptive segmentation, but they still suffer from two limitations in the remote sensing community due to the characteristics of RSIs. First, RSIs often cover a wider range of areas and have more complex structure than natural scene images, which results in various lighting conditions, imaging angles, and object sizes in different local regions of an image [50]. Due to the rich diversities of local features in RSIs, the issue of spatial discrepancies should be taken seriously. In addition, different geographic locations and different imaging sensors in cross-domain RSIs enlarge the diversities of category features (e.g., category layout, ratio, and color presentation), which brings more challenges compared with the UDA segmentation in computer vision field [51]. However, recent studies either perform various degrees of adaptation to different spatial regions or pay attention to the underlying semantic structure of categories, which neglects both the spatial domain-invariant features and semantic domain-invariant structure among categories are crucial to UDA segmentation of RSIs. Second, the cross-domain model produce dispersed category features in the target domain since there are no segmentation labels in the target domain and the improvement of feature transferability leads to the decrease of category separability. Moreover, the large intraclass variance and small interclass variance in RSIs further exacerbate this issue, resulting in unsatisfactory segmentation. However, most of the current category-level alignment ignored tackling the issue of dispersed features.

In order to handle the above-mentioned challenges of unsupervised domain adaptive semantic segmentation of RSIs, we expect that the fine-grained local alignment and the category-level alignment could be complementary to global alignment for improving the performance of UDA semantic segmentation. Hence, we propose a two-stage framework to effectively integrate both local and category-level alignment, and compact category features at the same time. Fig. 1 illustrates the main process of the proposed framework including global adaptation, local adaptation, and category-level adaptation. Specifically, in stage 1, we achieve both the local and global alignment by the game between segmentation network and two discriminators. Since global adaptation will lead to negative transfer in domain-invariant regions and unsatisfactory adaptation in hard patches, we use the confidence obtained from intermediate discriminator as the attention map to align various local regions in RSIs. We overcome the shortage of concentrating on boundary but ignoring interior in entropy-guided map by means of discriminator confidence map. However, the process in stage 1 does not explicitly incorporate category information, resulting in dispersed feature distribution in the target domain, which is difficult for the linear classifier to distinguish in semantic segmentation. Hence, we expect to pursue category-level alignment in stage 2, where the category feature compact module is used to relieve the issue of feature dispersion. Specifically, it utilizes prototypes within teacher-student network to maximize interclass variance and minimize intraclass variance. The main contributions of this study can be summarized as follows.

1) A two-stage framework for UDA semantic segmentation of RSIs is proposed, which achieves fine-grained local and category-level alignment on top of global alignment. It considers the spatial domain discrepancy and semantic domain discrepancy when eliminating the domain shift.
2) The category feature compact module is designed to relieve the issue of target domain feature dispersion which affects adaptive segmentation performance. We use prototypes that created by adaptive network in stage 1 as category anchors to learn the underlying and compact category structure within the teacher-student network.
3) The proposed unified UDA framework achieves outstanding performance in the cross-domain segmentation of RSIs.

II. Method

A. Method Overview

We are concerned with the problem of model adaptation across domains, where the source images x_s with the

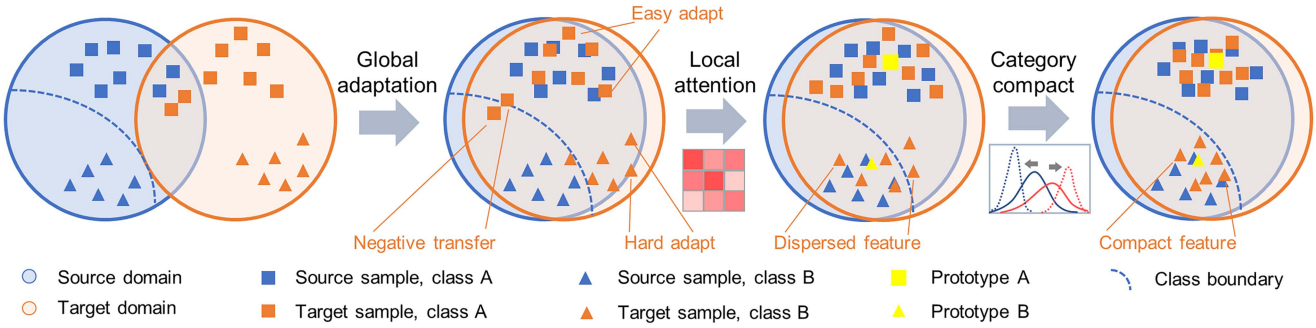


Fig. 1. Illustration of global adaptation, local adaptation, and category-level adaptation. Global adversarial learning aligns all regions in an image equally; local adaptation focuses on hard regions to alleviate negative transfer by discriminator attention; and prototypes are utilized to compact category feature in the target domain for better category-level alignment.

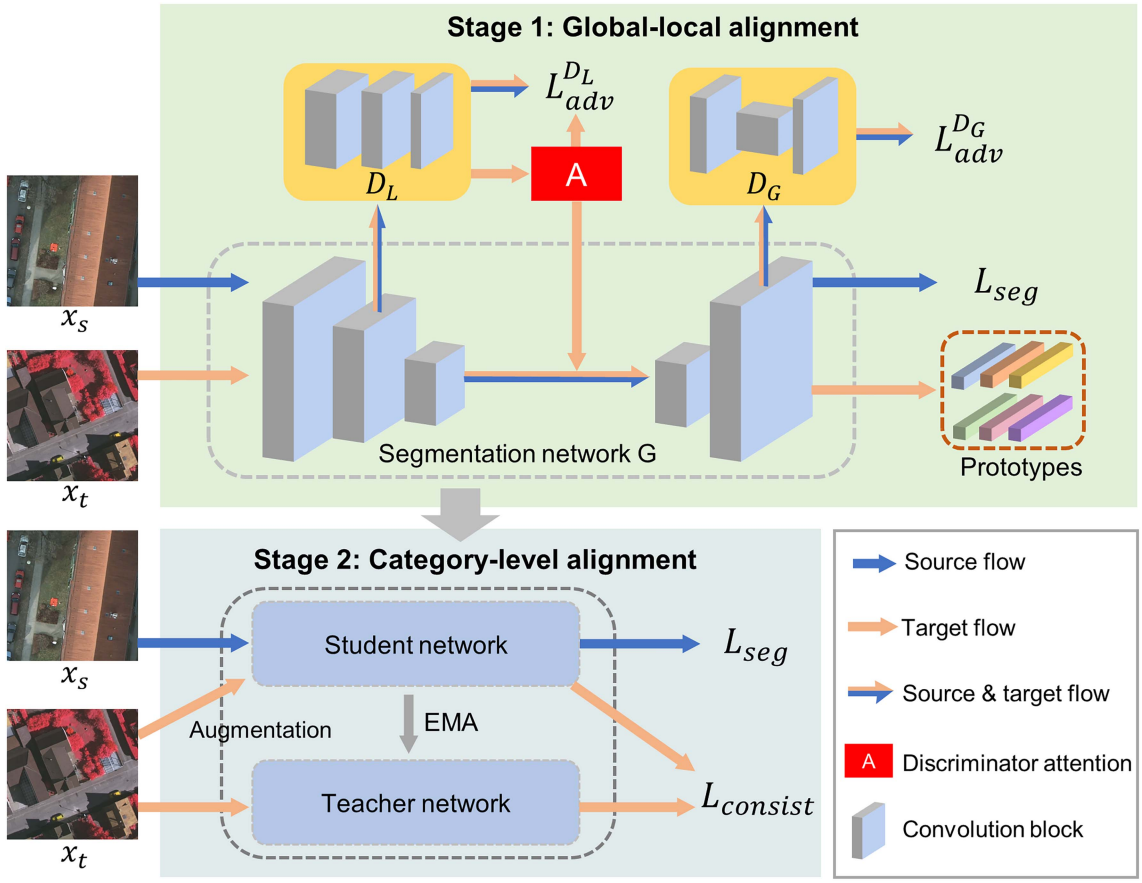


Fig. 2. Overview of the proposed two-stage UDA framework. In the first stage, attention map generated by discriminator scores tends to implement local alignment; and in the second stage, the teacher-student network updated by EMA (exponential moving average) is used to align category features.

corresponding ground truth y_s and the target images x_t are utilized during training, aiming at improving the transferability of the segmentation network G consisting of a feature extractor F and a classifier C.

For the challenge of adaptive segmentation of RSIs, we propose a two-stage framework to gradually align two different domains. The model framework is shown in Fig. 2, including the first stage of global-local alignment and the second stage of category-level alignment. According to previous studies [23], [24], it is known that global discriminator D_G at output level along with the segmentation network G serving as a generator can effectively reduce the marginal distribution discrepancies by adversarial learning. However, the insufficiently refined alignment ignores the local detail differences. We therefore go a step further by the local discriminator D_L, which outputs discriminator confidence to measure the domain discrepancies of local regions and pursue spatially local adaptation. The model has learned some domain-invariant features after global and local

alignment in stage 1. But the global-local adversarial learning neglects the explicit category-level adaptation, which incorporates semantic information of categories. In addition, the dispersed features from the target domain bring great challenges to category-level alignment. Inspired by [52] and [53], we generate prototypes, which are the centroids of categories, as the category anchors in stage 2 of category-level alignment. The teacher-student network in stage 2 constrains the feature-prototype distance between the target image x_t and it is augmented image by consistency loss, to learn the underlying category structure from compact category feature in the target domain.

B. Global-Local Alignment

The adversarial method for UDA is effective to alleviate domain shift by training the segmentation network *G* to fool the discriminator network *D*. Specifically, the adversarial game between two networks allows the *G* to learn domain-invariant feature representations which are confused with *D*. However, alignment from a global view would ignore the local distribution discrepancy and result in negative transfer. To address the problem, we design attention-based global-local alignment by adversarial learning of the global discriminator D_G and the local discriminator D_L.

1) Discriminator Attention: Considering that different local regions are various in color, texture, and context, which results in different spatial domain discrepancies, we apply an attention mechanism to the cross-domain alignment. Inspired by [43] and [54] that discriminator confidence can reflect the degrees of domain discrepancies, we utilize D_L to measure the cross-domain feature differences of local regions and focus on the alignment of hard-adapt patches. Specifically, the local features are input to D_L to obtain a spatially dense confidence score *S*. The confidence value close to the target domain label 1 means large cross-domain discrepancies. Conversely, the confidence value close to the source domain label 0 indicates easier cross-domain adaptation. Therefore, we calculate the attention map *A* as follows:

$$A = |tanh\,(S)\,|. \tag{1}$$

A $|tanh|$ activation is implemented to S as a normalization layer for preventing the gradient explosion at the early phase of the training. The original features of the target domain $F(x_t)$ are weighted to generate the updated features $F\ (x_t) = F(x_t) + A \odot F(x_t)$, by applying the residual attention mechanism to prevent gradient disappear in the attention map *A*. In the updated feature map, local features are enhanced in hard-adapt regions while weakened in easy-adapt regions.

2) Global-Local Adversarial Learning: Due to the domain discrepancies of global layout and local feature in RSIs, we separately design a global discriminator D_G and a local discriminator D_L. D_G is structured with four convolution layers and one classifier layer, with kernel size of 4×4, stride of 2, and each convolution layer is followed by a leaky ReLU. It is noted that spatially dense discriminator D_L has the similar structure to D_G, while the stride is set to 1. The discriminator D_L can align local features by preserving spatial resolution.

It is necessary to implement adaptation at the global level since cross-domain RSIs differ heavily in marginal distributions. Such differences mainly arise from spectral band composition variations, sensor variations, imaging time variations, and spatial layout variations, which degrade the performance of the cross-domain segmentation model but do not have an essential impact on the semantic features. Hence, we use the output space discriminator D_G along with *G* to globally decrease domain shift by learning the domain-invariant semantic information from predictions, and optimize them as follows:

$$L_{adv}^{D_G}(G) = \sum_{h=1}^{H}\sum_{w=1}^{W}\left[D_G(G(x_t)')^{(h,w)}\right]^2 \tag{2}$$

$$L_{adv}^{D_G}(D_G) = \sum_{h=1}^{H}\sum_{w=1}^{W}\left\{\left[D_G(G(x_t)')^{(h,w)} - 1\right]^2 + \left[D_2(G(x_s))^{(h,w)}\right]^2\right\} \tag{3}$$

where $G(x_t)\prime$ denotes updated predictions in the target domain and $G(x_s)$ denotes predictions in the source domain. The loss in (2) is designed to train *G* to fool D_G by minimizing the distance between the target predictions and source predictions. Then, D_G is trained to discriminate the target predictions and source predictions by the loss in (3).

However, global adversarial learning ignores various domain discrepancies in local regions, which causes negative transfer. In order to tackle the issue, we apply a local discriminator to the cross-domain model. Specifically, the spatially dense discriminator D_L, which has a fine-grained capability of distinguishing local features by preserving the spatial resolution, is used to align source feature and target updated feature in the way of adversarial learning. The local adversarial loss functions are defined as follows:

$$L_{adv}^{D_L}(F) = \sum_{h=1}^{H}\sum_{w=1}^{W}\left[D_L(F(x_t)')^{(h,w)}\right]^2 \tag{4}$$

$$L_{adv}^{D_L}(D_L) = \sum_{h=1}^{H}\sum_{w=1}^{W}\left\{\left[D_L(F(x_t)')^{(h,w)} - 1\right]^2 + \left[D_L(F(x_s))^{(h,w)}\right]^2\right\} \tag{5}$$

where $F(x_t)\prime$ denotes the new feature in the target domain updated by discriminator attention and $D_L(F(x_t)\prime)^{(h,w)}$ is the discriminator confidence located at (h, w) of the spatially dense confidence map. Note that the updated feature is strengthened in local regions with large domain shift, to which D_L pays more attention for achieving various degrees of local adaptation. By alternatively minimizing (4) and (5), *F* extracts domain-invariant local features to confuse D_L, and then the discriminatory ability of D_L is enhanced.

C. Category-Level Alignment

The segmentation network G pursues the spatial joint distribution alignment and the marginal distribution alignment by adversarial learning. However, the global-local alignment does not explicitly learn the multiple categories semantic structure in the target domain. Hence, we design the category-level alignment with the structure of teacher-student network to learn the shared category feature representations. Obtaining an effective student model is the purpose of teacher-student network training, which can transfer clean knowledge from teacher model to student model. The knowledge distillation in teacher-student network is able to learn cross-domain knowledge and improve the accuracy of student model [55]. Thus, we apply the consistency loss between teacher model and student model to learn the compact underlying category structure in the target domain. By minimizing the cross-entropy loss in source domain, student model can also learn the supervised category information in the source domain.

1) Prototype Initializing: Based on the observation that the features from the same category tend to cluster, we calculate the prototypes of each category, which are the centroids of each category features, to represent the category distribution. Note that the adapted network G from stage 1 can be used to acquire more exact target domain prototypes than the source-only training model. The target domain prototypes are defined as follows:

$$p_i = \frac{1}{\Lambda_i} \sum_{j=1}^{N} \sum_{k}^{H \times W} P_{ijk} f_j \tag{6}$$

where Λ_i is the number of pixels that are predicted as the ith category in all target training images; P_{ijk} denotes the prediction of jth target image at index k; and f_j is the feature vector at index j on the class-feature map f, which is generated by the classifier C.

2) Category Feature Compact: In the community of UDA for semantic segmentation, the qualities of the learned feature representations are usually judged by two key metrics, transferability and discriminability. However, the enhanced transferability by adversarial domain adaptation brings unexpected deterioration to the discriminability of category [56], and such a phenomenon affects the category-level label assignment. Thus, we work on improving the separability of category features, hoping to obtain more compact features in each category. During the process, we effectively learn the underlying category structure in the target domain, which plays a crucial role in category-level alignment. Considering that the consistency constraint of teacher-student network under divergent augmented views can learn compact category features without labels [57], we apply random augmentations to the image in the target domain and make it consistent with the original image in terms of feature-prototype distance. The L_{consist} is defined by the Kullback–Leibler (KL) divergence, which is

$$L_{\text{consist}} = KL\left(d_{\text{T}}, d_{\text{T}'}\right) \tag{7}$$

where d_{T} and $d_{\text{T}'}$ denote the Euclidean distance of feature prototypes under original and augmented views, which are generated from the teacher model and student model, respectively. Intuitively, this formula is applied to make the network assign the same prototypical label for neighboring feature points, thus clustering neighboring features together to compact target feature space. At the end of each iteration, the teacher model is updated by EMA with the help of student model to generate neighboring features.

D. Network Training

We implement a two-stage training strategy to gradually pursue cross-domain adaptation. The cross-entropy loss L_{seg} in the source domain is optimized in both stages to guide the network to learn category supervision information. Overall, the loss function of stages 1 and 2 is defined as follows:

$$L_{stage\,1} = L_{seg} + \lambda_1 L_{\text{adv}}^{D_G} + \lambda_2 L_{\text{adv}}^{D_L} \tag{8}$$

$$L_{stage\,2} = L_{seg} + w L_{\text{consist}} \tag{9}$$

where λ_1 and λ_2 denote the hyperparameters of global and local adversarial learning; and w is the hyperparameter to control the consistency constraint. In (8), λ_1 and λ_2 are used to balance the relative importance of global and local adaptation. The hyperparameter w in (9) reflects the degree of compacting category features. The training process of the proposed two-stage framework by minimizing (8) and (9) is shown in Algorithm 1.

III. Experiments

A. Datasets

We conduct a series of experiments on two high-resolution aerial datasets, the ISPRS Potsdam and Vaihingen datasets [58], to validate the proposed framework. The variations in geographic location and spectral band composition in the two datasets provide various experimental scenarios for cross-domain adaptation.

1) Potsdam Dataset: The Potsdam dataset contains 38 true orthophotos located in Potsdam city with a spatial resolution of 5 cm and a fixed size of 6000 × 6000 pixels. The images in the dataset have three different spectral band compositions: IR-R-G mode, R-G-B mode, and R-G-B-IR mode, in which IR, R, G, and B represent the near-infrared, red, green, and blue bands, respectively. All images are provided with their pixel-wise labels which comprise six classes of objects: clutter, impervious surfaces, car, tree, low vegetation, and building. In order to improve the operation efficiency, we crop all images and corresponding labels into the size of 512 × 512 pixels with the strides of 512 pixels in both horizontal and vertical directions, which is consistent with the work in [34].

2) Vaihingen Dataset: The Vaihingen dataset has 33 orthophotos with a spatial resolution of 9 cm, but only one spectral band composition: IR-R-G mode. The Vaihingen dataset with an image size near 2000 × 2000 pixels covers a smaller area than that of the Potsdam dataset. To train the semantic segmentation network, the images are also cropped to a size of 512 × 512 pixels but with both horizontal and vertical strides of 256 pixels to obtain enough samples. This data processing process remains consistent with the benchmark in [34].

Algorithm 1: Training Process of the Proposed Two-Stage UDA Framework.

Stage 1
Input:
Source sample, $(x_s,\ y_s)$
Target sample, x_t
Initialize:
G^0 , D_L^0, and D_G^0
Train:
for $k\ =\ 0$ to K do
train $G^{k+1} \leftarrow G^k$, $D_L^{k+1} \leftarrow D_L^k$, and $D_G^{k+1} \leftarrow D_G^k$ with L_{seg}, $L_{adv}^{D_L}$, and $L_{adv}^{D_G}$
end for
return G^{k+1}, D_L^{k+1}, and D_G^{k+1}
Initialize prototypes:
Input x_t into G^{k+1} and yield prototypes p_i^0 $(i = [1,\ 2,\ \ldots,\ N])$
Stage 2
Input:
Source sample, $(x_s,\ y_s)$
Target sample, x_t and augmentation target sample $x\prime_t$
Initialize:
T^0 (Teacher) $\leftarrow G^{k+1}$, S^0 (Student) $\leftarrow G^{k+1}$
Train:
for $k\ =\ 0$ to K do
train $T^{k+1} \leftarrow T^k$, $S^{k+1} \leftarrow S^k$ with L_{seg} and $L_{consist}$
update prototypes $p_i^{k+1} \leftarrow p_i^k + p_i$
update $T^{k+1} \leftarrow \alpha T^{k+1} + (1-\alpha)S^{k+1}$
end for
return T^{k+1}, S^{k+1}

B. Experimental Setup

We design four sets of experiments on Potsdam and Vaihingen datasets to verify the effectiveness of the proposed framework: 1) P(IR-R-G)_V(IR-R-G), Potsdam IR-R-G dataset serves as the source domain and Vaihingen IR-R-G dataset serves as the target domain, which considers the variations in geographic location; 2) P(R-G-B)_V(IR-R-G) considers the variations in both geographic location and spectral band composition; 3) V(IR-R-G)_P(IR-R-G), Vaihingen IR-R-G dataset serves as the source domain and Potsdam IR-R-G dataset serves as the target domain. Note that, the Vaihingen dataset has a lower resolution than the Potsdam dataset, which means a larger adaptation difficulty; and 4) V(IR-R-G)_P(R-G-B) is designed to explore the adaptive capacity of the proposed framework under a more challenging scenario including variations in geographic location and spectral band composition.

We employ the widely used Deeplab v2 [7] and Deeplab v3+ [8] with ResNet 101 [59] as the backbone. During the training of stage 1, the segmentation network is trained by the stochastic gradient descent (SGD) optimizer with an initial learning rate of 2.5×10^{-4}, a weight decay of 5×10^{-4}, and a momentum of 0.9. The Adam optimizer with an initial learning rate of 10^{-4} is utilized to optimize the two discriminators. Both optimizers are decayed by a poly learning rate policy with a power of 0.9. We train the network for 50 k iterations with a source and target batch size of 2 in a PyTorch environment. The hyperparameters λ_1 and λ_2 are both set to 0.01, then the parameters of the best segmentation network are transferred to stage 2. Similarly, SGD with an initial learning rate of 10^{-4}, a weight decay of 2×10^{-4}, and a momentum of 0.9 is used to optimize the network. We train a total of 20 k iterations in this stage of category-level alignment.

For comparison, the state-of-the-art methods are executed in the above-mentioned experimental setup. Deeplab v2 only and Deeplab v3+ only denotes baseline network training on only source images. Five advanced methods of UDA semantic segmentation are compared in the study, including AdaptSegNet [23], ADVENT [24], ProDA [53], Li's [34], and Zhang's [60]. The first two studies performed adversarial learning at the output level (prediction and entropy, respectively) for achieving global alignment while ProDA used prototypical denoising pseudo labels to align categories. The last two methods focused on cross-domain adaptation of RSIs, e.g., DualGAN was applied to generate fake source domain images similar to the target domain to improve the transferability of the segmentation model [34], and a local-to-global domain adaptation framework by the way of easy-to-hard curriculum was used to perform the cross-domain segmentation of RSIs [60]. To achieve a comprehensive comparison, we follow the network structure Deeplab v2 and hyperparameter settings in their paper for AdaptSegNet and ADVENT, while for ProDA, Li's, and Zhang's, we report their results which are realized by [60] in the framework of Deeplab v3+.

Following previous studies [34], [60], we use F1-score, intersection over union (IoU), mean F1-score, and mean IoU (mIoU) as metrics to evaluate the cross-domain adaptive segmentation performance.

C. Comparison Result

1) P(IR-R-G)_V(IR-R-G): P(IR-R-G)_V(IR-R-G) adaptive experiment is conducted to verify the effectiveness of the proposed framework in eliminating domain shift due to geographic location. Observing the quantification results in Table I, the proposed framework (based on Deeplab v2) achieves state-of-the-art accuracies with 53.63% and 67.21% separately on mIoU and mean F1-score, which reveals its capability to transfer knowledge between domains. The baseline suffers from serious degradation of performance due to a large domain shift. Thus, the mIoU and mean F1-score values of the baseline Deeplab v2 only are 32.33% and 44.11%, respectively. Segmentation accuracies are improved by nearly 11%–13% on mIoU in AdaptSegNet and ADVENT by aligning cross-domain marginal distribution. However, we find that IoU and F1-score of the category car decrease to some extent in AdaptSegNet because of the negative transfer by global alignment. The self-training method ProDA and generative method Li's obtain the nearing accuracies of global adversarial learning. Although Zhang's method achieves great performance, it is still overshadowed by the proposed framework due to the limited ability of category-level adaptation.

For individual categories, the outcomes in Table I suggest that the proposed framework (based on Deeplab v2) obtains the highest accuracy in the categories of impervious surfaces, low

TABLE I
COMPARATIVE RESULTS ON THE CROSS-DOMAIN ADAPTATION TASK OF P(IR-R-G)_V(IR-R-G)

Method	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	F1
Deeplab v2 only [7]	2.76	5.38	47.43	64.34	21.60	35.53	55.53	71.40	7.17	13.38	59.50	74.61	32.33	44.11
AdaptSegNet [23]	5.54	10.50	67.50	80.60	15.27	26.50	**60.82**	**75.64**	35.54	52.44	75.48	86.03	43.36	55.29
ADVENT [24]	11.57	20.74	67.31	80.46	23.07	37.48	58.72	73.99	35.38	52.27	74.26	85.23	45.05	58.36
Ours (Deeplabv2)	21.85	35.87	**76.58**	**86.73**	35.44	52.33	55.22	71.15	**49.97**	**66.64**	82.74	90.56	**53.63**	**67.21**
Deeplab v3+ only [8]	5.71	10.79	35.84	52.73	20.27	33.70	54.95	70.92	17.88	30.26	51.59	68.06	31.04	44.40
ProDA [53]	3.99	8.21	62.51	76.85	39.20	56.52	56.26	72.09	34.49	51.65	71.61	82.95	44.68	58.05
Li's [34]	**29.66**	**45.65**	49.41	66.13	34.34	51.09	57.66	73.14	38.87	55.97	62.30	76.77	45.38	61.43
Zhang's [60]	20.71	31.34	67.74	80.13	**44.90**	**61.94**	55.03	71.90	47.02	64.16	76.75	86.65	52.03	66.02
Ours (Deeplabv3+)	15.45	26.77	76.17	86.47	43.82	60.94	54.09	70.20	46.05	63.06	**84.37**	**91.52**	**53.33**	**66.50**

The bold entities represent the best results.

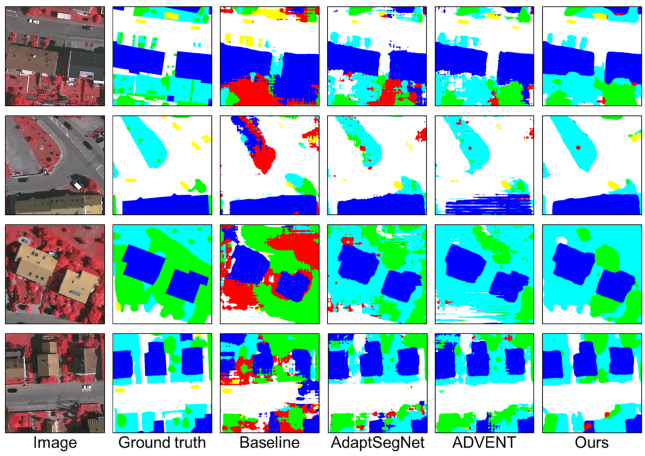


Fig. 3. Qualitative results of the proposed framework (based on deeplab v2) compared with other methods on the task of P(IR-R-G)_V(IR-R-G).

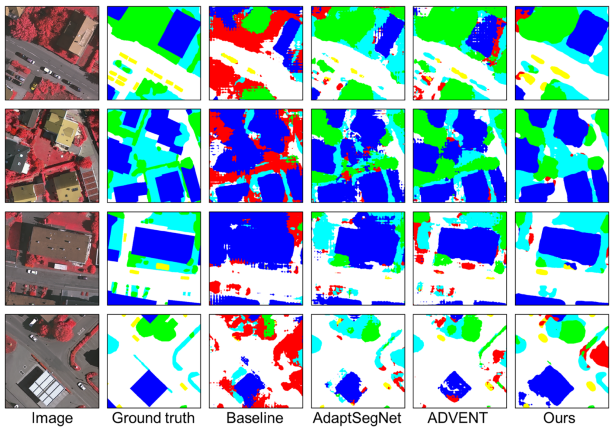


Fig. 4. Qualitative results of the proposed framework (based on deeplab v2) compared with other methods on the task of P(R-G-B)_V(IR-R-G).

vegetation, which has F1-score of 86.73% and 66.64%. These categories with conspicuous results usually occupy vast areas in RSIs and normally have homogeneous internal texture and color, especially for impervious surfaces. The proposed category-level alignment can effectively promote the performance of such categories in cross-domain segmentation by compacting feature space. Although the proposed framework is not the most accurate one for the car category, it still realizes improvements over Deeplab v2 only by 13.84% on mIoU and 16.80% on mean F1-score. The reason may be that our discriminator-attention mechanism can notice difficult patches despite the small size.

From the qualitative results in Fig. 3, we can note that the segmentation outputs of Deeplab v2 only are confusing, while after the global adversarial learning adaption, the performance in AdaptSegNet and ADVENT has been improved but the boundaries of objects are still unclear. In particular, the results of Deeplab v2 only contain much noise, which intuitively visualizes the degradation of segmentation model accuracy due to domain shift. AdaptSegNet and ADVENT learn domain-invariant features to alleviate the above-mentioned defect from a global perspective. The proposed framework achieves the best segmentation results by introducing local alignment and category-level alignment, in which there are fewer misclassifications than other methods. For example, it can extract more complete objects such as low vegetation and clearer object boundaries such as buildings. This is because, first, the attention module can focus on hard local regions; and second, the learning of category latent features makes indistinguishable feature close to the surrounding distinguishable feature, which results in less misclassification within objects.

2) P(R-G-B)_V(IR-R-G): We implement the experiment of P(R-G-B)_V(IR-R-G) to examine the effectiveness of the proposed framework on domain shift of various regional locations and spectral band compositions, which causes the large appearance discrepancy. Table II denotes the quantitative results. The proposed framework (based on Deeplab v3+) has the topgallant accuracy despite the setting of more challenging cross-domain adaptation, and improving the baseline of Deeplab v3+ only by 26.46% and 28.96% on mIoU and mean F1-score, respectively. It achieves more accurate prediction compared with other methods, as shown in Fig. 4. The segmentation results of global

TABLE II
COMPARATIVE RESULTS ON THE CROSS-DOMAIN ADAPTATION TASK OF P(R-G-B)_V(IR-R-G)

Method	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	F1
Deeplab v2 only [7]	1.21	2.38	26.31	41.65	7.03	13.13	13.94	24.47	10.71	19.34	55.83	71.66	19.17	28.77
AdaptSegNet [23]	5.07	9.65	56.10	71.88	14.77	25.74	**57.69**	**73.17**	34.37	51.15	67.77	80.79	39.30	52.06
ADVENT [24]	5.35	10.15	60.66	75.52	24.98	39.98	56.64	72.31	24.84	39.79	68.09	81.02	40.09	53.13
Ours (Deeplabv2)	10.63	19.22	70.37	82.61	34.85	51.68	56.08	71.86	42.84	59.99	79.61	88.64	49.06	62.33
Deeplab v3+ only [8]	1.03	2.03	46.39	63.37	27.33	42.93	13.58	23.73	4.61	8.82	49.39	66.12	23.72	34.50
ProDA [53]	2.39	5.09	49.04	66.11	31.56	48.16	49.11	65.86	32.44	49.06	68.94	81.89	38.91	52.70
Li's [34]	3.94	13.88	46.19	61.33	40.31	57.88	55.82	70.66	27.85	42.17	65.44	83.00	39.93	54.82
Zhang's [60]	12.38	21.55	64.47	77.76	**43.43**	**60.05**	52.83	69.62	38.37	55.94	76.87	86.95	48.06	61.98
Ours (Deeplabv3+)	**12.61**	**22.39**	**73.80**	**84.92**	43.24	60.38	44.41	61.50	**43.27**	**60.40**	**83.76**	**91.16**	**50.18**	**63.46**

The bold entities represent the best results.

TABLE III
COMPARATIVE RESULTS ON THE CROSS-DOMAIN ADAPTATION TASK OF V(IR-R-G)_P(IR-R-G)

Method	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	F1
Deeplab v2 only [7]	5.84	11.04	48.03	64.89	33.15	49.79	14.35	25.10	38.26	55.34	40.92	58.08	30.09	44.04
AdaptSegNet [23]	10.95	19.74	66.11	79.60	55.19	71.13	29.36	45.40	**50.23**	**66.87**	71.25	83.21	47.18	60.99
ADVENT [24]	11.00	19.82	66.33	79.76	55.58	71.45	11.42	20.49	48.79	65.58	70.64	82.79	43.96	56.65
Ours (Deeplabv2)	9.13	16.74	70.50	82.70	**64.22**	**78.21**	30.57	46.83	49.84	66.52	74.49	85.38	49.79	62.73
Deeplab v3+ only [8]	9.30	16.86	49.18	65.93	38.51	55.60	7.67	14.24	29.32	45.34	36.96	53.97	28.49	41.99
ProDA [53]	10.63	19.21	44.70	61.72	46.78	63.74	31.59	48.02	40.55	57.71	56.85	72.49	38.51	53.82
Li's [34]	11.48	20.56	51.01	67.53	48.49	65.31	34.98	51.82	36.50	53.48	53.37	69.59	39.30	54.71
Zhang's [60]	**12.31**	**24.59**	64.39	78.59	59.35	75.08	**37.55**	**54.60**	47.17	63.27	66.44	79.84	47.87	62.66
Ours (Deeplabv3+)	11.65	19.47	**73.43**	**84.55**	63.86	77.85	32.68	47.36	47.69	63.45	**76.32**	**87.43**	**50.94**	**63.31**

The bold entities represent the best results.

adaptation method, AdaptSegNet and ADVENT are unsatisfactory, especially in complex scenario due to the negligence of regional interdomain discrepancy. The proposed framework considers local adaptation which pursues strong alignment to hard-adapt patches such as low vegetation to obtain refined segmentation results. Furthermore, it provides more precise results of roofs with different spectral characteristics and tackles the problem of holes mainly by the category compact module.

3) V(IR-R-G)_P(IR-R-G): The V(IR-R-G)_P(IR-R-G) experiment is conducted to further validate the effectiveness and generalization ability of the proposed framework. It is noteworthy that the Vaihingen dataset has a lower spatial resolution than the Potsdam dataset, which brings more challenges for cross-domain adaptation. Nevertheless, it is obvious from Table III that the proposed framework (based on Deeplabv3+) is still superior to other comparing adaptation methods with mIoU of 50.94% and mean F1-score of 63.31%. ProDA achieves the lowest accuracy, possibly owing to that the quality of the pseudo labels is affected by serious domain shift. The results in Li's are also unsatisfactory because the image style translation is unable to fall off the domain discrepancies caused by geographic location variation. AdaptSegNet and ADVENT obtain more advanced results by output space adaptation.

Moreover, the proposed framework makes significant improvement in the categories of impervious surfaces, car, and building due to the consideration of local discrepancies and category-level structure. Small objects of car category in the target domain the Potsdam dataset is nearly twice the size of the source domain Vaihingen dataset, which makes the transfer of car category less challengeable and our feature compact module is more effective for larger scale objects, thus the proposed framework can segment cars accurately. We visualize comparison results in Fig. 5 which explicitly reveals the advantages of our framework. Specifically, the categories of impervious surfaces and building with similar color and material are easy to be confused by AdaptSegNet and ADVENT, while the proposed framework learns category-level semantic features to alleviate the confusion caused by low-level features.

4) V(IR-R-G)_P(R-G-B): The V(IR-R-G)_P(R-G-B) is the most difficult adaptation task and was conducted to further explore the availability of the proposed framework in pursuing adaptation for different geographic locations and spectral band

TABLE IV
COMPARATIVE RESULTS ON THE CROSS-DOMAIN ADAPTATION TASK OF V(IR-R-G)_P(R-G-B)

Method	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	F1
Deeplab v2 only [7]	0.33	0.66	33.29	49.95	25.26	40.33	3.60	6.96	32.28	48.80	38.60	55.70	22.23	33.73
AdaptSegNet [23]	4.65	8.88	61.35	76.04	50.38	67.00	17.73	30.13	41.52	58.68	67.15	80.35	40.46	53.51
ADVENT [24]	2.96	5.75	56.47	72.18	51.64	68.11	2.61	5.10	42.72	59.86	52.96	69.25	34.89	46.71
Ours (Deeplabv2)	10.06	18.28	64.11	78.13	58.08	73.48	25.29	40.38	**44.45**	**61.55**	63.99	78.04	44.33	58.31
Deeplab v3+ only [8]	6.99	13.04	42.98	60.12	38.01	55.08	0.53	1.06	1.59	3.13	29.09	45.05	19.86	29.58
ProDA [53]	11.13	20.51	44.77	62.03	41.21	59.27	30.56	46.91	35.84	52.75	46.37	63.06	34.98	50.76
Li's [34]	**13.56**	**23.84**	45.96	62.97	39.71	56.84	25.80	40.97	41.73	58.87	59.01	74.22	37.63	52.95
Zhang's [60]	13.27	23.43	57.65	73.14	56.99	72.27	**35.87**	**52.80**	29.77	45.88	65.44	79.11	43.17	57.77
Ours (Deeplabv3+)	10.84	17.49	**66.11**	**79.75**	**65.45**	**80.17**	28.64	43.51	35.47	51.85	**68.63**	**81.32**	**45.86**	**59.74**

The bold entities represent the best results.

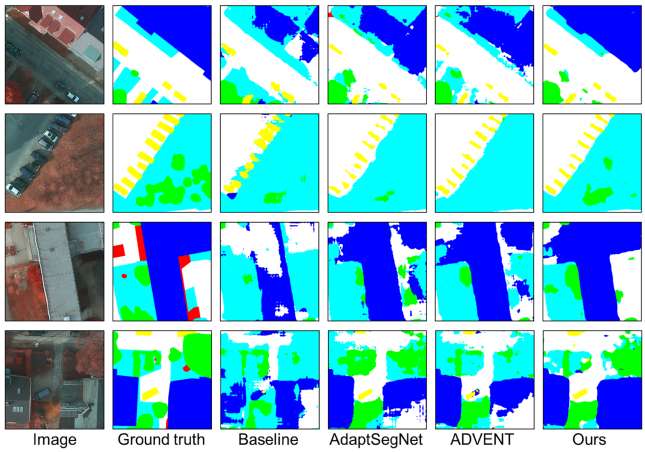


Fig. 5. Qualitative results of the proposed framework (based on deeplab v2) compared with other methods on the task of V(IR-R-G)_P(IR-R-G).

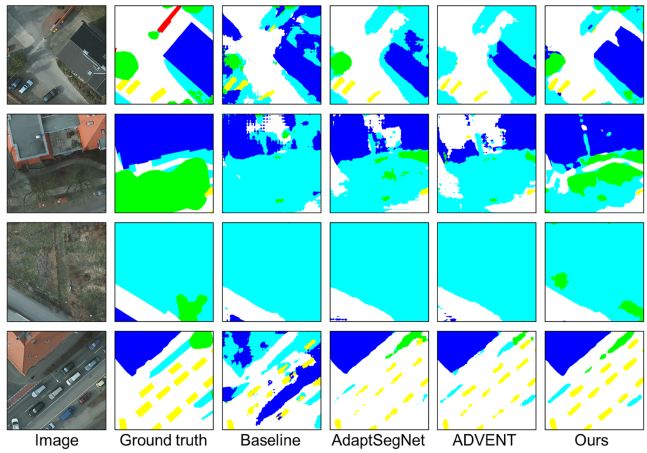


Fig. 6. Qualitative results of the proposed framework (based on deeplab v2) compared with other methods on the task of V(IR-R-G)_P(R-G-B).

compositions. As shown in Table IV, the proposed framework (based on Deeplabv3+) still attains the highest performance with the mIoU of 45.86% and mean F1-score of 59.74%, despite the decrease in performance compared with the previous experimental setup. For the categories of impervious surfaces, car, and building, the proposed framework outperforms Zhang's [60] on F1-score by 6.61%, 7.90%, and 2.21%, respectively. We exhibit the visualization results in Fig. 6. The spectral features of sparse trees are confused with other objects like low vegetation, which makes it difficult to align them only by global adversarial learning. Thus, the segmentation results of AdaptSegNet and ADVENT suffer from misclassification between similar categories while the designed framework achieves intradomain adaptation by attention mechanism. Moreover, the result of the proposed framework responds complete profile for car category which usually has a small size due to category-level alignment.

D. Component Analysis

1) Ablation Study: We dissect the contributions of each component by ablation study on the cross-domain adaptation of P(IR-R-G)_V(IR-R-G). The results are listed in Table V, including the performance of the baseline (Deeplab v2 only), the framework with only a global discriminator (G), adding a local discriminator (G + L), and adding a category-level alignment module (G + L + C). By conducting global adversarial adaptation, the performance is improved by 22.35% on mIoU and 25.55% on mean F1-score compared with the no-adaptation baseline, which suggests the alignment of the marginal distribution can largely mitigate the accuracy degradation. The attention-based local alignment module can refine global adaptation due to the consideration of regional distribution discrepancy. The complete model with the above-mentioned three modules achieves the best performance that has mIoU and mean F1-score up to 49.06% and 62.33%, respectively. It further verifies the contributions and complementarity of the two fine-grained local alignment and category-level alignment. Fig. 7 further visualizes the improvement on F1-score of the proposed framework variants compared with the baseline for each category.

2) Evaluation on Discriminator Attention: The discriminator attention maps and segmentation results on the adaptive task of P(IR-R-G)_V(IR-R-G) are exhibited in Fig. 8. The categories

TABLE V
ABLATION STUDY RESULTS ON THE DOMAIN ADAPTIVE TASK OF P(IR-R-G)_V(IR-R-G)

Method	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	F1
Deeplab v2 only [7]	1.21	2.38	26.31	41.65	7.03	13.13	13.94	24.47	10.71	19.34	55.83	71.66	19.17	28.77
Ours (G)	4.77	9.11	59.90	74.92	28.01	43.76	**59.78**	**74.83**	24.70	39.61	71.97	83.70	41.52	54.32
Ours (G+L)	7.65	14.22	66.06	79.56	32.56	49.13	54.12	70.23	35.26	52.14	74.87	85.63	45.09	58.49
Ours (G+L+C)	**10.63**	**19.22**	**70.37**	**82.61**	**34.85**	**51.68**	56.08	71.86	**42.84**	**59.99**	**79.61**	**88.64**	**49.06**	**62.33**

The bold entities represent the best results.

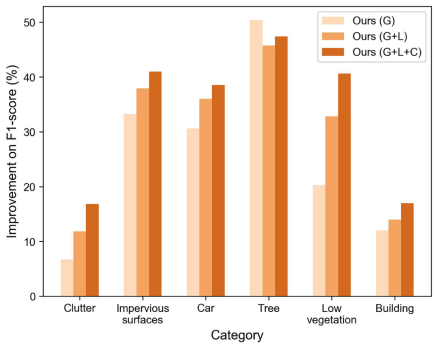


Fig. 7. Improvement on F1-score of each category compared with the baseline (deeplab v2 only). The framework with only a global discriminator is called ours (G), adding a local alignment module is called ours (G + L), and further adding a category-level alignment module is called ours (G + L + C).

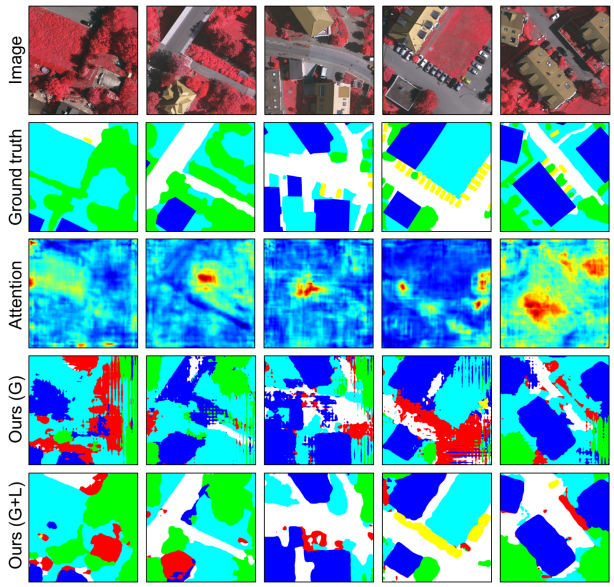


Fig. 8. Comparison of discriminator attention maps and segmentation results on the adaptive task of P(IR-R-G)_V(IR-R-G). The top two rows represent the input target images and corresponding ground truth. The third row is the discriminator attention maps where blue to red indicates low to high attention. The bottom two rows separately represent the segmentation results of the proposed framework with global adaptation and adding attention-based local adaptation.

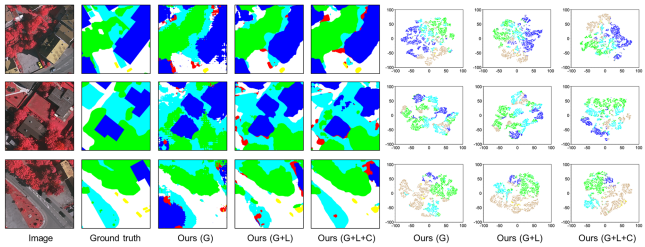


Fig. 9. Visualization of segmentation results which are shown by the third to fifth columns and feature maps which are displayed by the last three columns. In 2-D space feature maps, the brown, yellow, cyan, green, and blue represent the categories of impervious surfaces, car, tree, low vegetation, and building, respectively.

of low vegetation and tree have high attention values, due to the large appearance divergencies in vegetation regions that mostly appear green in the source domain while red in the target domain. The discriminator attention module measures the domain discrepancies of different regional features to guide the local adaptation, which obtains a refined segmentation result that has a significant improvement by the noisy result of global adaptation. Although it is subconsciously believed that the domain shift of impervious surface is small, some factors impact the alignment, such as easy confusion with roofs made of cement, often covered by shadows, and the existence of much co-occurrence. Global adversarial learning is unable to effectively pursue the alignment in such hard regions, but the discriminator attention module helps achieve a precise segmentation result by focusing on the internal regions with large domain discrepancies, which may be ignored in the boundary-emphasized entropy map. In addition, the designed attention module can respond to both small objects such as car, and large objects such as building to obtain a clearer boundary.

3) Evaluation on Category Feature Compact: The visualization results in Fig. 9 contain predicted outputs and the corresponding 2-D space feature maps, which are converted from high-dimensional features by t-SNE [61]. The complete framework obtains apparently better segmentation performance compared with global adaptation and global-local adaptation. In particular, global adaptation model captures domain-invariant features but misclassifies the domain-variant regions. By considering the intradomain distribution divergency, segmentation performance of global-local adaptation model is promoted. However, it still generates dispersed category features which significantly affects the segmentation accuracy. The proposed category feature compact module can decrease the confusion of categories with high feature similarities such as impervious

TABLE VI
HYPERPARAMETER ANALYSIS OF GLOBAL AND LOCAL ADAPTATION IN STAGE 1 ON THE DOMAIN ADAPTIVE TASK OF P(IR-R-G)_V(IR-R-G) AND P(R-G-B)_V(IR-R-G), RESPECTIVELY

Hypermeters		P(IR-R-G)_V(IR-R-G)		P(R-G-B)_V(IR-R-G)	
λ_1	λ_2	mIoU	F1	mIoU	F1
0.005	0.01	46.17	59.31	42.65	56.02
0.01	0.005	45.46	58.79	43.29	57.21
0.01	0.01	**47.50**	**60.47**	**45.09**	58.49
0.01	0.05	46.64	60.10	44.85	**58.73**
0.05	0.01	43.01	57.19	39.91	54.22

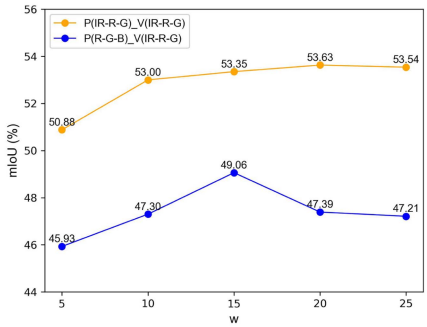


Fig. 10. Curves of mIoU of the proposed framework (deeplab v2) with the changes of w on P(IR-R-G)_V(IR-R-G) and P(R-G-B)_V(IR-R-G).

surfaces and building. In addition, the proposed framework can reduce inaccurate segmentation inside the object by explicitly considering the category feature structure.

E. Hyperparameter Analysis

We conducted experiments on two Potsdam-to-Vaihingen adaptive tasks to analyze the hyperparameters effects of global and local adaptation, and the results of stage 1 are shown in Table VI. P(IR-R-G)_V(IR-R-G) considers the variations in geographic location and P(R-G-B)_V(IR-R-G) considers the variations both in geographic location and spectral band composition. It can be seen from Table VI that model almost achieves the most outstanding performance when λ_1 and λ_2 are both set to 0.01, which means we should pay equal attention to global and local adversarial learning. In addition, we investigate the impact of the hyperparameter w which determines the degree of compacting category features, via conducting sensitivity experiments. The mIoU curves are presented in Fig. 10, in which the proposed framework achieves the best results when w is set to 20 and 15 on the task of P(IR-R-G)_V(IR-R-G) and P(R-G-B)_V(IR-R-G), respectively. A small w indicates that the feature compact constraint is weak, leading to dispersed features for each category. However, a large w suggests a strong feature compact constraint, which results in mistakenly compacting different categories of features that are easily confused due to the presence of domain shift and lacking supervision labels. The proposed framework is not particularly sensitive to the hyperparameter w as the curves in the figure are smooth, thus w is separately set to 20 and 15 on the adaptation of V(IR-R-G)_P(IR-R-G) and V(R-G-B)_P(IR-R-G).

IV. DISCUSSION

Based on the experimental results, it is necessary to implement local alignment and category-level alignment together in the cross-domain adaptation of RSIs, both of which are fine-grained auxiliaries to global alignment. Essentially, the local adaptation pursues spatial joint distribution alignment and the category-level adaptation pursues semantic joint distribution alignment. Enabling the domain adaptation model to spatially and semantically concentrate on domain discrepancies can significantly improve its cross-domain segmentation capability. The complementary of local alignment and category-level alignment can be proved in Fig. 7, in which the improvement on F1-score obtained by the two alignments is shown. With the help of local alignment, the accuracy of almost all categories can be improved compared with the global-only alignment, except for the category of tree. Discriminator attention in local adaptation usually focuses on spatial regions with large domain shift, such as low vegetation and building patches that exhibit different colors in the two domains, and car patches with various sizes due to the spatial resolution change. However, such a spatial refined adaptation does not explicitly incorporate category information which is crucial to semantic segmentation [45]. Hence, we further promote category-level alignment by learning the underlying category structure. As a result, it brings enhancements to the global-local alignment in all categories, which means semantic adaptation at the category level is essential.

The proposed two-stage framework gradually adapts different domains following the principles of coarse-to-fine and low-to-high. First, we coarsely align the marginal distribution by adversarial learning, which globally alleviates the domain shift. Such domain discrepancies are caused by global imaging conditions and sensor characteristics that seriously affect the transferability of segmentation model. Although global adaptation obtains advanced alignment performance, it neglects fine-grained domain-invariant feature representations. We then refine the coarse alignment process by applying the local adaptation and category-level adaptation, which focuses on low-level spatial features and high-level semantic features, respectively. In general, the deep learning segmentation network encodes the feature representations in the way of low-to-high, because the high-level features evolve from low-level features. In domain adaptation for the segmentation of RSIs, we also follow the idea to match the local distributions and then match the category-level distributions.

V. CONCLUSION

In this study, we propose a novel UDA framework for semantic segmentation of RSIs, which unifies the global-local alignment and category-level alignment to pursue a fine-grained adaptation. In particular, the attention-based local discriminator measures the domain discrepancies in various local regions, which alleviates the negative transfer caused by global-only adversarial learning. Moreover, we learn the underlying category structure from compact category features in the target domain

to semantically match the category-level joint distribution. The experimental results show the effectiveness of the proposed framework, which achieves the best performance in comparison with state-of-the-art methods, demonstrating that alleviating local and category-level domain shifts simultaneously is useful. Exploring the context-invariant relationship between two domains can also facilitate the cross-domain segmentation of RSIs, which will be investigated in our future work.

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.*, 2014, pp. 891–898.

[2] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[3] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2021, Art. no. 5609413.

[4] B. Du et al., "Landslide susceptibility prediction based on image semantic segmentation," *Comput. Geosci.*, vol. 155, Oct. 2021, Art. no. 104860.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[7] 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.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[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. Photogrammetry Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.

[10] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, "Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network," *Remote Sens.*, vol. 11, no. 15, Jul. 2019, Art. no. 1774.

[11] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[12] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 102–118.

[13] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4068–4076.

[14] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 136–144.

[15] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 99–105.

[16] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[17] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4500–4509.

[18] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6936–6945.

[19] F. Pizzati, R. de Charette, M. Zaccaria, and P. Cerri, "Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2979–2987.

[20] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv: 1612.02649*.

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

[22] R. Li, X. Jia, J. He, S. Chen, and Q. Hu, "T-SVDNet: Exploring high-order prototypical correlations for multi-source domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9971–9980.

[23] 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.*, 2018, pp. 7472–7481.

[24] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2517–2526.

[25] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1456–1465.

[26] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[27] J. Yang, W. An, C. Yan, P. Zhao, and J. Huang, "Context-aware domain adaptation in semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 514–524.

[28] J. Huang, D. Guan, S. Lu, and A. Xiao, "MLAN: Multi-level adversarial network for domain adaptive semantic segmentation," 2021, *arXiv: 2103.12991*.

[29] Y. Zhang and Z. Wang, "Joint adversarial learning for domain adaptation in semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 6877–6884.

[30] 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, pp. 1369–1378, Jun. 2019.

[31] X. Deng, H. L. Yang, N. Makkar, and D. Lunga, "Large scale unsupervised domain adaptation of segmentation networks with adversarial learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4955–4958.

[32] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.

[33] S. Ji, D. Wang, and M. Luo, "Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3816–3828, May 2021.

[34] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 20–33, May 2021.

[35] J. Li, S. Zi, R. Song, Y. Li, Y. Hu, and Q. Du, "A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5618515.

[36] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5400515.

[37] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5603518.

[38] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2868–2876.

[39] 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 Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2507–2516.

[40] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Category-level adversarial adaptation for semantic segmentation using purified features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 3940–3956, Aug. 2022.

[41] Q. Zhou et al., "Context-aware mixup for domain adaptive semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 804–817, Feb. 2023.

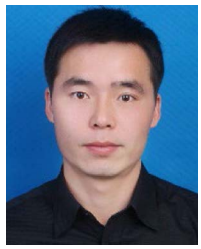
[42] Q. Zhou et al., "Uncertainty-aware consistency regularization for cross-domain semantic segmentation," 2020, *arXiv:2004.08878*.
[43] F. Yu, M. Zhang, H. Dong, S. Hu, B. Dong, and L. Zhang, "DAST: Unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 10754–10762.
[44] A. Zheng, M. Wang, C. Li, J. Tang, and B. Luo, "Entropy guided adversarial domain adaptation for aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5405614.
[45] H. Wang, T. Shen, W. Zhang, L.-Y. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 642–659.
[46] Z. Wang et al., "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12632–12641.
[47] Q. Xu, X. Yuan, and C. Ouyang, "Class-aware domain adaptation for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2022, Art. no. 4500317.
[48] J. Chen, J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng, "Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5616915.
[49] J. Huang, D. Guan, A. Xiao, S. Lu, and L. Shao, "Category contrast for unsupervised domain adaptation in visual tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1203–1214.
[50] W. Liu, K. Quijano, and M. M. Crawford, "YOLOv5-tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8085–8094, Sep. 2022.
[51] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[52] Q. Zhang, J. Zhang, W. Liu, and D. Tao, "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 435–445.
[53] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12409–12419.
[54] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018, *arXiv: 1802.07934*.
[55] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[56] X. Chen, S. Wang, M. Long, and J. Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1081–1090.
[57] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 596–608.
[58] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," ITC, Univ. Twente, Enschede, The Netherlands, Tech. Rep., 2014, doi: 10.13140/2.1.5015.9683.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2016, pp. 770–778.
[60] B. Zhang, T. Chen, and B. Wang, "Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2021.
[61] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Luhan Wang received the B.S. degree in geographical information science in 2021 from Nanjing University, Nanjing, China, where she is currently working toward the M.S. degree in cartography and geographic information system.

Her research interests include semantic segmentation, unsupervised domain adaptation, and deep learning for remote sensing.

Pengfeng Xiao (Senior Member, IEEE) received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018. Since 2019, he has been a Professor with Nanjing University. He was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Hesse, Germany, from 2011 to 2012, and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA, from 2014 to 2015. He has authored four books and more than 60 articles. His current research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.



Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information system and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a Visiting Student with Informatics Institute, University of Missouri, Columbia, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University. He is currently an Associate Professor with the Department of Geographic Information Science, Nanjing University. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.

Xinyang Chen received the B.S. degree in geographic information science from Jilin University, Jilin, China, in 2021. She is currently working toward the M.S. degree in cartography and geographical information system with Nanjing University, Nanjing, China.

Her research interests include semantic segmentation and deep learning for remote sensing.