remote sensing

MDPI

Article

Learning to Adapt Adversarial Perturbation Consistency for Domain Adaptive Semantic Segmentation of Remote Sensing Images

Zhihao Xi [1,2], Yu Meng [1,2], Jingbo Chen [1,2,*], Yupeng Deng [2], Diyou Liu [2], Yunlong Kong [2] and Anzhi Yue [2]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China; xizhihao19@mails.ucas.ac.cn (Z.X.); mengyu@aircas.ac.cn (Y.M.)
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China; dengyp@radi.ac.cn (Y.D.); liudiyou@aircas.ac.cn (D.L.); kongyl@radi.ac.cn (Y.K.); yueaz@aircas.ac.cn (A.Y.)
* Correspondence: chenjb@aircas.ac.cn

Abstract: Semantic segmentation techniques for remote sensing images (RSIs) have been widely developed and applied. However, most segmentation methods depend on sufficiently annotated data for specific scenarios. When a large change occurs in the target scenes, model performance drops significantly. Therefore, unsupervised domain adaptation (UDA) for semantic segmentation is proposed to alleviate the reliance on expensive per-pixel densely labeled data. In this paper, two key issues of existing domain adaptive (DA) methods are considered: (1) the factors that cause data distribution shifts in RSIs may be complex and diverse, and existing DA approaches cannot adaptively optimize for different domain discrepancy scenarios; (2) domain-invariant feature alignment, based on adversarial training (AT), is prone to excessive feature perturbation, leading to over robust models. To address these issues, we propose an AdvCDA method that guides the model to adapt adversarial perturbation consistency. We combine consistency regularization to consider interdomain feature alignment as perturbation information in the feature space, and thus propose a joint AT and self-training (ST) DA method to further promote the generalization performance of the model. Additionally, we propose a confidence estimation mechanism that determines network stream training weights so that the model can adaptively adjust the optimization direction. Extensive experiments have been conducted on Potsdam, Vaihingen, and LoveDA remote sensing datasets, and the results demonstrate that the proposed method can significantly improve the UDA performance in various cross-domain scenarios.

Keywords: unsupervised domain adaptation; adversarial perturbation consistency; self-training; semantic segmentation; remote sensing

Citation: Xi, Z.; Meng, Y.; Chen, J.; Deng, Y.; Liu, D.; Kong, Y.; Yue, A. Learning to Adapt Adversarial Perturbation Consistency for Domain Adaptive Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5498. https://doi.org/10.3390/rs15235498

Academic Editor: Mohammad Awrangjeb

Received: 20 October 2023
Revised: 19 November 2023
Accepted: 21 November 2023
Published: 25 November 2023

1. Introduction

Image segmentation has been widely researched as a basic remote sensing intelligent interpretation task [1–4]. In particular, semantic segmentation based on deep learning plays an important role as a pixel-level classification method in remote sensing interpretation tasks, such as building extraction [5], landcover classification [6] and change detection [7,8]. However, the prerequisite for good performance in existing fully supervised deep learning approaches is sufficiently annotated data. It is also essential that the training and test data follow the identical distributions [9]. Once applied to unseen scenarios with different data distributions, model performance can degrade significantly [10–12]. This means new data might be annotated and retrained for performance requirements, which requires considerable labor and time [13].

In practical applications, the domain discrepancy problem is prevalent in remote sensing images (RSIs) [14,15]. Different remote sensing platforms, payload imaging mechanisms, and photographic angles will induce variations in image spatial resolution and

object features [16]. Due to the variation in seasons, geographic locations, illumination, and atmospheric radiation conditions, the same source images may also show significant feature distribution differences [17]. The data distribution shift caused by the mix of these complex factors leads the segmentation network to behave poorly in the unseen target domain.

As a transfer learning paradigm [18], unsupervised domain adaptation (UDA) can improve the domain generalization performance of the model by transferring knowledge from the source domain data with annotations to the target domain [19]. This method has been extensively researched in computer vision to address the domain discrepancy issue in natural image scenes [20]. Domain adaptive (DA) methods have also gained intensive attention in remote sensing [21]. Compared with natural images, RSIs contain more complex spatial detail information and object boundary situation, and homogeneous and heterogeneous phenomena are more common in images. Additionally, the factors that generate domain discrepancies are more complex and diverse. Thus, solving the problem of domain discrepancies in RSIs became more challenging. Currently, existing research works focus on three main approaches: UDA based on image transfer [17,22], UDA based on deep adversarial training (AT), and UDA based on self-training (ST) [23,24]. Image transfer methods achieve image-level alignment based on generative adversarial networks. AT-based methods (as shown in Figure 1a) reduce the feature distribution in the source and target domains by minimizing the adversarial loss to achieve feature-level alignment [25]. The ST approach (as shown in Figure 1b) focuses on generating high-confidence pseudolabels in the target domain and then participating in the iterative training of the model to achieve the progressive transfer process [26,27].

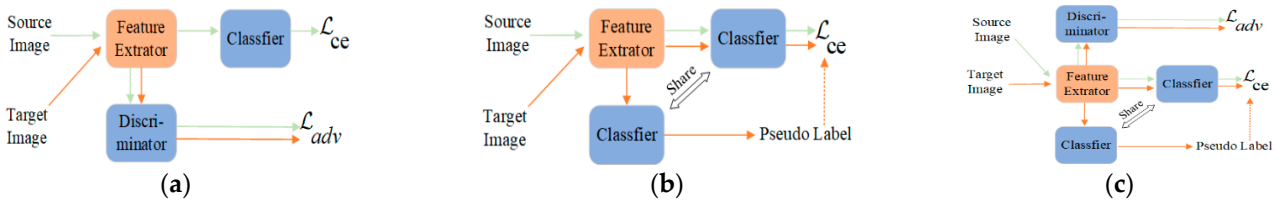

Figure 1. General paradigm description of existing DA training methods. (**a**) AT based DA approach. (**b**) Self-training (ST) based DA approach. (**c**) A combined ST and AT for DA methods.

One general conclusion about the DA performance of the model is: AT + ST > ST > AT [27]. However, as shown in Figure 1c, combining ST and AT methods typically requires strong coupling between submodules, which leads to a poorly stabilized model during training [28]. Therefore, fine-tuning the network structure and the submodules parameters is generally needed, so that model performance depends on specific scenarios and loses its scalability and flexibility. Recently, several studies have been conducted to optimize and improve the process, such as decoupling AT and ST methods functionally by constructing dual-stream networks [28], and using exponential moving average (EMA) techniques to construct teacher networks to smooth instable features in the training process [29]. However, it also complicates the network architecture, increasing the spatial computational complexity, and reducing training efficiency.

This paper combines the consistency regularization idea in semi-supervised learning and proposes a DA semantic segmentation method based on adversarial perturbation consistency to overcome the limitations of the aforementioned methods. Inspired by FixMatch [30], our approach first generates pseudolabels using weak augmentation to predict target domain images. The same images are strongly augmented with the RandAugment(RA) [31] and ClassMix [32] techniques and then fed into the model for training. The supervised information comes from generating higher quality pseudolabels using the weakly augmented branch, thus preserving output prediction consistency in the case of diverse input perturbations. This process is termed the weak-to-strong consistency stream. Critically, AdvCDA provides feature-level perturbations in the feature space by AT for

interdomain alignment, while leveraging the same weakly augmented branch to provide high-quality pseudolabels for supervised constraints. In this way, the model generalization is improved by reducing the interdomain discrepancies while maintaining model training stability through the supervisory constraints of the pseudolabel information. This process is termed the adversarial perturbation consistency stream. In addition, the confidence estimation mechanism is designed to assess the reliability of the two consistent perturbation processes, and thus the model can adaptively optimize the learning direction according to the training scenes.

In this paper, the main contributions are summarized as follows:

1. We propose an AdvCDA method for high-resolution RSIs based on adversarial perturbation consistency. The method combines AT and ST strategies to provide feature perturbation information through interdomain alignment in order to improve the domain generalization of the model during the ST process. Moreover, the ST method provides high-quality labels that maintain the predictive consistency of the model during AT, thus alleviating the over robustness that is prone to arise during domain alignment.
2. We propose a confidence estimation mechanism to determine the learning weights of the weak-to-strong consistency stream and the adversarial perturbation consistency stream so that the model can adaptively adjust the optimization direction according to different scenarios. Our method has been effectively demonstrated in various domain discrepancy scenarios of high-resolution RSIs.

2. Related Works

2.1. Image-Level Alignment for UDA

Image-level alignment reduces the data distribution shift between the source and target domains through image transfer methods [33,34]. This scheme generates pseudo images that are semantically identical to the source images, but whose spectral distribution is similar to that of the target images [17]. Cycle-consistent adversarial domain adaptation (CyCADA) improves the semantic consistency of the image transfer process through cycle consistency loss [35]. To preserve the semantic invariance of RSIs after being transferred, ColorMapGAN designs a color transformation method without a convolutional structure [17]. Many UDA schemes adopt GAN-based style transfer methods [36] to align data distributions in the source and target domains. ResiDualGAN [22] introduces scale information of RSIs based on DualGAN [37]. Some work also leverages non-adversarial optimization transform methods, such as Fourier transform-based FDA [38] and Wallis filtering methods [39], to reduce image domain discrepancies.

2.2. Feature-Level Alignment by AT

Adversarial-based feature alignment methods train additional domain discriminators [19,40] to distinguish target samples from source samples and then train the feature network to fool the discriminator, thus generating a domain-invariant feature space [41]. Many works have made significant progress using AT to align the feature space distribution to reduce the domain variance in RSIs. Wu et al. [42] focused on interdomain category differences and proposed class-aware domain alignment. Deng et al. [23] designed a scale discriminator to detect scale variation in RSIs. Considering regional diversity, Chen et al. [43] focused on difficult-to-align regions through a region adaptive discriminator. Bai et al. [20] leveraged contrast learning to align high-dimensional image representations between different domains. Lu et al. [44] designed global-local adversarial learning methods to ensure local semantic consistency in different domains.

2.3. Self-Training for UDA

Self-training acts as a kind of semi-supervised learning [45], which involves high-confidence prediction as easy-to-transfer pseudolabels, and participates in the next iteration of training together with the corresponding target images, progressively realizing the

knowledge transfer process [26,27]. Yao et al. [39] used the ST paradigm to improve the performance of the model for building extraction on unseen data. CBST [26] designs class-balanced selectors for pseudolabels to avoid the easy-to-predict classes becoming dominant. ProDA [46] computes representation prototypes that represent the centers of category features to correct pseudolabels. CLUDA [47] constructs contrast learning between different classes and different domains by mixing source and target domain images. Additionally, several works have attempted to combine ST and adversarial methods to improve domain generalization performance. However, these models are difficult to optimize and often require fine-tuning of the model parameters. Zhang et al. [48] established the two-stage training process of AT followed by ST. DecoupleNet [28] decouples ST and AT through two network branches to alleviate the difficulty of model training.

2.4. Consistency Regularization

Consistency regularization is generally employed to solve semi-supervised problems, where the essential idea is to preserve the output consistency of the model under different versions of input perturbations, thus improving the generalization ability of the model for test data [49,50]. FixMatch [30] establishes two network flows, which include weak perturbation augmentation and strong perturbation augmentation at the image level, using the weak perturbation to ensure the high quality of the output and using the strong perturbation to provide better training of the model. FeatMatch [51] extracts class representative prototypes for feature-level augmentation transformations. Liu et al. [52] constructed dual-teacher networks to provide more rigorous pseudolabels for unlabeled test data. UniMatch [50] provides an auxiliary feature perturbation stream using a simple dropout mechanism. Several recent regularization models have been designed under the ST paradigm, but fail to account for domain discrepancy scenes, which has led to the fact that pure consistency regularization has not behaved remarkably well in cross-domain scenes.

3. Materials and Methods

In this section, the general architecture of the proposed network is illustrated and each component of our approach is elaborated. We attempt to improve the domain generalization performance through a combination of AT and ST methods. However, distinguishing from existing work [28,29,53], we are devoted to leveraging the idea of consistency regularization [52,54,55] to preserve the output prediction consistency during the feature alignment process to mitigate the instability issues that are easily induced by the adversarial perturbation. Simultaneously, a confidence estimation mechanism is established to optimize the training direction for different complicated domain difference scenarios in RSIs. First, some preliminary work is introduced in Section 3.1. Then, the proposed adversarial perturbation with consistency is described in Section 3.2, and the proposed confidence estimation mechanism is described in Section 3.3.

3.1. Preliminaries

In the DA semantic segmentation task, the source domain images are defined as $\mathcal{X}_S = \{x_s^i\}_{i=1}^{N_s}$, where $x_s^i \subset \mathbb{R}^{H\times W\times 3}$, and its corresponding one-hot ground truth is $\mathcal{Y}_S = \{y_s^i\} \subset \mathbb{R}^{H\times W\times C}$. Let us define the target domain images as $\mathcal{X}_T = \{x_t^i\}_{i=1}^{N_s}$, where $x_t^i \subset \mathbb{R}^{H\times W\times 3}$, and the ground truth of the target domain cannot access the model. Typically, the annotated source domain data is used to train model G with parameters θ, and then the trained weights are directly applied to the target domain. Supervisory losses are formulaic as follows:

$$\mathcal{L}_S = \frac{1}{B_S}\sum_{i=1}^{B_S}\mathcal{H}(y_s^i, p(y|x_s^i;\theta)) \tag{1}$$

where $p(y|x_s^i;\theta) = G(x_s^i)$ and B_S is defined as the batch size of the source domain data input to the model at each iteration. $\mathcal{H}$ represents the loss entropy of minimizing the

ground truth with respect to the predicted probability distribution. This method is set up as the multicategory cross-entropy. In general, the generalization ability of the model tends to perform poorly if domain discrepancies exist between the source and target domains, resulting in the model performance in the target domain usually being suboptimal.

Several strategies and methods [25,28,41,56] have been proposed to address the domain shift problem, among which AT and ST have become the two dominant DA methods [57]. In ST, the model generates pseudolabels for the target domain images and iteratively transfers training for the model to be adapted to the target domain. The overall objective function is the linear combination of the supervised loss in the source domain and the unsupervised loss in the target domain $\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_T$.

$$\mathcal{L}_T = \frac{1}{B_T} \sum_{i=1}^{B_T} \mathbb{I}\Big(\max(p(y|x_t^i, \theta)) \geq \tau\Big) \mathcal{H}(y_t^i, p(y|x_t^i, \theta)) \tag{2}$$

$$y_t^i = \text{argmax}(p(y|x_t^i, \theta)) \tag{3}$$

where B_T is the batch size of the target domain data for the input model, τ is defined as the default confidence threshold, which is usually set at 0.9 to select high-quality pseudolabels for the target domain, and y_t^i represents the candidate pseudolabels from the target domain.

As a common concept in semi-supervised learning [30,51,58], consistency regularization [52,55] typically imposes random perturbation information on unannotated data while constraining the model to maintain output prediction consistency. FixMatch [30] uses weak-to-strong consistency regularization to assign different levels of perturbation augmentation, dubbed weak perturbation $\mathcal{A}^w$ and strong perturbation $\mathcal{A}^s$, to each unannotated target domain images $\mathcal{X}_t$. It is written as

$$p_t^w(y|x_t^i, \theta) = \hat{G}(\mathcal{A}^w(x_t^i)), p_t^s(y|x_t^i, \theta) = G(\mathcal{A}^s(\mathcal{A}^w(x_t^i))) \tag{4}$$

where the teacher network $\hat{G}$ generates higher-quality pseudolabels from weakly perturbed target images, and the student network G serves as a trainable segmentation network to apply stronger perturbations to the same images for optimizing the model. In our method, the teacher network $\hat{G}$ and the student network G are designed to share weights.

AT obtains the domain-invariant feature space of the source and target domains via aligning the interdomain global feature distributions, which provides another effective method to alleviate the domain discrepancy problem. It generally consists of segmentation network G and discriminative network D. The segmentation network can be divided into the feature extractor F and the classifier C, where $G = F \circ C$. AT depends on the discriminative network D to align the feature distributions extracted by the segmentation network in the source and target domains. Specifically, the segmentation network G and the discriminative network D are optimized alternately and iteratively by the following two steps [25,28,40]:

(1) First, F and C of the segmentation network are frozen, and only the determination network is optimized, which improves the domain discrimination ability of the discriminator D to distinguish the output features of different domains:

$$\min_D \mathcal{L}_D = -(1-d) \log P(d=0|f_s) - d \log P(d=1|f_t) \tag{5}$$

where f_s and f_t are feature extractors whose inputs are source images $\mathcal{X}_s$ and target images $\mathcal{X}_t$. d denotes the domain indicator, where 0 denotes the source domain, and 1 denotes the target domain. $P(d=0|f)$ and $P(d=1|f)$ denote the output probability that discriminator D determines; the input comes from the source and target domains, respectively.

(2) The segmentation network G not only conducts supervised training tasks with labeled source domains, but also participates in the AT process. Specifically, the adversarial

loss is as follows, and this process is achieved by fixing the discriminative network D and optimizing F and C of the segmentation network.

$$\mathcal{L}_{adv} = \log P(d = 0|f_t) \tag{6}$$

$$\min_{F,C} \mathcal{L}_s + \lambda_{adv}\mathcal{L}_{adv} \tag{7}$$

The main purpose of adversarial loss $\mathcal{L}_{adv}$ is to confuse the discriminator and encourage the segmentation network to perform interdomain alignment and learn domain invariant features.

In general, the ST method combined with consistency regularization shows better stability with small discrepancies in data distributions between source and target domains. However, in practical cases, the factors that cause the data distribution discrepancies in RSIs are often complicated. For complex domain discrepancy scenarios, the generalization performance of simple ST methods usually fails to meet the requirements due to the impact of pseudolabel noise. Deep AT methods aim to reduce domain discrepancies through feature space alignment. However, for the semantic segmentation task, fine-grained feature alignment in high-dimensional space is needed, which is prone to induce more noise disturbances causing the model to become over robust and affecting the stability of adversarial learning.

Based on the above issues, we propose a novel DA method for high-resolution RSIs based on adversarial perturbation consistency. We provide directional feature perturbation through AT and align the source domain features with the target domain to improve the domain generalization ability of the model. Additionally, combining consistency regularization and the ST paradigm maintains the output prediction consistency after feature perturbation and improves the stability of AT. Moreover, to adapt to the complex domain discrepancy scenarios in RSIs, based on the complementary advantages of weak-to-strong and adversarial perturbation consistency, we further develop a confidence estimation mechanism for pseudolabels to constrain the direction of the decision boundary.

3.2. Adversarial Perturbations Consistency

To combine the AT and ST paradigms to improve the domain transfer performance of the model, and simultaneously ensure model stability during the training process, inspired by the consistency regularization idea of semi-supervised learning, we propose an adversarial perturbation consistency-based DA semantic segmentation method. Consistency regularization has achieved significant effects in the semi-supervised domain. However, it is difficult to achieve breakthrough performance improvement when applied directly to scenarios where large data distribution shifts exist between the source and target domains, mainly due to the lack of an effective feature alignment mechanism to reduce the interdomain discrepancies. AT is an effective interdomain feature alignment method, but it relies on fine-grained alignment in high-dimensional feature space, which is prone to generating ineffective feature perturbations and causing instability in the training process. Hence, AdvCDA considers the AT process as a single directional feature perturbation stream in consistency regularization to reduce the interdomain variance. Simultaneously, the output consistency is constrained by consistency loss to maintain AT stability.

The framework of AdvCDA is shown in Figure 2. For source images with ground truth, we use supervised loss to train the segmentation network and improve the semantic discrimination performance of the model for each category. For the target domain, we set up three branches to achieve domain transfer between the source and target images to improve the generalization of the model: the weak augmentation branch, the strong augmentation branch, and the adversarial perturbation branch, respectively. Similar to some existing semi-supervised methods [30,50], we provide different versions of input perturbations at the input level through weak and strong augmentation to improve the generalization of the model. However, due to domain shifts, consistency learning [51] at only the input image

level is often insufficient and requires the model to maintain consistency at multiple levels under various perturbations to fully exploit the ability of the model to learn generalized features. In particular, it is important to note that the goal of UDA is to align the feature space between different domains to reduce domain discrepancies.

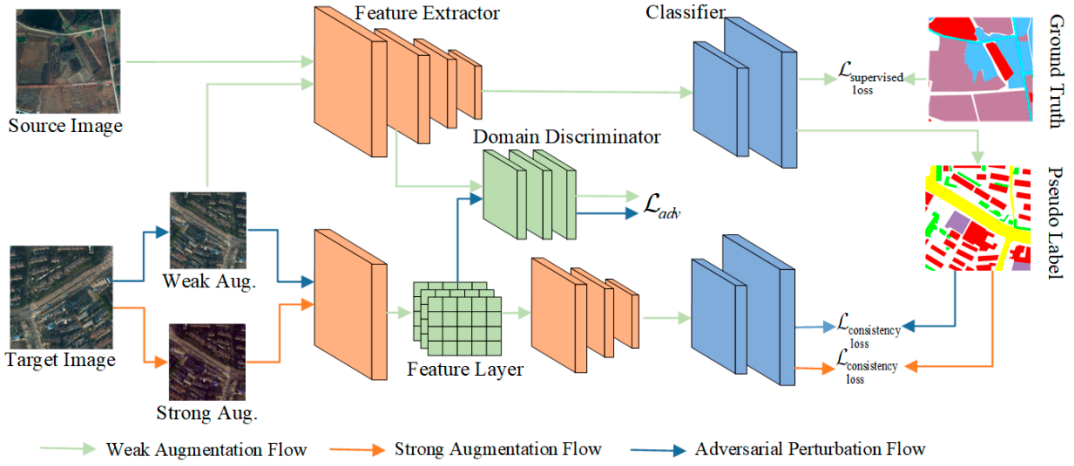

Figure 2. Overall framework of AdvCDA. The source images are fed into the feature extractor and classifier, and the supervised loss is computed using the source predictions and the corresponding ground truth to help the segmentation network learn task-specific knowledge. The target images pass through a weak augmentation flow to obtain high-quality pseudolabels. The same target images are put through a strong augmentation flow and an adversarial perturbation flow to obtain two target predictions, which are used to minimize the consistency loss. The two consistency training processes are weak-to-strong consistency and adversarial perturbation consistency. The domain discriminator is part of the AT to generate feature perturbations to the network layer. The feature alignment of the source and target domains is performed to minimize domain discrepancies.

Therefore, based on weak-to-strong perturbation consistency learning [30], as shown in Figure 3a, we propose injecting adversarial perturbation information to maintain the consistency of the output prediction with the adversarial perturbation. Specifically, as shown in Figure 3b, we separate image- and feature-level perturbations into individual network streams, allowing the model to directly achieve target consistency with each type of perturbation information.

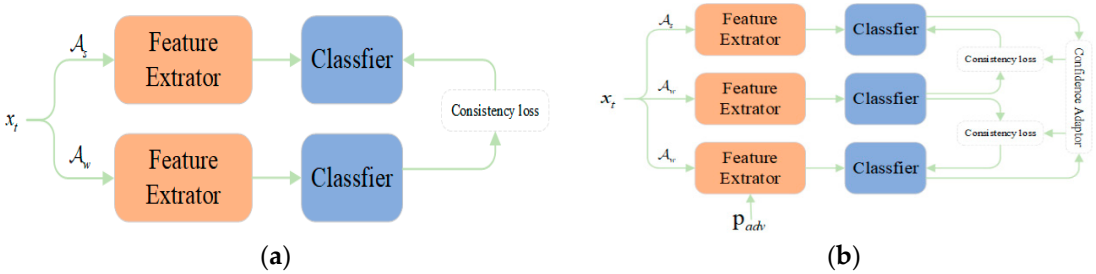

Figure 3. Comparison of consistency regularization pipelines. (**a**) Weak-to-strong consistency baseline framework. (**b**) The proposed adversarial perturbation consistency framework.

First, we divide the model encoder into f_{low} and f_{high} parts, that is, $f = f_{low} \circ f_{high}$. We attempt to align the shallow feature space of the model in the source and target domains. This design explains that the domain discrepancies between the source and target domains are represented in the low-level feature information, such as spectral and textural differences, because of the geographic location, atmospheric radiation conditions, or seasons. These features are generally captured by the shallow layer of the feature extractor, so we decided to inject adversarial perturbation information into the shallow network features,

which will capture domain-invariant features more accurately and simultaneously prevent excessive invalid perturbation information from affecting the stability of the model training process. The source and target domain images x_s^i and x_t^i are fed into the model to obtain the corresponding shallow features ϕ_s^i, ϕ_t^i and the predicted results:

$$\phi_s^i = f_{low}\Big(\mathcal{A}^w(x_s^i)\Big), p_s^w(y|x_s^i,\theta) = C(f_{high}(\phi_s^i)) \tag{8}$$

$$\phi_t^i = f_{low}\Big(\mathcal{A}^w(x_t^i)\Big), p_t^w(y|x_t^i,\theta) = C(f_{high}(\phi_t^i)) \tag{9}$$

Specifically, to reduce the domain discrepancies and improve the generalization performance of the model in the target domain, we attempt to align the feature distributions of the source and target domains through AT methods. Therefore, we apply a discriminator in the shallow feature space of the model for adversarial learning. The adversarial loss is as follows:

$$\mathcal{L}_{adv} = \frac{1}{B_T}\sum_{i=1}^{B_T}\left(D(\phi_t^i) - 0\right)^2 \tag{10}$$

Training the discriminator is also required to improve the discriminant performance on source and target domains. Discriminatory loss is described as follows:

$$\mathcal{L}_d = \frac{1}{B_S}\sum_{i=1}^{B_S}\left(D(\phi_s^i) - 0\right)^2 + \frac{1}{B_T}\sum_{i=1}^{B_T}\left(D(\phi_t^i) - 1\right)^2 \tag{11}$$

where 0 denotes the source domain and 1 denotes the target domain. The feature space of the target domain gradually converges to the source domain through AT to obtain the domain-invariant feature space. The alignment process in the source and target domains can be regarded as injecting a feature perturbation in the shallow feature space of the model and obtaining the new feature parameter, which is f_{low}^{fp}. Furthermore, we can obtain the predicted results after feature perturbation by AT:

$$\phi_t^{fp} = f_{low}^{fp}\Big(\mathcal{A}^w(x_t^i)\Big), p_t^{fp}(y|x_t^i,\theta) = C(f_{high}(\phi_t^{fp})) \tag{12}$$

Fine-grained feature alignment in high-dimensional space can be more prone to generate adversarial noise [59], leading to a lack of stability in training DA methods. Therefore, we constrain the model to maintain the consistency of the output predictions after noise perturbation based on the idea of consistency regularization, which helps to improve the stability of the model. Eventually, the unsupervised loss in the target domain is reformulated as $\mathcal{L}_w$ and $\mathcal{L}_{fp}$, where $\mathcal{L}_w$ denotes the weak-to-strong consistency loss and $\mathcal{L}_{fp}$ denotes the adversarial perturbation consistency loss.

$$\mathcal{L}_w = \frac{1}{B_T}\sum_{i=1}^{B_T}\mathbb{I}\Big(\max(p_t^w(y|x_t^i,\theta)) \geq \tau\Big)\mathcal{H}(y_t^i, p_t^s(y|x_t^i,\theta)) \tag{13}$$

$$\mathcal{L}_{fp} = \frac{1}{B_T}\sum_{i=1}^{B_T}\mathbb{I}\Big(\max(p_t^w(y|x_t^i,\theta)) \geq \tau\Big)\mathcal{H}(y_t^i, p_t^{fp}(y|x_t^i,\theta)) \tag{14}$$

$$y_t^i = \text{argmax}(p_t^w(y|x_t^i,\theta)) \tag{15}$$

To adapt to the complicated domain discrepancies in RSIs, one can find that our framework is designed with a weak-to-strong consistency stream and an adversarial perturbation consistency stream, which skillfully combines the ST and AT methods to improve domain transfer performance while guaranteeing training stability. Specifically, AT plays a crucial role in the network to conduct interdomain alignment to reduce domain discrepancies. On the one hand, AT provides feature-level perturbations to allow the model to learn various consistent features with more abundant perturbation information. On the other hand,

feature alignment is used to reduce the domain discrepancies between the source and target images to improve the domain generalization performance of the model. Meanwhile, consistency regularization enables the model to maintain strong stability during the co-learning process of ST and AT, which fully exploits the potential for domain generalization.

3.3. Confidence Estimation Mechanism

In general, for large domain discrepancy scenarios, feature alignment by AT plays the primary role in reducing the interdomain discrepancy and improving the generalization of the model. In contrast, ST methods are prone to pseudo-label noise that can lead to performance degradation [46]. For scenarios with small domain discrepancies, such as semi-supervised domains, the ST method can be sufficient to attain satisfactory results for the model in the target domain. Therefore, for the weak-to-strong consistency and adversarial perturbation consistency stream, it is better to allow the model to adaptively optimize the learned weights of the two streams to meet uncertain domain discrepancy scenarios.

The design key of this method is how to evaluate the confidence estimation of each stream to guide the model for better transfer training. As we know, it is especially critical for ST methods to design confidence thresholds for pseudolabels, where labels lower than the confidence threshold are generally considered incorrect labels for prediction. In contrast, labels higher than the threshold will be involved as candidate labels in the next iterative training process to improve the performance of the model in the target domain. Based on this, as shown in Figure 3b, we propose a confidence estimation mechanism that estimates the training confidence of the two streams by calculating the similarity of the outputs from the strong augmented branch, and the adversarial perturbation branch to the weakly augmented branch, thus constraining the model to assign more training weights to the higher-quality consistent network stream. In addition, it can be found that both of the proposed consistency regularization streams conduct consistently supervised learning based on weak augmentation. Intuitively, the weakly augmented branch is more prone to produce high-quality prediction results. We define the final target domain loss as:

$$\mathcal{L}_T = \lambda_1 \mathcal{L}_w + \lambda_2 \mathcal{L}_{fp} \tag{16}$$

where λ_1 and λ_2 are the key weights for estimating the confidence of the two streams. The weight values determine the influence level of the corresponding stream on the training and gradient optimization, guiding the optimization direction of the model. When $\lambda_2 = 0$, the model degenerates into a semi-supervised model, FixMatch [30]. Specifically, we use the similarity of the logit outputs from the strongly augmented branch and the adversarial perturbation branch, with the weakly augmented branch, respectively, as a confidence estimation for the two streams:

$$c_{ws}^i = \frac{1}{\mathcal{H}(y_t^i, p_t^s(y|x_t^i, \theta))}, \; c_{fp}^i = \frac{1}{\mathcal{H}(y_t^i, p_t^{fp}(y|x_t^i, \theta))} \tag{17}$$

where c_{ws}^i and c_{fp}^i are the confidence weights assigned to the two streams of weak-to-strong consistency and adversarial perturbation consistency, the higher weight value represents the higher confidence assigned to the corresponding stream, and the model tends to learn from the stream with high confidence. To avoid the instability problem caused by scale variation in weight values, we normalize the final weight values:

$$\lambda_1 = \frac{c_{ws}^i}{c_{ws}^i + c_{fp}^i}, \lambda_2 = \frac{c_{fp}^i}{c_{ws}^i + c_{fp}^i} \tag{18}$$

In this case, the final loss we use to train the segmentation network was $\mathcal{L} = \mathcal{L}_S + \mathcal{L}_T$, and $\mathcal{L}_{adv}$ as an adversarial loss will inject interdomain feature alignment perturbation information into the feature extractor before the gradient optimization of the segmentation

network. The data distribution shifts between source and target images mainly manifest in the shallow information, so $\mathcal{L}_{adv}$ focuses primarily on the domain-invariant features in the shallow feature space, and $\mathcal{L}_d$ is employed to individually train and optimize the discriminative network.

In addition, for weak-to-strong augmentation in consistency regularization learning, we leverage the ClassMix [32] augmentation strategy in the strongly augmented perturbations by mixing the foreground and background regions of the image to provide more diverse information about the perturbations, as illustrated in Figure 4. Compared to the commonly adopted CutMix [60] strategy, ClassMix has more advantages in maintaining the semantic integrity and the boundary information of each object in the images.

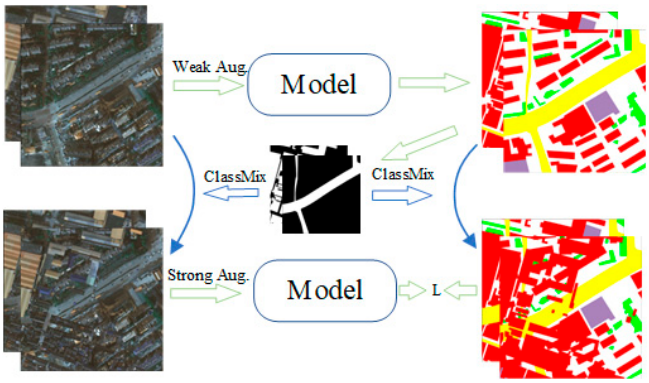

Figure 4. Weak-to-strong consistency with the introduction of ClassMix.

4. Experimental Results and Discussion

4.1. Dataset Description

To validate the segmentation performance of AdvCDA with different domain discrepancy scenarios in RSIs, three benchmark datasets are used: the Potsdam dataset, Vaihingen and LoveDA datasets.

Potsdam dataset: The Potsdam dataset consists of 38 pieces of 5 cm high-resolution RSIs with a size of 6000 × 6000, and annotated data include six interpretation categories: impervious surfaces, buildings, trees, cars, low vegetation, and background. The dataset has red, green, blue, and near-infrared bands, and we use both IRRG and RGB imaging modes in the experiments. In addition, we follow the same sample splitting method and crop the image to 512 × 512 patch size [24]. A total of 4598 samples are generated and divided into 2904 training sets and 1694 test sets [22,24,61].

Vaihingen dataset: The Vaihingen dataset contains the same interpretation categories as the Potsdam dataset, with an image resolution of 9 cm and only IRRG imaging modes. The dataset contains 33 VHR TOP images. During data preprocessing, we also crop the images to 512 × 512 size and divide 1296 images as training data and 440 images as test data [22,24,61].

LoveDA dataset: The LoveDA dataset provides both rural and urban land cover scenes and contains seven interpretation categories: building, road, water, barren, forest, agricultural land, and background. It contains 5987 0.3 m high-resolution images from three different cities, with a size of 1024 × 1024. The urban scene in this dataset contains 1156 training images, 677 validation images, and 820 test images, while the rural scene contains 2358 images, of which 1366 images are used for training and 976 are used for test data [62,63]. On the LoveDA dataset, we focused our experiments on the remote sensing cross-domain task for rural-to-urban scenes.

4.2. Experimental Settings and Evaluation Metrics

All the network architectures in our experiments were implemented using the PyTorch framework. We primarily leveraged SegFormer [64] as our typical baseline segmentation model. During the training process, the SegFormer model was optimized by AdamW [65]

with the momentum parameter set to 0.9 and the weight decay set to 10^{-2}. The initial learning rates for the encoder and decoder were set to 6×10^{-5} and 6×10^{-4}, respectively, and then the learning rate decayed linearly with iterations. We set horizontal flipping and random rotation as weak augmentation methods in consistency learning while adding RandAugment [31] and ClassMix [32] as strong augmentation methods for the weak-to-strong consistency branch.

We comprehensively evaluated the performance of the model using the mean intersection over union (mIoU), which was obtained by calculating the intersection over union (*IoU*) for each category and then averaging them. As follows, we computed the IoU for each category by a confusion matrix with three terms, true positive (*TP*), false positive (*FP*) and false negative (*FN*), in the formulation:

$$IoU = \frac{TP}{TP + FP + FN} \tag{19}$$

In addition, following the settings of [22,24,66], the *F1* score was used to further evaluate the proposed method, which is defined as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \tag{20}$$

4.3. Comparisons with Other Methods

To verify the effectiveness of AdvCDA, we performed experiments in three kinds of domain discrepancy scenarios that are commonly observed in RSIs, including cross-spectral discrepancy scenarios, cross-space discrepancy scenarios, and complex domain discrepancy scenarios.

4.3.1. Cross-Space Scenarios

We conducted experiments with the Potsdam (IRRG) dataset as the source domain and the Vaihingen (IRRG) dataset as the target domain. We focused primarily on practically meaningful goals, so five categories were evaluated: impervious surfaces, buildings, low vegetation, trees, and cars [23,53,67]. One can find that objects in the two datasets have significant characteristic differences, such that there are large buildings, narrow streets, and dense residential structures within the Potsdam dataset images. In contrast, the Vaihingen dataset images contain mostly free-standing structures and small buildings; the results are shown in Table 1. Compared to the existing state-of-the-art (SOTA) method, AdvCDA improves the mIoU performance by 2.84% and the mFscore performance by 2.01%. In terms of each category, our method significantly improved the results for all categories. In terms of categories, the best IoU and F1 performance of AdvCDA is achieved for impervious surfaces, cars, buildings, and trees, indicating that the proposed DA method has a more robust and stable domain transfer ability. Note that both ST-DASegNet and DAFormer, the best performance among the compared methods, used transformer (SegFormer) as the baseline, and an equally transformer-based model is used for the best performance of AdvCDA. Furthermore, as shown in Figure 5 for the qualitative visualization, it can be intuitively found that the proposed AdvCDA performed strongly in the Potsdam (IRRG) $\rightarrow$ Vaihingen (IRRG) cross-domain task.

Geographical discrepancies arising from urban and rural areas are also very common in practical remote sensing applications. Urban areas cover many building clusters and dense road grids compared to rural areas with more agricultural land, increasing the difficulty of generalization of the model. As shown in Table 2, we give the type of architecture for each method. Obviously, for this rural$\rightarrow$urban cross-domain task, one can find that the combination of ST and AT outperforms purely ST methods, while purely AT methods show the lowest performance. Furthermore, using SegFormer as the baseline model, the mIoU performance of our DA approach outperformed the baseline by 9.09%. Among the compared DA methods, DAFormer, ST-DASegNet, and our method are all transformer-

based networks, which obviously achieve a significant advantage over CNN-based DA methods, while our method achieved the optimal comprehensive performance.

Table 1. Comparison results of AdvCDA with existing DA methods. The mIoU performance is validated on the test set of the Potsdam (IRRG) → Vaihingen (IRRG) task. The best results are highlighted in bold.

Method	Architecture	Impervious Surfaces		Car		Tree		Low Vegetation		Building		Overall	
		IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	mFscore
AdaptSegNet [25]	ResNet-Based	54.39	70.39	6.40	11.99	52.65	68.96	28.98	44.91	63.14	77.40	41.11	54.73
FADA [41]		60.01	75.00	26.79	42.25	58.06	73.46	47.23	64.16	70.96	83.01	52.61	67.58
DualGAN [37]		49.41	66.13	34.34	51.09	57.66	73.14	38.87	55.97	62.30	76.77	48.52	64.62
ResiDualGAN [22]		72.29	83.89	57.01	72.51	63.81	77.88	49.69	66.29	80.57	89.23	64.67	77.96
Zhang et al. [66]		67.74	80.13	44.90	61.94	55.03	71.90	47.02	64.16	76.75	86.65	58.29	72.96
ST-DASegNet [24]	Transformer-based	74.43	85.36	43.38	60.49	67.36	80.49	48.57	65.37	85.23	92.03	63.79	76.75
DAFormer [56]		76.01	86.54	51.40	70.69	**68.43**	**80.62**	51.23	67.81	81.99	88.40	65.81	78.81
AdvCDA		**77.19**	**87.13**	**61.63**	**76.26**	65.78	79.36	**52.21**	**68.60**	**86.44**	**92.73**	**68.65**	**80.82**

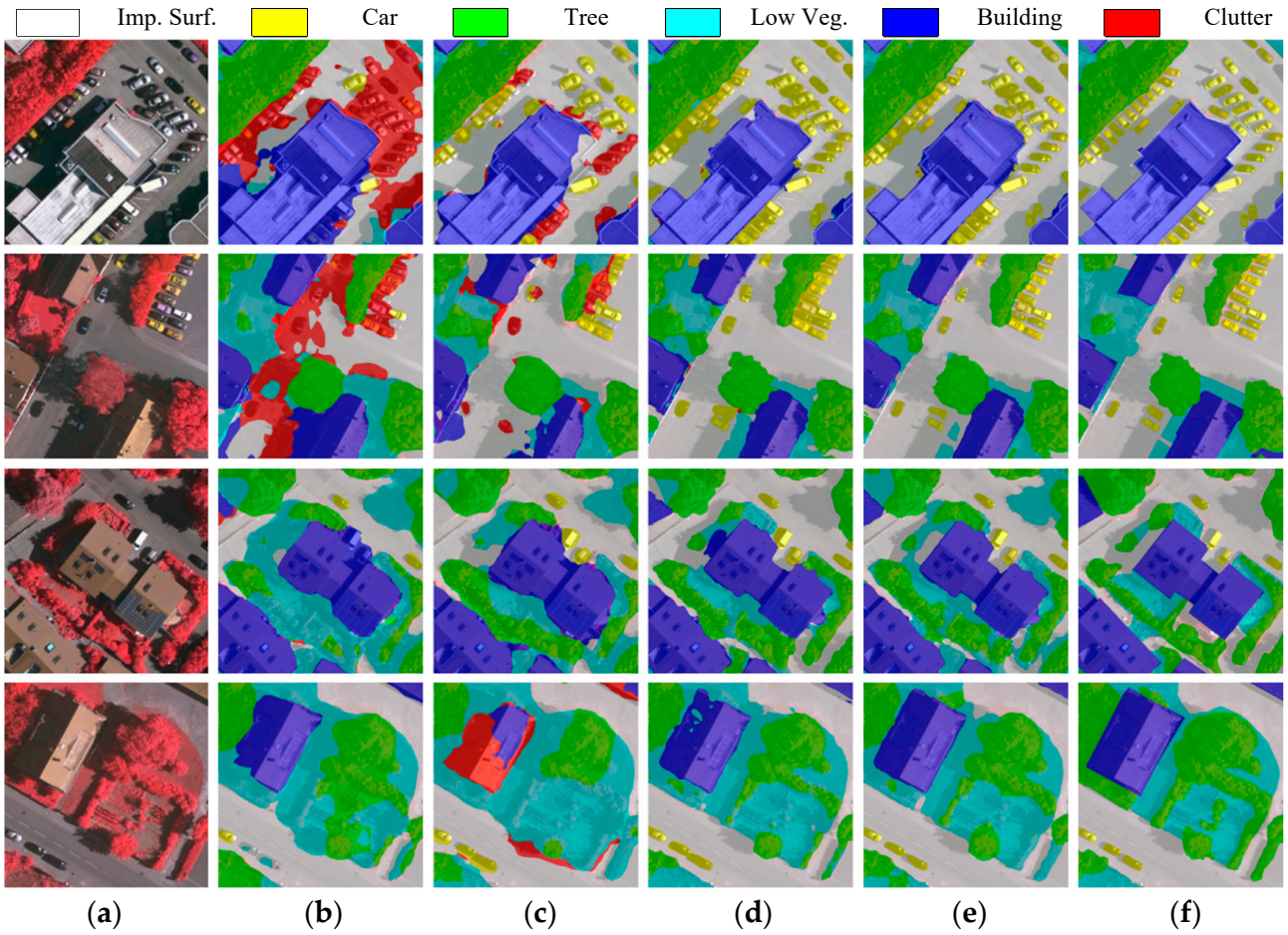

Figure 5. Qualitative comparison of AdvCDA with other methods in the Potsdam (IRRG) → Vaihingen (IRRG) task. (**a**) Target images. (**b**) AdaptSegNet. (**c**) FADA. (**d**) DAFormer. (**e**) AdvCDA. (**f**) Ground truth.

We provide the visualization results for the rural → urban task in Figure 6. Since the ground truth is not available for the test set, we show the validation data for the LoveDA dataset. It can be found that AdvCDA has more advantages in preserving the integrity and edge accuracy of the objects.

Table 2. Comparison results of AdvCDA with existing DA methods. The mIoU performance is validated on the test set of the rural → urban task. The best results are highlighted in bold.

Method	Arch.	Background	Building	Road	Water	Barren	Forest	Agriculture	mIoU
SegFormer [64]	baseline	47.14	53.28	55.50	52.93	18.52	35.37	28.97	41.67
AdaptSegNet [25]	AT	42.35	23.73	15.61	81.95	13.62	28.70	22.05	32.68
FADA [41]	AT	43.89	12.62	12.76	80.37	12.70	32.76	24.79	31.41
PyCDA [68]	ST	38.04	35.86	45.51	74.87	7.71	40.39	11.39	36.25
CBST [26]	ST	48.37	46.10	35.79	80.05	19.18	29.69	30.05	41.32
IAST [27]	ST	48.57	31.51	28.73	86.01	**20.29**	31.77	36.50	40.48
DCA [63]	ST	45.82	49.60	51.65	80.88	16.70	42.93	36.92	46.36
DAFormer [56]	ST	50.94	**56.66**	62.83	**89.41**	11.99	45.81	25.26	48.99
ST-DASegNet [24]	AT + ST	**51.01**	54.23	60.52	87.31	15.18	**47.43**	36.26	50.28
AdvCDA	AT + ST	50.81	56.12	58.38	87.87	15.85	41.88	**44.40**	**50.76**

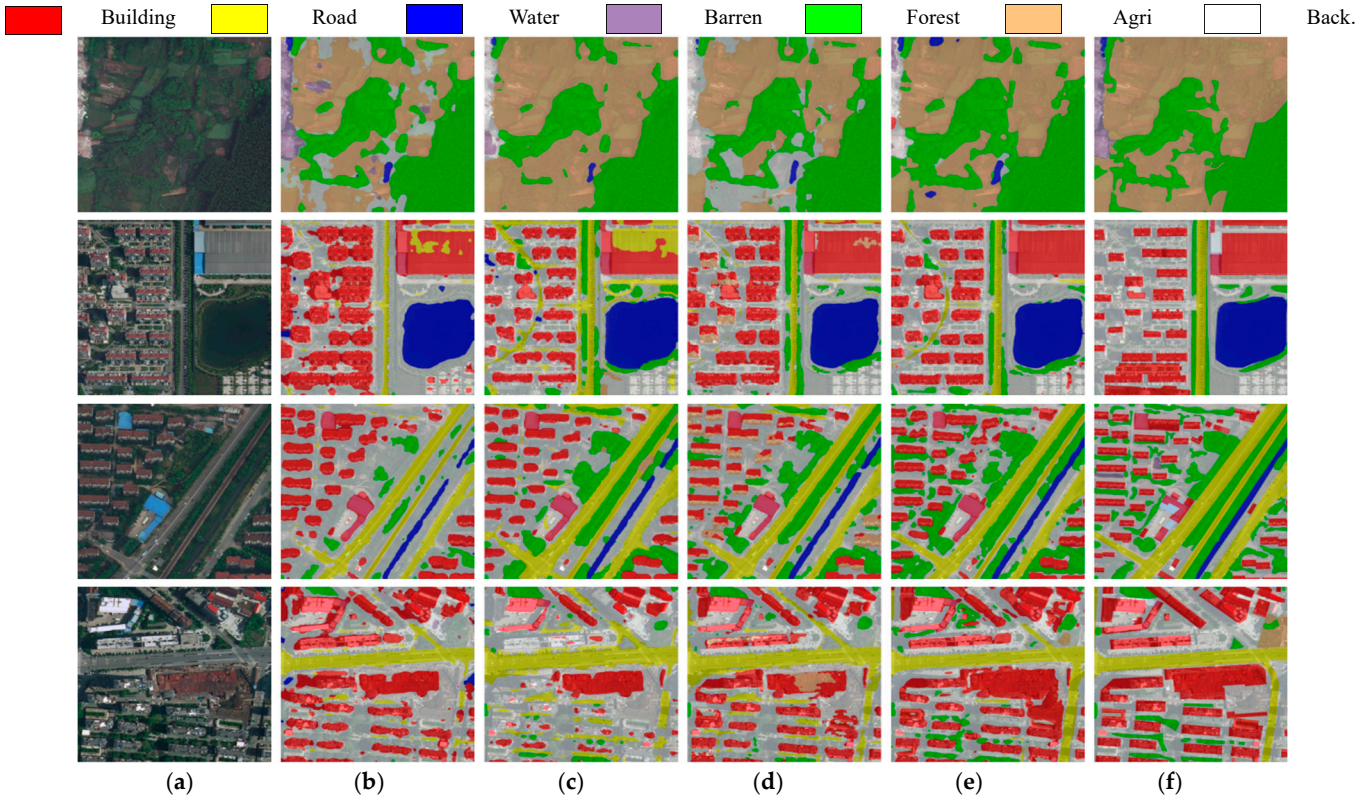

Figure 6. Qualitative comparison of AdvCDA with other methods in the rural→urban task. (**a**) Target images. (**b**) Baseline with SegFormer. (**c**) FADA. (**d**) DAFormer. (**e**) AdvCDA. (**f**) Ground truth.

4.3.2. Cross-Spectral Scenarios

The different spectral bands of RSIs also cause large data distribution shifts, and thus we conducted the Potsdam (RGB) → Potsdam (IRRG) cross-spectral scene task for comparison experiments to validate the effectiveness of the model. Table 3 shows that spectral variation produces large impacts on low vegetation and cars, with large performance differences between different DA methods, and small impacts on the performance of buildings. Figure 7 shows the visualization results of different DA methods. In terms of visual features, the differences in cross-band scenes are mainly shown in spectral features and color variation, with less impact on the shape and texture of the target objects. This indicates that the model focuses on different directions of feature learning for various classes, while purely ST methods, such as IAST, or purely AT methods, such as AdaptSegNet, do not perform well stably. In contrast, AdvCDA can constrain the model to adaptively optimize the learning direction according to different scenarios by estimating the confidence of the two streams, weak-to-strong consistency, and adversarial perturbation consistency, and

it achieves the best performance for buildings, impervious surfaces, low vegetation, and trees, with the overall performance of 1.61% and 1.13% for mIoU and mFscore better than that of the best comparative method.

Table 3. Comparison results of AdvCDA with existing DA methods. The mIoU performance is validated on the test set of the Potsdam (RGB) → Potsdam (IRRG) task. The best results are highlighted in bold.

Method	Architecture	Impervious Surfaces		Car		Tree		Low Vegetation		Building		Overall	
		IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	mFscore
AdaptSegNet [25]	ResNet-Based	73.80	84.92	69.56	82.05	67.18	80.37	51.19	67.71	80.81	89.39	68.51	80.89
FADA [41]		75.91	86.31	66.83	80.12	68.77	81.49	62.06	76.59	83.97	91.28	71.51	83.16
PyCDA [68]		76.41	86.62	73.69	84.85	69.31	81.87	63.49	77.67	82.70	90.53	73.12	84.31
IAST [27]		76.20	86.49	66.81	80.10	68.26	81.14	54.29	70.37	83.67	91.11	69.85	81.84
DACS [69]		74.09	85.12	71.16	83.15	66.83	90.11	63.44	77.63	81.14	89.59	71.33	85.12
DecoupleNet [28]		76.21	86.50	72.97	84.37	68.10	81.02	59.50	74.61	82.25	90.26	71.81	83.35
DAFormer [56]	Transformer-based	77.94	87.28	**86.59**	**90.02**	**71.57**	**83.80**	67.94	81.99	80.23	90.09	76.85	86.64
AdvCDA		**80.06**	**88.92**	81.72	89.94	68.66	81.42	**73.56**	**84.76**	**88.32**	**93.80**	**78.46**	**87.77**

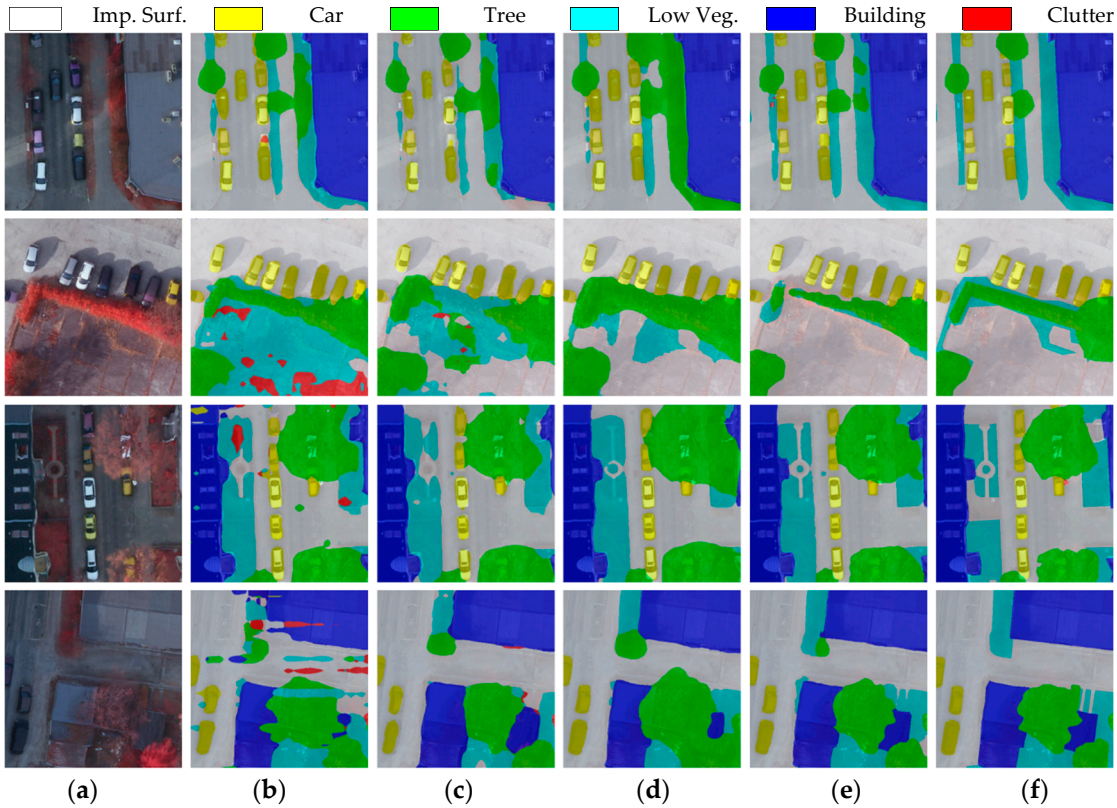

Figure 7. Qualitative comparison of AdvCDA with other methods in the Potsdam(RGB)→ Potsdam(IRRG) task. (**a**) Target images. (**b**) AdaptSegNet. (**c**) FADA. (**d**) DAFormer. (**e**) AdvCDA. (**f**) Ground truth.

4.3.3. Complex Domain Discrepancy Scenarios

To further validate the effectiveness of the model in complex domain discrepancy scenarios that represent more difficult and large data distribution shifts, we conducted experiments on the Potsdam (RGB) → Vaihingen (IRRG) task. Note that this task involves both cross-spectral and cross-spatial discrepancies, and the same classes also have large-scale variations, which pose a greater challenge to the generalization and stability of the model. As shown in Table 4, the quantitative comparison results between AdvCDA and

several existing DA methods are presented. Compared to the simple cross-spectral and cross-space scenarios, our approach has greater advantages in complex domain discrepancy scenarios. AdvCDA outperforms the best comparison method by 4.03% for mIoU and 4.26% for mFscore. The experimental results demonstrate that AdvCDA also achieves the best performance in cross-spectral and cross-spatial complex scenarios.

Table 4. Comparison results of AdvCDA with existing DA methods. The mIoU performance is validated on the test set of the Potsdam (RGB) → Vaihingen (IRRG) task. The best results are highlighted in bold.

Method	Architecture	Impervious Surfaces		Car		Tree		Low Vegetation		Building		Overall	
		IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	mIoU	mFscore
AdaptSegNet [25]	ResNet-Based	51.26	67.77	10.25	18.54	51.51	68.02	12.75	22.61	60.72	75.55	37.30	50.50
FADA [41]		56.66	72.34	27.36	42.97	34.39	51.18	36.34	53.31	65.89	79.44	44.13	59.85
ProDA [46]		49.04	66.11	31.56	48.16	49.11	65.86	32.44	49.06	68.94	81.89	46.22	62.22
Bai et al. [56]		62.40	76.90	38.90	56.00	53.90	70.00	35.10	51.90	74.80	85.60	53.02	68.08
DualGAN [37]		49.16	61.33	40.31	57.88	55.82	70.66	27.85	42.17	65.44	83.00	47.72	63.01
ResiDualGAN [22]		55.54	71.36	48.49	65.19	57.79	73.21	29.15	44.97	78.97	88.23	53.99	68.59
Zhang et al. [66]		64.47	77.76	43.43	60.05	52.83	69.62	38.37	55.94	76.87	86.95	55.19	70.06
DAFormer [56]	Transformer-based	58.85	75.50	46.33	65.54	62.94	79.49	18.89	27.46	74.20	86.50	52.24	66.90
ST-DASegNet [24]		68.36	81.28	43.15	60.28	**64.65**	**78.31**	**34.69**	47.08	84.09	91.33	58.99	71.66
AdvCDA		**72.31**	**83.93**	**61.69**	**76.31**	61.54	76.19	34.34	**51.12**	**85.25**	**92.04**	**63.02**	**75.92**

4.4. Ablation Study and Analysis

4.4.1. Design of Feature Alignment

In the comparison experiments, AdvCDA achieves significant advantages in various domain discrepancy scenarios that are common in RSIs, which proves the effectiveness of AdvCDA and the stability that can be adapted to different remote sensing scenario tasks. Intuitively, in contrast to the pure ST approach, the key component of the proposed joint ST and AT paradigm is the additional adversarial alignment idea to capture the domain-invariant feature space and promote the generalization ability of the model. Therefore, we investigate the impact of the feature alignment module in the AT process when it acts on different feature layers in the segmentation network architecture. With transformer-based SegFormer [64] as the backbone, stage-1 to stage-4 of the backbone and the output layers were used as inputs to the discriminative network. AT only updates the network layer gradients prior to the current feature layer for feature alignment. The results shown in Figure 8 indicate that conducting feature alignment at stage-2 achieves the best DA results, whereas the model performance tends to decrease when the feature alignment module is applied to the deep network, such as stage-3 and stage-4, which might be that the AT overly interferes with the feature parameters, resulting in the over robustness of the model. Feature alignment in the output space is commonly employed in AT methods to maintain the consistency of the output layouts of the source and target domains. However, experiments show that AdvCDA provides adversarial feature interference at stage-2 to achieve the best DA performance.

4.4.2. Effectiveness Analysis of Each Component

To validate the effectiveness of each component for the proposed AdvCDA, we conducted ablation experiments in Table 5. FixMatch leverages weak-to-strong consistency regularization ideas for ST, while our approach generalizes consistency regularization ideas to DA tasks. Therefore, the key idea is to leverage consistency regularization to improve the stability of AT, thus combining ST and AT methods to boost DA performance, which we dubbed adversarial perturbation consistency (AdvC). The adversarial perturbation consistency acts on the feature layer of the model to complement the advantages of weak-to-strong perturbation consistency at the input level, and the mIoU performance of the model is improved by 3.26%. In addition, confidence estimation (CB) on the two streams of

weak-to-strong and adversarial perturbation consistency from adaptive optimization learning is crucial for AdvCDA to maintain the stability of its performance in different domain discrepancy scenarios, where the mIoU performance of the model is further improved to 68.65% in this Potsdam (IRRG) → Vaihingen (IRRG) cross-domain task.

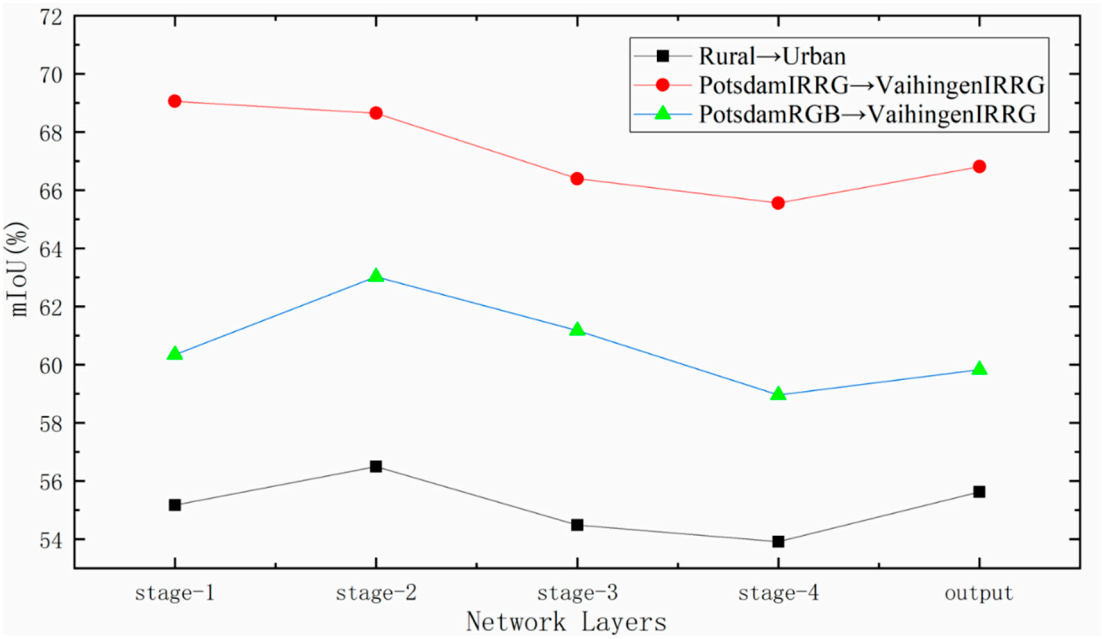

Figure 8. Performance of feature alignment modules on different network layers.

Table 5. Ablation experiments on the effectiveness of each component with the proposed approach.

Methods	FixMatch	Lce	Ladv	AdvC	CB	mIoU
SourceOnly		✔				56.30
FixMatch + ClassMix	✔	✔				61.43
AdvCDA (w/o AdvC)	✔	✔	✔			62.29
AdvCDA (w/o CB)	✔	✔	✔	✔		65.55
AdvCDA	✔	✔	✔	✔	✔	68.65
Target-only	-	-	-		-	76.10

4.4.3. Effectiveness of Augmentation Perturbation Strategies

Table 6 shows the performance obtained by imposing different augmentation perturbation strategies on strong augmentation branches for different cross-domain tasks. The baseline is augmented with horizontal flipping, rotation, and other common augmentation methods used in semantic segmentation models. One can find that although the CutMix strategy can effectively improve the generalization ability of the model in the semi-supervised learning task [52], it instead degrades the performance of the model in the cross-domain scenes. We assume that CutMix augmentation corrupts the local semantic integrity of the classes and that the loss of semantic information further enlarges the discrepancies between the source and target domains. In contrast, ClassMix provides complete object boundaries, which mix images from the source and target domains for augmentation, and the model performance is further improved. In addition, the RA, which is a commonly used strategy for weak-to-strong consistency learning, improves the mIoU performance by 2.48% and 1.02% in the Potsdam RGB → Vaihingen IRRG and rural → urban cross-domain tasks, respectively.

Table 6. The performance of applying different augmentation perturbation strategies to strong augmentation branches in the tasks Potsdam RGB → Vaihingen IRRG and rural → urban.

Augmentation Strategy	$\mathbf{mIoU}_{\mathbf{PotsdamRGB \rightarrow VaihingenIRRG}}$	$\mathbf{mIoU}_{\mathbf{rural \rightarrow urban(val)}}$
Baseline	60.35	54.58
Baseline (w/CutMix)	59.83	54.15
Baseline (w/ClassMix)	61.18	55.33
RA (w/o ClassMix)	62.83	55.60
RA (w/ClassMix)	63.02	56. 17

5. Conclusions

In this paper, we propose a novel DA semantic segmentation method based on adversarial perturbation consistency to solve the distribution discrepancies among different domains in RSIs. In the network architecture, we design a weak-to-strong consistency stream at the input level and an adversarial perturbation consistency stream at the feature level, aiming to further improve the domain generalization performance of the model through joint AT and ST. Crucially, considering the inherent instability problem of AT, we use consistency regularization to provide high-quality pseudolabels to prevent over robustness that can easily be induced by over-perturbation of the feature space for AT. Furthermore, we propose a confidence estimation mechanism to adaptively assign the optimization weights for each stream and thus guide the model to train better for domain transfer. The effectiveness of the proposed method is validated on three different remote sensing benchmark datasets with cross-space, cross-spectral, and complex domain difference scenarios. Extensive experiments demonstrate the performance superiority of AdvCDA compared to existing UDA methods. Notably, AdvCDA improves mIoU performance by 4.03% and mFscore performance by 4.26% in Potsdam (RGB) → Vaihingen (IRRG) complex domain discrepancy scenarios against existing SOTA methods, further demonstrating that the design of the adversarial perturbation consistency and confidence estimation mechanisms enables the model to obtain effectively adaptive optimization in complex unseen scenarios. Nevertheless, it remains the case that our approach focuses on specific target domains and mainly studies the transfer training process of domain-specific knowledge in known target domains. In future work, we will further explore domain generalized feature learning in the case of multi-target domains or unseen target domains.

Author Contributions: Conceptualization, Z.X.; Methodology, Z.X. and Y.D.; Software, Z.X. and J.C.; Validation, Z.X., Y.M. and J.C.; Formal analysis, Z.X., J.C. and Y.K.; Investigation, D.L. and Y.K.; Resources, Y.M. and Y.K.; Writing—original draft, Z.X.; Writing—review & editing, Y.M., J.C. and A.Y.; Visualization, Y.D. and D.L.; Supervision, Y.M. and J.C.; Project administration, Y.M. and J.C.; Funding acquisition, Y.M., J.C. and A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China under grant number 2021YFB3900504.

Data Availability Statement: The experiments in this article are based on open source data sets, and no new data is created.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhu, Q.; Sun, X.; Zhong, Y.; Zhang, L. High-Resolution Remote Sensing Image Scene Understanding: A Review. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3061–3064. [CrossRef]
2. Kotaridis, I.; Lazaridou, M. Remote Sensing Image Segmentation Advances: A Meta-Analysis. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 309–322. [CrossRef]
3. Zhao, C.; Qin, B.; Feng, S.; Zhu, W.; Sun, W.; Li, W.; Jia, X. Hyperspectral Image Classification with Multi-Attention Transformer and Adaptive Superpixel Segmentation-Based Active Learning. *IEEE Trans. Image Process.* **2023**, *32*, 3606–3621. [CrossRef] [PubMed]

4. Zhao, C.; Zhu, W.; Feng, S. Superpixel Guided Deformable Convolution Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2022**, *31*, 3838–3851. [CrossRef] [PubMed]
5. Yang, X.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Li, B.; Chen, P. An Attention-Fused Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 238–262. [CrossRef]
6. Marcos, D.; Volpi, M.; Kellenberger, B.; Tuia, D. Land Cover Mapping at Very High Resolution with Rotation Equivariant CNNs: Towards Small yet Accurate Models. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 96–107. [CrossRef]
7. Soto Vega, P.J.; da Costa, G.A.O.P.; Feitosa, R.Q.; Ortega Adarme, M.X.; de Almeida, C.A.; Heipke, C.; Rottensteiner, F. An Unsupervised Domain Adaptation Approach for Change Detection and Its Application to Deforestation Mapping in Tropical Biomes. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 113–128. [CrossRef]
8. Deng, Y.; Chen, J.; Yi, S.; Yue, A.; Meng, Y.; Chen, J.; Zhang, Y. Feature-Guided Multitask Change Detection Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9667–9679. [CrossRef]
9. Wilson, G.; Cook, D.J. A Survey of Unsupervised Deep Domain Adaptation. *arXiv* **2018**, arXiv:1812.02849. [CrossRef]
10. Zhao, C.; Qin, B.; Feng, S.; Zhu, W.; Zhang, L.; Ren, J. An Unsupervised Domain Adaptation Method Towards Multi-Level Features and Decision Boundaries for Cross-Scene Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
11. Jiang, Z.; Li, Y.; Yang, C.; Gao, P.; Wang, Y.; Tai, Y.; Wang, C. Prototypical Contrast Adaptation for Domain Adaptive Semantic Segmentation. *arXiv* **2022**, arXiv:2207.06654.
12. Yan, L.; Fan, B.; Xiang, S.; Pan, C. Adversarial Domain Adaptation with a Domain Similarity Discriminator for Semantic Segmentation of Urban Areas. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 1583–1587. [CrossRef]
13. Liu, W.; Su, F. Unsupervised Adversarial Domain Adaptation Network for Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1978–1982. [CrossRef]
14. Tong, X.Y.; Xia, G.S.; Zhu, X.X. Enabling Country-Scale Land Cover Mapping with Meter-Resolution Satellite Imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 178–196. [CrossRef] [PubMed]
15. Wang, D.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model. *arXiv* **2023**, arXiv:2305.02034.
16. Zhang, L.; Xia, G.S.; Wu, T.; Lin, L.; Tai, X.C. Deep Learning for Remote Sensing Image Understanding. *J. Sens.* **2016**, *2016*, 7954154. [CrossRef]
17. Tasar, O.; Happy, S.L.; Tarabalka, Y.; Alliez, P. ColorMapGAN: Unsupervised Domain Adaptation for Semantic Segmentation Using Color Mapping Generative Adversarial Networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7178–7193. [CrossRef]
18. Jiang, J.; Shu, Y.; Wang, J.; Long, M. Transferability in Deep Learning: A Survey. *arXiv* **2022**, arXiv:2201.05867.
19. Zhao, S.; Yue, X.; Zhang, S.; Li, B.; Zhao, H.; Wu, B.; Krishna, R.; Gonzalez, J.E.; Sangiovanni-Vincentelli, A.L.; Seshia, S.A.; et al. A Review of Single-Source Deep Unsupervised Visual Domain Adaptation. *IEEE Trans. Neural Networks Learn. Syst.* **2022**, *33*, 473–493. [CrossRef]
20. Bai, L.; Du, S.; Zhang, X.; Wang, H.; Liu, B.; Ouyang, S. Domain Adaptation for Remote Sensing Image Semantic Segmentation: An Integrated Approach of Contrastive Learning and Adversarial Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–3. [CrossRef]
21. Tuia, D.; Persello, C.; Bruzzone, L. Recent Advances in Domain Adaptation for the Classification of Remote Sensing Data. *IEEE Geosci. Remote Sens. Mag.* **2021**, *4*, 41–57. [CrossRef]
22. Zhao, Y.; Guo, P.; Sun, Z.; Chen, X.; Gao, H. ResiDualGAN: Resize-Residual DualGAN for Cross-Domain Remote Sensing Images Semantic Segmentation. *Remote Sens.* **2023**, *15*, 1428. [CrossRef]
23. Deng, X.; Zhu, Y.; Tian, Y.; Newsam, S. Scale Aware Adaptation for Land-Cover Classification in Remote Sensing Imagery. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 2159–2168. [CrossRef]
24. Zhao, Q.; Lyu, S.; Liu, B.; Chen, L.; Zhao, H. Self-Training Guided Disentangled Adaptation for Cross-Domain Remote Sensing Image Semantic Segmentation. *arXiv* **2023**, arXiv:2301.05526.
25. Tsai, Y.H.; Hung, W.C.; Schulter, S.; Sohn, K.; Yang, M.H.; Chandraker, M. Learning to Adapt Structured Output Space for Semantic Segmentation. *arXiv* **2018**, arXiv:1802.10349.
26. Zou, Y.; Yu, Z.; Vijaya Kumar, B.V.K.; Wang, J. Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. In Proceedings of the 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 297–313. [CrossRef]
27. Mei, K.; Zhu, C.; Zou, J.; Zhang, S. Instance Adaptive Self-Training for Unsupervised Domain Adaptation. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 415–430. [CrossRef]
28. Lai, X.; Tian, Z.; Xu, X.; Chen, Y.; Liu, S.; Zhao, H.; Wang, L.; Jia, J. DecoupleNet: Decoupled Network for Domain Adaptive Semantic Segmentation. In Proceedings of the 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 369–387. [CrossRef]
29. Chen, R.; Rong, Y.; Guo, S.; Han, J.; Sun, F.; Xu, T.; Huang, W. Smoothing Matters: Momentum Transformer for Domain Adaptive Semantic Segmentation. *arXiv* **2022**, arXiv:2203.07988.
30. Sohn, K.; Berthelot, D.; Zizhao, C.L.; Nicholas, Z.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv* **2020**, arXiv:2001.07685.

31. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical Automated Data Augmentation with a Reduced Search Space. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 3008–3017. [CrossRef]
32. Olsson, V. ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning. *arXiv* **2020**, arXiv:2007.07936.
33. Xu, Q.; Ma, Y.; Wu, J.; Long, C.; Huang, X. CDAda: A Curriculum Domain Adaptation for Nighttime Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2962–2971. [CrossRef]
34. Tasar, O.; Happy, S.L.; Tarabalka, Y.; Alliez, P. SEMI2I: Semantically Consistent Image-to-Image Translation for Domain Adaptation of Remote Sensing Data. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 1837–1840. [CrossRef]
35. Hoffman, J.; Tzeng, E.; Park, T.; Phillip, J.Z.; Kate, I.; Alexei, S.; Darrell, T.; Chang, W.G.W.L.; Wang, H.P.; Peng, W.H.; et al. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7346–7354.
36. Tasar, O.; Giros, A.; Tarabalka, Y.; Alliez, P.; Clerc, S. DAugNet: Unsupervised, Multisource, Multitarget, and Life-Long Domain Adaptation for Semantic Segmentation of Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1067–1081. [CrossRef]
37. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2868–2876. [CrossRef]
38. Yang, Y.; Soatto, S. FDA: Fourier Domain Adaptation for Semantic Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4084–4094. [CrossRef]
39. Peng, D.; Guan, H.; Zang, Y.; Bruzzone, L. Full-Level Domain Adaptation for Building Extraction in Very-High-Resolution Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
40. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *Adv. Comput. Vis. Pattern Recognit.* **2017**, *17*, 189–209. [CrossRef]
41. Wang, H.; Shen, T.; Zhang, W.; Duan, L.Y.; Mei, T. Classes Matter: A Fine-Grained Adversarial Approach to Cross-Domain Semantic Segmentation. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 1–17. [CrossRef]
42. Xu, Q.; Yuan, X.; Ouyang, C. Class-Aware Domain Adaptation for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 1–18. [CrossRef]
43. Chen, J.; Zhu, J.; Guo, Y.; Sun, G.; Zhang, Y.; Deng, M. Unsupervised Domain Adaptation for Semantic Segmentation of High-Resolution Remote Sensing Imagery Driven by Category-Certainty Attention. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
44. Lu, X.; Zhong, Y.; Zheng, Z.; Wang, J. Cross-Domain Road Detection Based on Global-Local Adversarial Learning Framework from Very High Resolution Satellite Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 296–312. [CrossRef]
45. Chapelle, O.; Schölkopf, B.; Zien, E.A. Semi-Supervised Learning. *IEEE Trans. Neural Netw.* **2009**, *20*, 2015975.
46. Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; Wen, F. Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2101.10979.
47. Vayyat, M.; Kasi, J.; Bhattacharya, A.; Ahmed, S.; Tallamraju, R. CLUDA: Contrastive Learning in Unsupervised Domain Adaptation for Semantic Segmentation. *arXiv* **2022**, arXiv:2208.14227.
48. Zhang, L.; Lan, M.; Zhang, J.; Tao, D. Stagewise Unsupervised Domain Adaptation With Adversarial Self-Training for Road Segmentation of Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
49. Chen, X.; Yuan, Y.; Zeng, G.; Wang, J. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2613–2622. [CrossRef]
50. Yang, L.; Qi, L.; Feng, L.; Zhang, W.; Shi, Y. Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation. *arXiv* **2022**, arXiv:2208.09910.
51. Kuo, C.W.; Ma, C.Y.; Huang, J.B.; Kira, Z. FeatMatch: Feature-Based Augmentation for Semi-Supervised Learning. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 479–495. [CrossRef]
52. Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; Carneiro, G. Perturbed and Strict Mean Teachers for Semi-Supervised Semantic Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4248–4257. [CrossRef]
53. Xi, Z.; He, X.; Meng, Y.; Yue, A.; Chen, J.; Deng, Y.; Chen, J. A Multilevel-Guided Curriculum Domain Adaptation Approach to Semantic Segmentation for High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
54. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised Data Augmentation for Consistency Training. *arXiv* **2020**, arXiv:1904.12848.
55. Kim, J.; Min, Y.; Kim, D.; Lee, G.; Seo, J.; Ryoo, K.; Kim, S. ConMatch: Semi-Supervised Learning with Confidence-Guided Consistency Regularization. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXII. Springer: Berlin/Heidelberg, Germany, 2022; pp. 674–690.
56. Hoyer, L.; Dai, D.; Van Gool, L. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2111.14887.

57. Hoyer, L.; Dai, D.; Van Gool, L. HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation. *arXiv* **2022**, arXiv:2204.13132.
58. Mittal, S.; Tatarchenko, M.; Brox, T. Semi-Supervised Semantic Segmentation with High- And Low-Level Consistency. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1369–1379. [CrossRef] [PubMed]
59. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014; pp. 1–10.
60. Yun, S.; Han, D.; Chun, S.; Oh, S.J.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6022–6031. [CrossRef]
61. Gao, H.; Zhao, Y.; Guo, P.; Sun, Z.; Chen, X.; Tang, Y. Cycle and Self-Supervised Consistency Training for Adapting Semantic Segmentation of Aerial Images. *Remote Sens.* **2022**, *14*, 1527. [CrossRef]
62. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. *arXiv* **2021**, arXiv:2110.08733.
63. Wu, L.; Lu, M.; Fang, L. Deep Covariance Alignment for Domain Adaptive Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
64. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
65. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
66. Zhang, B.; Chen, T.; Wang, B. Curriculum-Style Local-to-Global Adaptation for Cross-Domain Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
67. Liu, W.; Su, F.; Jin, X.; Li, H.; Qin, R. Bispace Domain Adaptation Network for Remotely Sensed Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
68. Lian, Q.; Lv, F.; Duan, L.; Gong, B. Constructing Self-Motivated Pyramid Curriculums for Cross-Domain Semantic Segmentation: A Non-Adversarial Approach. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6758–6767.
69. Tranheden, W.; Olsson, V.; Pinto, J.; Svensson, L. DACS: Domain Adaptation via Cross-Domain Mixed Sampling. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Virtual, 5–9 January 2021; pp. 1378–1388. [CrossRef]