Pseudo Features-Guided Self-Training for Domain Adaptive Semantic Segmentation of Satellite Images

Fahong Zhang, Yilei Shi, *Member, IEEE*, Zhitong Xiong, *Member, IEEE*, Wei Huang, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Semantic segmentation is a fundamental and crucial task that is of great importance to real-world satellite image-based applications. Yet a widely acknowledged issue that occurs when applying the semantic segmentation models to unseen scenery is that the model will perform much poorer than when it was applied to scenery similar to the training data. This phenomenon is usually termed as the domain shift problem. To tackle it, this article presents a self-training-based unsupervised domain adaptation (UDA) method. Different from the previous self-training approaches which focus on rectifying and improving the quality of the pseudo labels, we instead seek to exploit feature-level relation among neighboring pixels to structure and regularize the prediction of the adapted model. Based on the assumption that spatial topological relation is maintained despite the impact of the domain shift, we propose a novel self-training mechanism to perform DA by exploiting local relation in the feature space spanned by the teacher model, from which the pseudo labels are generated. Quantitative experiments on four different public benchmarks demonstrate that the proposed method can outperform the other UDA methods. Besides, analytical experiments also intuitively verify the proposed assumption. Codes will be publicly available at https://github.com/zhu-xlab/PFST.**

***Index Terms*— Self-training, semantic segmentation, transfer learning, unsupervised domain adaptation (UDA).**

I. Introduction

With the purpose of automatically classifying and segmenting different semantic targets in a single image at the pixel level, semantic segmentation [1], [2], [3], [4] has been serving as an important technique in satellite image processing-based applications such as urban planning [5], land use, and land cover mapping [6], automatic agriculture [7]. With the renaissance of deep learning, the performance of data-driven semantic segmentation algorithms has been pushed to a new era. However, such a performance boost largely relies on the emergence of large-scale manually annotated data.

This leads to a problem, that is when applying the semantic segmentation model in unseen scenario without sufficient labels, its performance may drop drastically compared to its performance on the source domain. In the field of remote sensing, since the satellite image data are highly diversified, biases and shifts widely exist between the source and the target domain (where we train and evaluate our model, respectively). Such shifts may result from the differences in the used sensors, different atmospheric conditions, seasonal changes, distributional biases of the ground objects, and so on. To tackle this issue, domain adaptation (DA) techniques [8], [9], [10] has been attracting more and more attention.

DA leverages the source and the target domain data at the same time to bridge the shifts between them. Unsupervised DA (UDA) is a common and practical DA setting where only the target data are available, without any target labels provided. One popular technique for UDA is self-training, which has consistently achieved state-of-the-art results [11], [12], [13]. The fundamental idea behind self-training is to generate pseudo labels for the target domain data using a source-trained model, and then fine-tune the UDA model using selected high-confidence pseudo labels. Many of these methods focus on evaluating the quality of pseudo labels and developing selection strategies to filter out noisy pseudo labels [14], [15].

However, previous self-training methods often overlook the potential benefits of utilizing feature-level knowledge from the teacher model, which is used to generate pseudo labels. Pseudo labels are susceptible to noise [illustrated in Fig. 1 (left)], possibly due to domain shifts that bias the distribution of target objects in the low-dimensional output space. This realization raises the question: can the higher-dimensional pseudo features generated by the teacher model (referred to as pseudo features) be more robust to domain shifts compared to the pseudo labels that lie in the low-dimensional output space?

Manuscript received 26 January 2023; revised 17 April 2023; accepted 1 May 2023. Date of publication 31 May 2023; date of current version 20 June 2023. This work was supported in part by the German Research Foundation (DFG GZ: ZH 498/18-1) under Project 519016653, in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Program under Grant ERC-2016-StG-714087 (Acronym: *So2Sat*), in part by the Helmholtz Association through the Framework of the Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100, in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001, and in part by the German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding author: Xiao Xiang Zhu.)*

Fahong Zhang, Zhitong Xiong, Wei Huang, and Xiao Xiang Zhu are with the Chair of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: fahong.zhang@tum.de; zhitong.xiong@tum.de; w2wei.huang@tum.de; xiaoxiang.zhu@tum.de).

Yilei Shi is with the Chair of Remote Sensing Technology, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: yilei.shi@tum.de).

Digital Object Identifier 10.1109/TGRS.2023.3281503

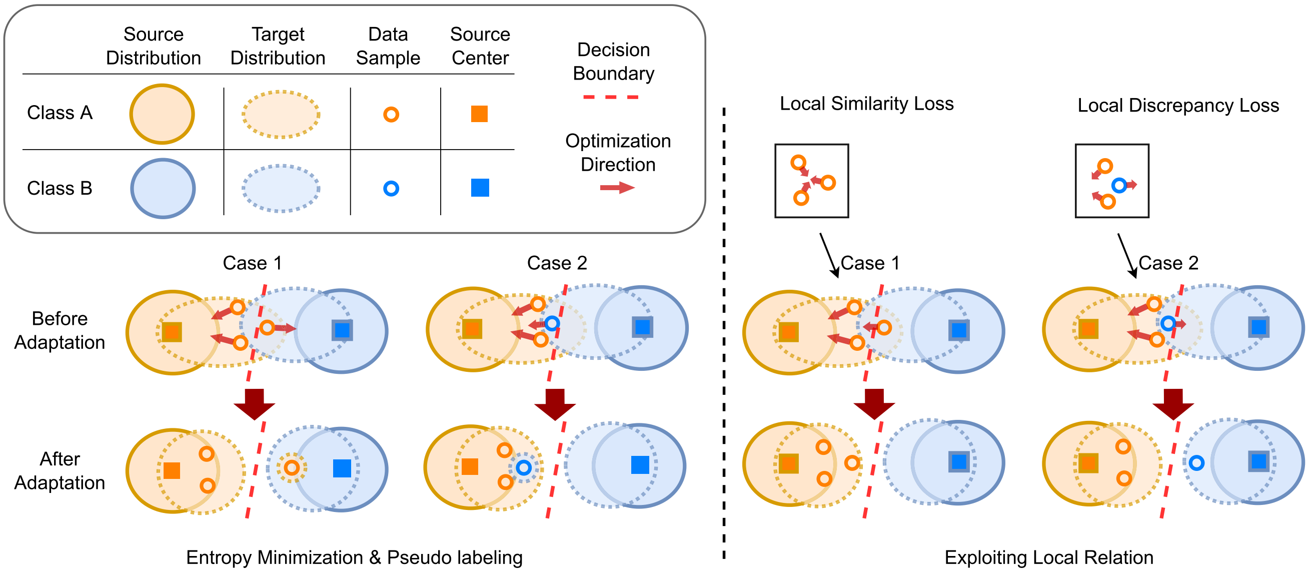


Fig. 1. Illustration of our motivation. Traditional pseudo-labeling or entropy minimization methods heavily rely on the correctness of initial predictions. We propose to model the relation between neighboring samples to counteract the negative effect of wrong initial predictions. Detailed discussion will be presented in Section III-A.

Intuitively, the answer seems to be positive. Based on such an assumption, we propose a pseudo features guided self-training (PFST) method that leverages feature-level relations in addition to pseudo labels. We assume that while domain shifts can affect the consistency and accuracy of target predictions in the output space, the spatial topological relation is relatively preserved in the high-dimensional feature space. Building on this insight, we propose to regularize target outputs using feature-level local relations between the pseudo features.

Specifically, we measure the similarity between neighboring feature pairs generated by the teacher model. For the most similar pairs within each local region, we strengthen the correlation of their corresponding target probability outputs. Conversely, for the most dissimilar pairs, we reduce their output-level correlation. By exploiting feature-level relations as a more robust and domain shift-insensitive source of information, we aim to improve the traditional self-training process. Our contributions can be summarized as follows.

1) We propose the assumption that the high dimensional local similarity structure in the pseudo feature space is less sensitive to domain shift than output space pseudo labels when applying the source-trained model on target data. Besides, we experimentally verify its correctness. Such an assumption is insightful to the development of further self-training-based UDA methods.
2) We develop a novel self-training approach named PFST. By exploiting feature-level local relation, PFST establishes the connection between the teacher model feature space and the student model output space, which further helps to counteract the negative effect caused by the noisy pseudo labels and improve the generalizability of the UDA model.
3) We establish and release a standard code library for UDA in remote sensing based on MMSegmentation [16] and EarthNets [17] framework. In this library, the data loading, augmentation and other factors related to the network training are standardized, which enables a fair comparison of different UDA methods. Compared to the results reported in the previous literature, our implemented baselines and other UDA methods outperform them by a large margin even under the same network architecture.
4) We conduct extensive comparative experiments on four different UDA settings, where different types of domain shift have been considered. The proposed PFST achieve the best and the most robust performance on all the settings, which verifies the proposed method is practical in real-world application.

II. Related Works

In this section, we will review some of the key techniques, strategies, or approaches that promote the development of UDA, both in computer vision society and in remote sensing society.

A. Adversarial Learning

Adversarial learning applies generative adversarial networks (GANs) [18] to perform adaptation between two or more domains. The philosophy is to train a discriminator together with the generator. While the discriminator is trained to be able to distinguish whether the generator's outputs come from the source or the target domain, the generator will be trained to confuse the discriminator to do so. In such a manner, the outputs from the generator conditioned on different domains will be undistinguishable and consistent, and the domain shifts will be reduced. According to the scale or the level of the networks where adversarial learning is applied, this line of work can be further categorized into image-level, feature-level, or output-level approaches.

1) Image-Level Adversarial Learning: In the earlier stage of the research toward adversarial learning, it is more widely used in the field of image generation [19] or image style transfer [20] instead of DA. Typical works like CycleGAN [21] and StarGAN [22], [23] apply GAN to transfer the image style, e.g., transfer oil painting to photos, images of zebra to horse, human faces of different gender and hair styles. Conditional GAN [24] extends these applications to perform transfer between images and their semantic annotations.

In the field of remote sensing, image-level adaptation or image-to-image translation are widely utilized to standardize the style of images from different domains. Compared to natural scene images, remote sensing data are usually multimodal [e.g., RGB, multispectral, hyperspectral, and synthetic aperture radar (SAR) data], multisensory, multitemporal, or geolocationally diversified. As a result, the difference in image appearances has a large impact on the generalizability of the downstream model. Previous works have demonstrated the usefulness of applying image-to-image techniques in remote sensing data. For example, Bidirectional Domain Adaptation Network (BiFDANet) [25] applies a CycleGAN architecture to perform image-to-image translation between the source and the target images, after which a semantic consistency loss is applied to the outputs of the original and the stylized images. StandardGAN [26], DAug [27] utilize StarGAN-like architectures and adaptive instance normalization (AdaIn) [28] to transfer the style of images captured from multiple cities, and enables multisource multitarget DA.

2) Feature- and Output-Level Adversarial Learning: Apart from the image-level domain shift, there are always biases between the source and the target domains that cannot be transferred solely by image-level style transfer. For example, difference in the spatial geometry or the unbalance distribution of source and target semantic objects. To mitigate those latent high-level domainwise biases, aligning the source and the target domain in feature- or output-level becomes necessary. AdaptSeg [29] highlights the importance of adopting adversarial learning in the output space and uses GAN to align the source and the target output. AdvEnt [30] discovers the effectiveness of entropy minimization in the target domain, and further proposes to align the source and the target entropy map in an adversarial manner.

In the field of remote sensing, full space domain adaptation network (FSDAN) [31] uses a CycleGAN structure to generate target-style source images to mitigate the domain shift problem. After that, they also apply feature-level and output-level adversarial learning to further improve the adaptation performance, which leads to a full-space alignment between the source and the target domain. Entropy-guided adversarial domain adaptation (EGA) [32] proposes an entropy-guided adversarial learning algorithm. While adversarial learning is conducted on the output level, a self-adaptive weight is calculated to reweight the prediction from the discriminator. Triplet adversarial domain adaptation (TriADA) [33] designs an output-level adversarial learning method based on the triplet loss, where an image triplet from the source and the target domain is input to the semantic segmentation network during training. Unlike the previous method, the discriminator is devised as a similarity metric to measure the domain-level similarity between two input images.

B. Self-Training

Adversarial learning-based methods are often characterized by their instability and difficulties in optimization. However, self-training [34], [35], [36], [37], which involves fine-tuning the UDA model using pseudo labels generated from target data, offers a more efficient way to leverage target information and is typically easier to optimize. In the field of computer vision, an important focus of self-training-based methods is to effectively filter out noisy pseudo labels. For example, CBST [14] points out that the training with pseudo labels suffers the risk of being overwhelmed by easy-to-transfer classes, and proposes to balance the distribution of pseudo labels by applying a classwise confidence threshold. To prevent the self-trained network from being over-confident during the learning toward hard pseudo labels, confidence regularized self-training (CRST) [15] argues to regularize the self-training process by using soft labels. In uncertainty reduction for model adaptation (URMA) [38], the prediction uncertainty is estimated via the variance of different network outputs, which is later used to automatically weigh the pseudo labels. Prototypical domain adaptation (ProDA) [39] maintains a set of prototypes for each class during training, and the relative distance between features and prototypes is used to rectify the false pseudo labels.

In remote sensing, it seems self-training receives less attention. Wang et al. [40] establish a benchmark for evaluating different domain adaptive semantic segmentation methods, where they study the effect of some classic DA methods proposed in the computer vision society. Zhang et al. [41] integrate the adversarial learning mechanism into the self-training pipeline to perform UDA in the task of road extraction. In remote sensing scene classification, there are also works like [42], which studies the influence of different strong augmentation applied to the student model branch in the self-training pipeline.

C. Data Augmentation and Other Techniques

Since self-supervised learning has achieved remarkable progress [43], the importance of data augmentation has been widely acknowledged. Among all different data augmentation methods, data mixing [44] has been demonstrated to be effective both in classification [44] and semantic segmentation [45]. Domain adaptation via cross domain mixed sampling (DACS) [45] studies the impact of data mixing in the field of UDA, where they use the ClassMix [46] strategy to cutout half of the classes in a single source image, and overlay the target image on top of the cut area.

In remote sensing, there are also works like randomized histogram matching (RHM) [47] pointed out that simple colorwise data augmentation strategy like RHM can produce comparable semantic segmentation results than complex image-to-image translation-based methods.

Other DA works focus on devising a better sampling strategy to reduce the influence of the domain shift. DAFormer [12] proposes a rare class sampling strategy to balance the distribution of different semantic classes. The oversampling of the rare classes tends to mitigate the long tail issue [48] and improve the generalizability of the semantic segmentation model. Curriculum-style local-to-global cross-domain adaptation (CCDA) [49] proposes a curriculum-style UDA method that rank the target patches from easy to hard according to the output entropy. Those patches are then fed to the network

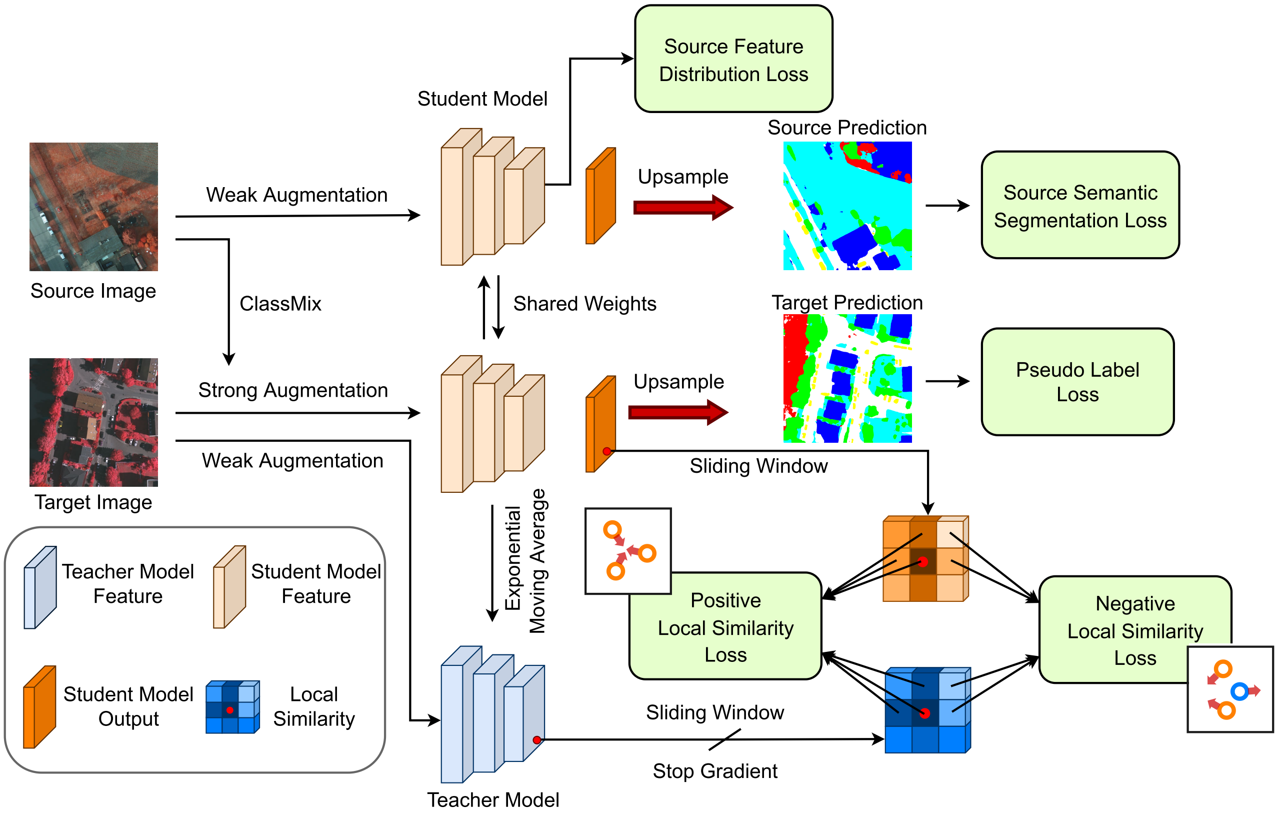


Fig. 2. Illustration of our proposed PFST. A teacher model and a student model will be maintained. The teacher model generates pseudo labels based on the target image to supervise the student model. The student model is trained on both the source labels and the pseudo labels. Its weights will be used to update the teacher model by exponential moving average [50]. The source and the target images are augmented via weak or strong data augmentation before input to the models (please refer to Section IV-B for more details). We calculate sliding windows over the teacher model features and the student model outputs simultaneously. For each corresponding window pair, we apply a local similarity loss on output probabilities according to their local feature similarity. In such a way, we incorporate a new regularization mechanism that connect the teacher model feature space and student model output space.

from easy to hard. This curriculum-based sampling strategy is reported to be effective.

III. Methods

The overall architecture of our approach is illustrated in Fig. 2. In Section III-A, we will first introduce and illustrate the DA problem and describe the motivation of the proposed method. Then in Section III-B, we formulate the UDA setting for semantic segmentation. Later on, we present the optimization objectives and other loss functions in Sections III-C–III-E, respectively.

A. Motivation

Fig. 1 provides an illustration of the UDA problem, highlighting two cases where conventional pseudo-labeling-based methods may fail. In the original source feature distributions, two different classes are easily separable based on the decision boundary. However, after applying the source model on the target domain, there could be overlaps between the target distributions of the two different classes, leading to incorrect predictions. In such cases, traditional pseudo-labeling approaches may push the target features toward the possibly incorrect predictions of the source model, thereby exacerbating the prediction errors. This can result in the adapted target distributions becoming indistinguishable, further hampering the performance of UDA methods.

In order to address the issue of optimization direction when such errors occur, we propose the utilization of local similarity and local discrepancy loss. Specifically, in case 1 where a sample belonging to class A is misclassified as class B, if we have correctly classified class A samples in its local neighborhood with high similarity, maximizing their output-level correlation can help redirect the optimization toward the expected direction. Similarly, in case 2 where a sample from class B is misclassified as class A, if it exhibits larger discrepancy with nearby class A samples in the source feature space, minimizing the output correlation between them can also aid the optimization process.

B. Problem Formulation

Let $D_s = \{\mathbf{x}_i^s\}_{i=0}^{N_s}$ and $D_t = \{\mathbf{x}_i^t\}_{i=0}^{N_t}$ be the source and the target domain data, and $\{\mathbf{y}_i^s\}_{i=0}^{N_s}$ and $\{\mathbf{y}_i^t\}_{i=0}^{N_t}$ be the corresponding labels. Here $\mathbf{x}_i^s, \mathbf{x}_i^t \in \mathbb{R}^{H\times W\times 3}$ denote the source and the target images, while $\mathbf{y}_i^s, \mathbf{y}_i^t \in \mathbb{R}^{H\times W}$ indicate their labels. H and W specify the height and width of the images. Note that the target domain labels $\{\mathbf{y}_i^t\}_{i=0}^{N_t}$ are only available during the evaluation time. N_s and N_t indicate the sizes of the source and the target datasets. With these notations defined, the UDA problem for semantic segmentation can be formulated as

$$\min_{\theta_{\mathcal{S}}} \sum_{\mathbf{x}^s\in D_s, \mathbf{x}^t\in D_t} \mathcal{L}(\mathbf{x}^s, \mathbf{x}^t, \mathbf{y}^s; \theta_{\mathcal{S}}, \theta_{\mathcal{T}}) \quad (1)$$

where $\mathcal{L}(\cdot)$ is the loss function, $\theta_{\mathcal{S}}$ and $\theta_{\mathcal{T}}$ are the parameters of the student and the teacher models, respectively. In self-training-based UDA methods, the teacher model $\mathcal{T}$ is usually used to generate pseudo labels to supervise the student model $\mathcal{S}$. $\mathcal{T}$ can be either pretrained on the source domain data in an offline manner [34] or updated according to the student model weights $\theta_{\mathcal{S}}$ via exponential moving average [12]. In our method, adopt the latter strategy. More specifically, weights of the teacher network $\theta_{\mathcal{T}}$ will be updated by

$$\theta_{\mathcal{T}}^{(t)} = \alpha\theta_{\mathcal{T}}^{(t-1)} + (1-\alpha)\theta_{\mathcal{S}}^{(t)} \tag{2}$$

where t denotes the current iteration step. The decay weight α is set to 0.999 following [12].

Besides, we denote h as the feature extractor, g as a classifier of the network model, and $f = h \circ g$ as their composition. Target feature and output probability are denoted by $h(\mathbf{x}^t) \in \mathbb{R}^{h\times w\times d}$ and $f(\mathbf{x}^t) \in \mathbb{R}^{h\times w\times c}$, where $h \times w$ corresponds to spatial dimension of extracted feature map. d indicates the feature dimension and c is the number of classes.

C. Objective Function

We define our optimization objective as

$$\mathcal{L}(\mathbf{x}^s, \mathbf{x}^t; \theta_{\mathcal{S}}, \theta_{\mathcal{T}}) = \mathcal{L}_{\text{src}} + \mathcal{L}_{\text{pse}} + \alpha\mathcal{L}_{\text{loc}} + \beta\mathcal{L}_{\text{feat}} \tag{3}$$

where $\mathcal{L}_{\text{src}}$ is the source domain semantic segmentation loss, defined as

$$\mathcal{L}_{\text{src}}(\mathbf{x}^s) = -\frac{1}{HW}\sum_{i=0}^{H\times W} \mathbb{1}_{\mathbf{y}_i^s}^T \log(f_{\theta_{\mathcal{S}}}(\mathbf{x}^s)_i). \tag{4}$$

Here $\mathbb{1}_{\mathbf{y}_i^s}$ is the one-hot encoding of the source label. $\mathcal{L}_{\text{pse}}$ is the pseudo label loss widely used for self-training approaches in the field of DA. We adopt a weighted pseudo label loss used in previous works [12], [45]

$$\mathcal{L}_{\text{pse}}(\mathbf{x}^t) = -\frac{q(\mathbf{x}^t)}{HW}\sum_{i=0}^{H\times W} \mathbb{1}_{\tilde{\mathbf{y}}_i^t}^T \log(f_{\theta_{\mathcal{S}}}(\mathbf{x}^t)_i). \tag{5}$$

Here $\mathbb{1}_{\tilde{\mathbf{y}}^t}$ is the one-hot encoding of the teacher prediction $\tilde{\mathbf{y}}^t$, where $\tilde{\mathbf{y}}^t = f_{\theta_{\mathcal{T}}}(\mathbf{x}^t)$ corresponds to the pseudo label generated by the teacher model on the target image. $q(\mathbf{x}^t)$ is a weighting factor that balances the loss based on the predicted confidence on each target image

$$q(\mathbf{x}^t) = \frac{1}{HW}\sum_{i=1}^{H\times W} [\max_c f_{\theta_{\mathcal{T}}}(\mathbf{x}^t)_{i,c} > \tau]. \tag{6}$$

It counts the number of pixels where the classwise maximum output probability is larger than a certain threshold τ. τ is fixed to 0.98 empirically in all our experiments.

D. Local Similarity Loss

Local similarity loss $\mathcal{L}_{\text{loc}}$ is used to supervise the student model by exploiting feature-level similarity implied in the teacher model

$$\mathcal{L}_{\text{loc}}(\mathbf{x}^t) = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}. \tag{7}$$

Specifically, it strengthens the correlation between two neighboring target outputs that share a strong similarity in the feature space defined by the teacher model, with a positive loss term $\mathcal{L}_{\text{pos}}$. Meanwhile, it increases the discrepancy between target outputs that have weak similarity in the feature space using a negative term $\mathcal{L}_{\text{neg}}$. The positive term is defined as

$$\mathcal{L}_{\text{pos}}(\mathbf{x}^t) = -\frac{1}{HW|\Omega|}\sum_{i=1}^{H\times W}\sum_{j\in\Omega_i^+\cup\{i\}} A_{\theta_{\mathcal{T}}}\left(\mathbf{x}_i^t, \mathbf{x}_j^t\right) I_{i,j}^+. \tag{8}$$

Here A_θ is a feature space similarity measurement for a pair of features extracted by a deep model θ. For simplicity, we adopt the cosine similarity

$$A_\theta(\mathbf{x}_i, \mathbf{x}_j) = \frac{h_\theta(\mathbf{x})_i \cdot h_\theta(\mathbf{x})_j}{\|h_\theta(\mathbf{x})_i\| \cdot \left\|h_\theta(\mathbf{x})_j\right\|}. \tag{9}$$

Ω_i defines a sliding window centered at position i (i itself excluded). Then Ω_i^+ contains the top ξ locations within Ω_i that yield the highest $A_\theta(\mathbf{x}_i^t, \mathbf{x}_j^t)$ value. ξ is set to 3 in all of our experiments. $I_{i,j}^+$ is a measurement in target output space evaluating the probability that two nearby located pixels produce the same prediction

$$I_{i,j}^+ = \sum_{k=1}^{c}\left(\mathbf{p}_i\mathbf{p}_j^T\right)_{k,k} \tag{10}$$

where $\mathbf{p} = f_{\theta_{\mathcal{T}}}(\mathbf{x}^t)$ is the target model's output. $\mathbf{p}_i\mathbf{p}_j^T \in \mathbb{R}^{c\times c}$ measures the joint probability distribution of $\mathbf{p}_i$ and $\mathbf{p}_j$ regardless of their dependence.

The negative loss term $\mathcal{L}_{\text{neg}}$ is defined as

$$\mathcal{L}_{\text{neg}}(\mathbf{x}^t) = -\frac{1}{HW|\Omega|}\sum_{i=1}^{H\times W}\sum_{j\in\Omega_i^-}\left(1 - A_{\theta_{\mathcal{T}}}\left(\mathbf{x}_i^t, \mathbf{x}_j^t\right)\right) I_{i,j}^-. \tag{11}$$

In contrast to Ω_i^+ in (8), here Ω_i^- defines the top ξ locations that have the lowest $A_{\theta_{\mathcal{T}}}$ value within a sliding window. Different from $I_{i,j}^+$, $I_{i,j}^-$ measures the probability of cases where $\mathbf{p}_i$ and $\mathbf{p}_j$ indicate different classes

$$I_{i,j}^- = \sum_{k=1}^{c}\sum_{l\neq k}^{c}\left(\mathbf{p}_i\mathbf{p}_j^T\right)_{k,l}. \tag{12}$$

Note that there is $I_{i,j}^+ + I_{i,j}^- = 1$ since $\mathbf{p}$ is probabilistic.

By imposing $\mathcal{L}_{\text{pos}}$ only to Ω_i^+ and $\mathcal{L}_{\text{neg}}$ to Ω_i^-, a relative local relation is considered in addition to the absolute one incorporated by $A_{\theta_{\mathcal{T}}}(\mathbf{x}_i^t, \mathbf{x}_j^t)$. The intuition behind this is that feature pairs of the same class are more likely to lie in Ω_i^+, while feature pairs of different class mostly lie in Ω_i^-.

E. Source Feature Distribution Loss

When applying the local similarity loss $\mathcal{L}_{\text{loc}}$ on the target domain, we hope the feature similarity between pixel pairs that share the same label is large, while the similarity between pixel pairs with different labels is small. To achieve this, we introduce a feature distribution loss on the source domain, aimed at increasing the separability of the two similarity distributions. In this context, let's consider a scenario where the similarity values between positive and negative feature pairs follow two unknown distributions denoted as $\mathcal{P}_{\theta_{\mathcal{S}}}$ and $\mathcal{N}_{\theta_{\mathcal{S}}}$, respectively. These distributions can be characterized by

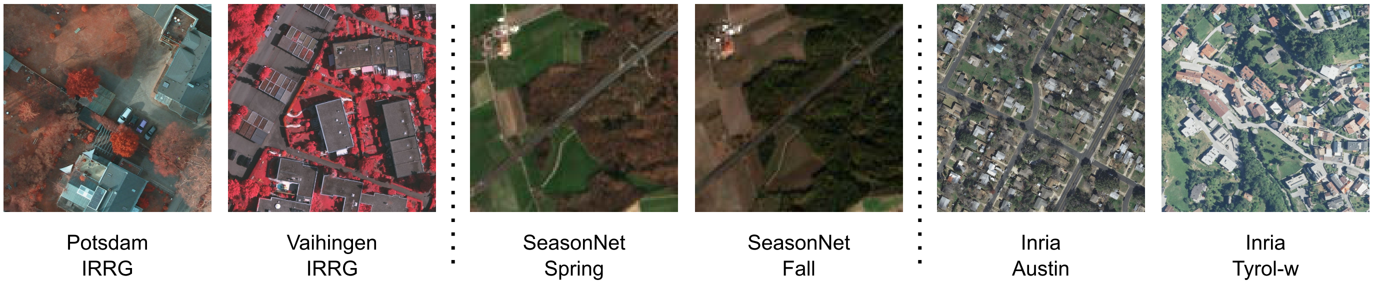


Fig. 3. Illustration of the datasets and the separated domains that are used to construct our UDA settings. For Inria datasets, we only chose one city from each of the source and the target domain.

their means, denoted as μ_{pos} and μ_{neg}, respectively, and their standard deviations, denoted as σ_{pos} and σ_{neg}.

$\forall i, \forall j \in \Omega_i$, there is

$$A_{\theta_S}\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) \sim \begin{cases} \mathcal{P}_{\theta_S}(\mu_{\text{pos}}, \sigma_{\text{pos}}), & \text{if } y_i^s = y_j^s \\ \mathcal{N}_{\theta_S}(\mu_{\text{neg}}, \sigma_{\text{neg}}), & \text{otherwise.} \end{cases} \tag{13}$$

We hope the two distributions can be as distinguishable as possible. To this end, we apply a feature distribution loss $\mathcal{L}_{\text{feat}}$ on the source domain using the labeled source data

$$\mathcal{L}_{\text{feat}}(\mathbf{x}^s) = -\tilde{\mu}_{\text{pos}} + \tilde{\mu}_{\text{neg}} + \tilde{\sigma}_{\text{pos}} + \tilde{\sigma}_{\text{neg}} \tag{14}$$

where $\tilde{\mu}_{\text{pos}}$ and $\tilde{\sigma}_{\text{pos}}$ are the mean and standard deviation of the similarity between all the positive pixel pairs within all the sliding window. They are used to approximate μ_{pos} and σ_{pos}

$$\begin{aligned} \tilde{\mu}_{\text{pos}} &= \frac{1}{HW|\Omega|} \sum_{i=1}^{H\times W} \sum_{j\in\Omega_i, y_i^s = y_j^s} A_{\theta_S}\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) \\ \tilde{\sigma}_{\text{pos}}^2 &= \frac{1}{HW|\Omega|} \sum_{i=1}^{H\times W} \sum_{j\in\Omega_i, y_i^s = y_j^s} \left(A_{\theta_S}\left(\mathbf{x}_i^s, \mathbf{x}_j^s\right) - \tilde{\mu}_{\text{pos}}\right)^2. \end{aligned} \tag{15}$$

Likewise, $\tilde{\mu}_{\text{neg}}$ and $\tilde{\sigma}_{\text{neg}}$ are the mean and standard deviation of the negative pixel pairs (given $j \in \Omega_i$, $y_i^s \neq y_j^s$).

By minimizing the value of $-\tilde{\mu}_{\text{pos}} + \tilde{\mu}_{\text{neg}}$, the difference between the means of the two distributions is maximized, leading to increased separation between them. Additionally, by minimizing $\tilde{\sigma}_{\text{pos}}$ and $\tilde{\sigma}_{\text{neg}}$, the standard deviations of the two distributions are reduced, further enhancing the distinction between them. As a result, the two distributions become more effectively separated from each other in the end.

IV. Experiments

A. Datasets and UDA Settings

We use four public datasets, including ISPRS Potsdam,[1] Vaihingen,[2] SeasonNet [51] and Inria [52] to evaluate the performance of different UDA methods. Some sample images of different UDA settings are shown in Fig. 3.

Potsdam and Vaihingen datasets consist of aerial images captured over the Potsdam and Vaihingen cities in Germany. Potsdam dataset contains 38 images with a size of 6000 × 6000 and a ground sampling distance (GSD) of 5 cm. Potsdam offers both RGB and near-infrared, red, and green (IRRG) images, yet in our experiments, we only use the IRRG ones. Vaihingen dataset has 33 images with a size of 2000 × 2000, and a GSD of 9 cm. Three IRRG channels are given. Both Potsdam and Vaihingen have six classes.

SeasonNet is a large-scale land cover and land use dataset captured over the whole Germany. It contains in total of 1 759 830 image patches from Sentinel-2 sensor, with a patch size of 120 × 120, annotated to 33 land cover classes. All those patches are categorized according to the season when they are captured. In total, there are four seasons plus an additional "Snow" domain where most of the land cover is covered by snow. This makes it a realistic and ideal setting for evaluating different UDA methods against the temporal and seasonal domain shift.

Inria dataset is an aerial image labeling dataset created for building footprint extraction [52]. It has a resolution of 0.3 m and a coverage of 810 km^2 captured over ten European and American urban settlements. Pixel-level annotations of two classes, including building and background are provided in the training set. 36 images with sizes of 5000 × 5000 are given for each city.

Based on the above public datasets, we organize four different UDA setting to evaluate the performance of our method. The detailed descriptions are given below.

1) ISPRS Potsdam IRRG to Vaihingen IRRG: In this setting, we consider Potsdam dataset with IRRG images as the source domain and Vaigingen dataset as the target domain. We split the images from both datasets to a patch size of 1024 × 1024. As the dataset provider gives official training and testing splits of these two datasets, we adopt the setting that we train the UDA model on labeled training split of the Potsdam IRRG dataset, as well as the training split of the Vaihingen dataset (without giving the label), validate the models on the Vaihingen train split and report the results on the Vaihingen test split.

2) ISPRS Vaihingen IRRG to Potsdam IRRG: This setting is similar to the previous setting, except that we switch the source and the target domain. As the result, we training the UDA model on the training split of Vaihingen, as well the training split of Potsdam IRRG (without providing the label), validate the models on Potsdam IRRG training split and report the results on Potsdam IRRG test split.

[1] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

[2] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

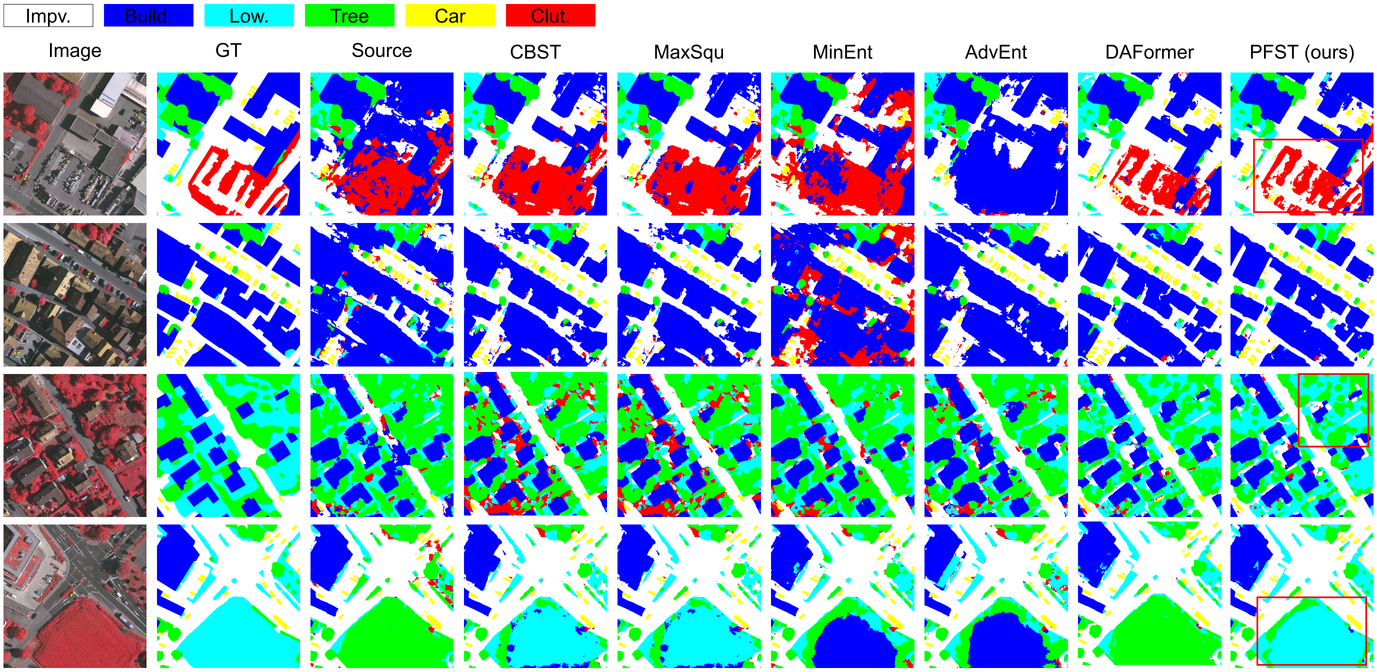


Fig. 4. Visualized semantic segmentation results of different UDA methods on ISPRS P2V setting.

3) SeasonNet Spring to Fall: To involve temporal and seasonal changes, we consider the spring season as the source domain and the fall season as the target domain. Note that in this case, we want to create a moderate domain shift so that it will not be neither too easy nor too hard for UDA method to effect. As the dataset provider gives official train, validation, and test splits, we train UDA methods on the train split of the spring season, and validation split of the fall season (without providing labels), validate them on the validation split of the fall season and report the results on the fall season test split.

4) Inria Intercity: By introducing this setting, we want to evaluate the generalizability of different UDA methods across different geo-locations, i.e., different cities. Since only the five cities in the training set of the Inria dataset are provided with labels, we only utilize these cities. As a result, Austin, Chicago, and Kitsap are considered as the source domain cities while Vienna and Tyrol-w are considered the target domain cities. We follow the suggestion from the dataset provider to use the first five images from each city as the validation set, while the others as the training set. To this end, we train the UDA methods on the training set of the source domain cities, validate on the target training set and report the results on the target validation set.

B. Implementation Details

We reimplemented several classic and state-of-the-art UDA methods for evaluation and comparison. These methods include class-balanced self-training (CBST) [14], MaxSqu [53], MinEnt [30], AdvEnt [30], and DAFormer [12]. All the implementations are under the same codebase from MMLab [16] and EarthNets [17], where the data loading pipelines, network architectures, optimizers and training flows are shared, making the comparison fairer. We use a classic network architecture for all the methods, where DeepLabV3+ [54] is used as the decoder and ResNet50 [55] is used as the encoder.

Regarding the data normalization and augmentation, for the SeasonNet spring to fall setting, the data are converted from 16 to 8 bits by cutting out the values beyond $\mu \pm \sigma$ for each channel, where μ and σ are the mean and the standard deviation of the whole datasets. The values within $\mu \pm \sigma$ are then normalized to [0, 1] and further converted to 8-bits data. For the other settings, data are normalized using the ImageNet statistics [56]. To perform the data augmentation, we include random resizing, cropping, random horizontal and vertical flipping, random rotation of 90°, 180°, or 270°, and random photometric distortion. For DAFormer and PFST which are based on online pseudo-label generation, these operations are considered as weak augmentation. ClassMix [46], color jittering, and random blurring are used as the strong augmentation. We observe that these data augmentations can largely improve the overall performances of different methods. About the hyperparameter settings of the proposed PFST, α and β in (3) are set as $\alpha = 0.1$, $\beta = 0.1$. The sliding window size is set to 3 with a dilation of 2. Such hyperparameter settings are applied for all the UDA settings.

For optimizing the networks, Adamw optimizer [57] with 0.01 weight decay and $6e-5$ learning rate is used to train all the approaches. The batch size is set to 2 for Potsdam IRRG to Vaihingen IRRG (P2V), Vaihingen IRRG to Potsdam IRRG (V2P), and Inria intercity settings, and set to 32 for SeasonNet spring to fall setting. The number of iterations is set to $40k$ for all the settings. Polynomial learning rate decay is applied for all the methods. All the experiments are conducted on a single NVIDIA RTX 3090 GPU with PyTorch library.

C. Quantitative Results

We list the quantitative comparison results of different UDA methods on four different settings on Tables I–IV. From

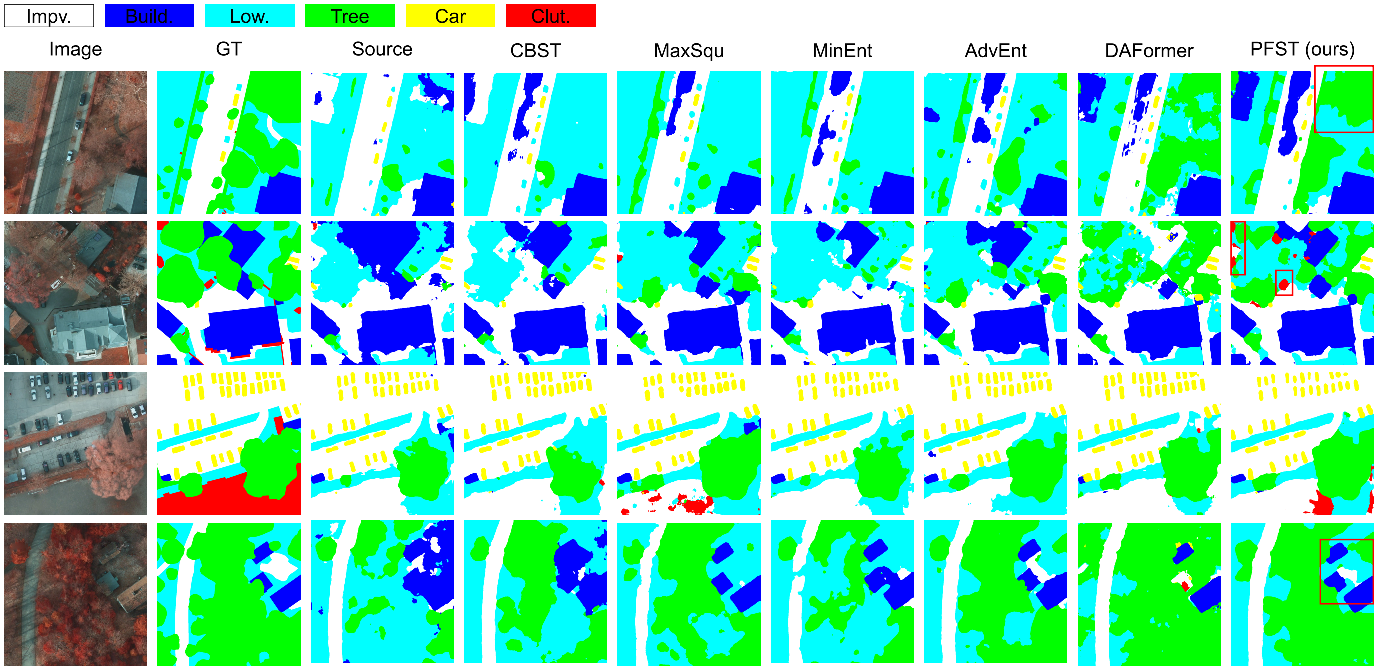


Fig. 5. Visualized semantic segmentation results of different UDA methods on ISPRS V2P setting.

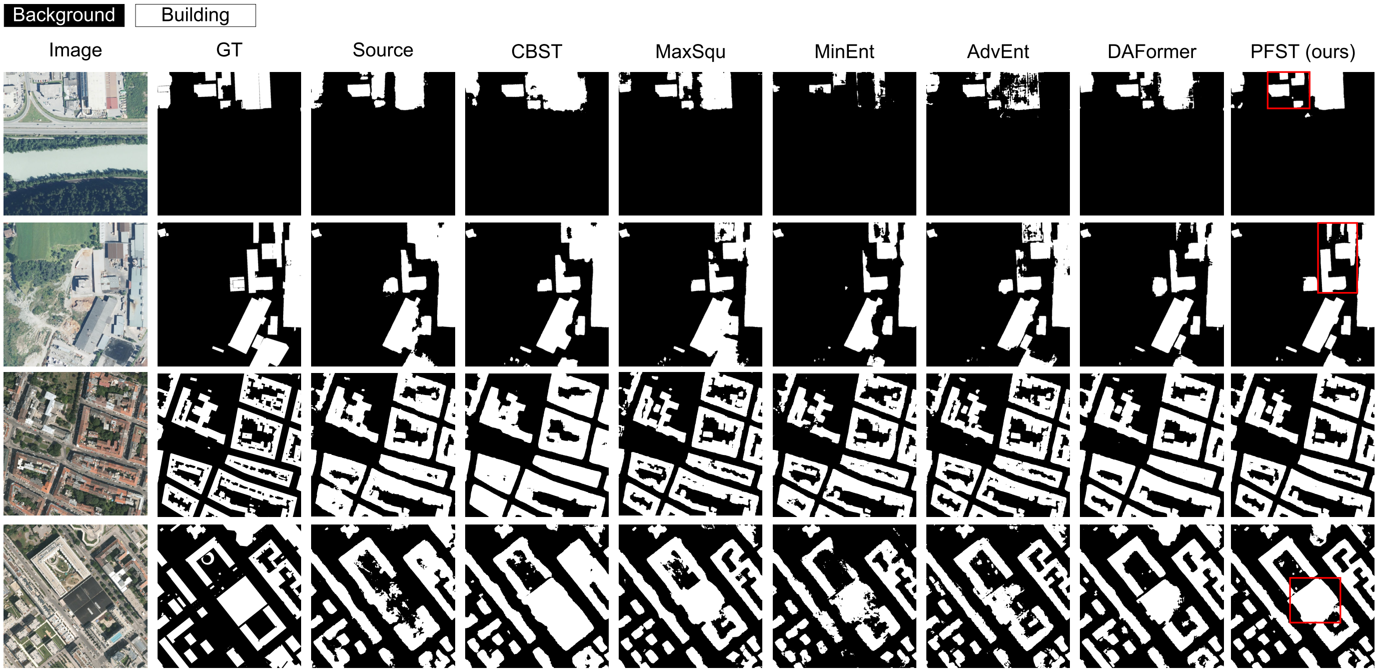


Fig. 6. Visualized semantic segmentation results of different UDA methods on Inria intercity setting.

Table I, one can observe that different UDA methods all make improvements over the baseline source trained model, especially on the foreground objects (categories except Clutter). Among all methods, DAFormer and the proposed PFST perform better on the "Clutter" classes, which may owe to the mix-up strategy that balance the distribution of rare classes. PFST performs the best at all categories except "Tree," demonstrating the effectiveness of mining the source domain feature-level similarity in distinguishing both foreground and background objects.

In the ISPRS V2P setting, shown in Table II, we notice that all UDA methods still improve over the baseline. One phenomenon to be noticed is that their performance variances on "Tree" class are larger. One possible reason is that in Potsdam-IRRG dataset, the "Tree" and the "Low Vegetation" classes are hard to be adapted, making them easily confused when making the prediction. The proposed PFST performs the best on these two classes, which could be explained by that the differences between these two types of objects can be better captured in the feature space.

TABLE I

Per Class IoU and mIoU (%) on ISPRS Potsdam-IRRG to Vaihingen IRRG Setting. "Source" Denotes the Source Model Which Is Only Trained With the Labeled Source Data. * Denotes the Results Are Cited From the Original Paper. All Methods Are Based on DeepLabV3+ [54] Architecture

Method	Impv.	Build.	Low.	Tree	Car	Clut.	mIoU
Source	58.20	73.51	35.81	59.83	48.76	16.61	48.79
CCDA* [49]	67.74	76.75	47.02	55.03	44.90	20.71	52.03
RDGAN* [58]	72.29	80.57	49.69	63.81	57.01	18.42	55.83
CSC* [59]	75.56	84.17	52.92	65.55	56.58	13.83	58.10
CBST [14]	61.45	84.36	42.36	63.28	53.32	6.87	51.94
MaxSqu [53]	72.60	82.14	43.00	64.77	54.59	7.99	54.18
MinEnt [30]	62.56	79.88	44.05	65.60	52.62	6.79	51.92
AdvEnt [30]	70.58	77.60	46.37	**65.95**	52.28	10.87	53.94
DAFormer [12]	77.18	86.45	52.44	65.06	60.25	29.92	61.88
PFST (ours)	**78.85**	**87.85**	**57.30**	62.99	**62.11**	**38.72**	**64.64**

TABLE II

Per Class IoU and mIoU (%) on ISPRS Vaihingen to Potsdam-IRRG Setting. "Source" Denotes the Source Model Which Is Only Trained With the Labeled Source Data. * Denotes the Results Are Cited From the Original Paper. All Methods Are Based on DeepLabV3+ [54] Architecture

Method	Impv.	Build.	Low.	Tree	Car	Clut.	mIoU
Source	65.87	70.42	50.86	19.46	72.78	3.34	47.12
CCDA* [49]	64.39	66.44	47.17	37.55	59.35	12.31	47.87
CBST [14]	69.19	79.85	53.96	21.47	74.54	3.76	50.46
MaxSqu [53]	71.14	79.86	52.81	34.45	72.30	6.73	52.88
MinEnt [30]	70.30	80.07	51.63	25.85	73.64	1.52	50.50
AdvEnt [30]	72.18	80.57	53.05	38.63	74.64	6.12	54.20
DAFormer [12]	**73.07**	**81.76**	53.96	47.70	66.56	2.26	54.22
PFST (ours)	71.77	81.59	**57.79**	**50.44**	66.84	**13.27**	**56.95**

TABLE III

Building IoU (%) on Inria Intercity Setting. "Source" Denotes the Source Model Which Is Only Trained With the Labeled Source Data. The Results on the Two Target Domain Cities and the Overall Results Are Reported

Method	Vienna	Tyrol-w	All
Source	73.26	69.02	72.40
CBST [14]	72.52	73.62	72.74
MaxSqu [53]	74.62	74.61	74.62
MinEnt [30]	71.34	63.30	69.69
AdvEnt [30]	73.93	73.34	73.80
DAFormer [12]	75.53	76.70	75.77
PFST (ours)	**75.54**	**77.15**	**75.87**

The results of the Inria intercity setting are shown in Table III. Compared to the previous setting, the improvements from different UDA methods are less obvious. Especially on Vienna city, methods include CBST, MinEnt, and AdvEnt cannot or can only slightly outperform the source model. Generally, the domain shift between Vienna and the source domain cities mainly lie on its larger and more complex building geometry, while the shift between Tyrol-w and the source domain cities are mainly on its color and appearance. This indicates the geometrywise differences between two domains are more difficult to be tackled. In this case, DAFormer and the proposed PFST can still provide stable improvements over both cities, which demonstrate their effectiveness.

The results on the SeasonNet spring to fall setting are shown in Table IV. This is a more challenging setting because seasonal changes usually have large impacts on some of the land cover types that related to agriculture, forests, natural landscape, and so on. From the per-class results and the averages results, one can tell that although DAFormer outperforms the others on some of the classes like C_4 and C_{29}, it fails drastically on classes like C_{30} and C_{33}. This may be because the utilized rare class sampling strategy [12] samples too many repeated image patches from rare classes, resulting in the underfitting of some of the major classes. In general, PFST performs very stable, and can make improvements against the baseline source model on almost all the classes, and achieves the highest mean IoU values. This further demonstrates its robustness.

D. Qualitative Results

We visualize the semantic segmentation results of different UDA methods in four different settings in Figs. 4–7. As can be observed in Fig. 4, only DAFormer and the proposed PFST can detect and segment the fine-grained clutter structure (in red color) in the first row. From the third and the fourth rows, all the other UDA methods except PFST perform not very well at distinguishing the differences between "Tree" and "Low Vegetation" categories, e.g., MinEnt, AdvEnt, and DAFormer confuse the large "Low Vegetation" area in the last row with the "Tree" class.

From Fig. 5, it is shown that the difference between "Low Vegetation" and "Tree" classes is still the main challenging issue. Among all the methods, DAFormer and PFST perform the best on this goal if we look into the first, second, and the fourth rows. Besides, PFST can also detect some tiny clutter objects in the second row, despite it's still hard to segment the large clutter area in the third row due to its similarity to the "Imprevious surface" class.

Fig. 6 shows the results on the Inria intercity setting. As highlighted in the red bounding boxes, PFST generally performs better at distinguishing the building structure that is easy to be confused with the background areas (like what is shown in the first and the fourth rows), and can also better captures the borders of separated building instances (highlighted in the second row).

The results on SeasonNet spring to fall setting are given in Fig. 7. Generally, all the methods can capture the overall land cover distribution in the image, yet they still tend to be confused when trying to distinguish between two similar classes. For example, in the first row, only CBST and PFST can distinguish the "Broad-leaved forest" (C_{16}) and the "Coniferous forest" (C_{17}) area highlighted with the red bounding box. In the second row, only PFST can recognize the "Fruit trees and berry plantations" (C_{14}) area inside the bounding box.

TABLE IV

aAcc, mAcc, and mIoU (%) on SeasonNet Spring to Fall Setting. "Source" Denotes the Source Model Which Is Only Trained With the Labeled Source Data. IoU Results on Some Selected Classes (Denoted by C_i), and the Averaged Results of All Classes Are Reported. For the Class Name of Each Class, Please Refer to [51]

Method	C_1	C_2	C_3	C_4	C_5	C_{12}	C_{13}	C_{15}	C_{17}	C_{20}	C_{21}	C_{27}	C_{29}	C_{30}	C_{32}	C_{33}	mIoU
Source	31.44	69.27	49.92	32.19	25.22	77.51	38.65	49.70	62.96	44.42	0.09	45.96	48.37	55.41	29.50	88.24	34.12
CBST [14]	31.13	69.08	50.71	36.05	24.40	76.95	**52.39**	52.01	66.21	**47.53**	**14.15**	49.38	55.15	41.56	33.42	75.28	35.07
MaxSqu [53]	**32.66**	70.07	50.36	33.58	26.34	77.85	40.57	51.19	63.83	44.41	12.03	43.01	48.76	52.32	28.88	86.68	34.81
MinEnt [30]	31.94	70.19	50.31	32.13	26.94	78.11	39.00	52.49	60.94	45.34	11.13	43.11	48.57	55.78	30.49	87.77	34.48
AdvEnt [30]	31.44	**70.62**	50.54	34.25	26.64	**78.33**	38.66	**53.07**	**67.53**	46.42	13.78	40.85	46.18	**57.06**	30.65	**88.75**	36.05
DAFormer [12]	26.54	67.73	48.90	**37.16**	23.30	75.45	51.38	45.03	62.01	27.19	8.46	33.03	**55.25**	16.16	**34.64**	13.66	29.70
PFST (ours)	31.79	70.23	**50.95**	36.83	**27.41**	78.22	48.08	52.19	64.21	43.99	12.25	**49.92**	51.38	55.17	32.92	88.32	**36.32**

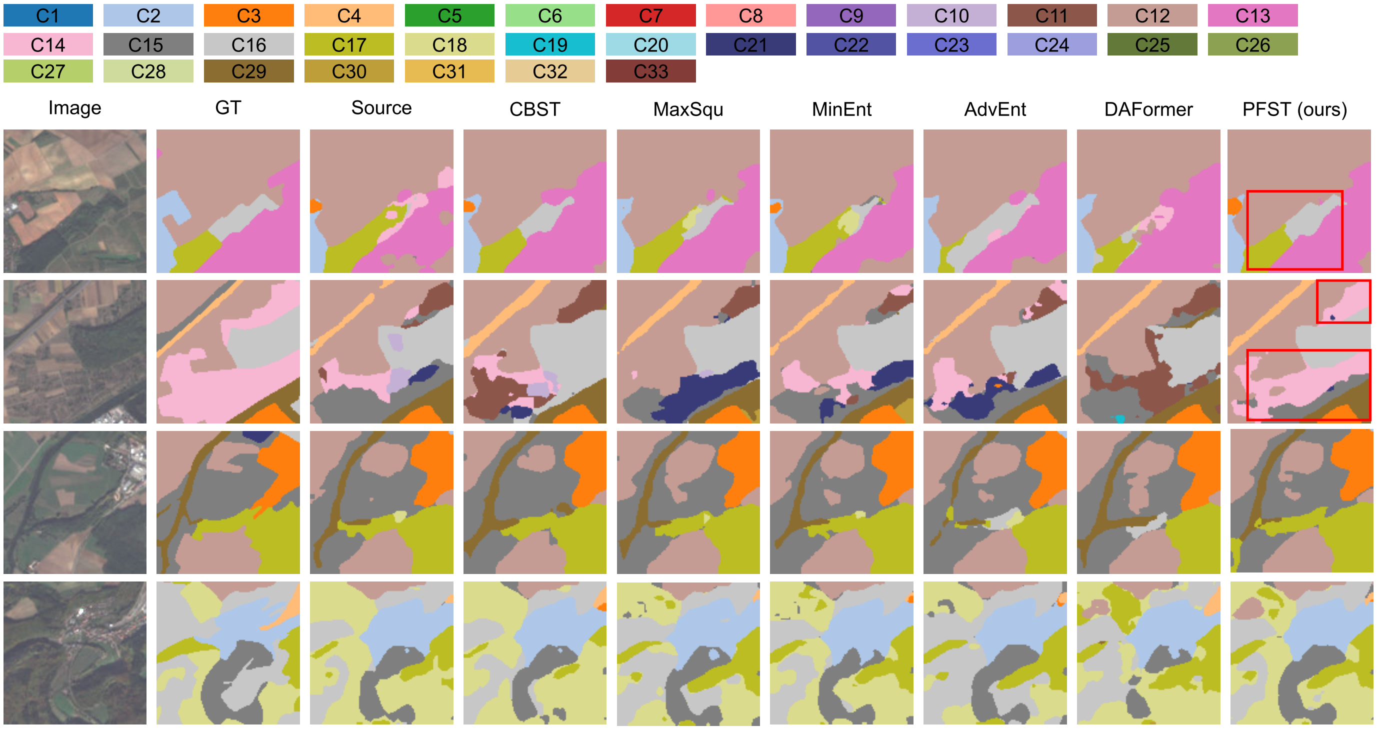


Fig. 7. Visualized semantic segmentation results of different UDA methods on SeasonNet spring to fall setting. For the sake of simplicity, we use "C1"–"C33" to denote the class labels. For the actual class names, please refer to [51].

TABLE V

Ablation Study of the Proposed Method on ISPRS P2V Setting and V2P Setting

Setting	ST w/o Strong Aug.	ST w/ Strong Aug.	$\mathcal{L}_{feat}+\mathcal{L}_{loc}$	mIoU
P2V	-	-	-	48.79
	✓	-	-	58.03
	✓	✓	-	62.26
	✓	✓	✓	64.64
V2P	-	-	-	47.12
	✓	-	-	53.25
	✓	✓	-	53.18
	✓	✓	✓	56.95

E. Ablation Study

To evaluate if the idea of mining the feature-level local similarity from the source domain model can really help the target model to generalize on the target domain, we ablate over the proposed local similarity losses and other components on the ISPRS P2V and V2P settings. As can be seen from Table V, self-training with exponential moving average plays an important and fundamental role in setting a strong baseline in our method, resulting in around 10% and 6% performance improvements on these two settings. If we apply strong augmentations on top of the student model branch during the self-training, we see a further performance increase on P2V setting, although the influence is not that obvious on V2P setting. In terms of the proposed local similarity loss, we can see it helps to further boost the performance of the UDA method on both settings, with relatively large margins (more than 2% and 3%) on top of an already very strong baseline. As for a nonparametric component, this result is promising and proves the effectiveness of leveraging feature-level local relation.

F. Interpreting the Local Feature Relation

To better explain the effectiveness of exploiting local relation, some verifying experimental results are presented in this

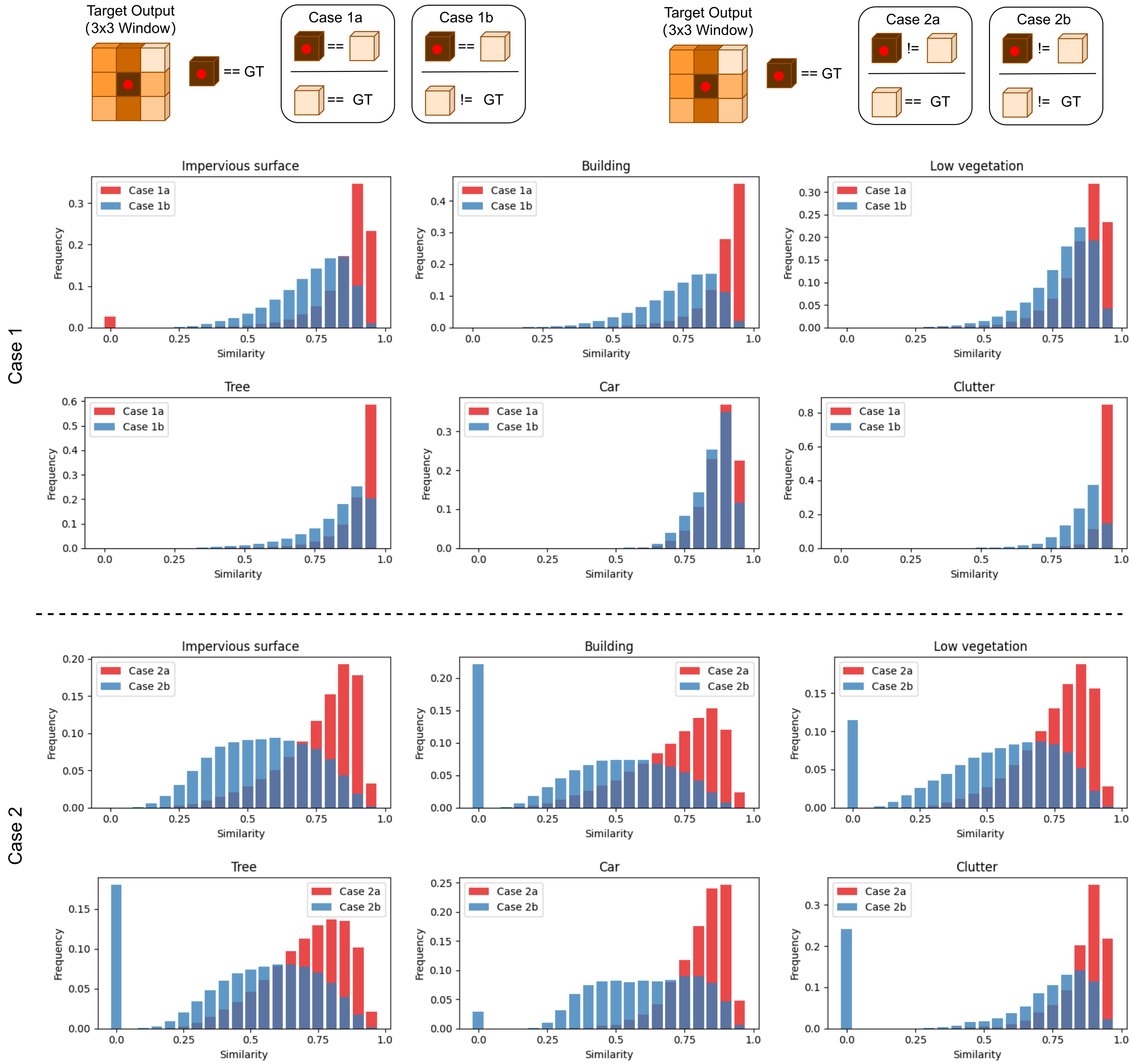


Fig. 8. Per-class feature similarity distribution for all the sliding windows on ISPRS P2V setting. The class is defined according to the label of the center pixel. For most of the classes, one can observe that the feature similarity of Case 1a and Case 2a are generally larger than the similarity of Case 1b and Case 2b, indicating that the similarity between local pseudo features are more accurate than the pseudo labels in revealing the true relation between neighboring target outputs.

section. Considering each local region defined by Ω_i that is sliding over the target image $\mathbf{x}^t$, we assume that the center pixel $\mathbf{x}_i^t$ is correctly classified by the semantic segmentation model, i.e., $argmax\ f_\theta(\mathbf{x}^t)_i = y_i^t$. To this end, by investigating whether the pseudo labels of $\mathbf{x}_i^t$ and its neighborhood $\mathbf{x}_j^t$ are of the same class or not, there will be two cases.

1) $argmax\ f_\theta(\mathbf{x}^t)_i = argmax\ f_\theta(\mathbf{x}^t)_j$: In this case, the two neighboring pixels have the same pseudo labels. There will be another two subcases depending on whether the prediction on $\mathbf{x}_j^t$ is correct or not, i.e., argmax $f_\theta(\mathbf{x}^t)_j = y_j^t$ or argmax $f_\theta(\mathbf{x}^t)_j \neq y_j^t$. We denote these two subcases as Case 1a and Case 1b.
2) $argmax\ f_\theta(\mathbf{x}^t)_i \neq argmax\ f_\theta(\mathbf{x}^t)_j$: In this case, the two neighboring pixels have different pseudo labels. Similarly, there will be two subcases according to whether there is argmax $f_\theta(\mathbf{x}^t)_j = y_j^t$ or argmax $f_\theta(\mathbf{x}^t)_j \neq y_j^t$. These two subcases are denoted as Case 2a and Case 2b.

With the above listed cases, we seek to verify the assumption that the pairwise feature similarities are more likely to reveal the true relationship between each pair of the neighboring pixels than the pseudo labels. In both Case 1a and Case 1b, the pseudo labels give the same predictions to the neighboring pixels, yet these predictions are correct in Case 1a, while incorrect in Case 1b. Hence our assumption can be supported if the pairwise similarity values in Case 1a are statistically larger than those in Case 1b. Likewise, since the pseudo labels give different predictions to the neighboring pixels both in Case 2a

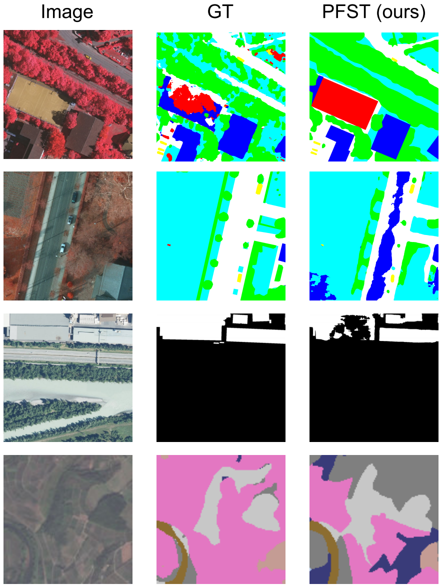


Fig. 9. Typical failure cases of the proposed methods on all the four UDA settings.

and Case 2b, our assumption can be supported if the similarity values in Case 2a are larger than those in Case 2b. As shown in Fig. 8, the expected phenomena can be observed, which verifies our assumption.

G. Limitations and Failure Cases

To analyze the limitations of the proposed method, we present some failure cases of the proposed methods in Fig. 9.

In the first row, it can be seen that PFST fails to recognize the central basketball field, which is a rare class that is not well-represented in the training set, and misclassifies it as the "Clutter" class. This suggests that the proposed local spatial layout (LSL) method may not be effective in recognizing out-of-distribution targets. In the second and third rows, PFST misclassifies the central "Impervious Surface" and the upper "Building" area, possibly due to the lack of spatial context. In such scenarios, the proposed method may not provide significant improvement. In the last row, PFST misclassifies the upper "Pastures" area as "Vineyards" area, which have similar appearances, indicating that the high-dimensional feature-level relation may not be sufficient to distinguish targets that exhibit only subtle differences.

These failure cases highlight the limitations of the proposed methods in handling rare classes, lack of spatial context, and subtle differences in appearance, indicating areas where further improvements may be needed.

V. Conclusion

In this article, we observe that domain-invariant knowledge can be better preserved within high dimensional featurewise topological relation than output space pseudo labels for UDA. Inspired by this, a novel self-training mechanism is developed to regularize target outputs using local relation within source feature space. The proposed method is evaluated on four standard UDA settings, and the results show that it achieves superior performance compared to the existing UDA methods.

While the proposed method has shown success in general cases, its performance may be limited when dealing with limited spatial contexts or out-of-distribution targets. Additional research is required to overcome these challenges and improve the method's robustness and effectiveness in such scenarios.

References

[1] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[2] 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.

[3] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12408–12417.

[4] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427161930259X

[5] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[6] 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.

[7] R. P. Sishodia, R. L. Ray, and S. K. Singh, "Applications of remote sensing in precision agriculture: A review," *Remote Sens.*, vol. 12, no. 19, p. 3136, Sep. 2020.

[8] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," in *Proc. Adv. Data Sci. Inf. Eng.*, Oct. 2021, pp. 877–894.

[9] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 59–69, Apr. 2019.

[10] W. Huang, Y. Shi, Z. Xiong, Q. Wang, and X. X. Zhu, "Semi-supervised bidirectional alignment for remote sensing cross-domain scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 192–203, Jan. 2023.

[11] X. Guo, J. Liu, T. Liu, and Y. Yuan, "SimT: Handling open-set noise for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7022–7031.

[12] 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.

[13] L. Hoyer, D. Dai, and L. Van Gool, "HRDA: Context-aware high-resolution domain-adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 372–391.

[14] Y. Zou, Z. Yu, B. V. K. V. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 297–313.

[15] 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.

[16] MMS Contributors. (2020). *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[17] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "EarthNets: Empowering AI in Earth observation," 2022, *arXiv:2210.04936*.

[18] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[19] P. Shamsolmoali et al., "Image synthesis with adversarial networks: A comprehensive survey and case studies," *Inf. Fusion*, vol. 72, pp. 126–146, Aug. 2021.

[20] P. Wang, Y. Li, and N. Vasconcelos, "Rethinking and improving the robustness of image style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 124–133.

[21] 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.

[22] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[23] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.

[24] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1335–1344.

[25] Y. Cai et al., "BiFDANet: Unsupervised bidirectional domain adaptation for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 14, no. 1, p. 190, Jan. 2022.

[26] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc, "StandardGAN: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 747–756.

[27] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "DAugNet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1067–1081, Feb. 2021.

[28] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] L. Yan, B. Fan, H. Liu, C. Huo, S. Xiang, and C. Pan, "Triplet adversarial domain adaptation for pixel-level classification of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3558–3573, May 2020.

[34] 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.

[35] M. N. Subhani and M. Ali, "Learning from scale-invariant examples for domain adaptation in semantic segmentation," in *Computer Vision—ECCV 2020*. Glasgow, U.K.: Springer, 2020, pp. 290–306.

[36] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Computer Vision—ECCV 2020*. Glasgow, U.K.: Springer, 2020, pp. 415–430.

[37] I. Shin, S. Woo, F. Pan, and I. S. Kweon, "Two-phase pseudo label Densification for self-training based domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 532–548.

[38] F. Fleuret et al., "Uncertainty reduction for model adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9613–9623.

[39] 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. (CVPR)*, Jun. 2021, pp. 12409–12419.

[40] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," 2021, *arXiv:2110.08733*.

[41] 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, 2022, Art. no. 5609413.

[42] Z. Yuan and C. Lin, "Research on strong constraint self-training algorithm and applied to remote sensing image classification," in *Proc. IEEE Int. Conf. Power Electron., Comput. Appl. (ICPECA)*, Jan. 2021, pp. 981–985.

[43] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[44] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[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] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "ClassMix: Segmentation-based data augmentation for semi-supervised learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1368–1377.

[47] C. Yaras, B. Huang, K. Bradbury, and J. M. Malof, "Randomized histogram matching: A simple augmentation for unsupervised domain adaptation in overhead imagery," 2021, *arXiv:2104.14032*.

[48] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," 2021, *arXiv:2110.04596*.

[49] 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, 2022, Art. no. 5611412.

[50] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15379–15389.

[51] D. Koßmann, V. Brack, and T. Wilhelm, "SeasoNet: A seasonal scene classification, segmentation and retrieval dataset for satellite imagery over Germany," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2022, pp. 243–246.

[52] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[53] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2090–2099.

[54] 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. 833–851.

[55] 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.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

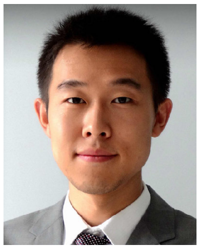
[57] L. Ilya and H. Frank, "Decoupled weight decay regularization," in *Proc. ICLR*, vol. 7, 2019, pp. 1–19.

[58] Y. Zhao, P. Guo, Z. Sun, X. Chen, and H. Gao, "ResiDualGAN: Resize-residual DualGAN for cross-domain remote sensing images semantic segmentation," 2022, *arXiv:2201.11523*.

[59] H. Gao, Y. Zhao, P. Guo, Z. Sun, X. Chen, and Y. Tang, "Cycle and self-supervised consistency training for adapting semantic segmentation of aerial images," *Remote Sens.*, vol. 14, no. 7, p. 1527, Mar. 2022.

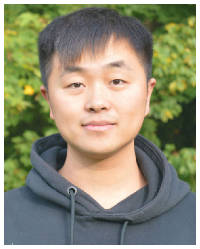
Fahong Zhang received the B.E. degree in software engineering from Northwestern Polytechnical University, Xi'an, China, in 2017, and the M.S. degree in computer science from the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, in 2020. He is currently pursuing the Ph.D. degree with the Department of Aerospace and Geodesy, Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany.

His research interests include computer vision and satellite image processing.



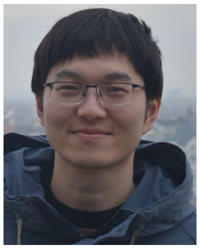
Yilei Shi (Member, IEEE) received the Dipl.Ing. degree in mechanical engineering and the Dr.-Ing. degree in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2019, respectively.

He is currently a Senior Scientist with the Chair of Remote Sensing Technology, TUM. His research interests include fast solver and parallel computing for large-scale problems, high-performance computing and computational intelligence, advanced methods on synthetic aperture radar (SAR) and InSAR processing, machine learning and deep learning for variety of data sources, such as SAR, optical images, and medical images, and partial differential equation (PDE)-related numerical modeling and computing.



Zhitong Xiong (Member, IEEE) received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2021.

He is currently a Senior Scientist and leads the ML4Earth Working Group, Chair of Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany. His research interests include computer vision, machine learning, Earth observation, and Earth system modeling.



Wei Huang received the B.E. degree in control theory and engineering from Northwestern Polytechnical University, Xi'an, China, in 2018, where he is currently pursuing the M.S. degree in computer science with the School of Artificial Intelligence, Optics and Electronics (iOPEN). He is also pursuing the Ph.D. degree with the Technical University of Munich, Munich, Germany.

His research interests include transfer learning, deep learning, and remote sensing.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was the Founding Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy; Fudan University, Shanghai, China; The University of Tokyo, Tokyo, Japan; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School (www.mu-ds.de), Munich. Since 2019, she has also been heading the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the PI and the Director of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has also been serving as the Director of the Munich Data Science Institute (MDSI), TUM, where she is currently the Chair Professor for Data Science in Earth Observation. She is also a Visiting AI Professor at the ESA's Phi-Laboratory, Frascati, Italy. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, UN's SDGs, and climate change.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves in the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of the IEEE Transactions on Geoscience and Remote Sensing and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.