SPGAN-DA: Semantic-Preserved Generative Adversarial Network for Domain Adaptive Remote Sensing Image Semantic Segmentation

Yansheng Li, *Senior Member, IEEE*, Te Shi, Yongjun Zhang, *Member, IEEE*, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*— Unsupervised domain adaptation for remote sensing semantic segmentation seeks to adapt a model trained on the labeled source domain to the unlabeled target domain. One of the most promising ways is to translate images from the source domain to the target domain to align the spectral information or imaging mode by the generative adversarial network (GAN). However, source-to-target translation often brings bias in the translated images causing limited performance, as semantic information is not well considered in the translation procedure. To overcome this limitation, we present an innovative semantic-preserved generative adversarial network (SPGAN), designed to mitigate the image translation bias and then leverage the translated images as well as unlabeled target images by class distribution alignment (CDA) module to train a domain adaptive semantic segmentation model. The above two stages are coupled together to form a unified framework called SPGAN-DA. Specifically, we first conduct semantic invariant translation from source to target domain, which is achieved by introducing representation-invariant and semantic-preserved constraints to the GAN model. To further narrow the landscape layout gap between the translated and target images, CDA semantic segmentation is proposed. CDA semantic segmentation consists of two aspects. At the model input level, object discrepancy is eliminated by introducing the ClassMix operation. At the model output level, boundary enhancement is proposed to refine the performance of object boundaries. Extensive experiments on three typical remote sensing cross-domain semantic segmentation benchmarks demonstrate the effectiveness and generality of our proposed method, which competes favorably against existing state-of-the-art methods.**

***Index Terms*— Class distribution alignment (CDA), domain adaptive semantic segmentation, generative adversarial network (GAN), semantic-preserved generative adversarial network (SPGAN), unbiased image translation.**

I. Introduction

SEMANTIC segmentation aims to assign each pixel of the image to a semantic label [1], [2], which is a fundamental task in remote sensing. It has been widely utilized in many applications, ranging from scene or land-cover classification [3], [4], [5], urban planning [6], [7], environment monitoring [8], intelligent traffic [9], and semantic-aware image fusion [10]. Deep learning-based approaches [1], [11], [12], [13], [14], [15], [16] have achieved great success at the price of large-scale densely annotated datasets, which are prohibitively expensive to collect in terms of time and money [17], especially in domains where experts are required. One potential solution is to use existing labeled datasets since it is possible to obtain their labels at a low cost. However, models trained on these datasets may not well generalize to the unlabeled target domain due to the large cross-domain appearance discrepancy, which is known as domain shift [18], [19].

The major challenges of the domain shift can be summarized as follows.

1) *Cross Imaging Mode:* In this situation, the source and target imaging modes are diverse, as shown in Fig. 1(a). For instance, the source images are captured in the R-G-B imaging mode, and the target imaging mode is IR-R-G or even more different. In terms of appearance, the image styles of the source and target domains can vary greatly. In terms of distribution, the spectral statistics of the source and target are totally different. This will cause the methods that learn from source annotated images to always lose effectiveness when applied to target images.
2) *Cross Geographic Location:* Under this circumstance, the source and target image are in the same imaging mode, as shown in Fig. 1(b). However, the spectral information has more or fewer differences caused by luminance, temporal difference, and geographic location. This will lead to different spectral distributions and consequently to domain shift problems.
3) *Cross Landscape Layout:* This usually happens between urban and rural dataset domain adaptation, as shown in Fig. 1(c). For urban and rural areas, in particular, the manifestation of the land cover is completely different, in the landscape layout and object style. For example, the buildings in the urban area are neatly arranged, with various shapes, while the buildings in the rural area are disordered, with simpler shapes. The roads are wide in the urban scenes. In contrast, the roads are narrow in the rural scenes. The inconsistent landscape layout between

Manuscript received 12 April 2023; revised 16 August 2023; accepted 5 September 2023. Date of publication 11 September 2023; date of current version 26 September 2023. This work was supported in part by the State Key Program of the National Natural Science Foundation of China under Grant 42030102, in part by the National Natural Science Foundation of China under Grant 41971284, in part by the Fundamental Research Funds for the Central Universities under Grant 2042022kf1201, and in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100032. *(Corresponding author: Te Shi.)*

Yansheng Li, Te Shi, and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Hubei Luojia Laboratory, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; te.shi@whu.edu.cn; zhangyj@whu.edu.cn).

Jiayi Ma is with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com).

Digital Object Identifier 10.1109/TGRS.2023.3313883

the urban and rural scenes increases the difficulty of model generalization.

To address the domain shift issue, unsupervised domain adaptation (UDA) approaches [20], [21], [22] are proposed to alleviate the domain shift problem by aligning the distributions of the source and the target domains. Following the advances of generative adversarial networks (GANs) [23], adversarial learning has been widely explored to match cross-domain representations by minimizing an adversarial loss [24], [25], [26], [27], [28], [29], [30] on the source and target feature representations or adapting structured output space across two domains. Recent studies further consider the pixel-level domain shift to enforce source and target images to be domain-invariant in terms of visual appearance [31]. This is achieved by translating images from the source domain to the target domain using image-to-image translation models such as CycleGAN [32]. Although these methods reduce the visual gap between source and target domains to some extent, overcoming the essential appearance disappearance, how to pursue semantic invariant content is still challenging. As well known, an ideal translation is to keep visual content invariant and make the style highly similar to the target domain. However, unsmooth and discordant areas often appear in the translated images as the yellow bounding boxes shown in Fig. 2. We define this phenomenon in the translated image as bias. Due to the lack of adequate consideration of semantic labels in the source domain, the GAN-based image-to-image translation inevitably introduces bias to the translated images. In the feature level, the bias means that the generator cannot thoroughly make features of the same class cluster together, i.e., some classes may be mixed up with other classes during the translation procedure. In the image level, the bias presents as the unsmooth and discordant areas in the translated images, which will cause the content of the translated images to be inconsistent with the original semantic labels. There is no doubt that the phenomenon will impair the following semantic segmentation training procedure. Some previous works [31], [33] have tried to use a source-pretrained segmentation model to compute the semantic consistency loss of the source image prediction and translated image prediction, which is utilized to achieve the goal of semantic-preserved source-to-target translation. Unfortunately, this kind of constraint highly relies on the pretrained model so that it cannot implement in an end-to-end way. However, once the pretrained model is fixed, there is no feedback from the segmentation model, causing limited performance gains. In addition, it is difficult for the existing methods to handle the three typical remote sensing cross-domain semantic segmentation tasks well at the same time.

Motivated by the above domain shift analysis and limitations of GAN, we propose a novel semantic-preserved generative adversarial network (SPGAN) that can conduct unbiased (i.e., sematic-preserved) source-to-target translation and further achieve to align the spectral information or even imaging mode. This is implemented by introducing representation-invariant and semantic-preserved constraints into the GAN-based translation model. It is beneficial to reduce the domain shift before training the semantic segmentation model. Second, there are some differences in landscape layout between datasets, and this is particularly evident in the domain shift between urban and rural. Thus, the ClassMix strategy is used in our proposed method to align the class distribution shift. We paste the translated image onto the target image to obtain a mixed image. In the mixed images, the style is quite similar to the target domain and the texture information of the image includes both source and target domains. The translated and mixed images are used to collaboratively train a domain adaptive semantic segmentation model. Finally, the boundary is essential in the semantic segmentation task, but this has not been given enough attention in previous UDA algorithms. For this reason, we propose the boundary enhancement module to constrain the segmentation boundary to obtain more accurate segmentation results. These modules are coupled together to form a unified framework called SPGAN-DA. The main contributions of this article are summarized as follows.

1) We propose a novel SPGAN, which conducts unbiased translation (i.e., visual content invariant translation) from source to target domain to align the spectral information or imaging mode. This is achieved by introducing representation-invariant and semantic-preserved constraints into the GAN framework, optimized in an end-to-end way.
2) One novel class distribution alignment (CDA) semantic segmentation module is proposed to further narrow the landscape layout gap between the different datasets. At the model input level, we first paste objects from the translated images onto the target images by ClassMix operation. At the model output level, boundary enhancement is proposed to refine the performance of object boundaries. These two aspects are utilized to collaboratively train a domain adaptive semantic segmentation model.
3) Our proposed SPGAN-DA can consistently work well on classic remote sensing cross-domain semantic segmentation benchmarks. The extensive experimental results demonstrate the remarkable effectiveness and generality of our proposed SPGAN-DA framework, which makes a new state-of-the-art performance.

The rest of this article is organized as follows. Section II discusses the related work. Section III introduces the proposed method in detail. Section IV reports the experiments and provides a discussion of the experimental results. Finally, the conclusion and potential future research directions are outlined in Section V.

II. Related Work

In this section, we briefly review the most relevant works in the literature that include cross-domain semantic segmentation. Here, we roughly group the existing algorithms into three categories: image-to-image translation methods, adversarial learning methods, and self-learning methods.

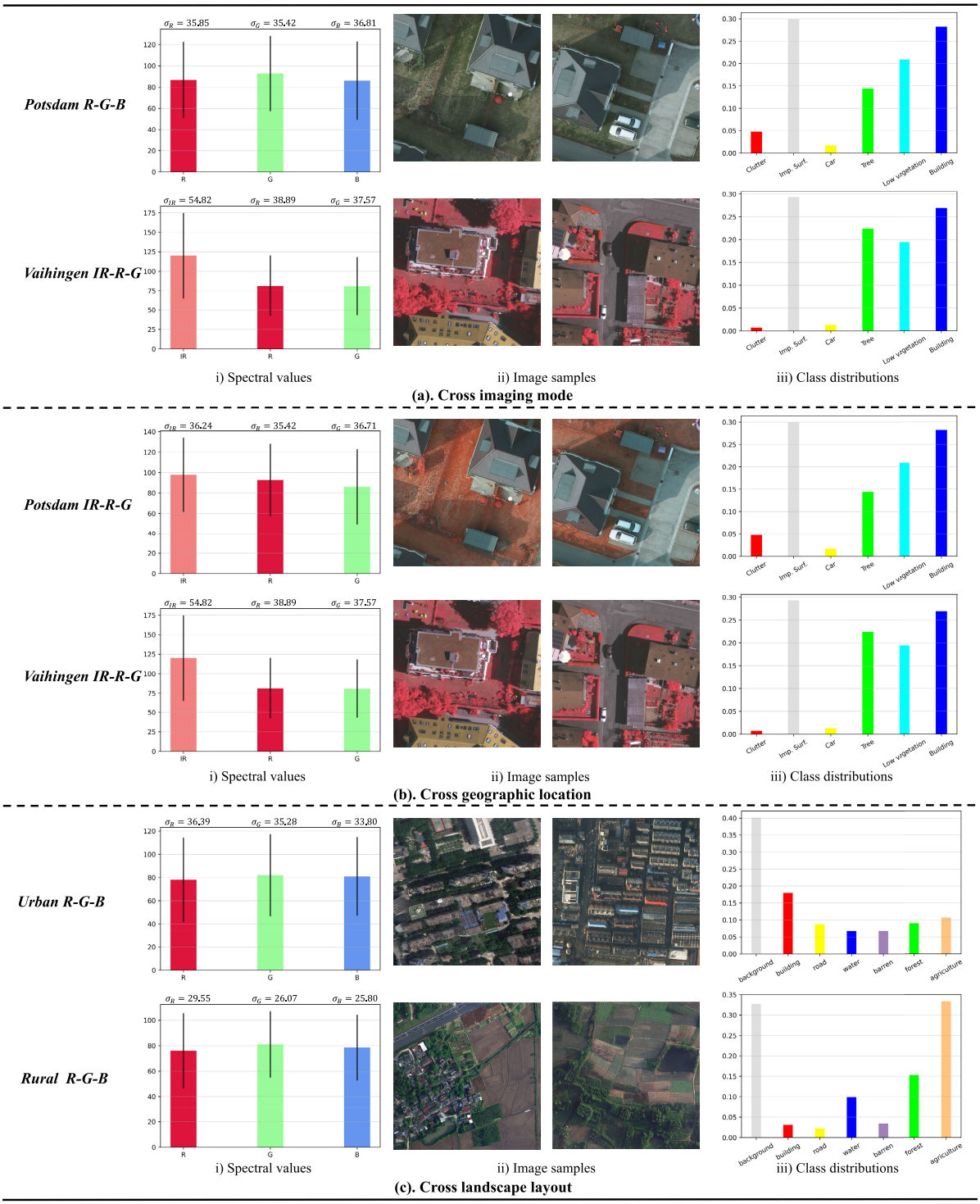


Fig. 1. Three typical unsupervised domain adaptation semantic segmentation tasks in the remote sensing field. (a) Cross imaging mode. (b) Cross geographic location. (c) Cross landscape layout.

A. Image-to-Image Translation

Benefiting from the recent advances in image translation (e.g., CycleGAN), a number of GAN-based methods are proposed to transfer the appearance of source image to make them visually similar to target, which can help reduce the domain discrepancy before training segmentation models. Hoffman et al. [31] first proposed CyCADA, in which they used CycleGAN to generate target-stylized images and achieved both feature-level alignment and pixel-level alignment. DCAN [34] explored channelwise feature alignment both in the generator and segmentation network. Choi et al. [35] raise a GAN-based self-ensembling data augmentation method by transferring source image style to facilitate domain alignment. CPN [36] and FDA [37] translated the style of source images via a simple Fourier transform

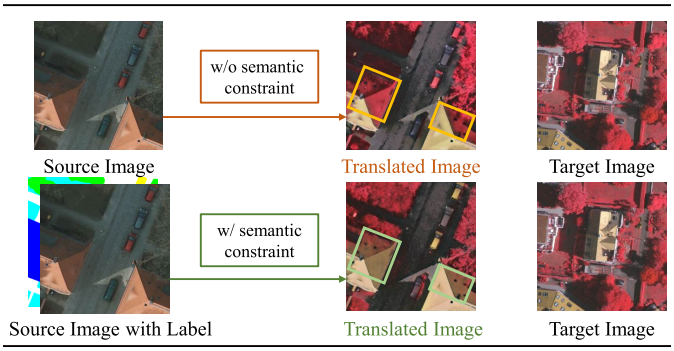


Fig. 2. Example images of source-to-target translation with or without the semantic constraint, which reveals the necessity of semantic constraint. For remote sensing images, the source images are in the R-G-B imaging mode, but the target domain images are in the IR-R-G imaging mode.

and its inverse. Gao et al. [33] proposed CRIN that utilized CycleGAN to generate content invariant images and further introduce an ancillary classifier module to focus on pixel-level divergences to boost the performance of domain adaptation. Unfortunately, these methods do not sufficiently consider semantic constraints and may bring bias to the translated images, which impairs the following semantic segmentation procedure.

B. Adversarial Learning

Adversarial learning-based UDA has been widely explored for semantic segmentation. Adversarial training aims to minimize the discrepancy between source and target feature distributions by introducing a discriminator network alongside the segmentation network. The discriminator takes the feature map from the semantic segmentation network and tries to distinguish the domain of the input. Meanwhile, the segmentation network is trained to fool the discriminator and produce good segmentations on both source and target domains. Hoffman et al. [24] were the first to apply adversarial learning to align the feature maps extracted by the semantic segmentation network between two different domains at a global scale. Tsai et al. [25] found that aligning directly the output space distribution is more effective for semantic segmentation. Domain adaptation in the output space enables the joint optimization for both prediction and representation. Furthermore, ADVENT [38] enforced high prediction certainty (low entropy) on target predictions by introducing an entropy adversarial loss to achieve domain adaptation, which provides an alternative way of output space alignment. Nevertheless, these methods usually focus on global feature alignment and therefore may suffer from negative transfer.

C. Self-Learning

Self-learning is widely explored in the field of unsupervised or semisupervised learning. The key idea is to utilize high-confidence prediction from an ensembled model or a previous model as pseudo-labels for the unlabeled data, which forces the model to learn the domain-invariant features in an implicit way. CBST [39] proposed an iterative self-training method that adjusts class weights to generate more accurate pseudo-labels on target data and retrains the model using these labels. Moreover, the authors proposed a confidence regularized self-training (CRST) [40] framework to regularize pseudo-label and model. Xu et al. [41] first utilized the self-ensembling attention network (SEANet) to extract attention-aware features for domain adaptation under the mean-teacher framework. PTMDA [42] is proposed, in which the authors construct a pseudo target domain to mimic a new domain in a group-specific subspace and align the remainder source domains with the pseudo target domain. Wang et al. [43] proposed a multiprototype clustering method, which enhances intraclass compactness and interclass separation for the target domain, making it easier to construct task-specific decision boundaries. BAFFT [44] proposed a multilevel UDA framework, which considers category homogeneity and diversity in the meantime. DACS [45] demonstrated strong results by combining self-training with ClassMix, which mixes source and target images during the training. Furthermore, DSP [46] not only softly pasted the source image onto the target image but also pasted the target image to the source image. However, this kind of approach is usually sensitive to the threshold.

All the above methods can mitigate the impact of domain gap to some extent, but they all have certain limitations. A combination of the above methods may be a promising solution for UDA in semantic segmentation.

III. Methodology

A. Framework Overview

We focus on the problem of UDA in semantic segmentation. In the source domain, we have N_s images and corresponding pixel-level labels marked as $I_S = \{X_S, Y_S\}$. Samples $x_s^i \in \mathbb{R}^{N\times H\times W}$ and $y_s^i \in \{0,1\}^{C\times H\times W}$ with H and W being the height and width, respectively, N standing for the channels, C denoting the number of classes, and $i = 1, \ldots, N_s$. For target domain, only N_t unlabeled images are available denoted as $I_T = \{X_T\}$. We aim to train a segmentation model to predict accurate label for I_T. Our proposed framework SPGAN-DA is shown in Fig. 3. As it can be seen, the framework is composed of two stages. First, we map the source domain to the target domain by transferring style with our proposed SPGAN. Second, we let the translated target-like images and target domain images to collaboratively train a robust segmentation model.

B. Semantic-Preserved Generative Adversarial Network

We first perform the sematic-preserved source-to-target translation to reduce the pixel-level discrepancy between source and target domains, which is done by our proposed SPGAN. The objective is to map the source domain images to mimic the ones in the target domain since ground-truth labels are only available in the source domain. As shown in Fig. 3, SPGAN has two direction mappings $G{:}S \rightarrow T$ and $F{:}T \rightarrow S$ and two adversarial discriminators D_S and D_T. Note that we make innovations in G architecture. It is intricately fashioned in three distinct components, namely, $G = \{G_{\text{enc}}, G_{\text{dec}}, G_{\text{seg}}\}$, wherein $G_{\text{img}} = G_{\text{enc}} \circ G_{\text{dec}}$ is trained

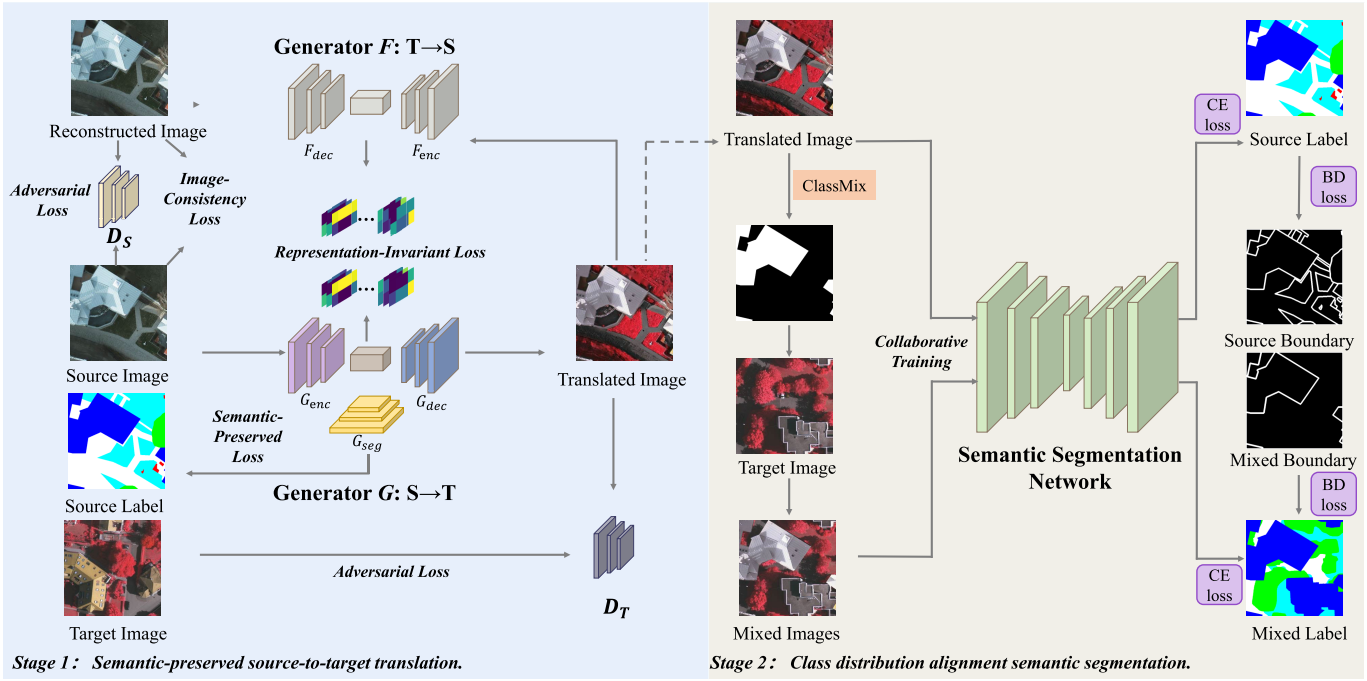


Fig. 3. Framework of our proposed SPGAN-DA. First, we conduct semantic-preserved source-to-target image translation by our proposed SPGAN, wherein G_{enc}, G_{dec}, and G_{seg} are represented in purple, blue, and yellow, respectively. Then, the translated images and mixed images are together fed into the segmentation network for collaborative training.

to produce target-stylized samples that fool an adversarial discriminator D_T and $G_{\text{sem}} = G_{\text{enc}} \circ G_{\text{seg}}$, a newly introduced branch, is utilized to conduct semantic segmentation so that G can preserve semantic information in a supervised way due to that the source labels are available.

We express the adversarial loss as

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, F, D_S, D_T, I_S, I_T) &= E_{x_t \sim X_T}\left[\log D_T(x_t)\right] \\ &+ E_{x_s \sim X_S}\left[\log\left(1 - D_T\left(G_{\text{img}}(x_s)\right)\right)\right] \\ &+ E_{x_s \sim X_S}\left[\log D_S(x_s)\right] \\ &+ E_{x_t \sim X_T}\left[\log(1 - D_S(F(x_t)))\right]. \end{aligned} \tag{1}$$

1) Image-Consistency Loss: To encourage the source content to be preserved during the conversion process, we impose an image-consistency constraint similar to the settings in [32], [47], and [48]. We then require that mapping a source sample from source to target and back to the source reproduces the original sample, thereby enforcing cycle consistency. This is done by introducing an ℓ_1 norm on the reconstruction error, which is referred to as the image-consistency loss

$$\begin{aligned}\mathcal{L}_{\text{img}}(G, F, I_S, I_T) = E_{x_s \sim X_S}\left[\left\|F\left(G_{\text{img}}(x_s)\right) - x_s\right\|_1\right] \\ + E_{x_t \sim X_T}\left[\left\|G_{\text{img}}(F(x_t)) - x_t\right\|_1\right]\end{aligned} \tag{2}$$

where $\|\cdot\|_1$ stands for ℓ_1 norm.

Existing GNA-based source-domain-target-domain image translation methods consider only the two aforementioned losses, thus giving rise to bias. This phenomenon can be described formally by $\lim_{\theta_G} |C(x_s) - C(G_{\text{img}}(x_s))| = b$, where θ_G stands for the parameters of the generator network, $C(\cdot)$ stands for abstract content extraction function, and $G_{\text{img}}(\cdot)$ represents the source-to-target generator.

2) Representation-Invariant Loss: The previous work just considers image-level consistency, but we further take high-level representation-invariant information into consideration since representations contain more high-frequency and abstract information. We observe that the intermediate representations of the two opposite generative networks are forced to be subjected to the same distribution. Eventually, the images generated by the generator are much closer in distribution to the target, and details will be well preserved. Formally, we want $G_{\text{enc}}(x_s) \approx F_{\text{enc}}(G_{\text{img}}(x_s))$ and $F_{\text{enc}}(x_t) \approx G_{\text{enc}}(F(x_t))$ simultaneously. Thus, the representation-invariant loss is formed as

$$\begin{aligned}\mathcal{L}_{\text{rep}}(G, F, I_S, I_T) &= E_{x_s \sim X_S}\left[\left\|G_{\text{enc}}(x_s) - F_{\text{enc}}\left(G_{\text{img}}(x_s)\right)\right\|_1\right] \\ &+ E_{x_t \sim X_T}\left[\left\|F_{\text{enc}}(x_t) - G_{\text{enc}}(F(x_t))\right\|_1\right].\end{aligned} \tag{3}$$

3) Semantic-Preserved Loss: Last but not least, we aim to generate target domain stylized images in which semantic contents are well preserved. The translated images with inconsistent semantic content will impair the following segmentation performance due to the pixel-level misalignment between the translated images and source labels. The previous works [31], [33] have tried to use a source-pretrained segmentation model and fix its weights to compute the semantic consistency loss of the source image prediction and translated image prediction, which is utilized to achieve the goal of semantic-preserved source-to-target translation. However, this kind of constraint highly relies on the pretrained model so that it cannot implement in an end-to-end way causing limited performance gains. In contrast, our approach does not require any pretrained models in the source domain, and semantic constraints can be implemented in an elegant and efficient

way during the training process. Recall that G is composed of three parts $\{G_{\text{enc}}, G_{\text{dec}}, G_{\text{seg}}\}$, where $G_{\text{img}} = G_{\text{enc}} \circ G_{\text{dec}}$ is utilized to perform source-to-target translation. We further explicitly encourage high semantic information preserved by training $G_{\text{sem}} = G_{\text{enc}} \circ G_{\text{seg}}$ in a supervised way since we have access to the source labels, which is regarded as an auxiliary task. Since the auxiliary task is coupled with high semantic information, they have been proved to be beneficial for our main image translation task [49], [50], [51], [52]. Due to this semantic constraint, our network can benefit from it and will preserve semantic information of the objects in images without distortion. We define the semantic-preserved loss as the cross-entropy loss

$$\mathcal{L}_{\text{sem}}(G, I_{\mathcal{S}}) = E_{(x_s, y_s)\sim(X_S, Y_S)}\left[\ell(G_{\text{sem}}(x_s), y_s)\right] \tag{4}$$

where $\ell(\cdot)$ indicates the commonly employed cross-entropy loss function, y_s is the label of the source domain, and $G_{\text{sem}}(x_s)$ is the predicted probability.

According to the above terms of loss, the overall loss function of our SPGAN is formed as

$$\begin{aligned}\mathcal{L}_{\text{SPGAN}}(G, F, D_S, D_T, I_S, I_{\mathcal{T}}) &= \mathcal{L}_{\text{GAN}}(G, F, D_S, D_T, I_{\mathcal{S}}, I_{\mathcal{T}}) \\ &+ \lambda_1 \mathcal{L}_{\text{img}}(G, F, I_{\mathcal{S}}, I_{\mathcal{T}}) + \lambda_2 \mathcal{L}_{\text{rep}}(G, F, I_{\mathcal{S}}, I_{\mathcal{T}}) \\ &+ \lambda_3 \mathcal{L}_{\text{sem}}(G, I_{\mathcal{S}})\end{aligned} \tag{5}$$

where λ_1 is typically set to a value within [10, 20], λ_2 is typically set to 1, and λ_3 is progressively larger with the training epochs.

By introducing representation-invariant and semantic-preserved constraints into the GAN-based translation model, the phenomenon of bias can be minimized, i.e., $\lim_{\theta_G \to \theta^*} b = 0$, where θ_G represents for the parameters of the generator network and θ^* represents for the optimal parameters of the generator network.

After SPGAN is well trained, it is utilized to conduct semantic-preserved source-to-target translation, and the translated images and their original source labels form a new dataset marked as $I'_{\mathcal{S}} = \{X'_S,\ Y_S\}$.

C. CDA Semantic Segmentation

To further reduce the gap between the translated and target images, multilevel refinement semantic segmentation is proposed. In detail, multilevel refinement semantic segmentation is composed of two aspects. On the one hand, object discrepancy is eliminated by introducing the ClassMix operation [45] in the model input stage. On the other hand, boundary enhancement is proposed to refine the performance of object boundaries during the model output period. We first utilize the ClassMix strategy to randomly paste half of the classes in a translated image and the corresponding pixels are cut out and pasted onto an image from the target domain. Then, the mixed images are obtained, whose style is almost highly similar to the target domain and the texture information of the object is both in the source and target domains. The translated images and mixed images are together fed into the segmentation network for cotraining, which will make the segmentation model robust enough on the target domain.

Formally, given a labeled translated image x'_s and an unlabeled target image x_t, let us denote M_{mask} as the selection indicator for the pixels of randomly selected half classes in x'_s, where $M_{\text{mask}}^{(h,w)} = 1$ if the pixel, located at the hth row and wth column, belongs to the selected classes, and $M_{\text{mask}}^{(h,w)} = 0$ otherwise. The mixed image can be formed as

$$x_m = M_{\text{mask}} \odot x'_s + (1 - M_{\text{mask}}) \odot x_t \tag{6}$$

where $\odot$ stands for elementwise product.

To get the labels of the mixed image x_m, a mean-teacher model [53] is employed to assign pseudo-labels to the target image. In particular, the target unlabeled image x_t is fed into the teacher segmentation network to obtain pseudo-label $\hat{y}_t$, and then, the labels for the mixed image x_m can be obtained by the same ClassMix operation. The details can be referred to [45]

$$y_m = M_{\text{mask}} \odot y_s + (1 - M_{\text{mask}}) \odot \hat{y}. \tag{7}$$

After obtaining $I_{\mathcal{M}} = \{X_m, Y_m\}$, both of the translated images and mixed images are utilized to train the semantic segmentation model with the cross-entropy loss

$$\begin{aligned}\mathcal{L}_{\text{seg}}\left(P, I'_{\mathcal{S}}, I_{\mathcal{M}}\right) = &E_{(x_{s'}, y_s)\sim(X_{S'}, Y_S)}\left[\ell(P(x_{s'}), y_s)\right] \\ &+ E_{(x_m, y_m)\sim(X_M, Y_M)}\left[\ell(P(x_m), y_m)\right]\end{aligned} \tag{8}$$

where $P(\cdot)$ represents the probability predicted by the segmentation model.

Boundary is an important factor in the semantic segmentation procedure, while existing methods usually pay attention to the overall performance but ignore the importance of object boundaries. Thus, boundary enhancement constraint is further proposed to achieve this goal. For the mixed images, the boundary weight map masks are obtained by the ClassMix masks and the nearest four pixels to cut-paste edge are kept. In other words, only pixels with distances smaller than four are considered to calculate the boundary weight. Thus, we can calculate the boundary enhancement loss $\mathcal{L}_{\text{mix\_b}}$ for mixed images as

$$\mathcal{L}_{\text{mix\_b}} = \frac{1}{HW} \sum_{h,w}^{H,W} M_{\text{mix\_b}} \odot \mathcal{L}_{ce}\left(X_m^{(h,w)}, Y_m^{(h,w)}\right). \tag{9}$$

For the translated images, it is easy to obtain boundary masks from their labels. It is worth noting that the number of pixel points located at the boundaries in the source domain image is more than that of the mixed image. For this reason, we consider a scaling factor γ that is used to balance the contribution of the source domain image and the mixed image to the boundary enhancement. The translated image boundary loss can be calculated as

$$\mathcal{L}_{\text{src\_b}} = \frac{\gamma}{HW} \sum_{h,w}^{H,W} M_{\text{src\_b}} \odot \mathcal{L}_{ce}\left(X_s^{(h,w)}, Y_s^{(h,w)}\right) \tag{10}$$

where $\gamma = |\{m_{i,j} = 1 | m \in M_{\text{mix\_b}}\}| / |\{n_{i,j} = 1 | n \in M_{\text{src\_b}}\}|$ represents for scaling factor to balance source boundary and mixed boundary pixel numbers.

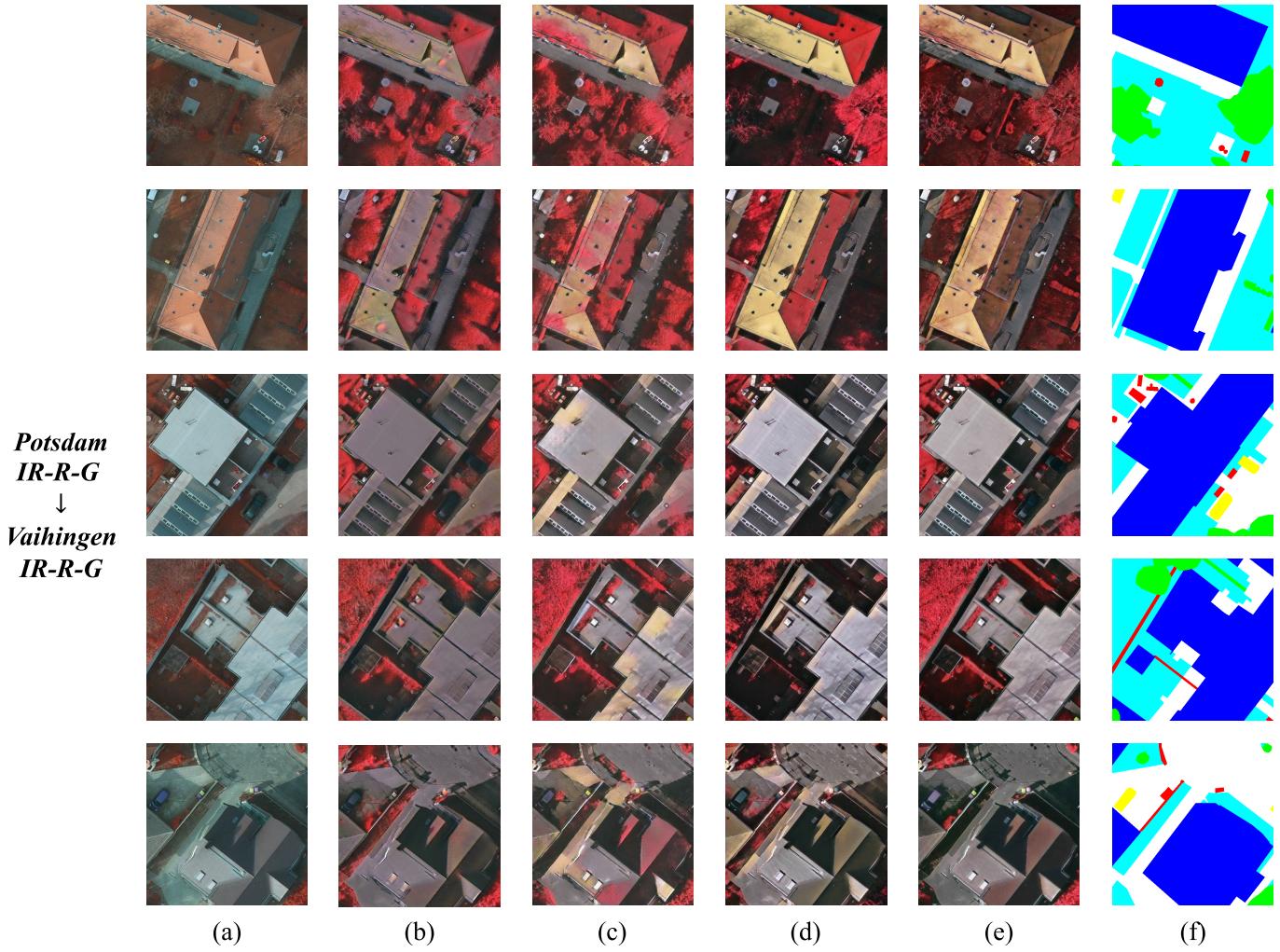


Fig. 4. Image translation results on Potsdam IR-R-G. (a) Source Images. (b) CycleGAN. (c) DualGAN. (d) DiscoGAN. (e) Our SPGAN. (f) Source labels.

Based on the above considerations, the overall objective of our collaboratively adaptation boundary enhancement semantic segmentation module can be written as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_b \cdot (\mathcal{L}_{\text{src\_}b} + \mathcal{L}_{\text{mix\_}b}) \tag{11}$$

where λ_b is a hyperparameter to control the weight of boundary enhancement module.

IV. Experiments

In this section, we first describe the dataset and experiment settings utilized in this work. Then, the experimental results are presented and analyzed in detail.

A. Task Settings

1) Experimental Settings and Evaluation Metrics:

a) ISPRS 2D: ISPRS 2D [54] is offered by the International Society for Photogrammetry and Remote Sensing 2-D Semantic Labeling Contest. It contains two subsets the Potsdam and the Vaihingen. Potsdam subset contains 38 aerial images covering 3.42-km^2 area of Potsdam city with a spatial resolution of 5 cm. The images are fixed with a size of 6000 × 6000 pixels in three channels: red, green, and blue, which spans six categories: building, tree, car, impervious surfaces, low vegetation, and clutter. The Vaihingen subset contains 33 aerial images covering 1.38-km^2 area of the

TABLE I
FID Score (Lower Is Better) on Potsdam R-G-B Translation Results

Task	Method	FID Score (↓)
Potsdam IR-R-G ↓ Vaihingen IR-R-G	CycleGAN	135.08
	DualGAN	133.17
	DiscoGAN	135.16
	Our SPGAN	**107.22**
Potsdam R-G-B ↓ Vaihingen IR-R-G	CycleGAN	120.71
	DualGAN	118.75
	DiscoGAN	125.86
	Our SPGAN	**84.24**
Urban R-G-B ↓ Rural R-G-B	CycleGAN	83.34
	DualGAN	81.55
	DiscoGAN	82.28
	Our SPGAN	**78.32**
Rural R-G-B ↓ Urban R-G-B	CycleGAN	96.87
	DualGAN	95.55
	DiscoGAN	95.68
	Our SPGAN	**89.75**

Fig. 5. Image translation results on Potsdam R-G-B. (a) Source Images. (b) CycleGAN. (c) DualGAN. (d) DiscoGAN. (e) Our SPGAN. (f) Source labels.

Vaihingen city with a spatial resolution of 9 cm. The size of each image is approximately 2000 × 2000 pixels in three different channels: near-infrared, red, and green, with the same categories as in Potsdam. The Potsdam dataset contains three different imaging modes—IR-R-G: three channels (IR-R-G), R-G-B: three channels (R-G-B), and RGBIR: four channels (R-G-B-IR). We use the first two kinds. The Vaihingen dataset contains only one imaging mode—IR-R-G: three channels (IR-R-G).

b) LoveDA: The LoveDA [55] dataset contains 5987 fine-resolution optical remote sensing images (a ground sample distance (GSD) of 0.3 m) at a size of 1024 × 1024 pixels and includes seven categories, i.e., building, road, water, barren, forest, agriculture, and background. The dataset encompasses urban and rural two scenes, which are collected from three cities (Nanjing, Changzhou, and Wuhan) in China. Therefore, considerable challenges are brought due to the multiscale objects, complex background, and inconsistent class distributions.

In detail, we provide three cross-domain experimental settings: 1) cross geographic location, i.e., Potsdam IR-R-G dataset serves as the source domain and the Vaihingen IR-R-G dataset serves as the target domain; 2) cross imaging mode, more precisely, the Potsdam R-G-B dataset serves as the source domain and the Vaihingen IR-R-G dataset serves as the target domain; and 3) cross landscape layout, Urban R-G-B in LoveDA serves as the source domain and Rural R-G-B in LoveDA serves as the target domain. In addition, we also have tried in the opposite direction.

2) Implementation Details: For the generator, it is configured with equal eight numbers of convolution layers (kernel size of 4 × 4, stride 2, and output channel {64, 128, 256, 512, 512, 512, 512, 512}) and deconvolution layers. In addition, we configure the generator with skip connections between mirrored convolution and convolution layers, making it a U-shaped net. The discriminator has five convolution layers with kernel 4 × 4 with channel numbers {64, 128, 256, 512, 1}. The model is optimized by RMSProp with a learning rate of 5×10^{-5} and a weight decay of 0.1. The model is trained for a total of 45 epochs with a batch size of 2. We also set λ_1 and λ_2 to 20 and 1, respectively. We empirically observe that weight run-up is necessary for enhancing the effectiveness of the semantic-preserved loss. Thus, λ_3 follows the formula $\lambda_3 = 5 \cdot e^{-5(1-x)^2}$, where $x \in [0, 1]$ denotes the ration between the current epoch and the whole epochs.

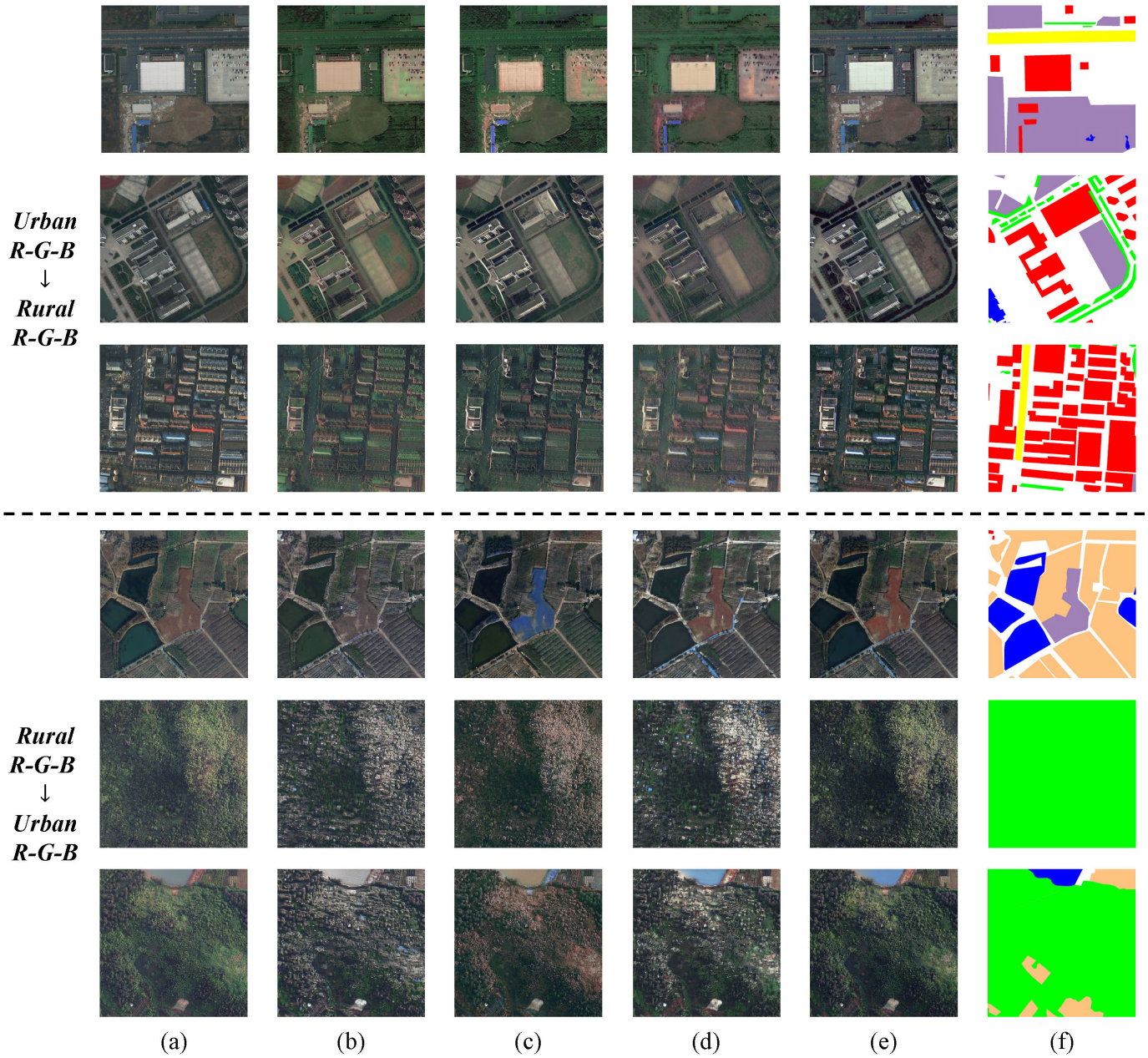


Fig. 6. Image translation results on Urban R-G-B ↔ Rural R-G-B. (a) Source Images. (b) CycleGAN. (c) DualGAN. (d) DiscoGAN. (e) Our SPGAN. (f) Source labels.

Following previous work [45], we employ DeepLab-v2 segmentation model [12] with a ResNet-101 backbone [56]. We train using SGD [57] with a learning rate of 5×10^{-4}, a weight decay of 5×10^{-4}, and a polynomial decay with exponent of 0.9. We also apply color jittering and Gaussian blurring for data augmentation. The model is trained for a total of 250 000 iterations with a batch size of 2. All the experiments are implemented by PyTorch and trained on a single Nvidia GeForce RTX 3090.

3) Evaluation Metrics: We adopt the evaluation metric from [58], aimed at assessing visual quality and discovered correspondence. For the first, we utilize the widely used Frechet inception distance (FID) metric, which empirically estimates the distribution of real and generated images in a deep network space and computes the divergence between them. Intuitively, if the generated images are realistic, they should have similar summary statistics as real images, in any feature space. In particular, we have the ground truth of paired label maps. If accurate correspondences are discovered, the algorithm should generate images that are recognizable as the correct class.

Moreover, we also used the intersection over union (IoU) to measure the efficiency of the segmentation. Since we have some different classes, IoU is calculated for every class separately. Then, the mean IoU of all classes is calculated. Equation (12) represents how to calculate IoU for two different data samples

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} \tag{12}$$

where A is the set of ground-truth pixels, B is the set of predicted pixels, $\cap$ and $\cup$ denote intersection and union, respectively, and $|\cdot|$ denotes calculating the number of pixels in the set.

B. Superiority Verification of the Presented SPGAN Model

The three typical tasks mentioned have more or less spectral differences or even imaging band differences between the source and target domains. According to our approach, the GAN will first be used to align the spectra and even the bands. In fact, all the GAN-based image translation methods (CycleGAN, DualGAN, DiscoGAN, and so on) can align the spectral or imaging mode differences. However, bias (unsmooth and discordant areas) often appears in the translated images. Thus, the advantages of our proposed SPGAN will be verified in this section.

We first conduct experiments on Potsdam IR-R-G to Vaihingen IR-R-G datasets. According to Fig. 1, it can be concluded that there are differences in spectral statistics between Potsdam IR-R-G and Vaihingen IR-R-G datasets. The image translation results are shown in Fig. 4. It can be found that in the comparison methods, all have different degrees of biased translation phenomenon, showing unsmooth or even distorted areas. For example, part of the building is mixed up with low vegetation in the translated images. The reason for this phenomenon is that GAN focuses on the alignment of global statistics and lacks attention to low-level semantic information. However, SPGAN presents more realistic and smooth translation results benefiting from introducing semantic constraints. Furthermore, we evaluate the quality of the translated images using FID, as shown in Table I. The FID score of SPGAN is 107.22 and significantly smaller than other methods, indicating that the distribution of the translated images is closer to the distribution of the target domain images.

Then, the experiments are implemented under the setting of Potsdam R-G-B to Vaihingen IR-R-G, which is more challenging with different imaging modes. The translated images are shown in Fig. 5. Similarly, we use FID to evaluate the quality of the translated images, as shown in Table I. The FID score of SPGAN is 84.24, which is significantly smaller than other methods, indicating that the distribution of the translated images is closer to that of the target domain images.

Finally, experiments on image translation between urban and rural are conducted. In this setting, there are small discrepancies in spectra between the source and target domains. However, the large differences in category distribution and the lack of attention to semantic information in GAN still result in bias in the translated images. There will be the phenomenon of creating something out of nothing. For instance, in the urban-to-rural translation, vegetation appears above buildings. In the urban-to-urban migration, vegetation appears on top of buildings, while in the rural-to-urban migration, scattered building patches appear on the vegetation, as shown in the last row of Fig. 6. On the contrary, the proposed SPGAN generates semantically consistent target-style images that appear to be noise-free. The FID score of SPGAN is 78.32 in urban to rural and 89.75 in rural to urban, as shown in Table I, which is significantly smaller than other methods. In general, our proposed SPGAN can well output semantically consistent and noise-free target-style images in different scenarios, proving its effectiveness.

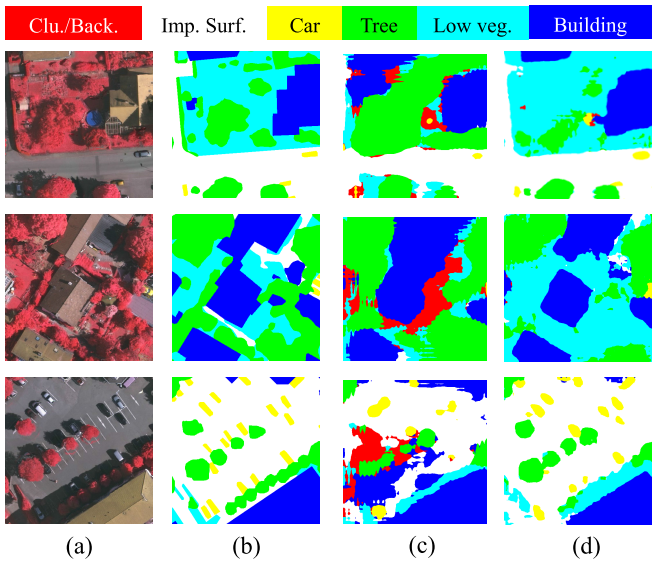


Fig. 7. Visualization results on Potsdam IR-R-G → Vaihingen IR-R-G. (a) Target image. (b) Ground truth. (c) Source only. (d) Ours.

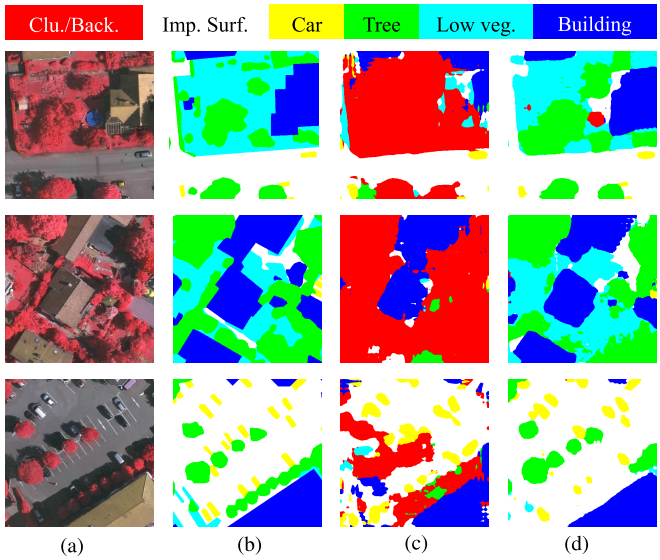


Fig. 8. Visualization results on Potsdam R-G-B → Vaihingen IR-R-G. (a) Target image. (b) Ground truth. (c) Source Only. (d) Ours.

C. Comparison With State-of-the-Art Methods

1) Comparison Under Cross Geographic Location: For the segmentation part, Table II reports the experimental results of our method compared with the advanced methods. Due to the serious domain shift between the two datasets, the source-only method presents poor performance. In general, the domain adaptation methods significantly boost the performance in overall metrics and individual categories, revealing their ability to transfer knowledge from the labeled source domain to the unlabeled target domain. Our best model SPGAN-DA achieves mIoU as high as 52.93%, thereby improving the baseline by

TABLE II
RESULTS (MIOU IN %) OF DIFFERENT DOMAIN ADAPTATION METHODS UNDER CROSS GEOGRAPHIC LOCATION

Task	Methods	Clutter	Impervious surfaces	Car	Tree	Low vegetation	Building	mIoU
Potsdam IR-R-G ↓ Vaihingen IR-R-G	Source Only	5.71	35.84	20.27	54.95	17.88	51.59	31.04
	Oracle	75.41	83.74	65.47	77.42	70.86	88.96	76.97
	Benjdirs's [59]	2.12	39.88	8.20	26.56	26.53	40.97	24.04
	MWCSS [60]	**29.66**	49.41	34.34	57.66	38.87	62.30	45.38
	AdaptSegNet [25]	4.60	54.39	6.40	52.65	28.98	63.14	35.02
	CsDA [61]	9.85	46.22	31.14	52.04	31.11	52.39	37.12
	AdvEnt [38]	10.18	57.03	35.28	59.02	33.65	67.41	43.76
	CLAN [62]	9.89	58.23	37.25	59.10	36.74	59.10	43.38
	LSR-EGA [63]	11.25	59.61	38.56	56.55	38.73	71.04	45.95
	MRNet [64]	3.81	55.02	34.37	54.79	26.46	76.39	41.80
	SPGAN-DA (Ours)	11.94	**65.33**	**48.25**	**66.04**	**45.20**	**83.11**	**53.31**

TABLE III
RESULTS (MIOU IN %) OF DIFFERENT DOMAIN ADAPTATION METHODS UNDER CROSS IMAGING MODE

Task	Methods	Clutter	Impervious surfaces	Car	Tree	Low vegetation	Building	mIoU
Potsdam R-G-B ↓ Vaihingen IR-R-G	Source Only	1.76	26.86	16.70	44.48	12.56	41.67	24.01
	Oracle	75.41	83.74	65.47	77.42	70.86	88.96	76.97
	Benjdirs's [59]	4.48	31.78	21.70	41.76	23.67	52.36	29.31
	MWCSS [60]	3.94	46.19	40.31	55.82	27.85	65.44	39.93
	AdaptSegNet [25]	1.08	50.05	14.18	56.45	20.73	62.61	34.18
	CsDA [61]	0.55	44.82	23.81	52.04	20.74	53.39	32.56
	AdvEnt [38]	0.73	55.43	28.28	59.02	20.73	68.49	38.78
	CLAN [62]	0.84	57.30	17.28	**59.10**	24.94	59.19	36.44
	LSR-EGA [63]	5.84	57.11	23.65	56.55	28.73	70.24	40.35
	MRNet [64]	0.81	54.11	29.39	54.99	16.16	75.39	38.47
	SPGAN-DA (Ours)	**23.45**	**67.88**	**49.79**	55.26	**47.27**	**82.83**	**54.41**

TABLE IV
RESULTS (MIOU IN %) OF DIFFERENT DOMAIN ADAPTATION METHODS UNDER CROSS LANDSCAPE LAYOUT

Task	Methods	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU
Urban R-G-B ↓ Rural R-G-B	Source Only	24.16	37.02	32.56	49.42	14.00	29.34	35.65	31.74
	Oracle	37.18	52.74	43.74	65.89	11.47	45.78	62.91	45.67
	DDC [65]	25.61	44.27	31.28	44.78	13.74	33.83	25.98	31.36
	AdaptSeg [25]	26.89	40.53	30.65	50.09	16.97	32.51	28.25	32.27
	FADA [29]	24.39	32.97	25.61	47.59	15.34	34.35	20.29	28.65
	CLAN [62]	22.93	44.78	25.99	46.81	10.54	37.21	24.45	30.39
	TransNorm [66]	19.39	36.30	22.04	36.68	14.00	**40.62**	3.30	24.62
	PyCDA [67]	12.36	38.11	20.45	57.16	**18.32**	36.71	41.90	32.14
	CBST [39]	25.06	44.02	23.79	50.48	8.33	39.16	49.65	34.36
	IAST [68]	29.97	49.48	28.29	64.49	2.13	33.36	**61.37**	38.44
	UDA-CL [69]	28.55	49.69	**35.74**	53.52	4.96	31.36	52.26	36.58
	SPGAN-DA (Ours)	**55.06**	**50.71**	33.80	**65.01**	9.07	25.43	55.44	**42.07**
Rural R-G-B ↓ Urban R-G-B	Source Only	43.30	25.63	12.70	76.22	12.52	23.34	25.14	31.27
	Oracle	48.18	52.14	56.81	85.72	12.34	36.70	35.66	46.79
	DDC [65]	43.60	15.37	11.98	79.07	14.13	33.08	23.47	31.53
	AdaptSeg [25]	42.35	23.73	15.61	81.95	13.62	28.70	22.05	32.68
	FADA [29]	43.89	12.62	12.76	80.37	12.70	32.76	24.79	31.41
	CLAN [62]	43.41	25.42	13.75	79.25	13.71	30.44	25.80	33.11
	TransNorm [66]	38.37	5.04	3.75	80.83	14.19	33.99	17.91	27.73
	PyCDA [67]	38.04	35.86	45.51	74.87	7.71	**40.39**	11.39	36.25
	CBST [39]	48.37	46.10	35.79	80.05	19.18	29.69	30.05	41.32
	IAST [68]	**48.57**	31.51	28.73	**86.01**	20.29	31.77	36.50	40.48
	UDA-CL [69]	48.15	37.44	45.05	84.29	16.68	26.66	34.12	41.77
	SPGAN-DA (Ours)	35.45	**50.31**	**51.02**	68.42	**42.68**	40.17	**51.90**	**48.56**

21.89%. Compared with the other competing methods, the model SPGNA-DA still has higher performance and makes strong enough predictions in some challenging categories such as "car," "tree," and "building," which evidences the robustness of our SPGNA-DA. Fig. 7 shows some qualitative segmentation examples obtained by the baseline and our proposed method on Potsdam IR-R-G and Vaihingen IR-R-G. Due to the severe domain shift problem between the Potsdam IR-R-G and the Vaihingen IR-R-G, the predictions of "Source Only" usually appear as noisy segmentation or wrong context and lose the object boundary and structure information. After adaptation by our SPGAN-DA, the results improved a lot and preserved more structure information and detail, especially on the edges.

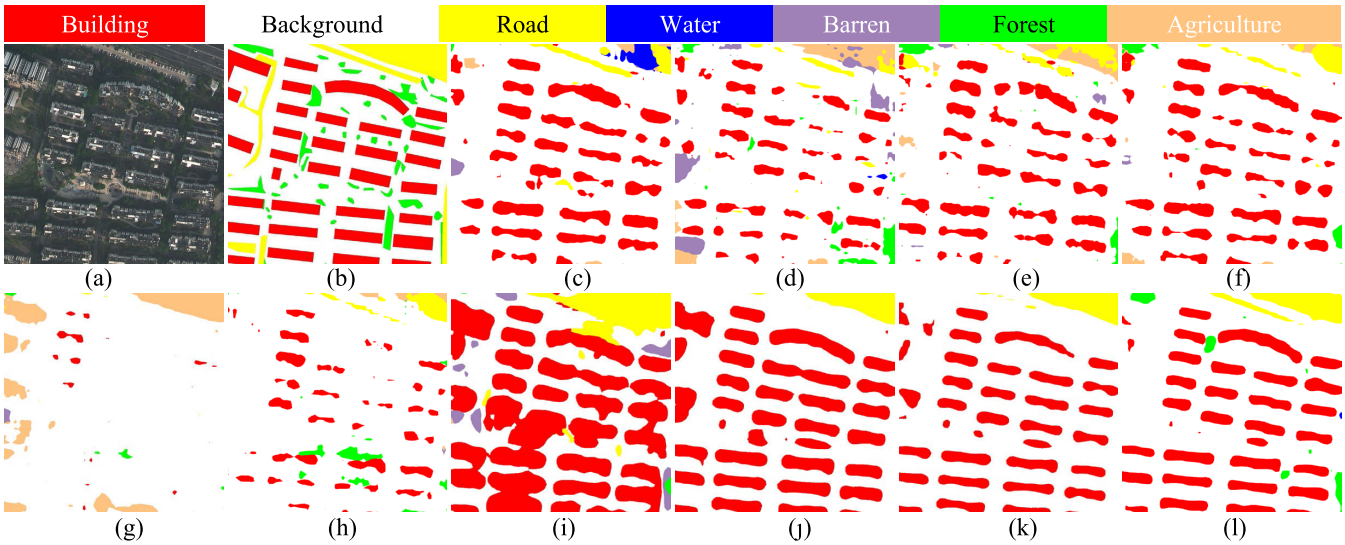


Fig. 9. Visualization results on Rural R-G-B → Urban R-G-B. (a) Image. (b) Ground truth. (c) Source Only. (d) DDC. (e) AdaptSeg. (f) CLAN. (g) TransNorm. (h) FADA. (i) PyCDA. (j) CBST. (k) IAST. (l) Ours.

2) Comparison Under Cross Imaging Mode: To further verify the generalization ability of our method, we conduct experiments on a classical remote sensing domain adaptation task, i.e., Potsdam R-G-B and Vaihingen IR-R-G datasets. The quality of the translated image is shown in Fig. 8 and Table III. Table III reports the semantic segmentation results of our method compared with the advanced methods on Potsdam R-G-B and Vaihingen IR-R-G datasets. Due to the serious domain shift between the two datasets, the source-only method presents extremely poor performance. Generally speaking, the domain adaptation methods significantly boost the performance in overall metrics and individual categories, revealing their ability to transfer knowledge from the labeled source domain to the unlabeled target domain. Our best model SPGAN-DA achieves mIoU as high as 51.89%, thereby improving the baseline by 27.88%. Compared with the other competing methods, the model SPGNA-DA still has higher performance and makes strong enough predictions in some challenging categories such as "low vegetation," "clutter," and "car," which evidences the robustness of our SPGNA-DA. The main reason is that, compared with natural images, there are huge discrepancies between the two datasets caused by different imaging bands so that the other approaches may overshadow than ours. As for our method, it first aligns the style (i.e., imaging bands in this task) by our proposed SPGAN and second uses the ClassMix strategy so that the model has seen both the source and target domain objects. Thus, it is robust enough to produce high-confidence predictions for the target domain even when dealing with such significant domain gaps. The sematic segmentation visual results are shown in Fig. 8. Similar to the previous experimental task setting, the predictions of "Source Only" usually appear as noisy segmentation or wrong context. Our proposed SPGAN-DA presents the results similar to the real labels.

3) Comparison Under Cross Landscape Layout: To evaluate the effectiveness of our method in more cases, we conduct experiments in the setting of urban and rural in both directions.

TABLE V
ABLATION STUDY OF OUR PROPOSED SPGAN-DA FRAMEWORK

	SO	AL + ICL	RIL	SPL	ClassMix	BEL	mIoU(%)	Gain(%)
Potsdam IR-R-G ↓ Vaihingen IR-R-G	✓						31.0	-
		✓					38.9	7.9
		✓	✓				43.3	12.2
		✓	✓	✓			45.4	14.4
		✓	✓	✓	✓		51.5	20.5
		✓	✓	✓	✓	✓	**53.3**	**22.3**
Potsdam R-G-B ↓ Vaihingen IR-R-G	✓						24.0	-
		✓					36.9	12.9
		✓	✓				41.5	17.5
		✓	✓	✓			44.2	20.2
		✓	✓	✓	✓		52.3	28.3
		✓	✓	✓	✓	✓	**54.4**	**30.4**
Urban R-G-B ↓ Rural R-G-B	✓						31.7	-
		✓					33.5	1.8
		✓	✓				34.1	2.4
		✓	✓	✓			36.1	4.4
		✓	✓	✓	✓		40.5	8.8
		✓	✓	✓	✓	✓	**42.0**	**10.3**
Rural R-G-B ↓ Urban R-G-B	✓						31.3	-
		✓					34.2	2.9
		✓	✓				35.9	4.6
		✓	✓	✓			38.2	6.9
		✓	✓	✓	✓		46.7	15.4
		✓	✓	✓	✓	✓	**48.6**	**17.3**

In this situation, the semantic segmentation model suffers from small spectral or imaging mode discrepancies but large class distribution discrepancies.

a) Urban R-G-B → Rural R-G-B: The results for this set of experiments are reported in Table IV. It can be concluded that the semantic segmentation network has lost its effectiveness due to the domain shift, referring to the result of the Source Only setting. It corroborates the complexity of the task due to a strong and inconsistent class distribution between the source and target domains, which is dominated by urban scenes with a mix of buildings and highways but few

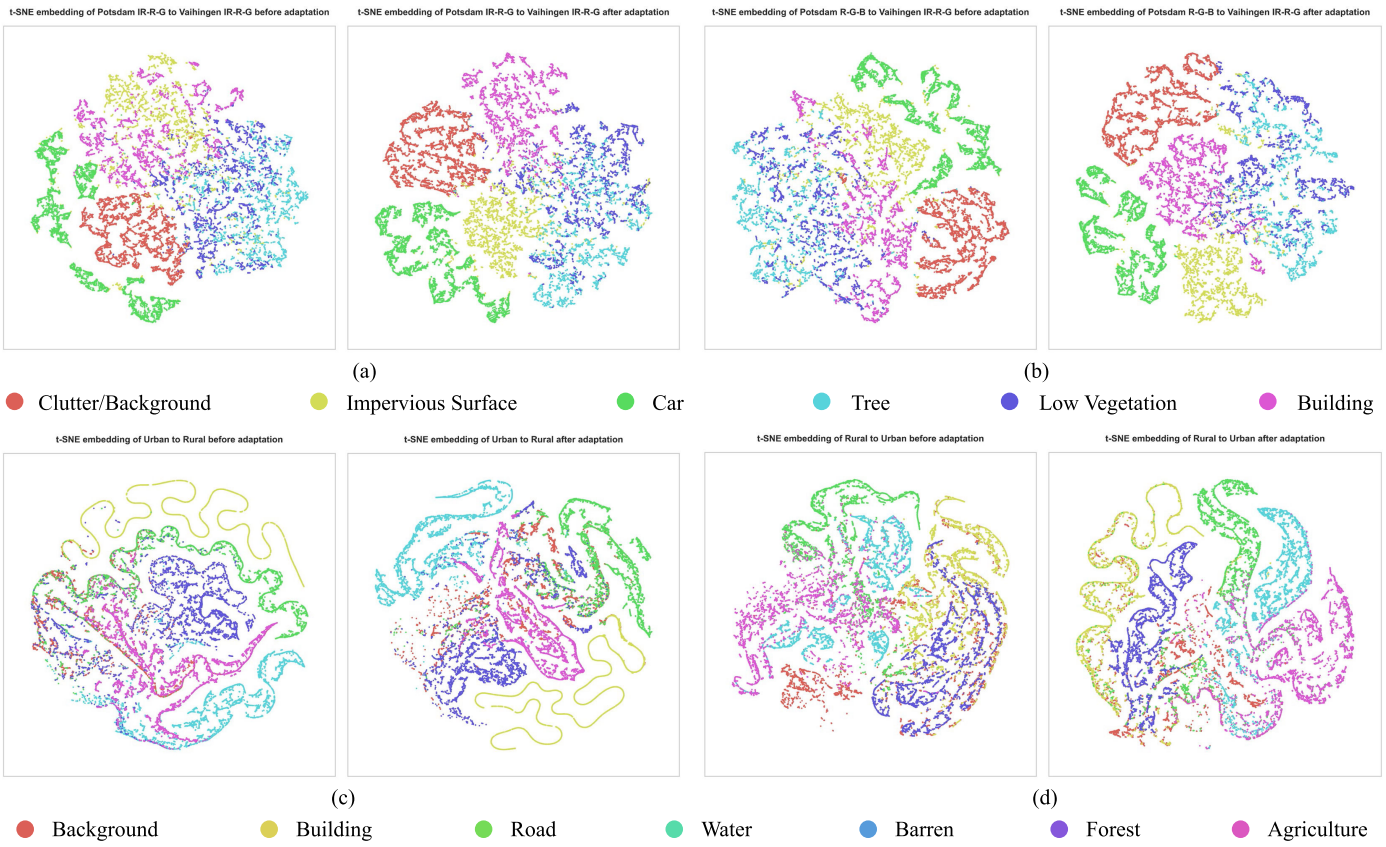


Fig. 10. Visualization of feature distribution. (a) Potsdam IR-R-G → Vaihingen IR-R-G. (b) Potsdam R-G-B → Vaihingen IR-R-G. (c) Urban R-G-B → Rural R-G-B. (d) Rural R-G-B → Urban R-G-B.

natural items. Table IV shows that adversarial methods may suffer from negative transfer and achieve overall performance equivalent to, if not worse than, the Source Only model. Self-training methods generate pseudo-labels on the target images. With the involvement of target samples, the class distribution divergence is eliminated to some extent during the training. Our method exhibits its ability to boost rural and underrepresented classes, such as agriculture. In addition, our method recognizes and classifies better contours and classes, such as water, despite their underrepresentation in the source domain.

b) Rural R-G-B → Urban R-G-B: The results for this set of experiments are summarized in Table IV The source domain in this scenario is dominated by large-scale natural objects and a few man-made samples. In terms of mIoU, our method achieves 48.56, which surpasses the Source Only model by 17.29. The qualitative results for the Rural to Urban experiments are shown in Fig. 9. Our method successfully recognizes the buildings and roads and is the closest to the ground truth.

D. Model Analyses

1) Ablation Study: In this section, we conduct the ablation experiments to validate the individual effects of different components under the setting of Potsdam IR-R-G/R-G-B → Vaihingen and Urban ↔ Rural, as shown in Table V.

a) SO: We consider Source Only as the baseline, and the mIoU is 34.83% in the setting of Potsdam RGB-to-Vaihingen IRRG. The mIoU is 40.29% on the Potsdam IRGB-to-Vaihingen IRRG case. In addition, the mIoU is 31.7 and 31.3 under the setting of Urban ↔ Rural, respectively.

b) AL + ICL.: "AL + ICL" means the adversarial loss and image-consistency loss used in the traditional GAN model. The AL + ICL improves the mIoU by 7.9% and 12.9% under Potsdam IR-R-G/Potsdam R-G-B and Vaihingen IR-R-G, respectively. This demonstrates that spectral or imaging mode differences are a dominant factor in this cross-domain semantic segmentation setting. However, AL + ICL does not boost as significantly as the two previous situations in the Urban ↔ Rural cases. This indicates that in this situation, the model is less affected by spectral differences but mainly by the discrepancy of the class distribution.

c) RIL: "RIL" means our proposed representation-invariant loss corresponding to (3).

d) SPL: "SPL" means semantic-preserved loss as shown in (4). With the participation of SPL, mIoU is improved by 14.4, 20.2, 4.4, and 6.9. It proves that our proposed SPGAN can well output semantically consistent and noise-free target-style images in different scenarios.

e) BEL: "BEL" stands for boundary enhancement loss in CDA corresponding to (6). CDA is designed to mitigate the differences in class distribution through the ClassMix operation, and the boundary enhancement loss refines the boundaries of objects, contributing to the model performance. Although there is little difference in category distribution in P2V, CDA is able to enhance the generalization of the model

TABLE VI
HYPERPARAMETER ANALYSIS OF BOUNDARY WEIGHT. (a) POTSDAM IR-R-G/R-G-B → VAIHINGEN IR-R-G. (b) URBAN R-G-B↔ RURAL R-G-B

(a)

Task	Weight	Clutter	Imp. Surf.	Car	Tree	Low veg.	Building	mIoU
Potsdam IR-R-G ↓ Vaihingen IR-R-G	$\lambda_b = 1$	**19.29**	63.67	44.38	62.61	44.43	81.78	52.69
	$\lambda_b = 2$	15.53	63.25	48.21	64.17	**46.35**	83.27	53.46
	$\lambda_b = 3$	13.66	**63.72**	48.98	**66.64**	44.25	**83.84**	**53.51**
	$\lambda_b = 4$	11.25	58.83	**51.89**	60.12	43.71	83.83	51.60
Potsdam R-G-B ↓ Vaihingen IR-R-G	$\lambda_b = 1$	12.64	**67.99**	**51.72**	61.01	48.38	84.99	54.45
	$\lambda_b = 2$	**29.06**	65.83	50.41	56.24	46.97	82.36	**55.14**
	$\lambda_b = 3$	14.06	66.93	51.49	**62.36**	**49.83**	**85.03**	54.94
	$\lambda_b = 4$	8.90	66.49	51.18	60.31	46.93	83.25	52.84

(b)

Task	Weight	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU
Urban R-G-B ↓ Rural R-G-B	$\lambda_b = 1$	57.08	**58.36**	39.05	63.50	2.31	26.19	**61.18**	43.95
	$\lambda_b = 2$	59.32	52.14	36.80	**66.33**	10.92	29.59	57.14	**44.60**
	$\lambda_b = 3$	58.70	43.16	**40.91**	59.91	**14.68**	**35.35**	54.12	43.83
	$\lambda_b = 4$	**61.05**	53.87	38.07	61.17	11.66	13.87	55.79	42.21
Rural R-G-B ↓ Urban R-G-B	$\lambda_b = 1$	**39.96**	49.64	51.18	70.17	44.01	40.63	53.27	49.83
	$\lambda_b = 2$	37.61	53.31	53.02	72.09	**45.00**	40.17	**54.90**	**50.87**
	$\lambda_b = 3$	39.36	52.57	**53.27**	**72.20**	44.39	38.89	51.87	50.36
	$\lambda_b = 4$	36.71	**53.33**	52.01	55.78	44.14	**49.95**	54.26	49.45

TABLE VII
EXPERIMENTS RESULTS (MIOU IN %) UNDER DIFFERENT SEMANTIC SEGMENTATION MODELS

	Model	Potsdam IR-R-G ↓ Vaihingen IR-R-G	Potsdam R-G-B ↓ Vaihingen IR-R-G	Urban R-G-B ↓ Rural R-G-B	Rural R-G-B ↓ Urban R-G-B
SPGAN-DA	UNet	48.93	48.42	37.46	46.78
SPGAN-DA	DeepLab V2	53.31	54.41	42.07	48.56
SPGAN-DA	DeepLab V3+	54.73	55.65	42.96	48.53
SPGAN-DA	SegFormer	**61.77**	**59.97**	**43.36**	**49.61**
DAFormer [72]	SegFormer	60.02	55.91	43.01	48.89
HRDA [73]	SegFormer	61.05	57.97	43.27	48.79

by aligning the architectural differences in the two domains. Specifically, Vaihingen has more densely spaced buildings and Potsdam has wider roads. As for Urban R-G-B ↔ Rural R-G-B, the urban areas always contain more man-made objects such as buildings and roads due to their high population density. In contrast, the rural areas have more agricultural land, that is to say, there is a big difference between the category distribution in urban and rural. CDA directly aligns the class distributions between urban and rural, significantly boosting the segmentation performance on the target domain.

In summary, the domain gap is bridged progressively under our proposed SPGAN-DA framework and each component/stage contributes to the improvement of overall performance.

2) Sensitive Analysis of the Boundary Enhancement Weight: In the collaboratively adaptive boundary enhancement segmentation module, λ_b is a vital hyperparameter that guides how much the network pays attention to the boundary. To analyze the sensitivity of lambda, we evaluate the performance of the whole proposed framework under different values of lambda on both Potsdam IR-R-G-to-Vaihingen IR-R-G and Potsdam R-G-B-to-Vaihingen IR-R-G task settings.

For boundary weight lambda, if the value of λ_b is too low, then boundary enhancement will play a very limited role and bring little performance improvement. On the contrary, if it is too high, it will affect the normal optimization of the network and even lead to performance degradation. Therefore, it is necessary to find a suitable λ_b value. The results of different weights on the two tasks are shown in Table VI. For the cross-domain semantic segmentation task Potsdam IR-R-G to Vaihingen IR-R-G, the method achieves an mIoU of 53.51% on the validation set under the weight of $\lambda_b = 3$. For the other task, it is deduced that the best performance is with an mIoU of 55.14% on the validation set under the weight of $\lambda_b = 2$. As for Urban R-G-B↔ Rural R-G-B, λ_b is set to 2.

3) Visualization of Feature Distribution: To illustrate the effectiveness of the proposed module more visually, Fig. 10 shows the feature distribution maps of the Source Only and our methods. The 2-D space feature distribution maps are obtained by the t-distributed stochastic neighbor embedding (t-SNE)

algorithm [72]. As for Potsdam IR-R-G/R-G-B to Vaihingen IR-R-G, the corresponding features of different categories are mixed, as shown in Fig. 10. In particular, the features of impervious surfaces, low vegetation, buildings, and trees are highly overlapping together, which undoubtedly leads to poor segmentation results. The urban and rural experiments also face the same problem, where different category features are confused, resulting in poor semantic segmentation results. With the help of domain adaptation, SPGAN-DA obtained more accurate segmentation results and attenuated the feature confusion between different categories when compared with the results from Source Only, as presented in Fig. 10. The method further boosts the introcategory feature compactness and intercategory separability. In particular, it increases the distances of feature distributions between category pairs that are easily confused.

4) Semantic Segmentation Method Comparisons: We further explore the cross-domain semantic segmentation experiments under different semantic segmentation models, including DeepLab V3+, UNet, and even SegFormer [73] (vision transformer-based). In addition, we also equipped our SPAGAN-DA with SegFormer and then compared with DAFormer [70] and HRDA [71]. The quantitative results are shown in Table VII. The results indicate that the performance of UNet [74] falls short of expectations, possibly attributed to its relatively simplistic model architecture. The efficacy of DeepLab V3+ surpasses that of DeepLab V2. Benefiting from the global modeling capability of the attention mechanism within Transformers, Segformer significantly outperforms CNN-based semantic segmentation models.

V. Conclusion

In this work, we propose a novel framework, called SPGAN-DA, to bridge the domain gap among different domains. First, SPGAN is proposed to translate source images to the style of the target domain with semantic information preserved, which will minimize the spectral or imaging mode discrepancy without bias. Furthermore, we propose a CDA module that leverages the translated target-like images and target domain images in the model input and output aspect to collaboratively train a domain adaptive semantic segmentation model. This provides an innovative paradigm to deal with UDA tasks. Experiments on remote sensing benchmark datasets demonstrate the effectiveness and generality of our proposed method, which achieves competitive performance compared with other state-of-the-art methods.

References

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

[2] Y. Li et al., "MFVNet: A deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation," *Sci. China Inf. Sci.*, vol. 66, no. 4, Apr. 2023.

[3] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[4] Y. Li, Y. Zhou, Y. Zhang, L. Zhong, J. Wang, and J. Chen, "DKDFN: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 186, pp. 170–189, Apr. 2022.

[5] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 145–158, Sep. 2021.

[6] H. Shi, L. Chen, F.-K. Bi, H. Chen, and Y. Yu, "Accurate urban area detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1948–1952, Sep. 2015.

[7] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.

[8] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018.

[9] Y. Zhang, Y. Lu, D. Zhang, L. Shang, and D. Wang, "RiskSens: A multi-view learning approach to identifying risky traffic locations in intelligent transportation systems using social and remote sensing," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 1544–1553.

[10] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[12] 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.

[13] Y. Li, S. Ouyang, and Y. Zhang, "Combining deep learning and ontology reasoning for remote sensing image semantic segmentation," *Knowl.-Based Syst.*, vol. 243, May 2022, Art. no. 108469.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

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

[16] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: A survey," *Inf. Fusion*, vol. 67, pp. 94–115, Mar. 2021.

[17] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[18] T. Adel, H. Zhao, and A. Wong, "Unsupervised domain adaptation with a relaxed covariate shift assumption," in *Proc. AAAI*, 2017, pp. 1691–1697.

[19] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Stat. Planning Inference*, vol. 90, no. 2, pp. 227–244, Oct. 2000.

[20] D. Zhang, M. Ye, Y. Liu, L. Xiong, and L. Zhou, "Multi-source unsupervised domain adaptation for object detection," *Inf. Fusion*, vol. 78, pp. 138–148, Feb. 2022.

[21] S. Sun, H. Shi, and Y. Wu, "A survey of multi-source domain adaptation," *Inf. Fusion*, vol. 24, pp. 84–92, Jul. 2015.

[22] H. Huang and Q. Liu, "Domain structure-based transfer learning for cross-domain word representation," *Inf. Fusion*, vol. 76, pp. 145–156, Dec. 2021.

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

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

[25] 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.

[26] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. ICCV*, Jun. 2019, pp. 6778–6787.

[27] L. Du et al., "SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 982–991.

[28] Y. Chen, W. Li, X. Chen, and L. Van Gool, "Learning semantic segmentation from synthetic data: A geometrically guided input–output adaptation approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1841–1850.

[29] 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. ECCV*, 2020, pp. 642–659.

[30] M. Kim, S. Joung, S. Kim, J. Park, I.-J. Kim, and K. Sohn, "Cross-domain grouping and alignment for domain adaptive semantic segmentation," in *Proc. AAAI*, 2021, pp. 1799–1807.

[31] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. ICML*, 2018, pp. 1989–1998.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[33] L. Gao, L. Zhang, and Q. Zhang, "Addressing domain gap via content invariant representation for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 9, pp. 7528–7536.

[34] Z. Wu et al., "DCAN: Dual channel-wise alignment networks for unsupervised scene adaptation," in *Proc. ECCV*, 2018, pp. 518–534.

[35] J. Choi, T. Kim, and C. Kim, "Self-ensembling with GAN-based data augmentation for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6829–6839.

[36] Y. Yang, D. Lao, G. Sundaramoorthi, and S. Soatto, "Phase consistent ecological domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9008–9017.

[37] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4084–4094.

[38] 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. (CVPR)*, Jun. 2019, pp. 2512–2521.

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

[40] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5981–5990.

[41] Y. Xu, B. Du, L. Zhang, Q. Zhang, G. Wang, and L. Zhang, "Self-ensembling attention networks: Addressing domain shift for semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 5581–5588.

[42] C.-X. Ren, Y.-H. Liu, X.-W. Zhang, and K.-K. Huang, "Multi-source unsupervised domain adaptation via pseudo target domain," *IEEE Trans. Image Process.*, vol. 31, pp. 2122–2135, 2022.

[43] S. Wang et al., "Cluster alignment with target knowledge mining for unsupervised domain adaptation semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 7403–7418, 2022.

[44] B. Yuan, D. Zhao, S. Shao, Z. Yuan, and C. Wang, "Birds of a feather flock together: Category-divergence guidance for domain adaptive segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 2878–2892, 2022.

[45] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "DACS: Domain adaptation via cross-domain mixed sampling," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1378–1388.

[46] L. Gao, J. Zhang, L. Zhang, and D. Tao, "DSP: Dual soft-paste for unsupervised domain adaptive semantic segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2825–2833.

[47] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. ICML*, 2017, pp. 1857–1865.

[48] 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. (ICCV)*, Oct. 2017, pp. 2868–2876.

[49] K.-H. Lee, G. Ros, J. Li, and A. Gaidon, "SPIGAN: Privileged adversarial learning from simulation," in *Proc. ICLR*, 2019, pp. 1–14.

[50] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 675–684.

[51] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3712–3722.

[52] Q. Wang, D. Dai, L. Hoyer, L. Van Gool, and O. Fink, "Domain adaptive semantic segmentation with self-supervised depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8495–8505.

[53] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NeurIPS*, 2017, pp. 1195–1204.

[54] (2014). *2D Semantic Labeling Contest: Potsdam and Vaihingen*. ISPRS. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling/

[55] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. NeurIPS*, 2021, pp. 1–16.

[56] 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.

[57] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.

[58] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1–12.

[59] 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, p. 1369, Apr. 2019.

[60] 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. Photogramm. Remote Sens.*, vol. 175, pp. 20–33, May 2021.

[61] B. Fang, R. Kou, L. Pan, and P. Chen, "Category-sensitive domain adaptation for land cover mapping in aerial scenes," *Remote Sens.*, vol. 11, no. 22, p. 2631, Nov. 2019.

[62] 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. CVPR*, 2019, pp. 2507–2516.

[63] 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, 2022, Art. no. 5405614.

[64] Z. Zheng and Y. Yang, "Unsupervised scene adaptation with memory regularization in vivo," in *Proc. IJCAI*, 2020, pp. 1076–1082.

[65] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[66] X. Wang, Y. Jin, M. Long, J. Wang, and M. I. Jordan, "Transferable normalization: Towards improving transferability of deep neural networks," vol. 32, 2019, pp. 1–11.

[67] Q. Lian, L. Duan, F. Lv, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6757–6766.

[68] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 415–430.

[69] L. Ran, C. Ji, S. Zhang, X. Zhang, and Y. Zhang, "An unsupervised domain adaption framework for aerial image semantic segmentation based on curriculum learning," in *Proc. 7th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2022, pp. 354–359.

[70] L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9914–9925.

[71] L. Hoyer, D. Dai, and L. Van Gool, "HRDA: Context-aware high-resolution domain-adaptive semantic segmentation," in *Proc. ECCV*. Cham, Switzerland: Springer, 2022, pp. 372–391.

[72] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Mar. 2008.

[73] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[74] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

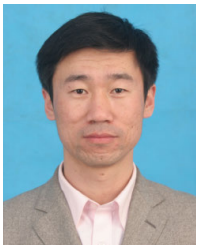
Yansheng Li (Senior Member, IEEE) received the B.S. degree in information and computing science from Shandong University, Weihai, China, in 2010, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015.

From 2017 to 2018, he was a Visiting Assistant Professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He is currently a Full Professor with the School of Remote Sensing and Information Engineering, Wuhan University (WHU), Wuhan. He has authored more than 100 peer-reviewed journal articles and conference papers. His research interests include knowledge graph, deep learning, and their applications in remote sensing big data mining.

Dr. Li was awarded the Young Surveying and Mapping Science and Technology Innovation Talent Award of the Chinese Society for Geodesy, Photogrammetry and Cartography in 2022. He received the recognition of the Best Reviewers of the IEEE TGRS in 2021 and the Best Reviewers of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2022. He is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), a Junior Editorial Member of *The Photogrammetric Record*, and a Lead Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and *Remote Sensing*.

Te Shi received the B.Eng. degree in software engineering from Anhui University, Hefei, China, in 2020, and the master's degree in pattern recognition and intelligent system from Wuhan University, Wuhan, China, in 2023.

His research interests include remote sensing image semantic segmentation, domain adaptation, and generative adversarial networks (GANs).



Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

From 2014 to 2015, he was a Senior Visiting Fellow with the Department of Geomatics Engineering, University of Calgary, Calgary, AB, Canada. From 2015 to 2018, he was a Senior Scientist at Environmental Systems Research Institute, Inc. (Esri), Redlands, CA, USA. He is currently the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. He holds 23 Chinese patents and 26 copyright-registered computer software. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence-driven remote sensing image interpretation, integration of light detection and ranging (LiDAR) point clouds and images, and 3-D city reconstruction.

Dr. Zhang was a Key Member of ISPRS Workgroup II/I from 2016 to 2020. He is the PI Winner of the Second-Class National Science and Technology Progress Award in 2017 and the PI Winner of the Outstanding-Class Science and Technology Progress Award in Surveying and Mapping (Chinese Society of Surveying, Mapping and Geoinformation, China) in 2015. In recent years, he has also served as the session chair for above 20 international workshops or conferences. He has been frequently serving as a referee for over 20 international journals. He is the Co-Editor-in-Chief of *The Photogrammetric Record*.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively.

He is currently a Professor with the School of Electronic Information, Wuhan University, Wuhan. He has authored or coauthored more than 300 refereed journal articles and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), *International Journal of Computer Vision* (IJCV), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), and European Conference on Computer Vision (ECCV). His research interests include computer vision, machine learning, and robotics.

Dr. Ma has been identified in the 2019–2022 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion*, an Associate Editor of IEEE/CAA JOURNAL OF AUTOMATICA SINICA, *Neurocomputing*, *Geo-Spatial Information Science*, and *Frontiers in Neuroscience*.