remote sensing

MDPI

Article

Cycle and Self-Supervised Consistency Training for Adapting Semantic Segmentation of Aerial Images

Han Gao [1], Yang Zhao [1], Peng Guo [1], Zihao Sun [1], Xiuwan Chen [1] and Yunwei Tang [2,3,*]

[1] Institute of Remote Sensing and Geographic Information System, Peking University, Beijing 100871, China; hgao@pku.edu.cn (H.G.); zy_@pku.edu.cn (Y.Z.); peng_guo@pku.edu.cn (P.G.); sunzihao211@gmail.com (Z.S.); xwchen@pku.edu.cn (X.C.)
[2] International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China
[3] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: tangyw@aircas.ac.cn; Tel.: +86-010-8217-8109

Abstract: Semantic segmentation is a critical problem for many remote sensing (RS) image applications. Benefiting from large-scale pixel-level labeled data and the continuous evolution of deep neural network architectures, the performance of semantic segmentation approaches has been constantly improved. However, deploying a well-trained model on unseen and diverse testing environments remains a major challenge: a large gap between data distributions in train and test domains results in severe performance loss, while manual dense labeling is costly and not scalable. To this end, we proposed an unsupervised domain adaptation framework for RS image semantic segmentation that is both practical and effective. The framework is supported by the consistency principle, including the cycle consistency in the input space and self-supervised consistency in the training stage. Specifically, we introduce cycle-consistent generative adversarial networks to reduce the discrepancy between source and target distributions by translating one into the other. The translated source data then drive a pipeline of supervised semantic segmentation model training. We enforce consistency of model predictions across target image transformations in order to provide self-supervision for the unlabeled target data. Experiments and extensive ablation studies demonstrate the effectiveness of the proposed approach on two challenging benchmarks, on which we achieve up to 9.95% and 7.53% improvements in accuracy over the state-of-the-art methods, respectively.

Keywords: unsupervised domain adaptation; semantic segmentation; self-supervision; remote sensing image

Citation: 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. https://doi.org/10.3390/rs14071527

Academic Editors: Mi Wang, Hanwen Yu, Jianlai Chen, Ying Zhu and Giuseppe Scarpa

Received: 28 January 2022
Accepted: 18 March 2022
Published: 22 March 2022

1. Introduction

Semantic segmentation of remote sensing (RS) images has attracted increasing attention and research interest. Many applications, such as environmental monitoring, crop production, and urban planning, need accurate and efficient segmentation mechanisms [1–5]. These demands coincide with the rise of deep learning methods in the RS field and application target-related RS image interpretation, including segmentation, object detection, and classification [6–9]. Semantic segmentation assigns a predefined ground truth label to each pixel in an image by clustering parts of an image that belong to the same object class [10], and usually applies end-to-end dense prediction networks to achieve pixelwise prediction. However, dense prediction architectures rely on pixel-level annotations of all categories to extract rich semantics and locate object boundaries accurately. Obtaining such dense annotations for semantic segmentation is notoriously laborious and expensive, and is the major limitation of semantic segmentation methods. Moreover, a trained model suffers from performance decreases in practical tasks due to the complexity of RS data, such as in the diversity of acquisition conditions, varied

geospatial locations, and different ground sampling distances [11]. The requirement to develop generalizable models with unlabeled data has motivated work on deep domain adaptation (DA) approaches for semantic segmentation. DA is a special case of transfer learning that uses labeled data in one or more relevant source domains to implement new tasks in the target domain [12]. It aims to improve the generalization ability of the model across domains and realize the cross-domain transfer reuse of domain-invariant knowledge. DA methods can be classified into supervised, semisupervised, and unsupervised DA according to whether the target domain has labels. Obviously, unsupervised domain adaptation (UDA) is a promising approach for semantic segmentation and other tasks for which large-scale data annotation is costly and time-consuming.

The UDA methods for semantic segmentation have been extended to address issues in the RS community. According to existing works in the RS fields, UDA methods can roughly be categorized into three methods: generative-based method, adversarial learning, and self-training [13]. Specifically, generative-based methods translate source data to the visual appearances of a target domain by performing distribution alignment in the image space by using generative adversarial networks (GANs). Adversarial training aims to minimize the discrepancy between source and target feature distributions by introducing a discriminator network alongside the main segmentation network. Self-training-based methods minimize the loss of latent variables by alternatively generating pseudolabels on target data and retraining the model with these labels. Recent works on UDA for RS image semantic segmentation have primarily used generative-based methods [14–22]. To the best of our knowledge, Benjdira et al. [14] first addressed the problem of DA for RS image semantic segmentation by GANs. This study performed image-level translation from the source domain to the target domain using CycleGAN [23], and the results showed that the method is capable of substantially improving the accuracy of the segmentation model. Most recently, Li et al. [15] found that the DualGAN [24] is recommended to conduct unsupervised image translation between the source and target domains to carry out a weakly supervised transfer invariant constraint (WTIC). Tazar et al. [16] designed a novel GAN architecture named color mapping generative adversarial networks (ColorMapGAN), which can generate fake training images that are semantically the same as training images but with a spectral distribution similar to that of test images. Due to the architecturally simple (the generator has no convolution and pooling layers) but powerful design, ColorMapGAN performs well in both accuracy and computational complexity. Most studies mentioned above were implemented on a single space (e.g., input, feature, or output). In the study by [18], an end-to-end GAN-based full-space DA for classification was proposed. In this framework, the source and target images are fully aligned in the image, feature, and output spaces by different methods. Mateo-Garcia et al. [20] used CycleGAN to train a domain adaptation method between Proba-V and upscaled Landsat images. This was applied to train a cloud detection algorithm and eventually achieved higher cloud detection accuracy. Some researchers also used the CycleGAN approach for DA in deforestation detection in the Amazon Forest [21]. Kou et al. [22] proposed progressive domain adaptation (PDA) by embedding a convolutional long short-term memory (ConvLSTM) network into a conditional generative adversarial network (cGAN [25]) for change detection using season-varying remote sensing images.

In addition to the dominant generative-based methods, other methods have been widely explored in RS fields. Zhang et al. [26] introduced a layer alignment method and a feature covariance loss function. They also adopted a self-training method to fine-tune the model with pseudolabels on target data and further improve segmentation performance. Zhang et al. [27] proposed a curriculum-style local-to-global cross-domain adaptation framework for the segmentation of very-high-resolution (VHR) RS images. The result showed that the curriculum-style cross-domain adaptation (CCDA) strategy and local-to global feature alignment (LGFA) module achieved better performance on common benchmarks. Similar to this idea, Shen et al. [28] introduced a spatial frequency, detached the target domain into an easy and hard split, and then adopted intradomain adaptation

by self-supervised training to improve the performance of the hard split. In addition, to be more efficient in addressing domain shifts on aerial datasets, Chen et al. [29] used entropy minimization and classwise distribution alignment to produce high-confidence prediction. In addition to extracting features from the image domain, Liu et al. [30] proposed a bispace alignment network for DA, named BSANet, which is able to extract features in the image and wavelet domains simultaneously. The dual-branch structure utilizing two discriminators minimizes the discrepancy in the fused feature space and output spaces.

While previous studies explored UDA for RS image semantic segmentation, challenges still exist: (i) the generative-based methods used above ignore the spatial scale difference of data across domains, and only one-sided translated images are used in the subsequent process, which results in a huge waste of computation; (ii) adversarial learning computationally involves adversarial objectives in different space that are highly sensitive to hyperparameters and are difficult to train; (iii) self-training methods do not sufficiently explore the intrinsic invariant characteristics of the target data caused by the data transformation. In this paper, we propose a cycle and self-supervised consistency (CSC) training framework. The three main contributions of our work can be summarized as follows:

1. We propose a novel UDA framework for RS image semantic segmentation that combines cycle-consistent image-to-image translation techniques and self-supervised transformation consistency training.
2. We investigate multiple transformation functions and enforce transformation consistency to provide supervision for self-supervised training of the model on unlabeled target data. These functions provide different perspectives on how to learn domain-invariant features for semantic segmentation.
3. Compared with previous methods, the framework we proposed achieves state-of-the-art performance on two challenging benchmarks of UDA for RS image semantic segmentation. Each consistency component independently outperformed some previous methods. This was verified by extensive ablation studies on our framework.
4. The proposed domain adaptation framework is easy to implement and is readily embedded in any existing semantic segmentation model to improve the prediction performance on unlabeled data.

2. Methods

In the following, we first give an overview of the proposed framework, and then introduce the ResiDualGAN to conduct unsupervised image translation between source and target domains to carry out cycle consistency training. Finally, we give the detailed descriptions of the proposed consistency training strategy.

2.1. Overview

In the setting of unsupervised domain adaptation for semantic segmentation, we consider a labeled source domain S and an unlabeled target domain T with the same set of semantic classes. In general, RS images from S and T are composed of different image attributes with different geographic locations, which leads to the cross-domain RS image semantic segmentation being more complex than other tasks in UDA fields. The goal of the purposed framework is to train a model on labeled domain S, and perform across-domain semantic segmentation tasks with minimal prediction error on the unlabeled target domain T. Intuitively, the overall workflow of the proposed consistency-based framework is shown in Figure 1, and is driven by consistency principles in different training stages. The framework consists of two consistency criteria: the cycle consistency across domains in the input space and the self-supervised transformation consistency in the training stage. A cycling generative adversarial network with scale change modules is trained in the input space to eliminate spectral- and spatial-scale discrepancies across domains by iteratively reconstructing and discriminating. A mapping of the data distribution between the source and target domains is established, and the target-stylized source domain data are

produced to train a semantic segmentation model in a supervised way. In the subsequent training stage, a self-supervised training module is incorporated to enforce the consistency of target semantic maps produced by the model across a serious image transformation. Such consistency training is based on the idea that a strong model's output should be consistent, with respect to the transformation of inputs, and enable the model to learn features from unlabeled target data in a self-supervised way.

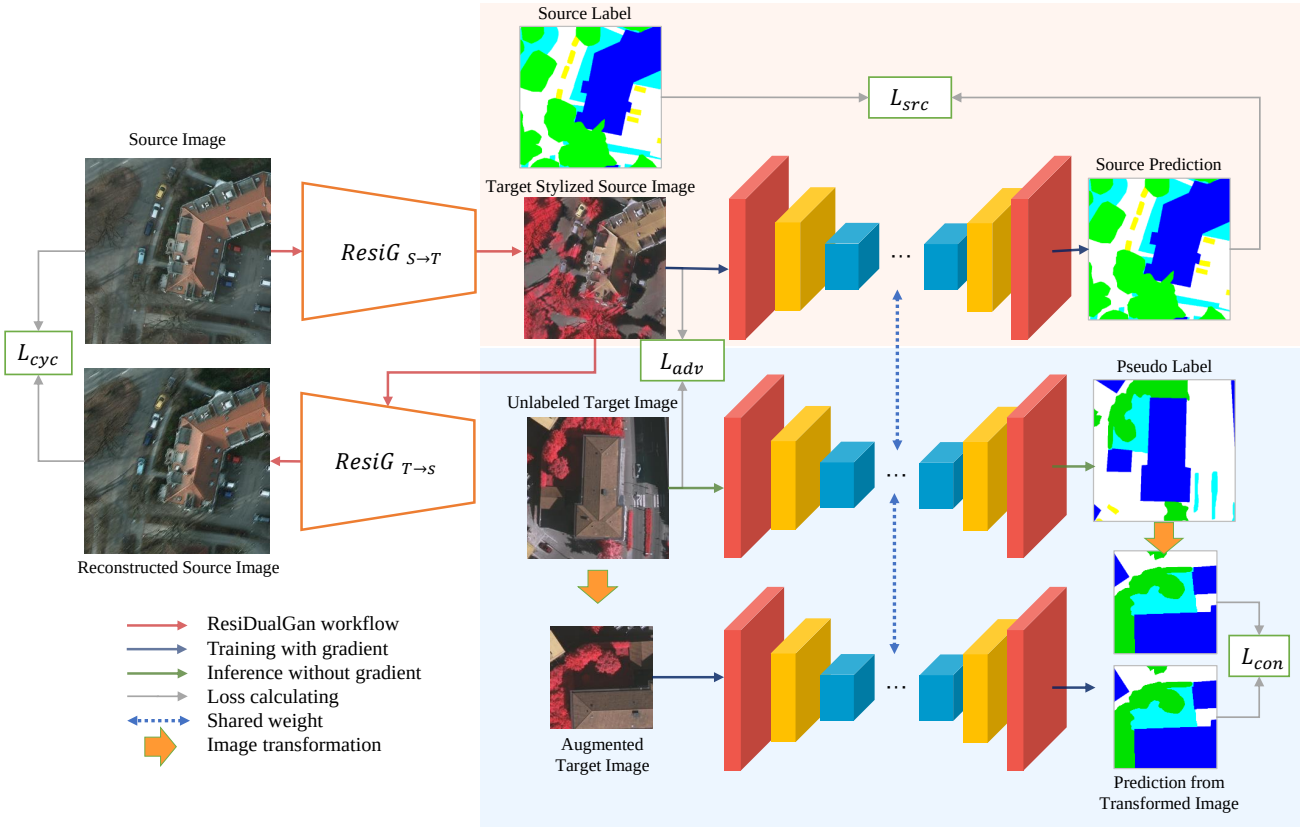

Figure 1. Overview of proposed framework. $ResiG_{S\to T}$ and $ResiG_{T\to S}$ are generators of the ResiDualGAN between two domains. L_{cyc} and L_{adv} are cycle loss function and adversarial loss function, respectively. L_{src} and L_{con} are segmentation loss function for source data and target data, respectively.

2.2. Cycle Consistency

As mentioned in Section 1, GAN-based unpaired image translation algorithms have been widely used in cross-domain adaptation tasks for semantic segmentation. The original GANs can teacha generator to capture the distribution of real data by introducing an adversarial discriminator that evolves to discriminate between the real data and the fake. Therefore, it usually includes two neural networks, in which generative network G generates candidates while the discriminate network D distinguishes candidates produced by G from the true data distribution. With such a strategy, a bidirectional GAN architecture enables cross-domain image-to-image translation. Suppose there are two collections of images X_S and X_T from source domain S and target domain T, respectively. The generator $G_{S\to T}$ maps an image $x_S \in X_S$ to an image $x_T \in X_T$, while the dual task is to train an inverse generator $G_{T\to S}$. Discriminators D_S and D_T are trained to measure how well the generated candidates fit in the corresponding domains. This idea of implementing inverse mapping and having two generators and two discriminators is common among a series of GAN-family methods for image translation, such as DiscoGAN [31], CycleGAN, and DualGAN. For RS image translation, Li et al. [15] quantitatively verified that DualGAN performed better than other unsupervised methods (e.g., CycleGAN and DiscoGAN) in across-domain semantic segmentation tasks.

However, this line of work on RS image translation suffers from two main limitations in practical applications. First, the spatial scale discrepancy of images between two domains is not considered, while the scale factor is very important in RS image interpretation tasks. Previous works have proven that RS images with similar ground sampling distance (GSD) are easier to adapt across domains [19,32]. Second, cycle translation in appearance may locally change the semantic information of certain ground types; hence such semantic consistency should be ensured during translation. To solve the above problems, we introduce ResiDualGAN [32] to align the distribution in the image space. Based on DualGAN, ResiDualGAN redesigned the structure of generators and named it Resi-Generator, where a residual module and a resize module were incorporated (Figure 2). It is proposed specifically for optimizing the translation results of RS images with different spatial resolutions. For translation of $X_S \rightarrow X_T$, the Resi-Generator can be denoted as $ResiG_{S\rightarrow T}$, containing a regular generator $G_{S\rightarrow T}$ and a resize module $Resize_{S\rightarrow T}$. As Figure 2 shows, the design of $ResiG_{S\rightarrow T}$ was inspired by the structure of skip connections (i.e., shortcuts) from the residual network [33]. The outputs of $G_{S\rightarrow T}$ are added as an identity mapping X_S to enhance the original semantic information and then fed into $Resize_{S\rightarrow T}$ to decrease the spatial resolution:

$$X_{S\rightarrow T} = ResiG_{S\rightarrow T}(X_S) = Resize_{S\rightarrow T}(G_{S\rightarrow T}(X_S) + X_S). \quad (1)$$

The mapping of $X_S \rightarrow X_T$:image $x_s \in X_S$ is translated to domain X_T using $ResiG_{S\rightarrow T}$. How well translation $ResiG_{S\rightarrow T}(x_S)$ fits in X_T is evaluated by D_T. $ResiG_{S\rightarrow T}(x_S)$ is then translated back to domain X_S using $ResiG_{T\rightarrow S}$, which outputs $ResiG_{T\rightarrow S}(ResiG_{S\rightarrow T}(x_S))$ as the reconstructed version of x_S. Similarly, $x_T \in X_T$ is translated to X_S as $ResiG_{T\rightarrow S}(x_T)$ and then reconstructed as $ResiG_{S\rightarrow T}(ResiG_{T\rightarrow S}(x_T))$. Note that the reverse mapping $X_T \rightarrow X_S$ and both D_T and D_S are not illustrated in Figure 1 for simplification.

The objective functions of ResiDualGAN consist of a feature-level adversarial loss and an image-level cycle loss. The adversarial loss function is integrated from loss functions used in D_T and D_S, which can be defined as

$$L_{adv} = (-D_S(ResiG_{T\rightarrow S}(x_T)) - D_T(ResiG_{S\rightarrow T}(x_S))). \quad (2)$$

Loss functions for generators $ResiG_{S\rightarrow T}$ and $ResiG_{T\rightarrow S}$, however, are the same, as both share the same objective. The cycle loss function measured by the L_1 distance can be defined as

$$\begin{aligned} L_{cyc} = \psi_{cyc}(||x_S - ResiG_{T\rightarrow S}(ResiG_{S\rightarrow T}(x_S))|| + ||x_T - ResiG_{S\rightarrow T}(ResiG_{T\rightarrow S}(x_T))||) \\ + \psi_{adv}(-D_S(ResiG_{T\rightarrow S}(x_T)) - D_T(ResiG_{S\rightarrow T}(x_S))), \end{aligned} \quad (3)$$

where ψ_{cyc} and ψ_{adv} are two hyperparameters. By alternately minimizing L_{cyc} and L_{adv}, cycle consistency enforces cross-domain consistent predictions in the context of image-to-image translation. The translated source and target data are used to train a semantic segmentation model in the next stage.

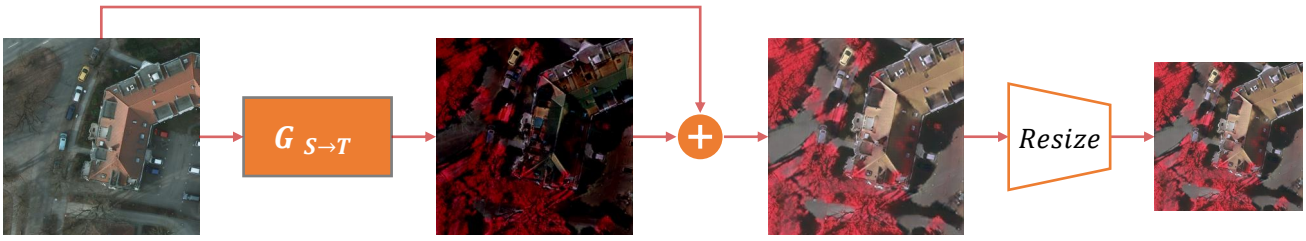

Figure 2. Resi-Generator of ResiDualGAN.

2.3. Self-Supervised Consistency

Our ultimate goal is to train a semantic segmentation model that is capable of high performance on unlabeled target domains. Cycle consistency reduces the distribution of data between the source domain and target domain. The semantic segmentation model can be trained on labeled, target-stylized source data and learn the features transferred from the target domain. However, the transferred features are insufficient for mapping real target data to ground truth. Inspired by approaches from self-training learning and semisupervised learning communities [34–40], a consistency training strategy is designed in this stage. This strategy is based on the assumption that the outputs from a well-trained model should be consistent across input transformations. For example, let f:$X \to Y$ represent a pixelwise mapping from images $x \in X$ to semantic output $y \in Y$. Denote t as a transformation function. We denote t_p:$X \to X'$ as a photometric image transformation such as whitening or style transfer, and t_q:$X \to X'$ as a spatial image transformation such as flipping or scaling. The consistency training is under the following hypothesis that for any image $x \in X$, f is invariant under t_p:$f(t_p(x)) = f(x)$ and equivariant under t_q:$f(t_q(x)) = t_q(f(x))$. If $x \in X_T$, then the $f(x)$ can be seen as pseudo labeling and the prediction inconsistency cases provide self-supervision for target data. As shown in Figure 3, the training step is composed of a standard pipeline of a supervised segmentation network driven by labeled target-stylized source domain data and a self-supervised consistency training branch with pseudo labeling. In one training loop, a batch of target-stylized source images are fed in the model f and the predicted semantic maps are used to calculate supervised loss with labels. After that, a batch of unlabeled target images are fed into the model and pseudolabels are created by selecting confident pixels from the averaged map using thresholds. A transformation function t mentioned above is used to obtain a transformed version of the same pair of target images and pseudolabels. The consistency between these two version predictions compels the model to learn the features from the target domain in a self-supervised way. From another perspective, minimizing the unsupervised consistency loss progressively propagates semantic information from labeled data to unlabeled ones.

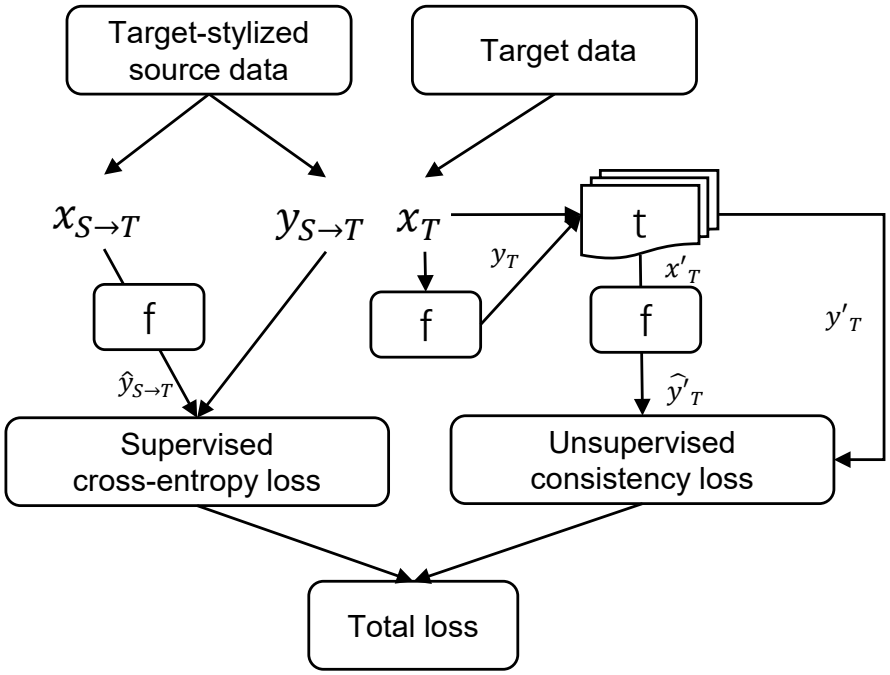

Figure 3. Flowchart of self-supervised consistency training: $x_{S\to T}$ and $y_{S\to T}$ are target-stylized source image and label; x_T and y_T are target image and pseudolabel. f denotes the pixelwise mapping from images to semantic output; t is the transformation function; x'_T and y'_T are the transformed target image and pseudolabel; $\hat{y}_{S\to T}$ and $\hat{y}'_T$ are the semantic outputs of the f.

Concisely, consistency training methods simply regularize model predictions to be invariant to perturbations applied to either input samples or hidden states [40]. Under this framework, we are interested in what perturbations or transformations to apply are beneficial for cross-domain semantic segmentation of RS images. In this paper, we explore three types of transformation functions for such self-supervised consistency training:

- **Translation consistency:** The GAN-based image translation outputs are bidirectional. However, in previous generative-based methods, source-stylized target images are useless in subsequent procedures, which results in a huge waste of computation. With the help of ResiDualGAN, the spatial scale and semantic information are well preserved during target-to-source stylization. Therefore, we regard such translation as transformation and enforce consistency between the model's outputs on the original target images and the translated target image (Figure 4b).
- **Augmentation consistency:** Data augmentation in model training is a technique to increase the amount of data by adding modified copies of already existing data. It acts as a regularizer and helps reduce overfitting when training a deep learning model. We randomly augment the target data by flipping, cropping, and brightness-changing. The corresponding augmentation is also performed on pseudolabels (Figure 4c).
- **CutMix consistency:** CutMix is a regularization method designed for image classification and transfer learning [41]. We leverage this strategy to randomly cut and paste patches from source data to target data. The ground truth labels or pseudolabels are also mixed proportionally. The added patches on target data can be seen as a transformation for consistency training. Pasting patches from the source domain further increases confidence in pseudo labeling (Figure 4d).

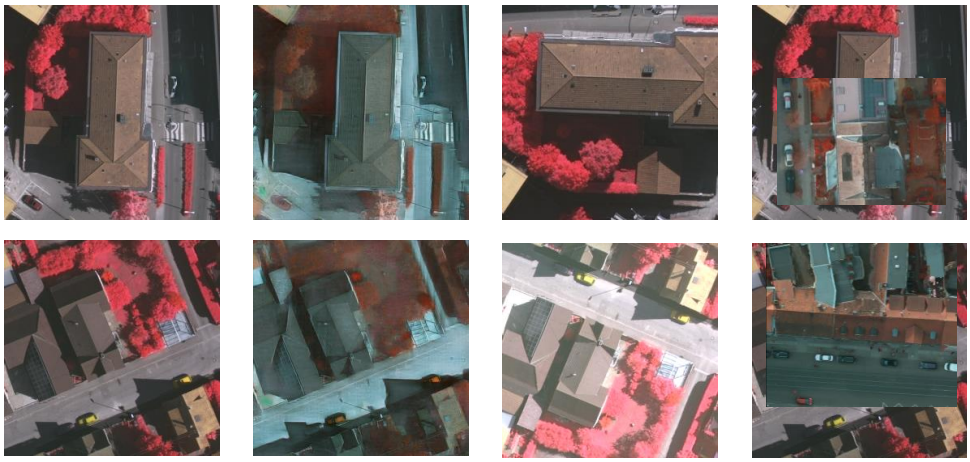

Figure 4. Transformation of images from target domain for consistency training. (**a**) Target domain images without transformation. (**b**) Source-stylized target images. (**c**) Augmented target images. (**d**) Target images mixed by source image patches.

In this paper, we adopt a combination objective function that combines a distribution-based cross-entropy loss and a region-based Dice loss. In this way, we can combine local with global information to improve the segmentation results. The loss function is defined as follows:

$$L_{CE\_DC}(y, \hat{y}) = (-\sum_{i}^{c}\sum_{j}^{N} y_i^j \log \hat{y}_i^j) + (1 - \frac{1}{c}\sum_{i=0}^{c}\frac{\sum_j^N 2y_i^j\hat{y}_i^j + \epsilon}{\sum_j^N y_i^j + \sum_j^N \hat{y}_i^j + \epsilon}), \quad (4)$$

where y_i^j and $\hat{y}_i^j$ denote the ith channel at the jth pixel location of the reference labels and neural network softmax output, respectively. We use c to denote the total channel count, N

to denote the total pixel count in a mini-batch, and ϵ as a small constant plugged to avoid numerical problems. In addition to the supervised loss, we use transformation functions to obtain transformed pairs of target data and pseudolabels to calculate consistency loss by the same loss function. Therefore, the final loss L_{seg} involves the source supervised loss and the consistency training loss, which can be defined as follows:

$$L_{seg} = (1-\lambda)L_{CE\_DC}(y_{S\to T}, \hat{y}_{S\to T}) + \lambda L_{CE\_DC}(y'_T, \hat{y'}_T), \quad (5)$$

where $y_{S\to T}$ and $\hat{y}_{S\to T}$ are labels and predictions of target-stylized source data, respectively. Meanwhile, y'_T and $\hat{y'}_T$ are transformed pseudo labels and predictions of transformed target data, respectively. λ is a hyperparameter for weighting the two loss terms.

3. Experiments

In this section, we describe the datasets and experimental implementation details first and then we give the experimental results.

3.1. Datasets

We evaluate our framework on the two standard large-scale RS image semantic segmentation datasets Potsdam and Vaihingen [42]. These two common datasets are published by the International Society for Photogrammetry and Remote Sensing (ISPRS) as benchmarks of a 2D semantic labeling contest, providing airborne images with very high-resolution true orthophotos and labeled ground truth. The semantic categories from both datasets are the same and have been defined in the ranges of impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. Both areas cover urban scenes captured from different geographic locations (the city of Potsdam and Vaihingen) with different geospatial distributions. Potsdam shows a typical historic city with large building blocks, narrow streets, and dense settlement structures, while Vaihingen is small with many detached buildings and small multistory buildings. There are also differences in the imaging modes, including the channels and resolution. Potsdam datasets contain subsets with different spectral band combinations such as IR-R-G and R-G-B, while images from Vaihingen datasets have only IR-R-G channels. Moreover, the ground sampling distance or spatial resolution of the Potsdam datasets is 5 cm, while that of the Vaihingen datasets is 9 cm.

The similarities and differences mentioned above imply that these two datasets are quite suitable to evaluate the performance of UDA methods because, as two domains, they are the same in task objective and label space but different in data distribution and feature space. Li et al. [15] first developed them as new benchmarks of unsupervised domain adaptation for RS image semantic segmentation, evaluating the performance of the UDA model from the perspective of variation in geographic location and imaging modes. Since then, some work has been evaluated using this benchmark [15,27,32]. In this paper, we continue to adopt this established evaluation protocol. Images and labels from the Potsdam datasets serve as the source data, and the images from the Vaihingen datasets serve as the target data without available semantic labels. The two domain adaptation scenarios in the following are denoted as Potsdam IRRG → Vaihingen IRRG and Potsdam RGB → Vaihingen IRRG, respectively. The former represents the scenario in which domain adaptation tasks are across domains with different geographic regions. The latter crosses domains with different geographic regions plus different spectral combinations. Note that both adaptation scenarios involve the difference in ground sample distance.

3.2. Implementation Details

We implement the proposed framework in PyTorch [43] and adopt DeepLabv3+ [44] with ResNet-34 backbone as the segmentation architecture. Backbone initializes from the models pretrained on ImageNet. In the first stage, we implement cycle consistency using ResiDualGAN to translate the data from both domains. The loss parameters ψ_{cyc} and ψ_{adv} are set to 10 and 5 for training. Based on the ratio of the spatial resolution of the data

in the two domains, Potsdam IRRG and Potsdam RGB are cropped into the size of 896×896 while images from the target domain are cropped into the size of 512×512. Finally, a total of 2368 Potsdam images and labels serve as source data for translation and training. The Vaihingen datasets contain 1697 images as target data for self-supervised training stage, 440 of which are used for validation. ResiDualGAN applies the color rendering and spatial resolution style of Vaihingen datasets (target domain) to higher resolution images of Potsdam datasets (the reverse translation occurred at the same time). We obtained the Vaihingen stylized Potsdam IRRG and Potsdam RGB images with downscale size of 512×512. In the self-supervised transformation consistency training stage, we train the model with labeled target-stylized source data in a supervised way. Joint training proceeds in a self-supervised manner with unlabeled target data. The total gradient is calculated after alternating source-target forward passes to update the weights of the model. The batch size of 16 comprises eight target-stylized images and eight real target images at a size of 512×512, which is a common practice. The optimizer is Adam and the initial learning rate is set to 1×10^{-4} and is scaled down by a factor of 0.5, according to the patience, which is a number of epochs with no improvement. To investigate the effects of each component and hyperparameter of our framework on performance, we present an extensive set of ablation studies. All experiments are conducted on a machine equipped with an Intel Core i7-7800X CPU, 16 GB of RAM, and one NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of VARM to accelerate the training process. We measure the segmentation accuracy with per-class intersection-over-union (IoU) and F1-score, following recent works [15,27]. We denote the number of true positives, false positives, and false negatives by TP, FP, and FN respectively. The formulation of IoU and *F1-score* can be written as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{6}$$

$$F1\text{-}score = \frac{2 \times TP}{2TP + FP + FN}. \tag{7}$$

The *F1-score* tends to measure something closer to average performance, while the IoU score measures something closer to the worst-case performance.

3.3. Performance of the CSC

We present our experimental results of the proposed CSC training framework on the Potsdam RGB → Vaihingen IRRG and Potsdam IRRG → Vaihingen IRRG benchmarks in Tables 1 and 2. Depending on the transformation functions involved, our methods in Tables 1 and 2 are denoted as CSC-Trans, CSC-Aug, and CSC-CutMix, respectively. The qualitative results of these approaches can be seen in Figures 4 and 5. In both domain adaptation scenarios, augmentation-based consistency performs best, followed by CutMix-based and translation-based consistency. We adopt DeepLabv3+ as the baseline segmentation model to show the practical performance of the segmentation model in the presence of a domain gap. The baseline model is trained on labeled source data and tested on the unlabeled target datasets. Compared to the baseline, the CSC improves the accuracy of the segmentation result on target domain by up to 83% on Potsdam RGB → Vaihingen IRRG and 65% on Potsdam IRRG → Vaihingen IRRG. The solid performance of augmentation-based consistency suggests that the simple geometric or photometric transformations provide more significant equivariant constraints to the model than CutMix and image translation. CutMix-based consistency involves patches from source domain data, which provide high-confidence pseudo labeling while facilitating model-learning domain-invariant features. Note that in the scenario of the Potsdam IRRG → Vaihingen IRRG, the CutMix-based method has a significantly higher segmentation accuracy for the category of cars than the other methods. Translation-based consistency performs the worst relative to the other two transformation methods of our framework. As shown in column (d) of Figures 5 and 6, the results of the translation-based method misclassified many regular objects into the class of clutter. Such poor performance is particularly

evident in the average accuracy of the Potsdam RGB → Vaihingen IRRG. This implies that the reverse appearance translation as a transformation function for self-supervised consistency training is too intense for domains with vastly different imaging modes. By comparing the results of the baselines, methods on the domains with the same spectral band combination can bring an 18% performance improvement. This is another verification that aligning the domains directly in image space is effective for adapting domains with different imaging modes.

Table 1. Quantitative results (%) of cross-domain semantic segmentation on Potsdam RGB → Vaihingen IRRG benchmark.

Method	Metrics	Clutter/ Background	Impervious Surface	Car	Tree	Low Vegetation	Building	Overall
Baseline (DeepLabv3+ [44])	IoU	2.67	40.24	18.35	53.14	12.88	52.63	29.98
	F1-score	4.65	56.93	30.40	69.19	22.68	68.74	42.10
GAN-RSDA [14]	IoU	2.29	48.27	25.73	42.16	23.34	64.33	34.35
	F1-score	3.50	64.79	40.20	59.03	37.55	78.13	47.20
AdaptSegNet [45]	IoU	6.26	55.91	34.09	47.56	23.18	65.97	38.83
	F1-score	9.55	71.44	50.34	64.17	37.22	79.36	52.01
MUCSS [15]	IoU	5.87	54.21	27.95	43.73	26.94	68.76	37.91
	F1-score	8.77	70.04	42.89	60.53	42.09	81.26	50.93
CCDA [27]	IoU	**12.38**	64.47	43.43	52.83	38.37	76.87	48.06
	F1-score	**21.55**	77.76	60.05	69.62	55.94	86.95	61.98
RDG-OSA [32]	IoU	9.84	62.59	54.22	56.31	37.86	79.33	50.02
	F1-score	14.55	76.81	70.00	71.92	54.55	88.41	62.71
CSC-Trans	IoU	5.79	57.32	52.93	51.78	30.61	74.31	45.46
	F1-score	9.16	72.69	69.02	68.11	46.40	85.13	58.42
CSC-Aug	IoU	8.12	**68.91**	**57.41**	**65.47**	**48.33**	**81.78**	**55.00**
	F1-score	11.23	**81.48**	**72.76**	**79.04**	**64.78**	**89.94**	**66.54**
CSC-CutMix	IoU	10.21	10.21	53.89	56.43	37.29	78.32	49.94
	F1-score	14.81	14.81	69.74	72.00	54.02	87.74	62.64

Table 2. Quantitative results (%) of cross-domain semantic segmentation on Potsdam IRRG → Vaihingen IRRG benchmark.

Method	Metrics	Clutter/ Background	Impervious Surface	Car	Tree	Low Vegetation	Building	Overall
Baseline (DeepLabv3+ [44])	IoU	2.99	47.88	20.82	58.74	19.57	61.37	35.23
	F1-score	5.18	64.40	33.93	73.88	32.47	75.83	47.61
GAN-RSDA [14]	IoU	7.26	57.32	20.04	44.27	35.47	65.35	38.28
	F1-score	10.32	72.60	32.53	61.04	51.99	78.84	51.22
AdaptSegNet [45]	IoU	6.32	62.50	29.31	55.74	40.30	70.41	44.10
	F1-score	9.67	76.66	44.81	71.36	57.01	82.50	57.00
MUCSS [15]	IoU	**11.16**	65.94	26.30	50.49	39.85	69.07	43.80
	F1-score	14.70	79.15	40.77	66.76	56.55	81.53	56.58
CCDA [27]	IoU	\	58.64	28.17	53.28	30.39	60.60	46.22
	F1-score	\	75.13	45.81	69.52	47.62	76.89	62.99
RDG-OSA [32]	IoU	10.70	70.31	54.04	59.22	49.03	81.20	54.08
	F1-score	15.48	82.43	69.85	74.22	65.52	89.57	66.18
CSC-Trans	IoU	8.42	65.67	54.75	61.72	42.69	75.88	51.52
	F1-score	12.77	79.15	70.56	76.23	59.46	86.20	64.06
CSC-Aug	IoU	13.83	**75.56**	56.58	**65.55**	**52.92**	**84.17**	**58.10**
	F1-score	**19.59**	**86.01**	72.01	**79.09**	**68.96**	**91.38**	**69.50**
CSC-CutMix	IoU	10.46	73.31	**57.91**	63.58	51.58	81.25	56.35
	F1-score	14.91	84.50	**73.05**	77.57	67.86	89.56	67.91

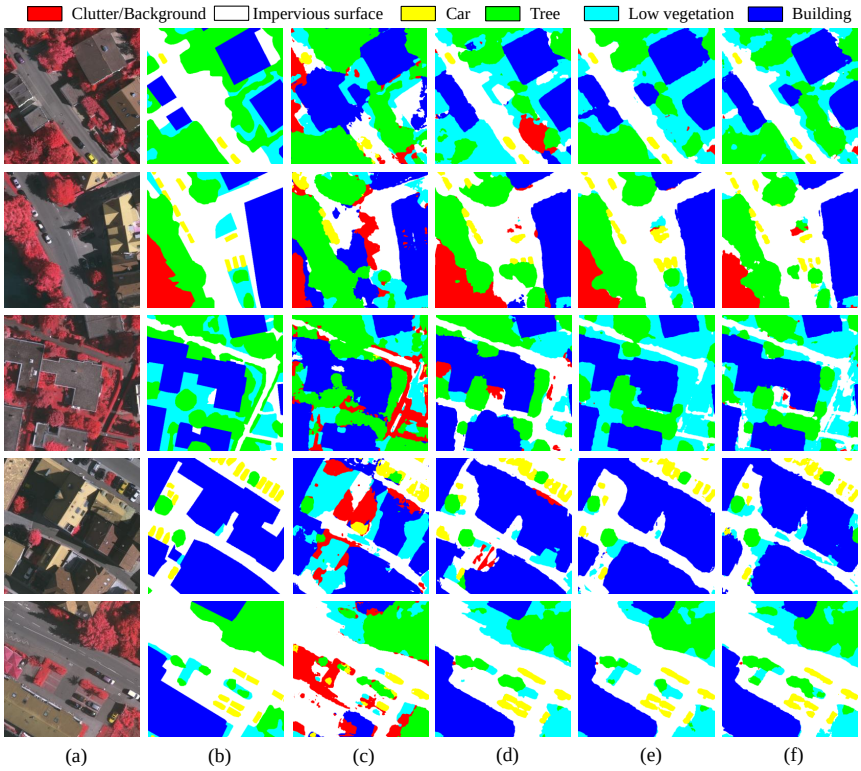

Figure 5. Qualitative results of cross-domain semantic segmentation on Potsdam RGB → Vaihingen IRRG. (**a**) Target images. (**b**) Ground truth. (**c**) Baseline. (**d**) CSC-Trans. (**e**) CSC-Aug and (**f**) CSC-CutMix.

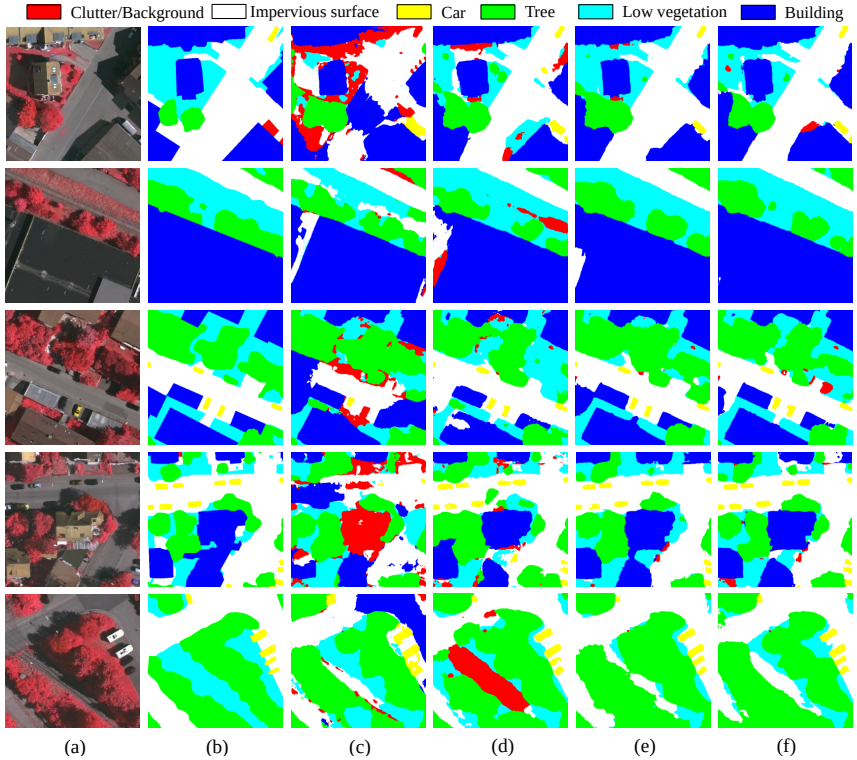

Figure 6. Qualitative results of the cross-domain semantic segmentation on Potsdam IRRG → Vaihingen IRRG. (**a**) Target images. (**b**) Ground truth. (**c**) Baseline. (**d**) CSC-Trans. (**e**) CSC-Aug and (**f**) CSC-CutMix.

3.4. Comparison to State-of-the-Art Methods

Since we adopt an established evaluation protocol, including the same test datasets and metrics, we compare our framework to previous work from the benchmark list. The methods involved in the comparison include an adversarial learning method (AdaptSegNet [45]), an image-to-image translation-based method (GAN-RSDA [14]), a curriculum learning method (CCDA [27]), and two hybrid methods (MUCSS [15] and RDG-OSA [32]). Among them, GAN-RSDA, MUCSS, CCDA, and RDG-OSA are specifically designed for RS image semantic segmentation across domains, and the AdaptSegNet is designed for natural images. AdaptSegNet aligns the distribution of features in the output space using a discriminator network, which is a representative adversarial learning in the UDA field. GAN-RSDA is the pioneer in the use of image-to-image translation techniques in UDA for RS images and introduces the CycleGAN for appearance adaptation in image space. MUCSS goes a step further from the GAN-RSDA by using the more effective DualGAN and involving a self-training strategy and consistency regularization. CUSS designs a curriculum-style local-to-global cross-domain adaptation method to learn domain-invariance features. RDG-OSA combines the ResiDualGAN with an output space adaptation method and claims the best published results. For fairness, all the methods we reproduce in experiments employ the same segmentation model and training setup as our proposed framework.

As shown in Tables 1 and 2, our proposed framework substantially outperforms the other methods and sets a new standard in terms of IoU and F1-score. Compared with the state-of-the-art methods, our framework increases the IoU and F1-score of segmentation results by 9.95% and 6.61% on Potsdam RGB → Vaihingen IRRG, respectively, while by 7.53% and 5.02% on Potsdam IRRG → Vaihingen IRRG. From the perspective of the adaptation scenario, the average performance improvement on the Potsdam RGB → Vaihingen IRRG is superior to that on the Potsdam IRRG → Vaihingen IRRG. For example, the mIoU is improved by 37% compared to MUCSS on Potsdam RGB → Vaihingen IRRG and shows an up to 28% improvement on Potsdam IRRG → Vaihingen IRRG. This phenomenon does not include the CCDA, which ignores the class of clutter on the Potsdam IRRG → Vaihingen IRRG. Specific to the accuracy comparison of each class, CCDA achieves the best accuracy of clutter on the Potsdam RGB → Vaihingen IRRG while MUCSS achieves the best accuracy on the Potsdam IRRG → Vaihingen IRRG. In addition, augmentation-consistency-based CSC achieves the best accuracy in other categories in both scenarios. The proposed framework shows strong robustness and great generalization capability. Figures 7 and 8 present a few qualitative examples, comparing our framework to the baseline and previous methods. Particularly prominent are the refinements of the small-scale elements such as cars and clutters, which may benefit from spatial resolution alignment in the cycle consistency. Moreover, the boundaries of the segments in our results are more precise by enforcing self-supervised consistency with transformation function, which makes our framework less prone to the contextual bias, leading to coarse boundaries.

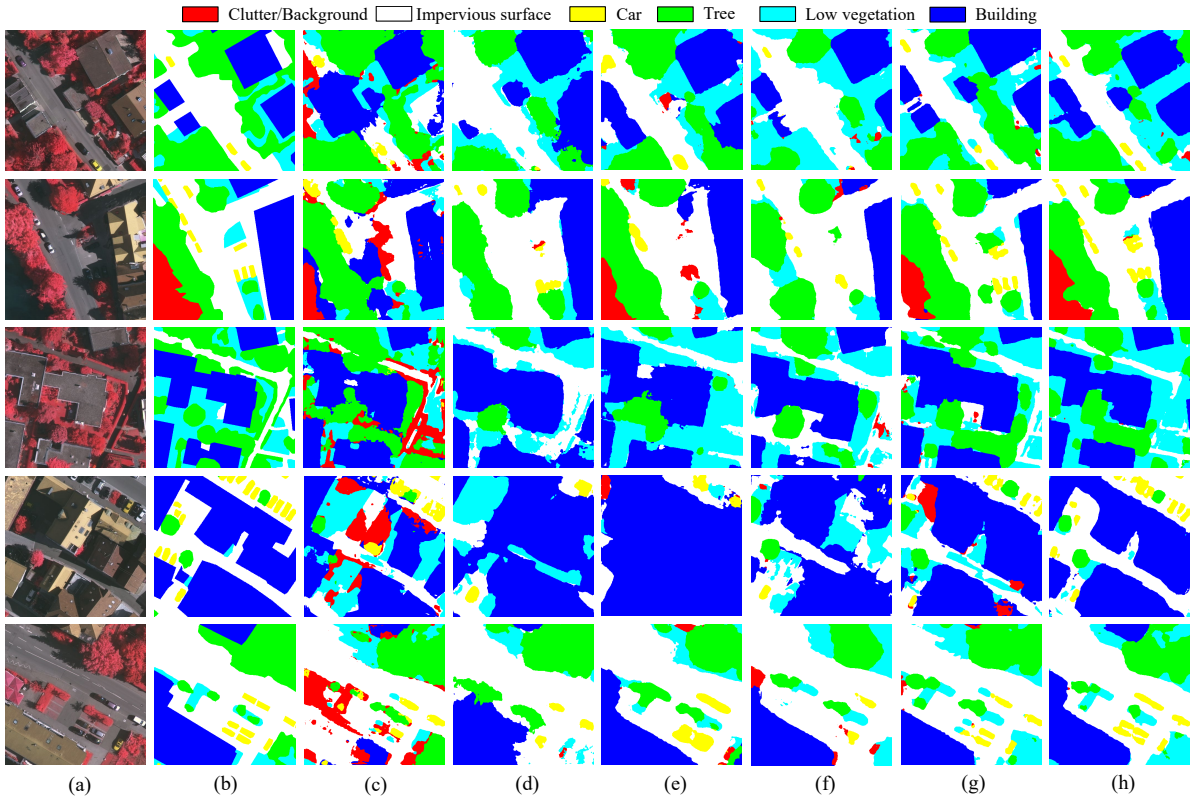

Figure 7. Qualitative results of the cross-domain semantic segmentation on Potsdam RGB → Vaihingen IRRG. (**a**) Target images. (**b**) Ground truth. (**c**) Baseline. (**d**) GAN-RSDA. (**e**) AdaptSegNet. (**f**) MUCSS. (**g**) RDG-OSA and (**h**) CSC-Aug.

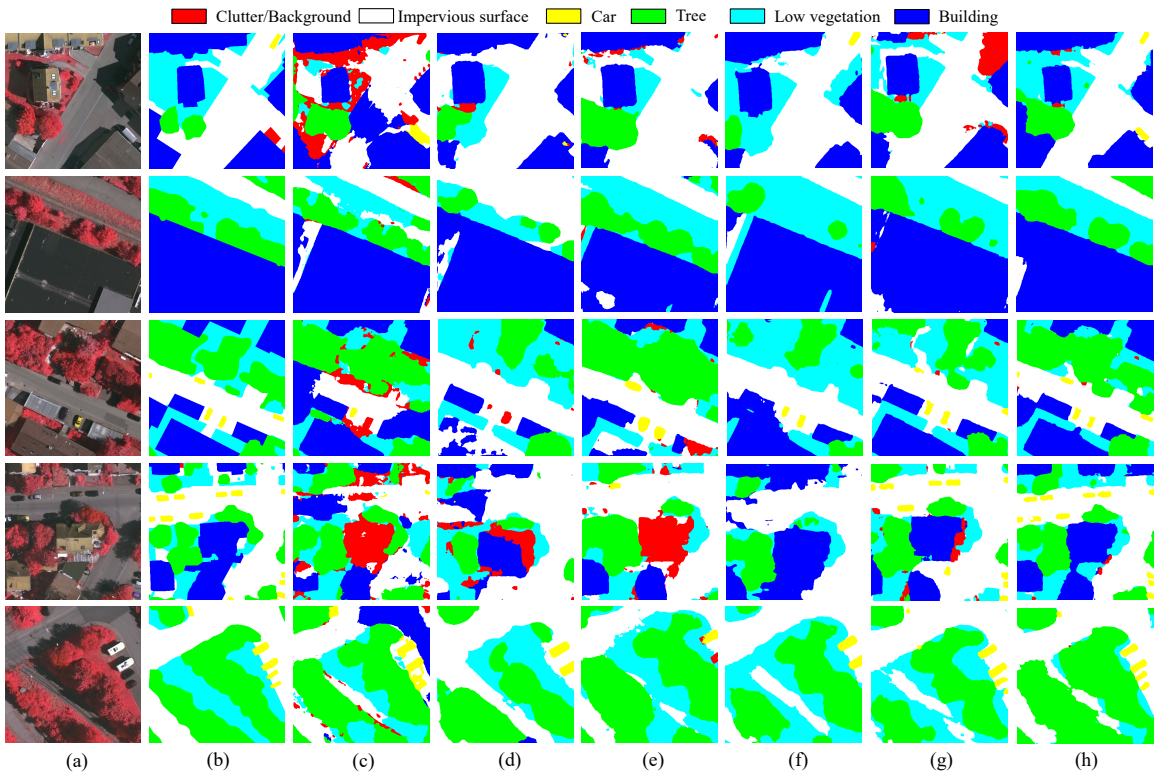

Figure 8. Qualitative results of the cross-domain semantic segmentation on Potsdam IRRG → Vaihingen IRRG. (**a**) Target images. (**b**) Ground truth. (**c**) Baseline. (**d**) GAN-RSDA. (**e**) AdaptSegNet. (**f**) MUCSS. (**g**) RDG-OSA and (**h**) CSC-Aug.

4. Discussion

In this section, we first discuss the effects of components and hyperparameters sensitivity of the proposed CSC training framework. Then, we discuss the limitations of the proposed method and possible further improvements.

4.1. Ablation Study

In this subsection, we examine an extensive set of ablation studies to determine what makes our framework effective. These studies concern each component of the proposed framework and hyperparameters, such as the loss function ratio and pseudolabel threshold. Note that the self-supervised module of the full framework (cycle + self-supervised consistency) in this subsection is augmentation-consistency-based. First, we analyze the contribution of two components: cycle consistency and self-supervised consistency. We independently switch off each component and report the results in Table 3. On the Potsdam RGB → Vaihingen IRRG, compared to the full framework, disabling the self-supervised consistency module (transformation functions) leads to a 12% IoU decrease, while abolishing cycle consistency (without ResiDualGAN) leads to a drop of 34% IoU. For similar cases in the Potsdam IRRG → Vaihingen IRRG scenario, the mIoU decreases by 7% and 17%, respectively.

Table 3. Ablation study of effects of components of our framework.

Method	Potsdam RGB → Vaihingen IRRG		Potsdam IRRG → Vaihingen IRRG	
	mIoU	Δ	mIoU	Δ
Cycle + Self-supervised	55.00	0.00	58.10	0.00
Cycle consistency only	48.50	↓ 12%	54.09	↓ 7%
Self-supervised consistency only	36.56	↓ 34%	48.10	↓ 17%
Source only	29.98	↓ 45%	35.23	↓ 39%

These statistics suggest that both cycle consistency and self-supervised consistency play a critical role in the entire framework. The contribution of cycle consistency is larger, especially in domain adaptation scenarios where there are vast differences in appearance or imaging mode. It is worth mentioning that the cycle-consistency-only cases in our methods significantly outperform CycleGAN-based GAN-RSDA and DualGAN-based MUCSS. This demonstrates the improvement in accuracy brought by the resize-residual module in the ResiDualGAN. To demonstrate the effectiveness of cycle consistency in reducing differences in data distribution, we randomly selected 500 images in each of two adaptation scenarios and extracted the high-dimensional features in the latent space of the backbone for semantic segmentation. These high-dimensional features are reduced to the two-dimensional space for visualization. As shown in Figure 9, the ResiDualGAN well matches the data distribution of the source domain data with the target domain data. The feature distribution of most translated images is similar to the target domain data feature distribution. In addition, it can be seen from Figure 9 that the matching difficulty of the data distribution between Potsdam IRRG and Vaihingen IRRG is less than the matching difficulty of Potsdam RGB and Vaihingen IRRG, which again illustrates the impact of imaging mode differences on domain adaptation.

Figure 10 illustrates the contribution of each component to the accuracy of individual classes on the Potsdam IRRG → Vaihingen IRRG. For most categories, the contribution of the different components to their accuracy is similar to that of the overall accuracy. However, for the impervious surface and clutter, there is little difference between the cycle-consistency-only-based method and the self-supervised-only-based method for accuracy improvement. In addition, the combination of cycle consistency and self-supervision gives the car an almost negligible improvement in accuracy.

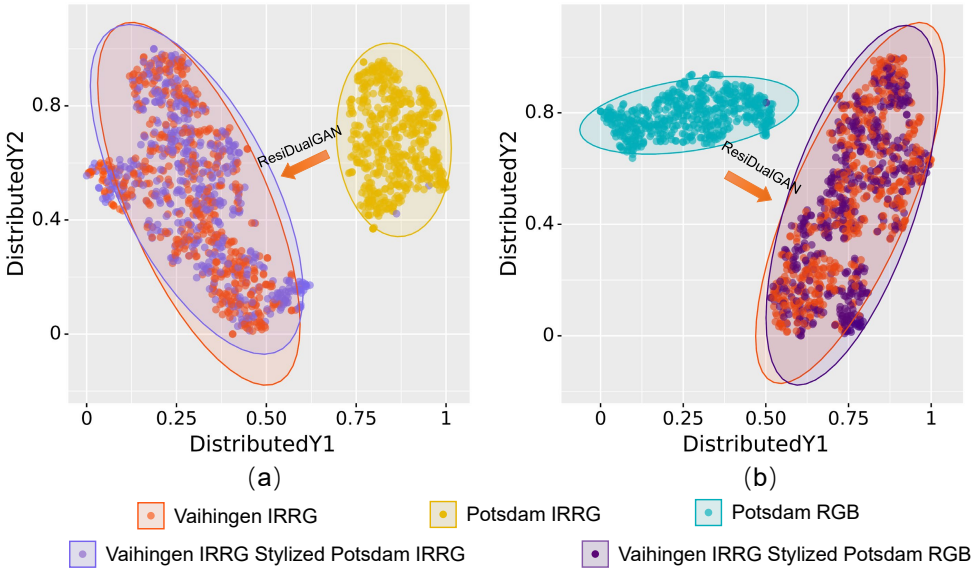

Figure 9. The visualization of distribution matching in latent space, where we map features to 2D space with t-SNE [46]: (**a**) Potsdam IRRG → Vaihingen IRRG; (**b**) Potsdam RGB → Vaihingen IRRG.

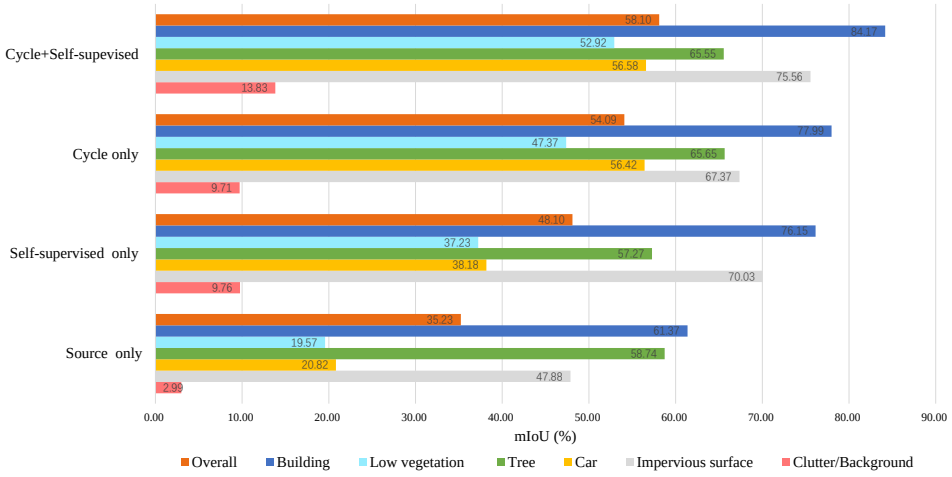

Figure 10. Ablation study of contribution of each component to accuracy of individual classes on Potsdam IRRG → Vaihingen IRRG.

Next, we analyze the sensitivity of hyperparameters λ and τ (Figure 11). Our proposed framework involves tuning the single hyperparameter λ, which trades off between the strength of the source supervised loss and self-supervised consistency loss. We present results for λ from the range of 0.1–0.9 and find that the performance of the model decreases gradually as λ grows. Similarly, we present results for the pseudo label threshold τ from the range of 0.1–0.9. We find that the accuracy starts to decrease significantly when the threshold $\tau > 0.5$. It should be noted that due to the datasets consist of 6 classes, $\tau = 0.1$ means there is no thresholding in the training process. We think that the decrease in accuracy with increasing pseudo label thresholds τ is due to the loss being dominated by easy classes when a high threshold is set.

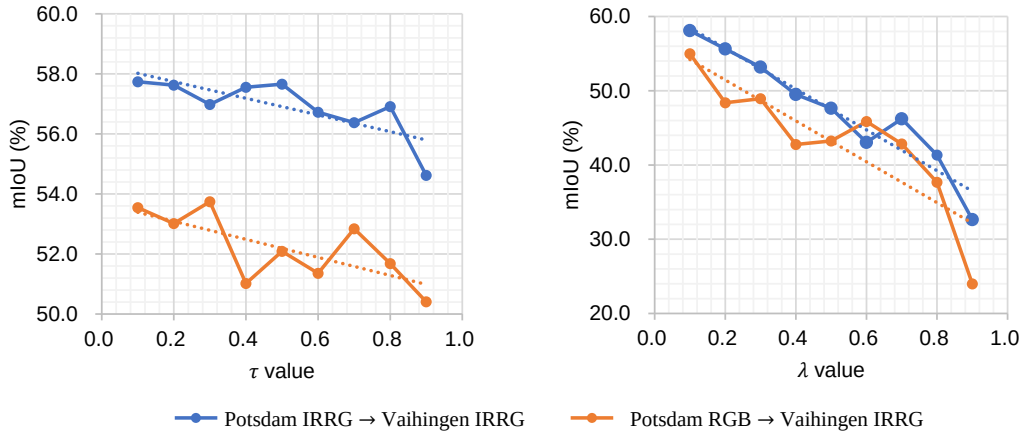

Figure 11. Ablation study of hyperparameter sensitivity.

4.2. Computational Complex Analysis

In Section 3, related domain adaptation methods involved in the comparison include an adversarial learning method (AdaptSegNet), a generative-based method (GAN-RSDA), a curriculum learning method (CCDA), and two hybrid methods (MUCSS and RDG-OSA).The complete framework of CSC can also be viewed as a hybrid of generative methods and self-supervised training. Since the objective of domain adaptation methods is to improve the performance of a semantic segmentation model on unlabeled target domain, the computational complexity of all methods involved in the comparison is consistent in the model inference. In this paper, the adapted semantic segmentation model in the experiments is DeepLabv3+ with with ResNet-34 backbone. However, the computational complexity of various domain adaptation frameworks has significant differences in the training step. Generative-based methods and hybrid methods that include generative methods are much more computationally complex than other methods due to the existence of cycle generative adversarial networks. For example, the network structure of ResiDualGAN contains two generators and two discriminators. The numbers of parameters for each generator and discriminator are 41.82 M and 6.96 M, respectively, and the multiply-accumulate operations (MACs) for each generator and discriminator are 144.44 G and 7.74 G, respectively. Similarly, domain adaptation methods based on adversarial learning and curriculum learning introduce additional neural network structures, resulting in higher computational complexity than self-supervised methods with shared parameters. In summary, the complete CSC training framework is similar to other generative methods in terms of the number of parameters and the computational complexity during the training step. When the self-supervised module of CSC is used alone, since there is no additional network structure, the computational complexity and the amount of parameters are smaller than other methods.

4.3. Limitations

We have verified that the proposed cycle and self-supervised consistency training framework with different transformation functions performs well on UDA of RS image semantic segmentation, but there are many issues that remain unexplored and undeveloped in our proposed framework. Although we reduced the loss of semantic information of images during reconstruction by ResiDualGAN, we still cannot completely avoid the error of local semantic information. Such errors are exacerbated when the source and target domains have different data imaging modes. The overall adaptation effect on the Potsdam IRRG → Vaihingen IRRG is better than that on the Potsdam RGB → Vaihingen IRRG. This phenomenon arises because the latter needs to adapt not only the domain gap posed by geographical location but also the gap posed by differences in imaging modes. For example, in the Potsdam RGB to Vaihingen IRRG translation, parts of the building roofs in the Potsdam datasets that appear close to red and have complex textures are generated

as trees or low vegetation. This semantic error affects the supervised training of the source domain data in the subsequent process. In addition, the differences in the label space of domains can lead to additional problems in the unsupervised domain adaptation process. The limitations of the segmentation accuracy for all methods in this paper occur in the class of clutter/background. The official documents define this class as water, clutters, and others. However, there exists a perception difference of "clutter or background" in the manual labeling of the two datasets. Many previous studies of UDA of semantic segmentation for RS images excluded this class from the training and validation procedure [19]. Finally, we do not analyze the individual contribution of the transformation function in augmentation-based consistency or whether there exists an optimal combination of target data transformation functions when faced with certain domain adaptation scenarios. The problem of class imbalance still exists in that our loss function does not include parameters to weight classes, nor does the self-supervised consistency training step set classwise thresholds. These are issues that need to be addressed and will be explored in our subsequent work.

5. Conclusions

We addressed the task of unsupervised domain adaptation for RS image semantic segmentation by proposing a simple but effective training framework (CSC) that unifies cycle consistency with self-supervised consistency. The CSC leverages cycle-consistent generative adversarial networks to reduce distribution discrepancies between the source and target domains. To further improve the performance on the target domain, self-supervision is embedded into a supervised learning framework by consistency training, which forces the model predictions from various transformed target images to be consistent. Compared to other UDA methods for RS image semantic segmentation, our framework achieved state-of-the-art performance on the two representative benchmarks. Based on an extensive set of ablation studies, we believe that each consistency component can work independently for UDA, and the ensemble of the two consistency components further improves the performance. Moreover, this method can be embedded in various types of semantic segmentation domain adaptation methods to solve the problem of performance degradation of semantic segmentation models between datasets with large differences in spectrum, resolution, and geographic distribution. Future work will focus on different combinations of transformation functions in the self-supervised step and exploring other transformation functions in certain UDA scenarios.

Author Contributions: Conceptualization, H.G.; methodology, H.G., Y.Z.; software, H.G.; validation, H.G. and Y.Z.; formal analysis, H.G. and Y.T.; investigation, Z.S.; resources, H.G.; data curation, P.G.; writing—original draft preparation, H.G. and P.G.; writing—review and editing, Z.S. and Y.T.; visualization, H.G.; supervision, X.C.; project administration, X.C.; funding acquisition, Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Key R&D Program of China, grant number 2021YFB3900503; the National Natural Science Foundation of China, grant number 41972308, 42071312, 42171291.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Potsdam and Vaihingen datasets are published by the International Society for Photogrammetry and Remote Sensing (ISPRS) and can be accessed at https://www.isprs.org/education/benchmarks/UrbanSemLab/Default.aspx, accessed on 27 January 2022.

Acknowledgments: Our sincere gratitude goes to the anonymous reviewers for the constructive comments and suggestions that have helped improve this paper substantially.

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

References

1. Audebert, N. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
2. Zhou, K.; Chen, Y.; Smal, I.; Lindenbergh, R. Building segmentation from airborne VHR images using Mask R-CNN. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 155–161. [CrossRef]
3. Chen, K.; Fu, K.; Yan, M.; Gao, X.; Sun, X.; Wei, X. Semantic Segmentation of Aerial Images With Shuffling Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 5. [CrossRef]
4. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. [CrossRef]
5. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
6. Zhang, L.; Zhang, L.; Du, B. Advances in Machine Learning for Remote Sensing and Geosciences. *IEEE Geosci. Remote Sens. Mag.* **2016**, *19*, 22–40. [CrossRef]
7. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
8. Audebert, N.; Le Saux, B.; Lefèvre, S. Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sens.* **2017**, *9*, 368. [CrossRef]
9. Zhao, W.; Du, S.; Wang, Q.; Emery, W.J. Contextually guided very-high-resolution imagery classification with semantic segments. *ISPRS J. Photogramm. Remote Sens.* **2017**, *132*, 48–60. [CrossRef]
10. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.
11. Tuia, D.; Persello, C.; Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
12. Wang, M.; Deng, W. Deep Visual Domain Adaptation: A Survey. *arXiv* **2018**, arXiv:1802.03601.
13. Toldo, M.; Maracani, A.; Michieli, U.; Zanuttigh, P. Unsupervised Domain Adaptation in Semantic Segmentation: A Review. *Technologies* **2020**, *8*, 35. [CrossRef]
14. Benjdira, B.; Bazi, Y.; Koubaa, A.; Ouni, K. Unsupervised Domain Adaptation Using Generative Adversarial Networks for Semantic Segmentation of Aerial Images. *Remote Sens.* **2019**, *11*, 1369. [CrossRef]
15. Li, Y.; Shi, T.; Zhang, Y.; Chen, W.; Wang, Z.; Li, H. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 20–33. [CrossRef]
16. 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]
17. 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]
18. Ji, S.; Wang, D.; Luo, M. Generative Adversarial Network-Based Full-Space Domain Adaptation for Land Cover Classification From Multiple-Source Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3816–3828. [CrossRef]
19. Wittich, D.; Rottensteiner, F. Appearance based deep domain adaptation for the classification of aerial images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *180*, 82–102. [CrossRef]
20. Mateo-García, G.; Laparra, V.; López-Puigdollers, D.; Gómez-Chova, L. Cross-Sensor Adversarial Domain Adaptation of Landsat-8 and Proba-V Images for Cloud Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 747–761. [CrossRef]
21. Soto, P.J.; Costa, G.A.O.P.; Feitosa, R.Q.; Happ, P.N.; Ortega, M.X.; Noa, J.; Almeida, C.A.; Heipke, C. Domain adaptation with cyclegan for change detection in the Amazon forest. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *43*, 1635–1643. [CrossRef]
22. Kou, R.; Fang, B.; Chen, G.; Wang, L. Progressive Domain Adaptation for Change Detection Using Season-Varying Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3815. [CrossRef]
23. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv* **2020**, arXiv:1703.10593.
24. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. *arXiv* **2018**, arXiv:1704.02510.
25. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
26. Zhang, Z.; Doi, K.; Iwasaki, A.; Xu, G. Unsupervised Domain Adaptation of High-Resolution Aerial Images via Correlation Alignment and Self Training. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 746–750. [CrossRef]
27. Zhang, B.; Chen, T.; Wang, B. Curriculum-Style Local-to-Global Adaptation for Cross-Domain Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**. [CrossRef]
28. Shen, W.; Wang, Q.; Jiang, H.; Li, S.; Yin, J. Unsupervised domain adaptation for semantic segmentation via self-supervision. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 2747–2750. [CrossRef]

29. Chen, Y.; Ouyang, X.; Zhu, K.; Agam, G. Domain Adaptation on Semantic Segmentation for Aerial Images. *arXiv* **2020**, arXiv:2012.02264.
30. Liu, W.; Su, F.; Jin, X.; Li, H.; Qin, R. Bispace Domain Adaptation Network for Remotely Sensed Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 1–11. [CrossRef]
31. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. *arXiv* **2017**, arXiv:1703.05192.
32. Zhao, Y.; Gao, H.; Guo, P.; Sun, Z. ResiDualGAN: Resize-Residual DualGAN for Cross-Domain Remote Sensing Images Semantic Segmentation. *arXiv* **2022**, arXiv:2201.11523.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
34. Melas-Kyriazi, L.; Manrai, A.K. PixMatch: Unsupervised Domain Adaptation via Pixelwise Consistency Training. *arXiv* **2021**, arXiv:2105.08128.
35. Araslanov, N.; Roth, S. Self-supervised Augmentation Consistency for Adapting Semantic Segmentation. *arXiv* **2021**, arXiv:2105.00097.
36. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual Event, 12–18 July 2020; pp. 1597–1607.
37. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. *arXiv* **2019**, arXiv:1905.02249.
38. Berthelot, D.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. *arXiv* **2020**, arXiv:1911.09785.
39. Sohn, K.; Berthelot, D.; Li, C.L.; Zhang, Z.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
40. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6256–6268.
41. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6023–6032.
42. Gerke, M. *Use of the Stair Vision Library within the ISPRS 2D Semantic Labeling Benchmark (Vaihingen)*; University of Twente: Enschede, The Netherlands, 2015. [CrossRef]
43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. Available online: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (accessed on 2 January 2022).
44. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:1802.02611.
45. Tsai, Y.H.; Hung, W.C.; Schulter, S.; Sohn, K.; Yang, M.H.; Chandraker, M. Learning to adapt structured output space for semantic segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7472–7481. [CrossRef]
46. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.