The Photogrammetric Record

ORIGINAL ARTICLE

SeConDA: Self-Training Consistency Guided Domain Adaptation for Cross-Domain Remote Sensing Image Semantic Segmentation

Bin Zhang[1] | Yongjun Zhang[2] | Chengdu Cao[1] | Yi Wan[2] | Yongxiang Yao[1] | Liang Fei[1]

[1]China Railway Siyuan Survey and Design Group CO., LTD., Wuhan, China | [2]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

Correspondence: Yongjun Zhang (zhangyj@whu.edu.cn) | Chengdu Cao (003095@crfsdi.com)

Received: 23 May 2024 | **Revised:** 26 November 2024 | **Accepted:** 31 December 2024

Funding: This work was supported in part by the National Natural Science Foundation of China under project number 42030102, and the Postdoctoral Research Project for China Railway Siyuan Survey and Design Group Co., LTD. under Grant KY2023127S, and the China Postdoctoral Science Foundation under Grant 2024M753810.

Keywords: consistency | domain adaptation | remote sensing | self-training | semantic segmentation

ABSTRACT

Well-trained remote sensing (RS) deep learning models often encounter a considerable decline in performance when applied to images that differ from the training data. This decline can be attributed to variations in imaging sensors, geographic location, imaging time, and radiation levels during image acquisition. Consequently, the widespread application of these models has been greatly impeded. An envisioned resolution to confront this challenge encompasses formulating a cross-domain RS image semantic segmentation network integrated with self-training consistency. This approach involves the generation of high-quality pseudo-labels for images in the target domain, which are then used to guide the training of the network. To enhance the model's ability to learn the data distributions of both the source and target domains, highly perturbed mixed samples are created by blending images from these domains. Additionally, adversarial training is incorporated to reduce the entropy of the model's predicted results, thereby mitigating the influence of noise present in the pseudo-labels. As a result, this approach effectively extracts domain-invariant features and minimizes the disparities between the distributions of the different domains. By employing the ISPRS and LoveDA datasets in a series of experiments conducted across varied scenarios, our empirical investigations evince the capacity of the proposed methodology to generalize the model to target domain data, which is achieved through the mitigation of disparities between domain distributions. It effectively alleviates the domain shift issues caused by differences in imaging locations and band combinations in RS image data and achieves state-of-the-art results and validates its effectiveness.

1 | Introduction

The primary objective of traditional machine learning is to construct a model that minimizes the potential risks associated with test data, utilizing a provided training sample. The prevalent and effective approach to train such a model is supervised learning. The effectiveness of this approach heavily relies on the availability of abundant labeled training data. Furthermore, it assumes that both the training and test datasets are derived from the same distribution and possess similar joint probability distributions. Nevertheless, in practical scenarios, it is frequently challenging to meet this condition, resulting in the test data originating from distinct feature spaces or distributions. When the training data fail to accurately represent the distribution of the test data (Csurka, Volpi, and Chidlovskii 2021; Liu, Yoo, et al. 2022), specifically when

encountering out-of-distribution data, the performance of the trained model significantly deteriorates when applied to the test data.

In the realm of remote sensing (RS), the performance of trained models often suffers when applied to images that differ from the training data (Tuia, Persello, and Bruzzone 2016; Peng et al. 2022). This is primarily due to variations in imaging sensors, image resolution, geographic location, imaging time, angle, and radiation. As a result, satisfactory results are not achieved, hindering the widespread use of deep learning models in RS (Peng et al. 2023). To overcome this obstacle, it is imperative to take these factors into account during model development and ensure their robustness in handling such variations.

Researchers have introduced a novel research area in machine learning known as domain adaptation (DA) to address the domain shift problem between the source domain (SD) and the target domain (TD) (Liu, Yoo, et al. 2022). Notably, within the unsupervised domain adaptation (UDA) paradigm, data in the TD often lack labels. The primary objective of DA is to train a model by leveraging both SD and TD data, with the intention of reducing the disparity between the feature distributions of these domains and aims to facilitate effective generalization of the model to the TD.

Domain adaptation tasks aim to learn representations that remain invariant across domains, thus facilitating cross-domain generalization, that is, learning domain-invariant features. To enhance the precision of the TD, numerous DA methodologies incorporate a style transfer network. This network is utilized to transform the image style of the TD to resemble that of the SD. This transformation enables the generation of corresponding labels (Hoffman et al. n.d.; Murez et al. n.d.; Wu et al. n.d.; Chen et al. n.d.; Li, Yuan, and Vasconcelos n.d.; Cheng et al. n.d.). While these methods effectively address the domain shift issue caused by radiometric disparities and variations in band combinations, they often introduce additional networks that lack elegance and stability during the training phase. Moreover, they struggle to handle discrepancies in geographic locations. In order to better tackle the domain shift problem, feature-level DA methods focus on aligning features in the hidden space rather than at the input level. Through adversarial training, these methods mitigate differences in features, allowing the network to concentrate on extracting domain-invariant features (Tsai et al. n.d.; Vu et al. n.d.; Pan et al. n.d.). Additionally, some other DA methodologies capitalize on unlabeled data sourced from the TD to augment its accuracy through self-training as an adjunctive approach (Zou et al. n.d.; Mei et al. n.d.; Araslanov and Roth n.d.; Melas-Kyriazi and Manrai n.d.; Tranheden et al. n.d.; Wang et al. n.d.-a; Zheng and Yang 2021; Hoyer, Dai, and Van Gool n.d.). Self-training, initially employed within the domain of semi-supervised learning, entails the generation of pseudo-labels for the TD data, subsequently utilized in training the network. However, these methods are not sufficiently robust due to the influence of noise in the pseudo-labels.

In this research, drawing inspiration from the aforementioned methodologies, we present a novel approach for cross-domain RS image semantic segmentation, which we refer to as SeConDA. The overall network architecture employs a teacher–student network. To leverage the knowledge from the TD, we propose a pseudo-label supervision strategy based on self-training consistency. This strategy supervises the network training by calculating losses on high-quality pseudo-labels generated from the TD images. Simultaneously, we address the domain shift issue between the SD and TD by introducing a consistency regularization method based on mixed samples. This method learns the SD and TD data concurrently by utilizing highly perturbed mixed samples. Furthermore, we propose an entropy-based adversarial training technique to position the decision boundary of the model in a low-density region. This helps reduce the entropy of predictions from both the SD and TD and facilitates the extraction of domain-invariant features. We conducted experiments on the ISPRS dataset (Rottensteiner et al. 2013) under four different migration scenarios and on the LoveDA dataset (Wang et al. 2021) under two different migration scenarios. The experimental results validate the effectiveness of our proposed method in generalizing to the TD data by minimizing the discrepancy between domain distributions. These results demonstrate that our approach achieves state-of-the-art performance.

The main contributions of this thesis are as follows:

1. To alleviate the problem of the existence of domain shift between the SD and TD, a consistent regularization method based on mixed samples is proposed to implicitly learn the SD and TD data simultaneously by obtaining the highly perturbed mixed samples.
2. The proposed entropy-based adversarial training aims to extract domain-invariant features, enhancing the transferability of the network model. This approach effectively addresses the domain shift problem arising from variations in imaging positions and band combination methods within RS image data.

To maintain a coherent structure, the rest of this paper is arranged as follows. Section 2 provides a literature review on semantic segmentation and DA. Section 3 describes the methodology proposed in this paper. Section 4 analyzes the methodology experimentally. Section 5 serves as the discussion section, delving into the methodology. Lastly, Section 6 concludes the paper.

2 | Related Work

2.1 | Supervised Semantic Segmentation

For a considerable duration, semantic segmentation has presented itself as a complex undertaking, prompting the exploration of various approaches employing supervised learning to tackle this challenge (Garcia-Garcia et al. 2017). By harnessing the power of convolutional operations, the FCN family of networks amalgamates dilated convolution and context modules to continuously refine accuracy (Long, Shelhamer, and Darrell n.d.; Ronneberger, Fischer, and Brox n.d.; Chen

et al. 2014; Zhao et al. n.d.). Furthermore, attention-based mechanisms further enhance the network's capacity to comprehend contextual information (Fu et al. n.d.). More recently, the integration of transformer-based networks in semantic segmentation, inspired by natural language processing, has facilitated the establishment of longer term dependencies between pixels (Dosovitskiy et al. 2020; Zheng et al. n.d.).

2.2 | DA For Semantic Segmentation in Computer Vision

Despite achieving commendable results on their test sets, certain models trained on accurately annotated datasets exhibit subpar performance when confronted with out-of-distribution images. UDA endeavors to alleviate the distribution misalignment between the SD and TD. Research in the realm of deep learning-based DA methods is commonly categorized into three main groups: image-style transformation-based methods, adversarial training-based methods, and self-training-based methods.

Methods that rely on image style transformation typically utilize generative adversarial networks (GANs) to transfer the SD image to the TD image. Subsequently, the segmentation network can be trained on the target-style SD data along with its corresponding labels (Hoffman et al. n.d.; Murez et al. n.d.; Wu et al. n.d.; Chen et al. n.d.; Li, Yuan, and Vasconcelos n.d.; Cheng et al. n.d.). In order to enhance the transfer performance of the segmentation model, many approaches integrate image style transformation with additional regularizations. Among these regularizations, cyclic consistency loss and semantic consistency loss, as proposed by (Hoffman et al. n.d.), are the most commonly employed. The proposed method in their study involves an image-to-image transformation and utilizes CycleGAN to calculate the consistency loss.

Adversarial training-based methods typically aim to minimize the disparity in network features or outputs between SD and TD images using GANs (Tsai et al. n.d.; Vu et al. n.d.; Pan et al. n.d.). Given the complex structure of the image segmentation task, aligning the potential features of both domains solely proves to be a difficult endeavor. Consequently, domain alignment is commonly conducted at various network layers to reduce the discrepancy in feature distribution. To enhance the adaptability of the model, Tsai et al. (n.d.) developed a multilevel adversarial network that minimizes the divergence in output distribution across different layers of the network.

Unlike approaches based on adversarial training which aim to reduce domain variance, self-training-based approaches achieve DA by utilizing unlabeled data from the TD (Zou et al. n.d.; Mei et al. n.d.; Araslanov and Roth n.d.; Melas-Kyriazi and Manrai n.d.; Tranheden et al. n.d.; Wang et al. n.d.-a; Zheng and Yang 2021; Hoyer, Dai, and Van Gool n.d.). Self-training-based methods employ a multi-round training scheme initially devised for semi-supervised learning, but they have recently found application in UDA as well. The procedure of self-training deep network-based UDA comprises two main stages: initially, the generation of pseudo-labels for the TD data, followed by training the network using both the TD data and the pseudo-labels derived from it. However, a significant challenge with self-training-based approaches lies in the fact that the pseudo-labels in the TD might be noisy, making a significant portion of them unreliable. To address this issue, it is crucial to select predictions with high confidence to refine the pseudo-labeling process. Zheng and Yang (2021) introduced a technique aimed at explicitly estimating prediction uncertainty throughout the training process, which involves modeling prediction variance and integrating it into the optimization objective. Additionally, to mitigate the impact of low-quality pseudo-labeling caused by domain bias, Tranheden et al. (n.d.) trained a model by ensuring prediction consistency across domain-mixed images. This was achieved by mixing images from both domains along with their corresponding labels and pseudo-labels.

2.3 | DA For Semantic Segmentation in RS

In the RS, data bias poses a significant concern, arising from variations in imaging sensors, geographic locations, and atmospheric conditions. This bias adds complexity to the problem of DA in the RS, affecting various RS tasks. Previous research on DA in the RS has primarily concentrated on scene classification. However, there has been a recent trend toward exploring DA in the context of semantic segmentation for RS images.

Yan et al. (2019) introduced a ternary adversarial DA technique that takes into account two domains to train domain invariant feature classifiers using a domain similarity discriminator. Their approach leverages information from both domains to minimize the distribution gap between them and generates confident predictive labels for the target data through the discriminator, which are then used as pseudo-labels for re-training. Tasar et al. (2020) proposed an alternative method leveraging a color mapping GAN. This method generates synthetic training images that retain a semantic resemblance to the original training data while aligning their spectral distributions with those of the test images. These synthetic images are then employed to refine the performance of the trained classifiers. Liu, Su, et al. (2022) introduced a novel approach involving a two-branch structured network capable of extracting features from both image and wavelet domains simultaneously. Furthermore, they presented a dual-space adversarial learning strategy employing two discriminators operating in distinct spaces. One discriminator aligns the feature distributions between the SD and TD, while the other contributes to generating a coherent spatial layout for the classification output. Zheng et al. (2021) proposed an entropy-guided adversarial learning method that utilizes adaptive weights learned from the target prediction probability maps. These weights facilitate local feature alignment between domains, enabling the measurement of interdomain differences. Chen, Zhu, et al. (2022) introduced a UDA framework utilizing adversarial learning to align high-level features, aiming to reduce the semantic gap between the SD and TD. They incorporated an attention module to direct the classifier's attention toward features aligned at the category level. Chen, Pan, and Chong (2022) also presented domain discriminators adapted to regions and categories, aiming to emphasize variations between different regions and categories during the alignment process. Cai et al. (2022) suggested an iterative intra-DA framework with a generator selection strategy to enhance image-to-image conversion performance using GANs. Additionally, they enhance

the quality of pseudo-labels by filtering out high-entropy and low-confidence pixels from the prediction map. Chen, Zhang, et al. (2022) developed a DA network based on mutual information, which incorporates multi-task learning within an adversarial network to focus on domain invariant information by simultaneously learning segmentation and height. Li et al. (2021) proposed an objective function incorporating multiple constraints for semantic segmentation DA.

3 | Method

3.1 | Overall Framework

In this research, we present SeConDA (self-training consistency guided DA network), a cross-domain RS image semantic segmentation network, illustrated in Figure 1. The network architecture follows a teacher–student framework, where both the student network S (SN) and the teacher network T (TN) share the same structure. The parameters of the SN are denoted as θ_S, while those of the TN are denoted as θ_T. During the training phase, our method focuses on enhancing domain-invariant features by incorporating three key strategies: pseudo-label supervision, mixed sample supervision based on self-training consistency, and uncertainty-based adversarial training. As a result, the entire framework consists of four components: a SD image supervision branch, a TD pseudo-label supervision branch, a cross-domain image mixing branch, and an adversarial training branch. In the testing phase, only the SN is utilized, while the TN and discriminator network are not involved in the inference process.

3.2 | SD Image Supervision Branch

Within the source-domain image supervision branch, the SN receives the source-domain data x_S for processing, as shown in Figure 1. As is common with semantic segmentation methodologies, the SN undergoes supervised training, utilizing a pixel-level cross-entropy (CE) function denoted as $\mathcal{L}_{\mathrm{CE}}$ for loss calculation:

$$\mathcal{L}_{\mathrm{CE}}(x_S, y_S; \theta_S) = -\frac{1}{HW}\sum_i \sum_c \mathbb{1}_{[y_S=c]} \log P^c_{s,i} \tag{1}$$

where the SD image is $x_S \in \mathcal{R}^{H\times W\times 3}$, the corresponding label is $y_S \in \mathcal{R}^{H\times W\times C}$. H, W, C, and c are the length, width, number of categories, and the corresponding label category, respectively. $\mathbb{1}_{[y_S=c]}$ is the one-hot label vector, and $P^c_{s,i}$ stands the Softmax probability of the SN to predict the SD image, where index $i \in \{1, 2, \cdots, H \times W\}$.

3.3 | Target Domain Pseudo-Label Supervision Branch

To mine the knowledge of the TD, the image x_T from the TD is input into the TN, yielding pseudo-label $\hat{y}_T$ during the training process:

$$\hat{y}_T = \underset{c}{\mathrm{argmax}}\, T(x_T)^c_i \tag{2}$$

In the absence of labeled data in the TD, the SN relies on pseudo-labels generated by the TN to facilitate the training process. Consequently, the loss function $\mathcal{L}_{\mathrm{PLCE}}$, which measures the CE of the pseudo-labels, can be mathematically represented as:

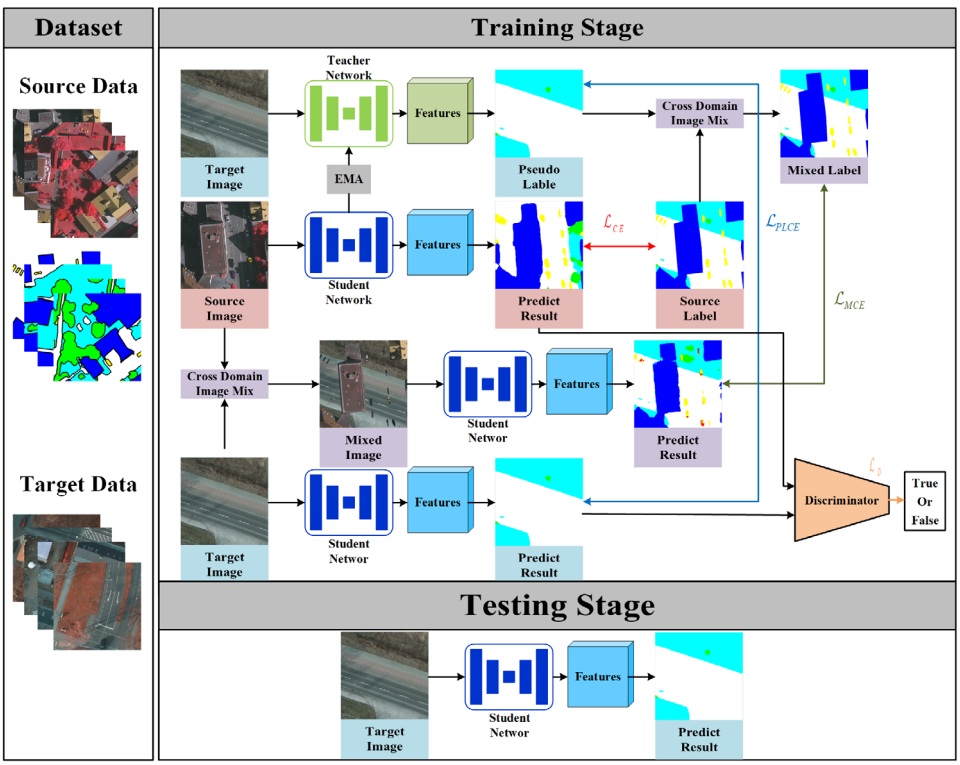


FIGURE 1 | Schematic diagram of our semantic segmentation method for cross-domain RS images.

$$\mathcal{L}_{\mathrm{PLCE}}(x_T;\theta_s) = -\frac{1}{HW}\sum_i\sum_c \mathbb{1}_{[\hat{y}_T=c]}\log P^c_{T,i} \tag{3}$$

where the TD image $x_T \in \mathcal{R}^{H\times W\times 3}$, $\mathbb{1}_{[\hat{y}_T=c]}$ represents the one-hot vector of the pseudo-label $\hat{y}_T$, and $P^c_{T,i}$ denotes the Softmax probability of SN to predict the TD image.

Due to the difference in data distribution between the SD and TD, noise inevitably creeps into the pseudo-labels. Hence, a pseudo-label supervision style centered on self-training is embraced. In this context, pixels with predicted confidence levels surpassing a designated threshold are utilized for loss computation. Consequently, the loss function $\mathcal{L}_{\mathrm{PLCE}}$ is adjusted to:

$$\mathcal{L}_{\mathrm{PLCE}}(x_T;\theta_s) = -\frac{1}{HW}\sum_i\sum_c \mathbb{1}_{\left[\max_c P^c_{T,i}>t\right]} \bullet \mathbb{1}_{[\hat{y}_T=c]}\log P^c_{T,i} \tag{4}$$

where t denotes the threshold parameter. We follow the settings of DACS and DAFormer and set the threshold parameter t to 0.968, and only pixel regions larger than this threshold are used as pseudo-labels to participate in the loss calculation.

By employing an exponential moving average (EMA) approach, the parameters of the TN are iteratively adjusted according to those of the SN.

$$\theta'_T = \alpha_{\mathrm{EMA}}\theta_T + (1-\alpha_{EMA})\theta_S \tag{5}$$

where α_{EMA} represents a hyperparameter representing the smoothing coefficient, set to 0.999.

3.4 | Cross-Domain Image Mixing Branch

The technique of mixed sample-based data augmentation combines pixels from two training images to create a new sample that is highly perturbed. This technique has demonstrated efficacy in endeavors like image classification and semantic segmentation. Zhang et al. (2023) conducted a study that showcases the effectiveness of mixed sample-based data augmentation in semi-supervised semantic segmentation. In the context of UDA, there are two types of data: SD images with corresponding labels, and TD images with pseudo-labels predicted by a TN. Nonetheless, owing to the discrepancy between the SD and TD, it may not be optimal for the network to exclusively depend on acquiring these two varieties of data autonomously. Therefore, this method employs the ClassMix algorithm (Yun et al. n.d.) to introduce fresh data through cross-domain mixing. Initially, ClassMix selects half of the classes from a given prediction output to create a mask M, and then transfers the corresponding pixels to a second image to create a significantly perturbed sample, as illustrated in Figure 2. Once the mixed image x_M and the corresponding label y_M are obtained, the mixed image is inputted into the SN, and the mixed label is utilized to calculate the CE loss $\mathcal{L}_{MCE}$ on the network's output:

$$\mathcal{L}_{MCE}(x_M;\theta_s) = -\frac{1}{HW}\sum_i\sum_c \mathbb{1}_{[y_M=c]}\log P^c_{M,i} \tag{6}$$

where $\mathbb{1}_{[y_M=c]}$ is the one-hot vector of mix label y_M, and $P^c_{M,i}$ is the Softmax probability of the SN predicting the mix image.

As with the TD pseudo-label supervision branch, the cross-domain image mixing branch implements a self-training strategy based on pseudo-labels. By calculating the proportion w of pixels predicted by the target-domain image with a trust level above a specific threshold, we generate a weight mask map w_M for the loss, that is,

$$w = \frac{1}{HW}\sum_i \mathbb{1}_{\left[\max_c P^c_{T,i}>T\right]} \tag{7}$$

$$w_M = M \odot \mathbb{1} + (1-M) \odot (\mathbb{1} \bullet w) \tag{8}$$

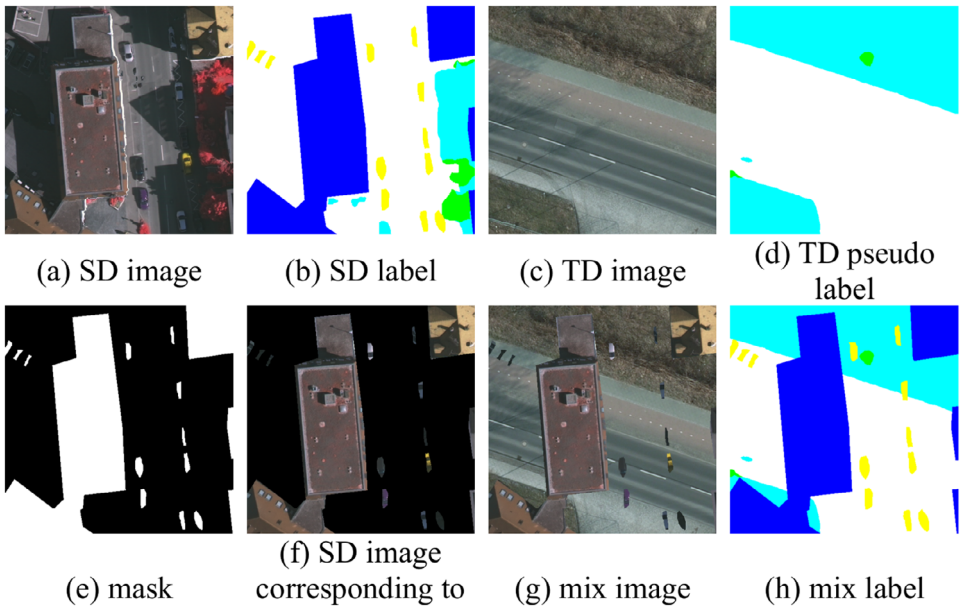


FIGURE 2 | Schematic of cross-domain image mix.

Therefore, the loss function $\mathcal{L}_{MCE}$ becomes

$$\mathcal{L}_{MCE}(x_M;\theta_s) = -\frac{1}{HW}\sum_i\sum_c w_M \mathbb{1}_{[y_M=c]}\log P^c_{M,i} \quad (9)$$

Through the online generation of pseudo labels from mixed samples, this method enables the network to learn features from both the SD and TD as training progresses. Despite the potential for artifacts from pseudo-labels, the network can effectively minimize the negative impact of domain shift by acquiring domain invariant features through pseudo-label filtering.

3.5 | Adversarial Training Branch

The SN tends to make overly confident predictions when trained solely on data from the SD, resulting in prediction outcomes with low entropy. Conversely, predictions made on data from the TD are more likely to yield chaotic results, leading to prediction outcomes with high entropy. Analyzing the entropy map of the prediction outcomes simplifies the process of discerning whether the input image pertains to the SD or the TD.

Our approach also incorporates entropy-driven adversarial training to guarantee that the decision boundary of the model is situated in a region of low density. As a consequence, there is a decrease in entropy for predictions within both the SD and TD, concurrently extracting features invariant across domains. Within the adversarial training component, a completely convolutional discriminator is introduced. Throughout the training process, the SN is presented with input images originating from either the SD or TD, while the fully convolutional discriminator produces feature maps from the SN and calculates their entropy. Subsequently, it attempts to categorize the input feature map as pertaining to either the SD or TD. The SN, on the other hand, aims to deceive the discriminator by ensuring that the feature distributions from both domains exhibit similarity.

Specifically, Shannon entropy serves as a metric for uncertainty in network predictions. When analyzing an image x, its entropy is computed and then normalized to a range of [0,1].

$$E_x^{(h,w)} = \frac{-1}{\log(C)}\sum_{c=1}^{C} P_x^{(h,w,c)}\log P_x^{(h,w,c)} \quad (10)$$

where $P_x^{(h,w,c)}$ is the Softmax probability of the network prediction result.

The entropy map is subsequently employed as input for a fully convolutional discriminator D. Consisting of five convolutional layers, each with a kernel size of 4 and a stride of 2, this network has channel counts of 64, 128, 256, 512, and 1 for the respective layers. Following each convolutional layer except the final one, a Leaky ReLU serves as the activation function. The primary objective of the discriminator network is to ascertain if the input pertains to the SD or TD. In other words, labels belonging to the SD are designated a value of 1, whereas those from the TD are assigned a value of 0. Consequently, the training loss $\mathcal{L}_D$ of the discriminator network is calculated as follows:

$$\mathcal{L}_D(x_S,x_T;\theta_D) = \frac{1}{HW}\sum_{x_S}\mathcal{L}_{CE}(D(E_{x_S}),1) + \frac{1}{HW}\sum_{x_T}\mathcal{L}_{CE}(D(E_{x_T}),0) \quad (11)$$

where discriminator parameters are denoted as θ_D. The entropy maps E_{x_S} and E_{x_T} correspond to predictions made by the SN for images from the SD and TD, respectively.

The adversarial training loss $\mathcal{L}_{D-S}$ for the SN is as follows:

$$\mathcal{L}_{D-S}(x_T;\theta_s) = \frac{1}{HW}\sum_{x_T}\mathcal{L}_{CE}(D(E_{x_T}),0) \quad (12)$$

To summarize, the overall loss incurred by the SN is as follows:

$$\mathcal{L}_S(x_S,x_T;\theta_D) = \mathcal{L}_{CE} + \lambda_{PLCE}\mathcal{L}_{PLCE} + \mathcal{L}_{MCE} + \lambda_{adv}\mathcal{L}_{D-S} \quad (13)$$

where λ_{PLCE} and λ_{adv} are hyperparameters used to balance the weight of the TD pseudo-label supervision branch and the adversarial training branch in the total loss.

4 | Experimental Results

4.1 | Experimental Dataset

Extensive experiments were carried out on the Vaihingen dataset and Potsdam dataset of ISPRS (Rottensteiner et al. 2013) to validate the proposed semantic segmentation DA method. The Vaihingen dataset images are situated in villages with sparse building layout and a resolution of 9 cm, while the Potsdam dataset images are in cities with dense buildings and a resolution of 5 cm. The Vaihingen dataset has three bands (NIR, R, G), whereas the Potsdam dataset has four bands (NIR, R, G, B), leading to two band selection methods for the Potsdam dataset: false color images (NIR, R, G) and true-color images (R, G, B). Considering four different DA scenarios in Table 1.

1. Both the false color image serves as the SD and TD. Despite differences in resolutions and imaging positions, the band combination remains consistent across the images.
2. The SD and TD are swapped from Case 1. Unlike Case 1, the Vaihingen dataset in Case 2 has a smaller sample

TABLE 1 | Four different domain adaptation scenarios for ISPRS dataset.

Domain case	Source domain	Target domain
1	Potsdam, NIR, R, G	Vaihingen, NIR, R, G
2	Vaihingen, NIR, R, G	Potsdam, NIR, R, G
3	Potsdam, R, G, B	Vaihingen, NIR, R, G
4	Vaihingen, NIR, R, G	Potsdam, R, G, B

size in the SD data, making Case 2 more challenging to migrate.

3. The SD in Case 3 is the true color image of the Potsdam dataset, while the TD is the false color image of the Vaihingen dataset. With varying resolutions, imaging positions, and image band combinations, transferring data in Case 3 become more complex.
4. Case 4 involves exchanging the SD and TD from Case 3. The Vaihingen dataset, used as the SD in Case 4, has a limited sample size, making this case more challenging than Case 3.

When utilizing the Vaihingen dataset as the TD data, the training set consists of 16 unlabeled images from the Vaihingen dataset and all images from the Potsdam dataset, while the test set comprises the remaining 17 images from the Vaihingen dataset. In the case of using the Potsdam dataset as the TD data, the training set includes 24 unlabeled images from the Potsdam dataset and all images from the Vaihingen dataset, with the test set composed of the remaining 14 images from the Potsdam dataset.

The LoveDA dataset (Wang et al. 2021) was subjected to experiments to delve deeper into its potential applications. This dataset comprises 5987 Google Earth images collected from Nanjing, Changzhou, and Wuhan. Due to the distinct planning strategies employed in each city, the arrangement of buildings varies significantly. The LoveDA dataset encompasses a total of 18 distinct areas, with nine urban areas located in densely populated districts and the remaining nine areas classified as rural. The dataset consists of 2713 images for urban areas and 3274 images for rural areas. The data have a spatial resolution of 0.3 m, with the images having been geometrically aligned and preprocessed. Additionally, the images for each area have been cropped into 1024×1024 image blocks. The labels assigned to the LoveDA dataset are categorized into seven classes: building, road, water, barren, forest, agriculture, and background. To ensure comprehensive evaluation, the dataset is partitioned into training, validation, and test sets, all of which are spatially independent from one another. The LoveDA dataset encompasses two DA tasks:

1. Rural→Urban. The SD data contain 1366 images, the TD data contain 1156 images, and the validation set data contain 677 images. The test set data contain 820 images.
2. Urban→ Rural. The SD data contain 1156 images, the TD data contain 1366 images, and the verification set data contain 992 images. The test set data contain 976 images.

4.2 | Implementation Details

In this experiment, DeepLab V2 (Chen et al. 2017) was employed as both the SN and TN, with ResNet-50 serving as the backbone network. The training was done using the AdamW optimizer. The learning rates were 6×10^{-5} for the backbone network, 6×10^{-4} for other parts, and weight decay set at 0.01. The batch size is 4. The learning rate adjustment strategy was poly, with a total of 40,000 iterations, and the learning rate for the first 1500 iterations increased slowly from 1×10^{-6}. Data augmentation included random flipping, and training set images were cropped to 512×512 pixels. Loss hyperparameters λ_{PLCE} and λ_{adv} were 0.01 each, threshold t was set to 0.968, and EMA hyperparameter α_{EMA} was 0.999. Each experiment was repeated three times for accurate evaluation using mIoU as the metric.

4.3 | Ablation Study

To validate the efficacy of the proposed method, the NIRRG data from the Potsdam dataset were employed as the SD, while the Vaihingen dataset was utilized as the TD. A series of ablation experiments were conducted on the method's hyperparameters, including λ_{PLCE} and λ_{adv}, as well as the combination method of different branches and the EMA hyperparameter. Additionally, the lower bound (referred to as Source only) and upper bound (referred to as Oracle) of accuracy were evaluated for comparison. When solely utilizing images from the SD for training, the lower bound achieved a mIoU of 48.63. Conversely, utilizing images from the TD for training resulted in an upper bound mIoU of 79.31.

4.3.1 | The Impact of Loss Hyperparameters

The loss analysis involves the selection of weight hyperparameters, λ_{PLCE} and λ_{adv}. Due to the impracticality of testing all possible values, this experiment utilized seven different values for λ_{PLCE} (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0) and nine different values for λ_{adv} (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0). The experimental results for each value were evaluated and the experimental setup and implementation details are as described

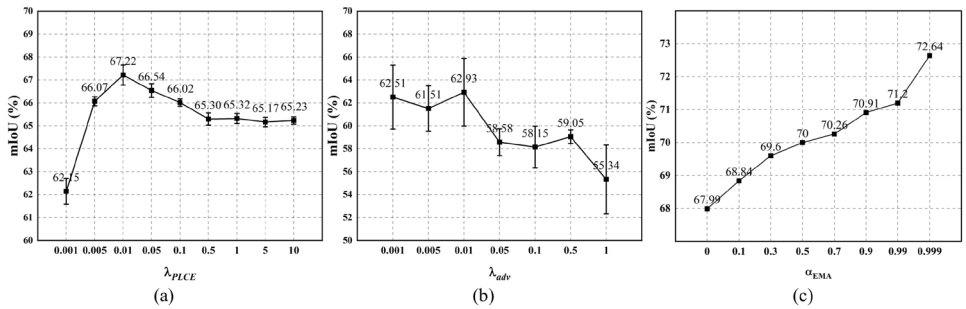


FIGURE 3 | The impact of different weight hyperparameters λ_{PLCE} and λ_{adv}, and different EMA hyperparameters α_{EMA} on accuracy.

earlier. The experimental results are shown in Figure 3a,b. By incorporating both the TD pseudo-label supervision branch and the adversarial training branch, the method demonstrates significant improvement in comparison to training solely with SD images. For the weight hyperparameter λ_{PLCE} of the target-domain pseudo-labeling supervised branch, the performance of the method tends to increase and then decrease with the increase of λ_{PLCE}. Notably, the method achieves the highest accuracy of 67.22 mIoU when $\lambda_{PLCE}=0.01$. Similarly, for the weight hyperparameter λ_{adv} of the adversarial training branch, the highest accuracy of 62.93 mIoU can be achieved when $\lambda_{adv}=0.01$. Therefore, for the subsequent experiments, λ_{PLCE} and λ_{adv} were both set to 0.01.

4.3.2 | The Impact of Different Branches

The integration of the TD pseudo-label supervision branch, the cross-domain image mixing branch, and the adversarial training branch yields more efficient domain-invariant features compared to using a single branch. To validate the approach, different ways of combining the above three branches were tried for evaluation. The results are reported in Table 2. When employing a single branch, namely the TD pseudo-label supervision branch, the cross-domain image mixing branch, and the adversarial training branch, the achieved mIoU values are 67.22, 70.22, and 62.93, respectively. These values represent improvements of 18.59, 22.19, and 14.30 compared to the lower bound of accuracy, respectively. By combining two of the transformations, the TD pseudo-label supervision branch and the cross-domain image mixing branch yield a mIoU of 71.64. Similarly, the combination of the target-domain image pseudo-label supervised branch and the adversarial training branch results in a mIoU of 69.25, while the combination of the cross-domain image mixing branch and the adversarial training branch achieves a mIoU of 72.09. Furthermore, when all three branches are combined simultaneously, the accuracy is further enhanced to 72.64 mIoU. This indicates that the concurrent utilization of these three branches enables the extraction of more effective domain-invariant features. Consequently, the default choice for subsequent experiments is to employ the combination of these three branches.

TABLE 2 | The impact of different branch combinations on accuracy.

Pseudo-label supervised ranch	Cross-domain image mixing branch	Adversarial training branch	mIoU
Source only			48.63
Oracle			79.31
√			67.22 (+18.59)
	√		70.82 (+22.19)
		√	62.93 (+14.30)
√	√		71.64 (+23.01)
√		√	69.25 (+20.62)
	√	√	72.09 (+23.46)
√	√	√	**72.64** (+24.01)

*Note:*The best value for each metric evaluated is shown in bold.

TABLE 3 | Quantitative transfer results from Potsdam NIRRG to Vaihingen NIRRG.

Method	Imp. Surf.	Building	Low Veg.	Tree	Car	mIoU
Source only	47.66	66.31	33.06	58.14	37.97	48.63
AdaptSeg (Tsai et al. n.d.)	67.60	74.66	45.90	63.55	3.98	51.14
AdvEnt (Vu et al. n.d.)	73.11	83.39	51.09	65.49	12.30	57.08
Seg-Uncert (Zheng and Yang 2021)	79.72	85.99	41.07	62.32	41.98	62.22
DACS (Tranheden et al. n.d.)	78.55	80.90	57.97	64.02	60.59	68.41
DaFormer (Hoyer, Dai, and Van Gool n.d.)	77.12	81.66	56.31	66.75	59.49	68.27
UemDA	69.97	80.43	53.90	65.63	35.87	61.16
Oracle	84.33	90.45	69.08	79.77	72.93	79.31
SeConDA	**81.26**	**88.57**	**59.82**	**67.42**	**66.15**	**72.64**

*Note:*The best value for each metric evaluated is shown in bold.

4.3.3 | The Impact of EMA Hyperparameters

EMA hyperparameter α_{EMA} was subjected to ablation experiments to assess its impact. The results, illustrated in Figure 3 (c) clearly demonstrate the gradual improvement in the model's performance as α_{EMA} increases. When α_{EMA} is set to 0, the proposed method achieves a mIoU of 67.99. At this point, the TN and the SN have identical parameters, resulting in no guidance from the TN. Consequently, the enhancement in model performance primarily stems from the cross-domain image mixing branch and the adversarial training branch. As α_{EMA} increases, the teacher model averages the student model's parameters throughout the training phase, leading to a more accurate model. When α_{EMA} reaches 0.999, the accuracy reaches 72.64 mIoU. Therefore, subsequent experiments utilized a α_{EMA} value of 0.999.

4.4 | Comparison With State-of-the-Art Methods in ISPRS Data

4.4.1 | Transfer Results From Potsdam-NIRRG to Vaihingen-NIRRG

The effectiveness of the proposed method is validated by utilizing the NIR-R-G dataset of Potsdam as the SD and the NIR-R-G dataset of Vaihingen as the TD. In Table 3 the quantitative evaluation of the proposed method, along with other methods and the upper and lower bounds of DA, is presented. Additionally, Figure 4 showcases the qualitative results.

In contrast to the upper bound, the accuracy of the lower bound significantly decreases, resulting in a substantial drop in mIoU from 79.31 to 48.63. This decrease highlights the significant

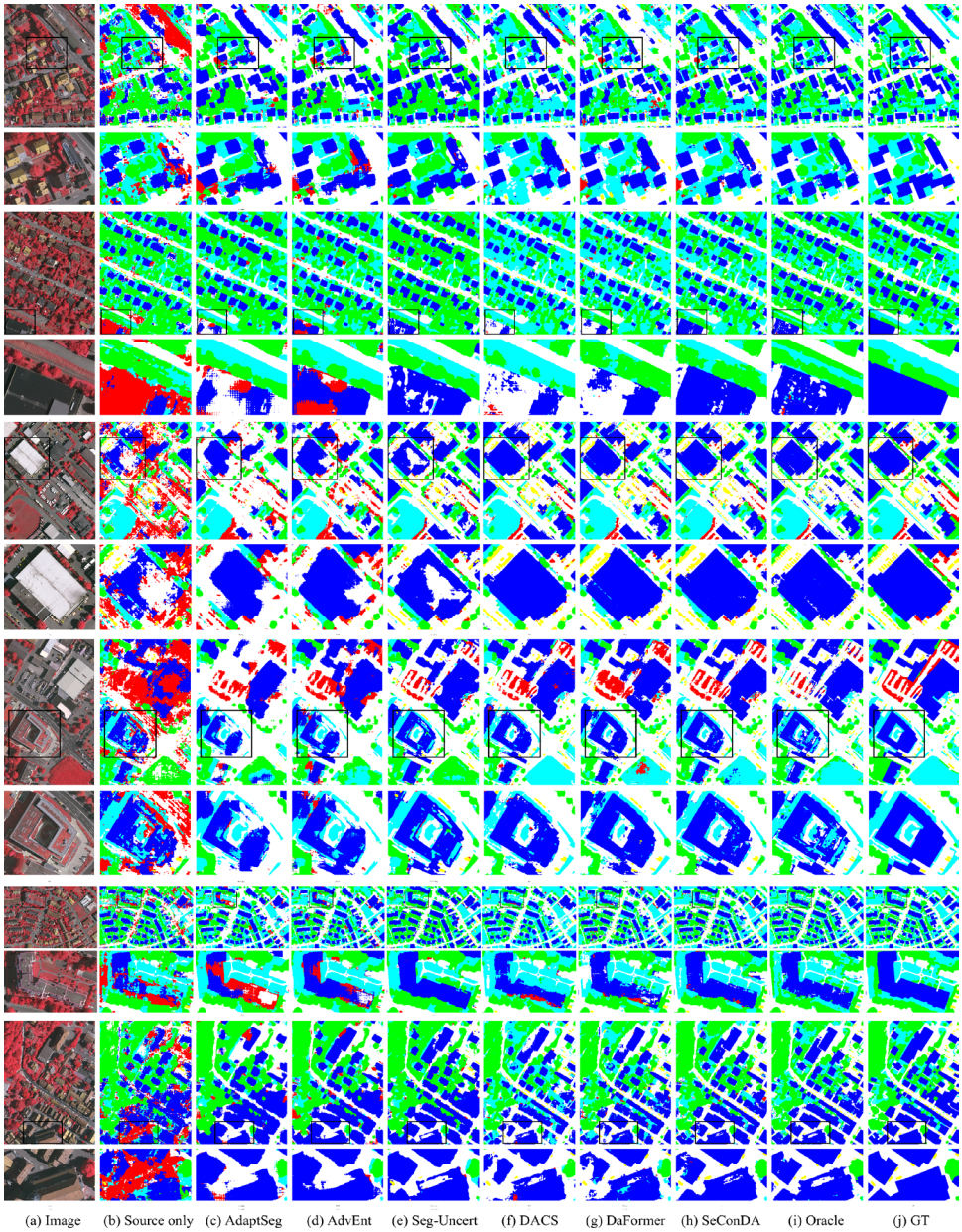


FIGURE 4 | Visualization results of transfer from Potsdam NIRRG data to Vaihingen NIRRG data.

TABLE 4 | Quantitative transfer results from Vaihingen NIRRG to Potsdam NIRRG.

Method	Imp. Surf.	Building	Low Veg.	Tree	Car	mIoU
Source only	54.44	56.53	50.18	18.81	58.02	47.59
AdaptSeg (Tsai et al. n.d.)	63.11	61.96	51.49	33.84	55.23	53.13
AdvEnt (Vu et al. n.d.)	68.50	**77.00**	**56.08**	29.68	63.81	59.01
Seg-Uncert (Zheng and Yang 2021)	66.39	65.88	52.74	24.80	**74.64**	56.89
DACS (Tranheden et al. n.d.)	55.80	56.34	46.70	46.37	62.01	53.44
DaFormer (Hoyer, Dai, and Van Gool n.d.)	57.97	53.66	32.09	39.76	61.70	49.04
UemDA	57.52	67.71	55.16	46.30	66.61	58.66
Oracle	84.76	91.68	75.00	78.64	90.09	84.04
SeConDA	**69.31**	73.76	45.80	**47.53**	66.39	**60.56**

*Note:*The best value for each metric evaluated is shown in bold.

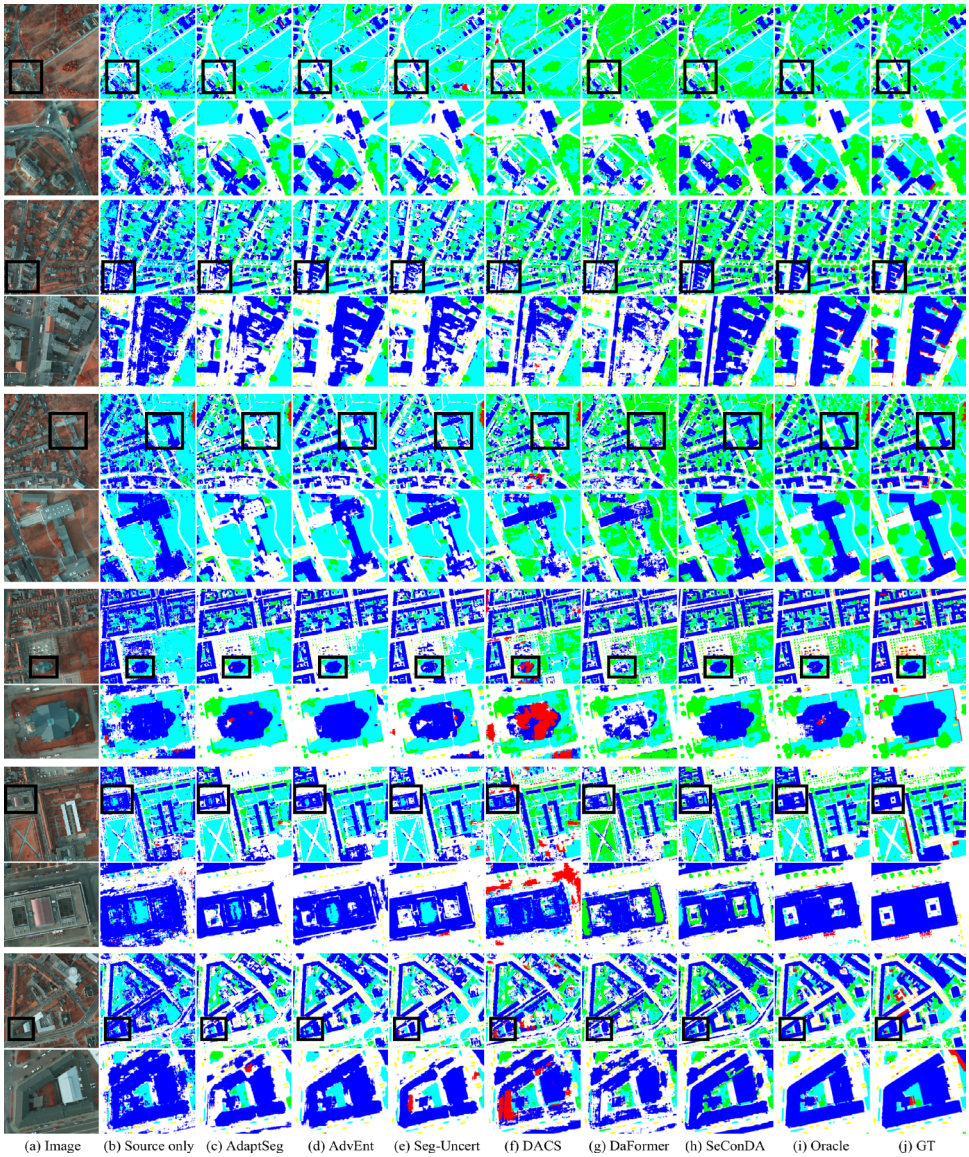


FIGURE 5 | Visualization results of transfer from Vaihingen NIRRG data to Potsdam NIRRG data.

TABLE 5 | Quantitative transfer results from Potsdam RGB to Vaihingen NIRRG.

Method	Imp. Surf.	Building	Low Veg.	Tree	Car	mIoU
Source only	43.77	52.39	6.63	52.82	25.72	36.26
AdaptSeg (Tsai et al. n.d.)	56.48	58.16	12.12	58.16	16.33	40.25
AdvEnt (Vu et al. n.d.)	64.30	73.19	24.45	57.11	14.20	52.79
Seg-Uncert (Zheng and Yang 2021)	61.46	71.37	11.08	57.73	20.57	44.44
DACS (Tranheden et al. n.d.)	77.24	79.76	44.84	29.03	58.88	57.95
DaFormer (Hoyer, Dai, and Van Gool n.d.)	68.51	74.97	38.40	48.33	52.97	56.64
UemDA	60.39	75.06	19.62	**58.67**	37.84	50.32
Oracle	84.33	90.45	69.08	79.77	72.93	79.31
SeConDA	**77.64**	**83.46**	**47.44**	49.09	**61.06**	**63.73**

*Note:*The best value for each metric evaluated is shown in bold.

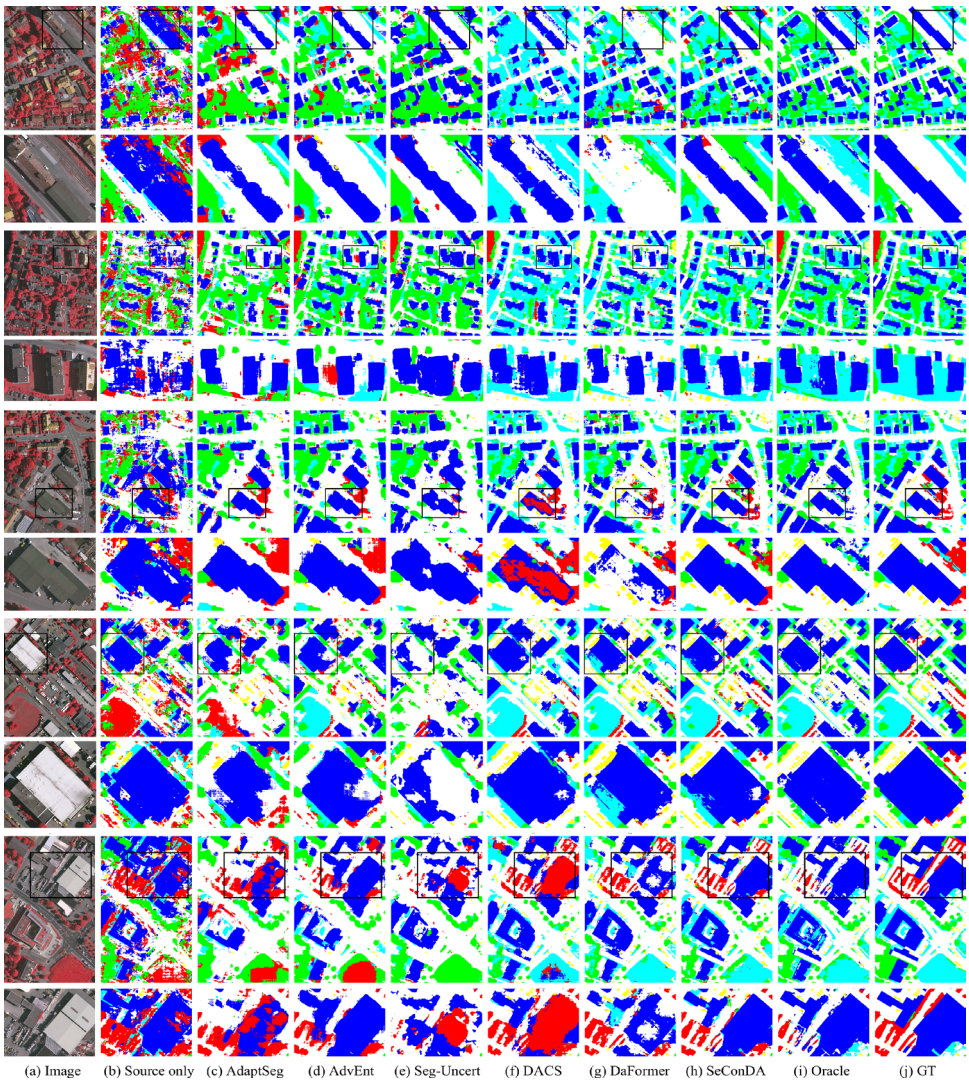


FIGURE 6 | Visualization results of transfer from Potsdam RGB data to Vaihingen NIRRG data.

domain bias problem caused solely by the difference in geographic location. Furthermore, the visualization results demonstrate that the predictions from the lower bound approach are considerably noisier compared to the upper bound results, leading to numerous false predictions. On the other hand, various DA methods exhibit improved accuracy on the TD compared to

TABLE 6 | Quantitative transfer results from Vaihingen NIRRG to Potsdam RGB.

Method	Imp. Surf.	Building	Low Veg.	Tree	Car	mIoU
Source only	42.57	39.21	27.04	12.52	53.6	34.99
AdaptSeg (Tsai et al. n.d.)	55.44	47.12	31.91	18.76	50.21	40.69
AdvEnt (Vu et al. n.d.)	56.02	50.81	26.09	29.32	47.11	41.87
Seg-Uncert (Zheng and Yang 2021)	**66.90**	**71.63**	**45.68**	2.09	71.32	51.52
DACS (Tranheden et al. n.d.)	50.72	64.98	25.89	34.62	63.69	47.98
DaFormer (Hoyer, Dai, and Van Gool n.d.)	53.38	59.55	21.84	27.36	56.32	43.69
Oracle	85.11	91.86	74.88	78.44	90.09	84.07
SeConDA	57.15	67.97	33.26	**38.19**	**73.20**	**53.95**

*Note:*The best value for each metric evaluated is shown in bold.

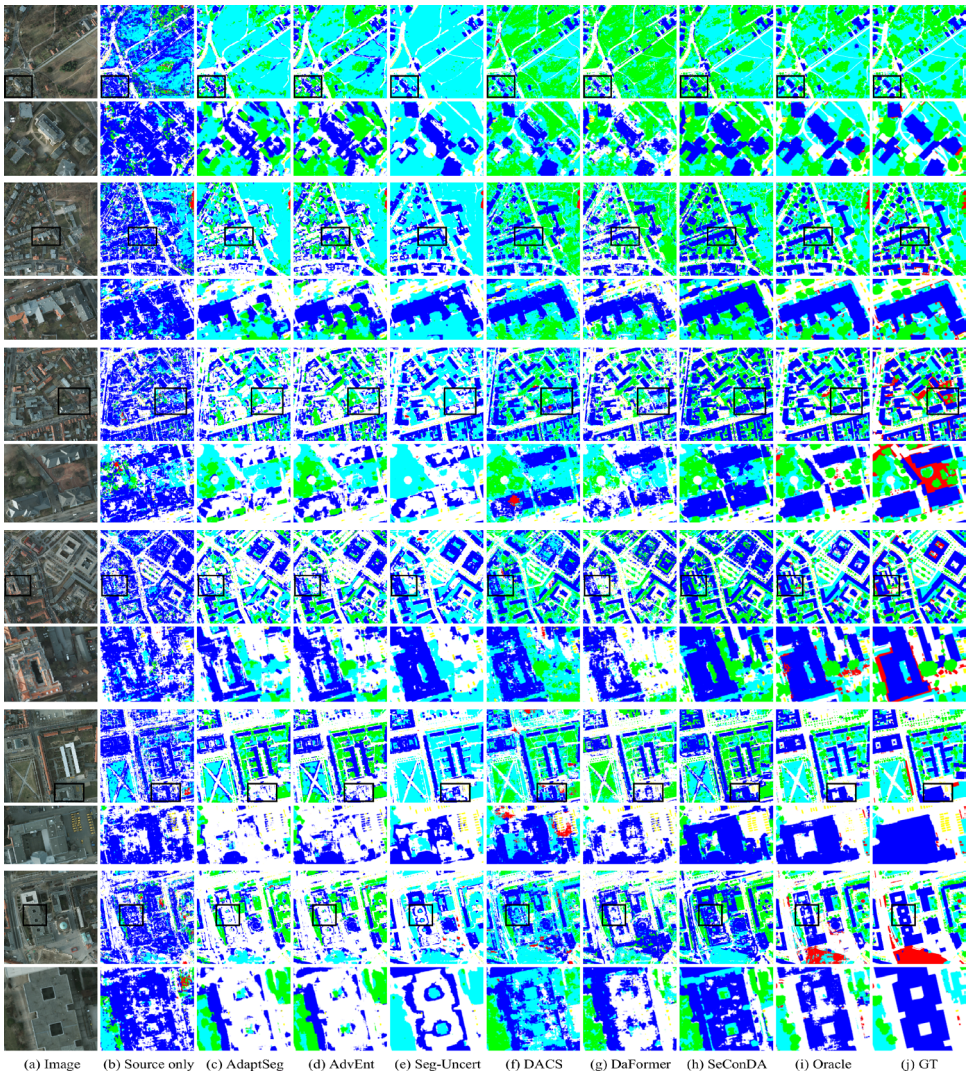


FIGURE 7 | Visualization results of transfer from Vaihingen NIRRG data to Potsdam RGB data.

the lower bound results. Specifically, AdaptSeg, AdvEnt, Seg-Uncert, DACS, DaFormer, and UemDA achieve mIoU scores of 51.14, 57.08, 62.22, 68.41, 68.27, and 61.16, respectively. The proposed approach stands out with a mIoU of 72.64, a significant 49.37% improvement over the lower bound results. Moreover, when compared to other methods, the proposed approach consistently outperforms them, surpassing the previous best-performing method by 4.23 mIoU. These findings indicate that

TABLE 7 | Quantitative transfer results from rural to urban.

Method	IoU							mIoU
	Background	Building	Road	Water	Barren	Forest	Agriculture	
Oracle	48.18	52.14	56.81	85.72	12.34	36.70	35.66	46.79
Source only	43.30	25.63	12.70	76.22	12.52	23.34	25.14	31.27
DDC	43.60	15.37	11.98	79.07	14.13	33.08	23.47	31.53
AdaptSeg	42.35	23.73	15.61	81.95	13.62	28.70	22.05	32.68
FADA	43.89	12.62	12.76	80.37	12.70	32.76	24.79	31.41
CLAN	43.41	25.42	13.75	79.25	13.71	30.44	25.80	33.11
TransNorm	38.37	5.04	3.75	80.83	14.19	33.99	17.91	27.73
PyCDA	38.04	35.86	45.51	74.87	7.71	40.39	11.39	36.25
CBST	48.37	46.10	35.79	80.05	19.18	29.69	30.05	41.32
IAST	48.57	31.51	28.73	86.01	20.29	31.77	36.50	40.48
LCGDM	47.09	49.91	48.16	84.23	18.05	32.06	35.49	44.99
PCEL	**54.19**	51.54	47.83	77.99	**37.99**	23.80	38.35	47.38
MTA	45.72	50.29	51.98	81.66	13.54	44.15	**41.77**	47.01
UemDA	46.91	48.28	49.40	83.89	15.58	41.26	34.29	45.66
SeConDA	44.70	**52.14**	**55.79**	**84.91**	17.33	**45.20**	37.88	**48.28**

*Note:*The best value for each metric evaluated is shown in bold.

the SeConDA method is superior at addressing the challenges posed by domain shift. Additionally, the visualization results consistently show that our proposed method outperforms alternative approaches.

4.4.2 | Transfer Results From Vaihingen-NIRRG to Potsdam-NIRRG

Subsequently, the effectiveness of the SeConDA method is validated by utilizing Vaihingen's NIR-R-G dataset as the SD and Potsdam's NIR-R-G dataset as the TD. The quantitative evaluation results of the SeConDA method, along with other methods and the upper and lower bounds of DA, can be found in Table 4. Additionally, the qualitative results are illustrated in Figure 5.

The accuracy of the results significantly decreased when training solely on SD data and testing on the TD, similar to the findings in Experimental Case 1. The mIoU dropped from 84.04 to 47.59 when compared to training directly on TD data. The visualization results also reveal noisy lower bound predictions with numerous misclassifications. However, DA methods like AdaptSeg, AdvEnt, Seg-Uncert, DACS, DaFormer, and UemDA show improved accuracy on the TD, achieving mIoU values of 53.13, 59.01, 56.89, 53.44, 49.04, and 58.66, respectively. The proposed method outperforms these methods with a mIoU of 60.56, representing a 27.54% improvement over the lower bound results. Additionally, the proposed method surpasses previous best-performing methods by 1.55 mIoU. Due to the limited training samples in the SD, there is less improvement in the accuracy of the transfer results compared to the experimental case 1. From the visualization results, the proposed method produces better segmentation results.

4.4.3 | Transfer Results From Potsdam-RGB to Vaihingen-NIRRG

The effectiveness of the proposed approach is subsequently confirmed by employing the Potsdam R-G-B dataset as the SD and the Vaihingen NIR-R-G dataset as the TD. The quantitative evaluation of the proposed method, along with other methods, and the upper and lower bounds of DA, are presented in Table 5. Additionally, the qualitative results can be observed in Figure 6.

In contrast to the previous two scenarios, this case involves different geographic locations and band combinations. Compared to the results of the upper bound, the accuracy of the lower bound significantly decreases, with mIoU dropping from 79.31 to 36.26. The visualization of prediction outcomes also reflects this trend, with lower bound results showing more misclassifications compared to upper bound results. Various DA methods, such as AdaptSeg, AdvEnt, Seg-Uncert, DACS, DaFormer, and UemDA, demonstrate improved performance on the TD, achieving mIoU scores of 40.25, 52.79, 44.44, 57.95, 56.64, and 50.32, respectively. The proposed method stands out with a mIoU of 63.73, marking a 75.76% enhancement over the lower bound results. Furthermore, compared to other methods, the proposed approach achieves superior performance, surpassing the previous best-performing method by 5.78 mIoU. The visualization results confirm that the proposed method delivers more accurate segmentation outcomes compared to other methods.

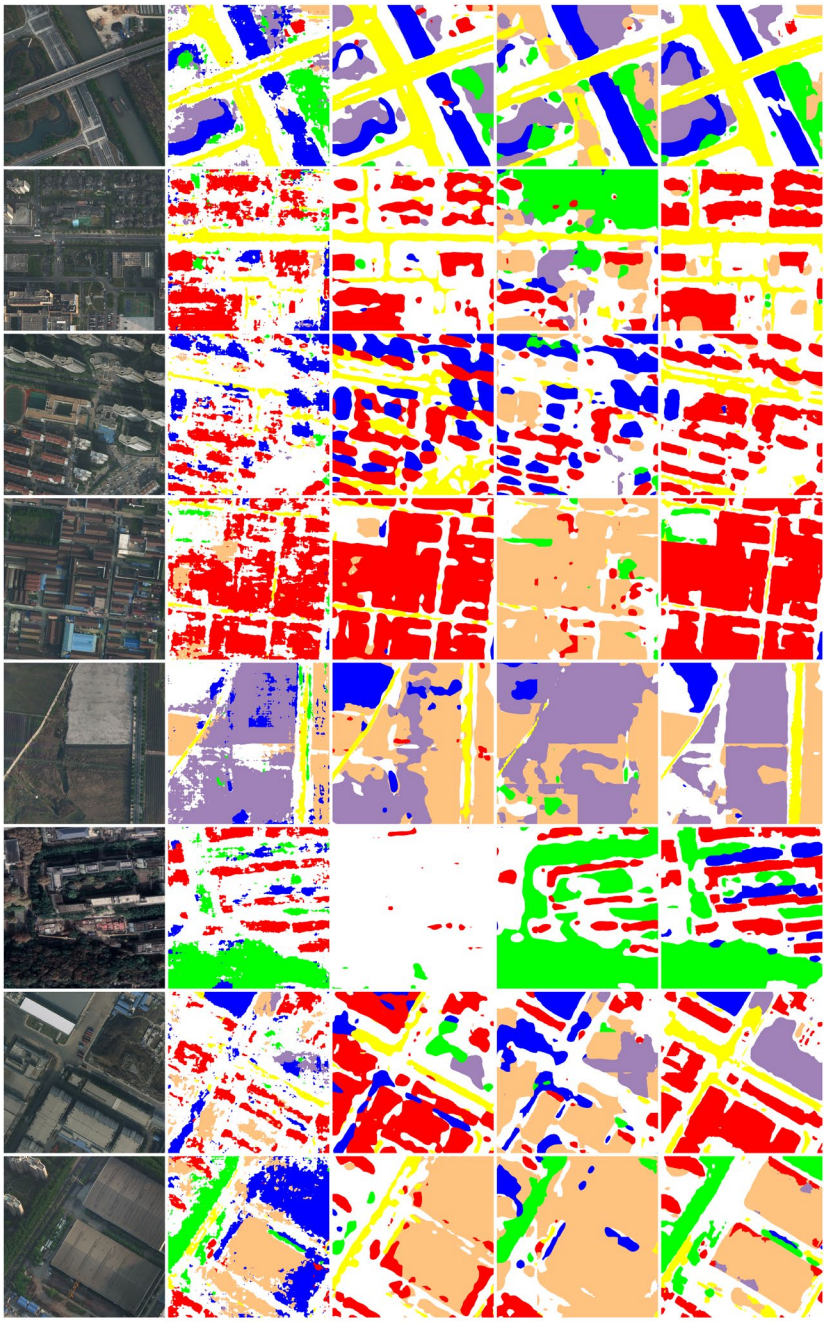


FIGURE 8 | Visualization results of transfer from LoveDA rural data to urban data.

4.4.4 | Transfer Results From Vaihingen-NIRRG to Potsdam-RGB

Finally, The Vaihingen NIR-R-G dataset serves as the SD, while the Potsdam R-G-B dataset is utilized as the TD to assess the efficacy of the proposed method. The quantitative evaluation results, along with those of other methods and the upper and lower bounds of DA, can be found in Table 6. Additionally, the qualitative results are illustrated in Figure 7.

As compared to the result of the upper bound, the accuracy of the lower bound decreases very much, with the mIoU decreasing from 84.07 to 34.99. The visualization shows that the lower bound predictions are often misclassified compared to the upper bound. Other DA methods achieve higher accuracy on the TD. The proposed methods achieve 53.95 mIoU, a 54.19% improvement over the lower bound. Compared to other methods, the SeConDA approach achieves a 2.43 mIoU improvement over the previous best-performing method. The visualization results demonstrate that the SeConDA method delivers superior segmentation results.

4.5 | Comparison With State-of-the-Art Methods in LoveDA Data

To assess the efficiency of our approach, we compare it with various established methods on the LoveDA dataset. These methods include DDC (Tzeng et al. 2014), AdaptSeg (Tsai

TABLE 8 | Quantitative transfer results from urban to rural.

Method	IoU							mIoU
	Background	Building	Road	Water	Barren	Forest	Agriculture	
Oracle	37.18	52.74	43.74	65.89	11.47	45.78	62.91	45.67
Source only	24.16	37.02	32.56	49.42	14.00	29.34	35.65	31.74
DDC	25.61	44.27	31.28	44.78	13.74	33.83	25.98	31.36
AdaptSeg	26.89	40.53	30.65	50.09	16.97	32.51	28.25	32.27
FADA	24.39	32.97	25.61	47.59	15.34	34.35	20.29	28.65
CLAN	22.93	44.78	25.99	46.81	10.54	37.21	24.45	30.39
TransNorm	19.39	36.30	22.04	36.68	14.00	40.62	3.30	24.62
PyCDA	12.36	38.11	20.45	57.16	18.32	36.71	41.90	32.14
CBST	25.06	44.02	23.79	50.48	8.33	39.16	49.65	34.36
IAST	29.97	49.48	28.29	64.49	2.13	33.36	**61.37**	38.44
LCGDM	24.88	52.50	26.15	66.79	24.16	33.04	58.49	40.86
MTA	29.40	55.72	37.86	61.28	18.65	37.69	54.97	42.22
UemDA	30.54	52.42	33.07	61.09	**24.75**	33.67	56.87	41.77
SeConDA	**34.36**	**57.25**	**45.59**	**68.15**	7.65	**41.05**	56.23	**44.33**

*Note:*The best value for each metric evaluated is shown in bold.

et al. n.d.), FADA (Wang et al. n.d.-b), CLAN (Luo et al. n.d.), TransNorm (Wang et al. 2019), PyCDA (Lian et al. n.d.), CBST (Zou et al. n.d.), IAST (Mei et al. n.d.), LCGDM (Ma et al. 2023), PCEL (Gao et al. 2023), MTA (Zeng et al. 2024), and UemDA (Liu et al. 2024). Furthermore, we also present the results of the lower bound and upper bound for domain-adapted segmentation.

4.5.1 | Transfer Results From Rural to Urban

The experimental results of this method and other methods transferred from rural to urban on the LoveDA dataset are shown in Table 7. The dataset exhibits significant differences in feature distribution due to the vast geographical variations between rural and urban images. Consequently, the model's performance trained on rural images suffers a notable decline when tested on urban images, resulting in a drop in accuracy from 46.79 to 31.27 mIoU. Notably, man-made structures (e.g., buildings and roads) experience a more substantial accuracy decrease compared to natural features (e.g., water bodies, forests, farmlands), creating a scenario where the deep learning model fails to meet performance standards in practical applications. Existing DA methods generally enhance accuracy beyond the lower bound results, with the CBST method achieving 41.32 mIoU and the TransNorm method yielding lower accuracy. The four methods LCGDM, PCEL, MTA and UemDA in RS obtained 44.99, 47.38, 47.01, and 45.66 mIoU, respectively. The proposed method surpasses the previous best method PCEL, achieving 48.28 mIoU and significantly improving accuracy across various categories, including buildings, roads, water bodies, forests, and farmlands. We also show the comparison of our method with the results of Source only, LCGDM, and UemDA in Figure 8, where it can be seen that the present method is superior in terms of correctness and completeness of predicted categories. This demonstrates the ability of our method to extract domain-invariant features effectively for generalizing TD data amidst challenges posed by substantial geographic differences.

4.5.2 | Transfer Results From Urban to Rural

The results of the experimental method proposed in this study, when applied to the LoveDA dataset for transferring from urban to rural settings, are presented in Table 8. Similar to previous findings, the model trained on urban images experiences a significant decrease in performance when tested on rural images due to the substantial geographical differences. The accuracy drops from 45.67 to 31.74 mIoU. While many existing DA methods do not show significant performance improvements, the top-performing method MTA achieves 42.22 mIoU. Negative transfer is observed in the road category across almost all methods, with some methods like FADA, CLAN, and TransNorm even showing an overall decrease in accuracy. The proposed method, however, achieves a mIoU of 44.33, approaching the upper limit of DA segmentation. It demonstrates notable accuracy improvements in building, road, water body, and forest categories by 20.23, 13.03, 18.73, and 11.71, respectively, further confirming its effectiveness. Similarly, the results of our method and other methods are shown in Figure 9, where our method has better cross-domain transfer capability.

5 | Discussion

To further verify the efficacy of the proposed method, this section employs the Potsdam NIR-R-G dataset as the SD and the Vaihingen NIR-R-G dataset as the TD. It conducts experiments with different semantic segmentation networks and encoder networks, calculates the information entropy of the prediction results, and provides visualizations for analysis.

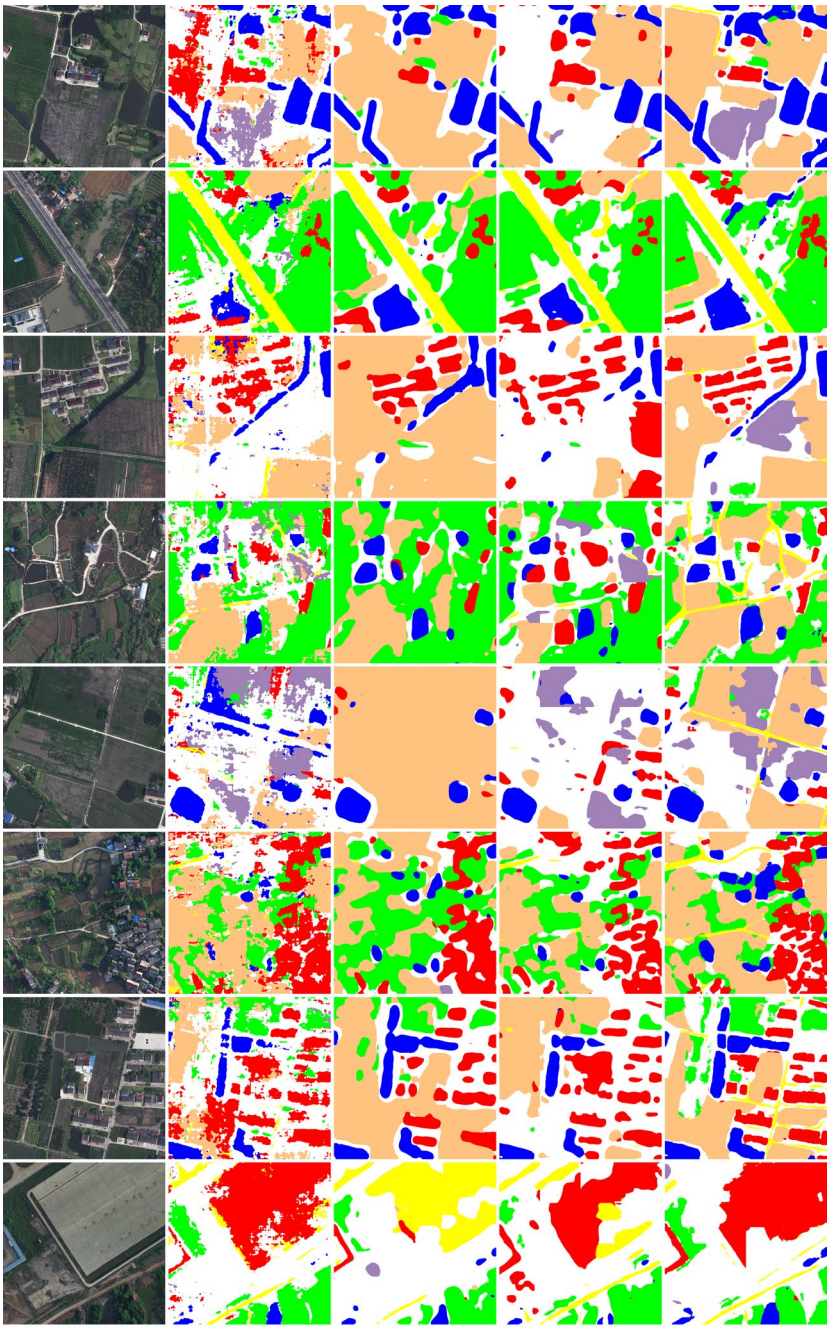


FIGURE 9 | Visualization results of transfer from LoveDA urban data to rural data.

5.1 | The Impact of Different Semantic Segmentation Networks on Accuracy

DeepLab V2 in the proposed method has been replaced with DeepLab V3 based on ResNet 50 and SegFormer based on a transformer (Xie et al. 2021). The experimental parameters remain unchanged, and the results of the proposed method, upper and lower limits, are displayed in Table 9. Both methods show relatively good accuracy when utilizing DeepLab V3 and SegFormer semantic segmentation networks. DeepLab V3 outperforms DeepLab V2 in terms of both lower and upper bounds, although its DA accuracy is not as high due to a tendency to overfit the SD data. The SegFormer semantic segmentation network demonstrates significantly higher results for both lower and upper bounds compared to DeepLab V2 and DeepLab V3, with the lower bound achieving 57.82 mIoU and the upper bound reaching 82.05 mIoU. The DA accuracy also reaches a high level of 76.20 mIoU, which is close to the upper bound result.

TABLE 9 | The impact of different semantic segmentation networks on accuracy.

Model	Source only	SeConDA	Oracle
DeepLab V2	48.63	72.64	79.31
DeepLab V3	46.16	72.03	81.06
SegFormer	57.82	76.20	82.05

5.2 | The Impact of Different Encoder Networks on Accuracy

The encoder ResNet 50 from DeepLab V2 was substituted with VGG-16 and ResNet 101. The experimental results of different encoders are presented in Table 10. When VGG-16 and ResNet 101 are utilized as the encoder, all proposed methods demonstrate relatively high accuracy. When VGG-16 is employed as the encoder, the experimental results show lower and upper bounds compared to using ResNet 50, with a DA accuracy of 68.85 mIoU. On the other hand, when ResNet 101 is used as the encoder, the experimental results indicate improved lower and upper bounds, with a DA accuracy of 68.85 mIoU.

TABLE 10 | The impact of different encoder networks on accuracy.

Model	Encoder	Source only	SeConDA	Oracle
DeepLab V2	VGG-16	46.15	68.85	78.87
	ResNet 50	48.63	72.64	79.31
	ResNet 101	50.50	73.52	79.96

5.3 | Visualization of Uncertainty

The effectiveness of the proposed method is demonstrated through the visualization of information entropy, as depicted in Figure 10. The entropy maps of the lower bound exhibit a high level of noise, resulting in low accuracy in recognizing building categories, poor edges and completeness, and frequent confusion between tree and vegetation categories. Additionally, the prediction results often display "checkerboard-like" grid blocks, leading to a significant number of incorrect predictions. The entropy map of the lower bound also reveals that incorrectly predicted pixels have a very high entropy value, indicating a high level of

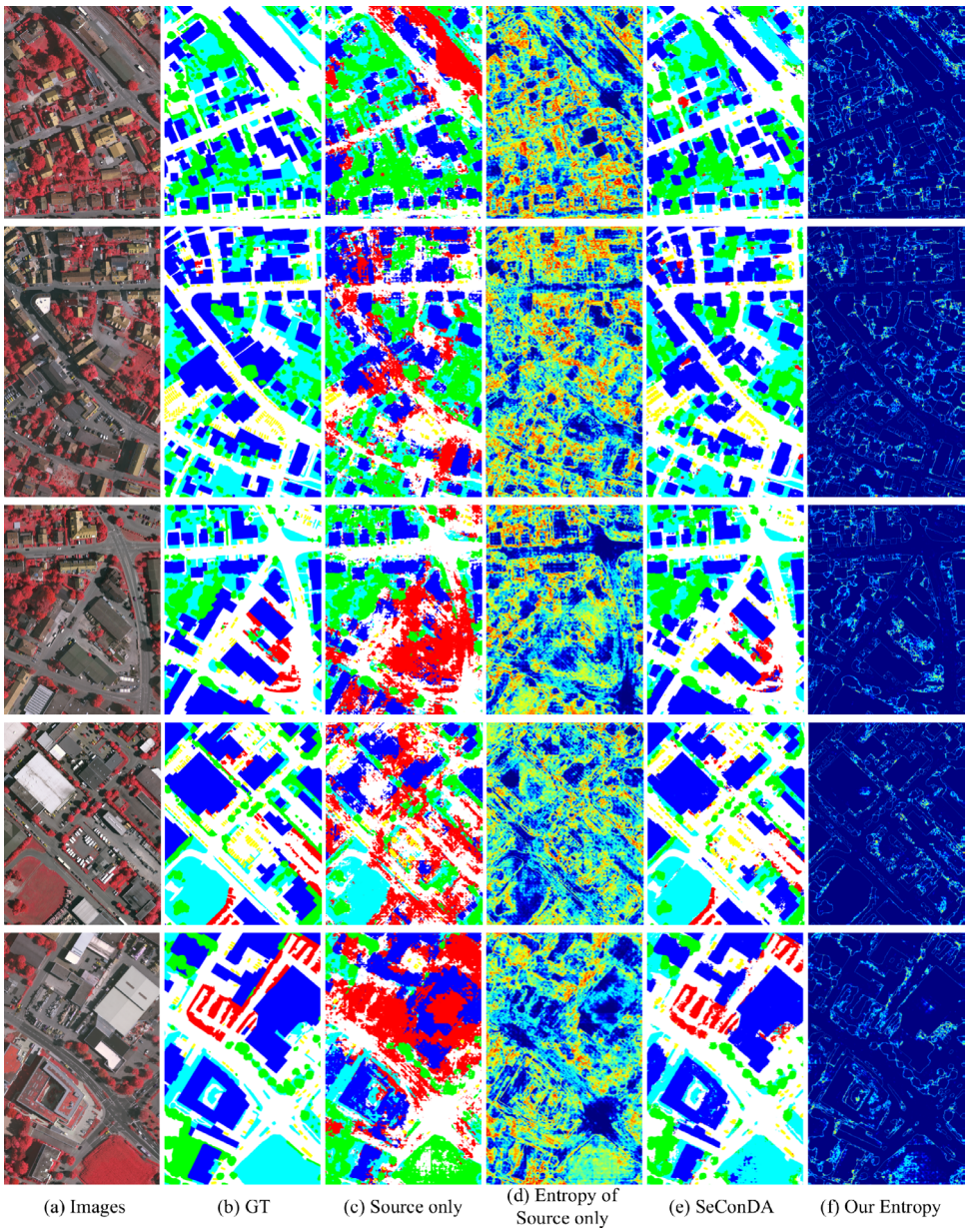
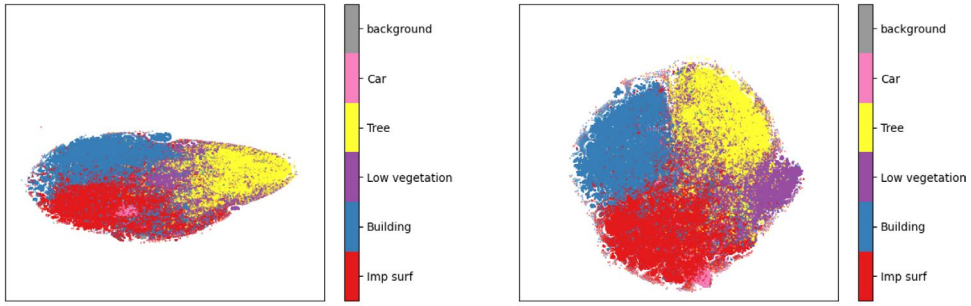


FIGURE 10 | The visualization results of entropy map.



(a) w/o adversarial training branch (b) with adversarial training branch

FIGURE 11 | t-SNE dimensionality reduction results.

uncertainty in the network's predictions for these pixels. On the other hand, the proposed method yields improved category edges and completeness in the prediction results, mitigating intercategory confusion to some extent. Moreover, the prediction results do not exhibit the presence of "checkerboard" grid blocks. In comparison to the lower bound results, the entropy maps of the proposed method tend to have lower values, with higher entropy values primarily concentrated at the category edges. This suggests that the proposed method enables the network to make predictions with greater certainty, further validating its effectiveness.

Furthermore, we employ t-SNE (Van der Maaten and Hinton 2008) to conduct a dimension reduction and visualization experiment on feature maps. In Figure 11a, when the entropy-based adversarial training branch is not used, it can be seen that the direct overlap of the classes is obvious, with the low vegetation mixed with the trees, and the car class completely mixed with the impervious surface. In Figure 11b, when the entropy-based adversarial training branch is added, the overlap between the different classes is greatly reduced although they remain close, the sample points within the classes exhibit a relatively concentrated pattern, and the low vegetation can be better separated from the trees, and the car class is completely with the surface as well. This can explain the reason why entropy-based adversarial training achieves better performance in segmentation performance, verifying that the proposed method can better extract domain-invariant features and accurately accomplish the cross-domain segmentation task.

6 | Conclusion

In this study, a cross-domain semantic segmentation method for RS images based on self-training consistency is proposed. To mine the knowledge of the TD, the whole network framework proposes a self-training consistency-based pseudo-label supervision strategy, which supervises the network training by calculating the loss of high-quality pseudo-labels generated from the TD images. Meanwhile, to alleviate the problem of domain bias between the SD and TD, a consistency regularization method based on mixed samples is proposed, in which highly perturbed mixed samples are obtained to implicitly learn the SD and TD data at the same time. In addition, entropy-based adversarial training is further proposed to make the decision boundary of the model located in low-density regions, so that the entropy of the SD and TD predictions becomes smaller, and further domain-invariant features are extracted. The state-of-the-art results are obtained through comprehensive experiments on two datasets, verifying that the SeConDA method can alleviate the domain offset problem caused by the differences in imaging locations and the differences in the way the bands are combined by minimizing the differences between the domain distributions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under project number 42030102, and the Postdoctoral Research Project for China Railway Siyuan Survey And Design Group Co., LTD. under Grant KY2023127S, and the China Postdoctoral Science Foundation under Grant 2024M753810.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All data that support the findings of this study are included in this manuscript.

References

Araslanov, N., and S. Roth. "Self-Supervised Augmentation Consistency for Adapting Semantic Segmentation." pp. 15384–15394.

Cai, Y., Y. Yang, Y. Shang, Z. Chen, Z. Shen, and J. Yin. 2022. "IterDANet: Iterative Intra-Domain Adaptation for Semantic Segmentation of Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–17.

Chen, H., H. Zhang, G. Yang, S. Li, and L. Zhang. 2022. "A Mutual Information Domain Adaptation Network for Remotely Sensed Semantic Segmentation." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–16.

Chen, J., J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng. 2022. "Unsupervised Domain Adaptation for Semantic Segmentation of High-Resolution Remote Sensing Imagery Driven by Category-Certainty Attention." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–15.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2017. "Deeplab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs."

IEEE Transactions on Pattern Analysis and Machine Intelligence 40, no. 4: 834–848.

Chen, L.-C., G. Papandreou, I. Kokkinos, et al. 2014. "Semantic Image Segmentation With Deep Convolutional Nets and Fully Connected crfs." arXiv preprint arXiv:1412.7062.

Chen, X., S. Pan, and Y. Chong. 2022. "Unsupervised Domain Adaptation for Remote Sensing Image Semantic Segmentation Using Region and Category Adaptive Domain Discriminator." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–13.

Chen, Y.-C., Y.-Y. Lin, M.-H. Yang, et al. "Crdoco: Pixel-Level Domain Transfer With Cross-Domain Consistency." pp. 1791–1800.

Cheng, Y., F. Wei, J. Bao, et al. "Dual Path Learning for Domain Adaptation Of Semantic Segmentation." pp. 9082–9091.

Csurka, G., R. Volpi, and B. Chidlovskii. 2021. "Unsupervised Domain Adaptation for Semantic Image Segmentation: A Comprehensive Survey." arXiv preprint arXiv:2112.03241.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al. 2020. "An Image is Worth 16×16 Words: Transformers for Image recognition at Scale." arXiv Preprint arXiv:2010.11929.

Fu, J., J. Liu, H. Tian, et al. "Dual Attention Network for Scene Segmentation." pp. 3146–3154.

Gao, K., A. Yu, X. You, C. Qiu, and B. Liu. 2023. "Prototype and Context-Enhanced Learning for Unsupervised Domain Adaptation Semantic Segmentation of Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–16.

Garcia-Garcia, A., S. Orts-Escolano, S. Oprea, et al. 2017. "A Review on Deep Learning Techniques Applied to Semantic Segmentation." arXiv preprint arXiv:1704.06857.

Hoffman, J., E. Tzeng, T. Park, et al. "Cycada: Cycle-Consistent Adversarial Domain Adaptation." pp. 1989–1998.

Hoyer, L., D. Dai, and L. Van Gool. "Daformer: Improving network Architectures and Training Strategies For Domain-Adaptive Semantic Segmentation." pp. 9924–9935.

Li, Y., T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li. 2021. "Learning Deep Semantic Segmentation Network Under Multiple Weakly-Supervised Constraints for Cross-Domain Remote Sensing Image Semantic Segmentation." *ISPRS Journal of Photogrammetry and Remote Sensing* 175: 20–33.

Li, Y., L. Yuan, and N. Vasconcelos. "Bidirectional Learning For Domain Adaptation of Semantic Segmentation." pp. 6936–6945.

Lian, Q., F. Lv, L. Duan, et al. "Constructing Self-Motivated Pyramid Curriculums For Cross-Domain Semantic Segmentation: A Non-Adversarial Approach." pp. 6758–6767.

Liu, W., P. Duan, Z. Xie, X. Kang, and S. Li. 2024. "Uncertain Example Mining Network for Domain Adaptive Segmentation of Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 62: 1–14.

Liu, W., F. Su, X. Jin, H. Li, and R. Qin. 2022. "Bispace Domain Adaptation Network for Remotely Sensed Semantic Segmentation." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–11.

Liu, X., C. Yoo, F. Xing, et al. 2022. "Deep Unsupervised Domain Adaptation: A Review of Recent Advances and Perspectives." *APSIPA Transactions on Signal and Information Processing* 11, no. 1.

Long, J., E. Shelhamer, and T. Darrell. "Fully Convolutional Networks For Semantic Segmentation." pp. 3431–3440.

Luo, Y., L. Zheng, T. Guan, et al. "Taking a Closer Look at Domain Shift: Category-Level Adversaries For Semantics Consistent Domain Adaptation." pp. 2507–2516.

Ma, A., C. Zheng, J. Wang, and Y. Zhong. 2023. "Domain Adaptive Land-Cover Classification via Local Consistency and Global Diversity." *IEEE Transactions on Geoscience and Remote Sensing* 61: 1–17.

Mei, K., C. Zhu, J. Zou, et al. "Instance Adaptive Self-Training For Unsupervised Domain Adaptation." pp. 415–430.

Melas-Kyriazi, L., and A. K. Manrai. "Pixmatch: Unsupervised Domain Adaptation via Pixelwise Consistency Training." pp. 12435–12445.

Murez, Z., S. Kolouri, D. Kriegman, et al. "Image to Image Translation For Domain Adaptation." pp. 4500–4509.

Pan, F., I. Shin, F. Rameau, et al. "Unsupervised Intra-Domain Adaptation For Semantic Segmentation Through Self-Supervision." pp. 3764–3773.

Peng, D., C. Zhai, Y. Zhang, and H. Guan. 2023. "High-Resolution Optical Remote Sensing Image Change Detection Based on Dense Connection and Attention Feature Fusion Network." *Photogrammetric Record* 38, no. 184: 498–519.

Peng, J., Y. Huang, W. Sun, N. Chen, Y. Ning, and Q. du. 2022. "Domain Adaptation in Remote Sensing Image Classification: A Survey." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15: 9842–9859.

Ronneberger, O., P. Fischer, and T. Brox. "U-Net: Convolutional Networks For Biomedical Image Segmentation." pp. 234–241.

Rottensteiner, F., G. Sohn, M. Gerke, et al. 2013. "The ISPRS Benchmark on Urban Object Classification and 3D Building Reconstruction." In *ISPRS Test Project on Urban Classification and 3D Building Reconstruction*, vol. I-3, 293–298. https://isprs-annals.copernicus.org/articles/I-3/293/2012/.

Tasar, O., S. Happy, Y. Tarabalka, et al. 2020. "ColorMapGAN: Unsupervised Domain Adaptation for Semantic Segmentation Using Color Mapping Generative Adversarial Networks." *IEEE Transactions on Geoscience and Remote Sensing* 58, no. 10: 7178–7193.

Tranheden, W., V. Olsson, J. Pinto, et al. "Dacs: Domain Adaptation Via Cross-Domain Mixed Sampling." pp. 1379–1389.

Tsai, Y.-H., W.-C. Hung, S. Schulter, et al. "Learning to Adapt Structured Output Space For Semantic Segmentation." pp. 7472–7481.

Tuia, D., C. Persello, and L. Bruzzone. 2016. "Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances." *IEEE Geoscience and Remote Sensing Magazine* 4, no. 2: 41–57.

Tzeng, E., J. Hoffman, N. Zhang, et al. 2014. "Deep Domain Confusion: Maximizing For Domain Invariance." arXiv preprint arXiv:1412.3474.

Van der Maaten, L., and G. Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9, no. 11: 2579–2605.

Vu, T.-H., H. Jain, M. Bucher, et al. "Advent: Adversarial Entropy Minimization For Domain Adaptation in Semantic Segmentation." pp. 2517–2526.

Wang, H., T. Shen, W. Zhang, et al. "Classes Matter: A Fine-Grained Adversarial Approach to Cross-Domain Semantic Segmentation." pp. 642–659.

Wang, J., Z. Zheng, A. Ma, et al. 2021. "LoveDA: A remote Sensing Land-Cover Dataset For Domain Adaptive Semantic Segmentation." arXiv Preprint arXiv:2110.08733.

Wang, X., Y. Jin, M. Long, et al. 2019. "Transferable Normalization: Towards Improving Transferability of Deep Neural Networks." *Advances in Neural Information Processing Systems* 32: 7834–7844.

Wang, Y., J. Peng, and Z. Zhang. "Uncertainty-Aware Pseudo Label Refinery For Domain Adaptive Semantic Segmentation." pp. 9092–9101.

Wu, Z., X. Han, Y.-L. Lin, et al. "Dcan: Dual Channel-Wise Alignment Networks For Unsupervised Scene Adaptation." pp. 518–534.

Xie, E., W. Wang, Z. Yu, et al. 2021. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." *Advances in Neural Information Processing Systems* 34: 12077–12090.

Yan, L., B. Fan, H. Liu, et al. 2019. "Triplet Adversarial Domain Adaptation for Pixel-Level Classification of VHR Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 58, no. 5: 3558–3573.

Yun, S., D. Han, S. J. Oh, et al. "Cutmix: Regularization Strategy to Train Strong Classifiers With Localizable Features." pp. 60236032.

Zeng, W., M. Cheng, Z. Yuan, et al. 2024. "Domain Adaptive Remote Sensing Image Semantic Segmentation With Prototype Guidance." *Neurocomputing* 580: 127484.

Zhang, B., Y. Zhang, Y. Li, et al. 2023. "Semi-Supervised Deep Learning via Transformation Consistency Regularization for Remote Sensing Image Semantic Segmentation." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16: 5782–5796.

Zhao, H., J. Shi, X. Qi, et al. "Pyramid Scene Parsing Network." pp. 2881–2890.

Zheng, A., M. Wang, C. Li, J. Tang, and B. Luo. 2021. "Entropy Guided Adversarial Domain Adaptation for Aerial Image Semantic Segmentation." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–14.

Zheng, S., J. Lu, H. Zhao, et al. "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective With Transformers." pp. 6881–6890.

Zheng, Z., and Y. Yang. 2021. "Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation." *International Journal of Computer Vision* 129, no. 4: 1106–1120.

Zou, Y., Z. Yu, B. Kumar, et al. "Unsupervised Domain Adaptation For Semantic Segmentation via Class-Balanced Self-Training." pp. 289–305.