remote sensing

MDPI

Article

Active Bidirectional Self-Training Network for Cross-Domain Segmentation in Remote-Sensing Images

Zhujun Yang [1,2,3,4,†], Zhiyuan Yan [1,2,*,†], Wenhui Diao [1,2], Yihang Ma [5], Xinming Li [1,2] and Xian Sun [1,2,3,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; yangzhujun19@mails.ucas.ac.cn (Z.Y.); diaowh@aircas.ac.cn (W.D.); 13911729321@139.com (X.L.); sunxian@mail.ie.ac.cn (X.S.)
[2] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[5] School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China; mayhaircas@mail.nwpu.edu.cn
* Correspondence: yanzy@aircas.ac.cn
† These authors contributed equally to this work.

Abstract: Semantic segmentation with cross-domain adaptation in remote-sensing images (RSIs) is crucial and mitigates the expense of manually labeling target data. However, the performance of existing unsupervised domain adaptation (UDA) methods is still significantly impacted by domain bias, leading to a considerable gap compared to supervised trained models. To address this, our work focuses on semi-supervised domain adaptation, selecting a small subset of target annotations through active learning (AL) that maximize information to improve domain adaptation. Overall, we propose a novel active bidirectional self-training network (ABSNet) for cross-domain semantic segmentation in RSIs. ABSNet consists of two sub-stages: a multi-prototype active region selection (MARS) stage and a source-weighted class-balanced self-training (SCBS) stage. The MARS approach captures the diversity in labeled source data by introducing multi-prototype density estimation based on Gaussian mixture models. We then measure inter-domain similarity to select complementary and representative target samples. Through fine-tuning with the selected active samples, we propose an enhanced self-training strategy SCBS, designed for weighted training on source data, aiming to avoid the negative effects of interfering samples. We conduct extensive experiments on the LoveDA and ISPRS datasets to validate the superiority of our method over existing state-of-the-art domain-adaptive semantic segmentation methods.

Keywords: semantic segmentation; domain adaptation; active learning; self-training network; remote-sensing images

Citation: Yang, Z.; Yan, Z.; Diao, W.; Ma, Y.; Li, X; Sun, X. Active Bidirectional Self-Training Network for Cross-Domain Segmentation in Remote-Sensing Images. *Remote Sens.* **2024**, *16*, 2507. https://doi.org/10.3390/rs16132507

Academic Editor: Jon Atli Benediktsson

Received: 9 April 2024
Revised: 5 June 2024
Accepted: 19 June 2024
Published: 8 July 2024

1. Introduction

Semantic segmentation plays a crucial role in interpreting remote-sensing images (RSIs) by assigning semantic labels to individual pixels. Its applications span various fields, including urban planning, disaster monitoring, road extraction, agricultural estimation, etc. [1–4]. Recently, popular deep neural networks [5] (DNNs) have achieved remarkable progress in this task, relying mainly on training on specific annotated datasets and testing [6–11]. However, these supervised learning methods require expensive and laboriously labeled images to obtain satisfactory performance. Meanwhile, the diversity and variability of remote-sensing scenes limit the cross-scene scalability of DNN models. Specifically, the differences in geographic locations and sensors lead to mismatched color textures and spatial layouts among remote-sensing data, i.e., domain shift. Proposed as a

solution to this issue, the domain adaptation (DA) method aims to bridge the gap between the source and target domains, thereby enhancing the adaptive capability of the model.

Recently, researchers have been actively engaged in extensive efforts on unsupervised domain adaptation (UDA) for semantic segmentation, employing methods such as adversarial training [12–17] and self-training [18–24]. Adversarial training (AT) methods typically leverage a discriminative network to facilitate the model learning domain-shared knowledge, thereby aligning the source and target domains in the image-level space or feature-level space. In addition, self-training (ST) methods involve training an initial adaptation model across domains and subsequently assigning pseudo-labels for target data based on the model's prediction results. Nevertheless, the training of discriminative networks for feature distribution alignment based on AT is challenging, and ST is often susceptible to noise introduced by pseudo-labels. These shortcomings pose challenges for existing UDA methods in achieving fully supervised performance with labeled target-domain data. We attribute these limitations in UDA to the potential distortion of underlying structural information in the target-domain data when unconditionally mapping its distribution to that of the source domain.

Active learning (AL) aims to maximize the efficiency of model training with minimal annotation effort. Active domain adaptation (ADA) presents an avenue for the model to grasp the realistic distribution of the target domain. Consequently, ADA is regarded as a semi-supervised domain adaptation task, improving cross-domain segmentation performance through the assistance of active samples. Figure 1 illustrates the typical domain shifts in RSIs and the differentiation of feature visualization between UDA-based and ADA-based approaches. On the left side of Figure 1, there are significant variations in object characteristics and class distribution between urban and rural scenarios. On the upper right of Figure 1, the UDA method attempts to mitigate these domain differences by aligning features of the same category. However, the outcomes are not satisfactory, with features corresponding to the "Tree" category appearing scattered, showing that the model struggles to keep uniform feature representations on this category. In contrast, our ADA method enhances the precise modeling of the target-domain distribution with fewer target-domain annotations. This refinement results in clearer classification boundaries for the segmentation task over target data in the lower right of Figure 1.

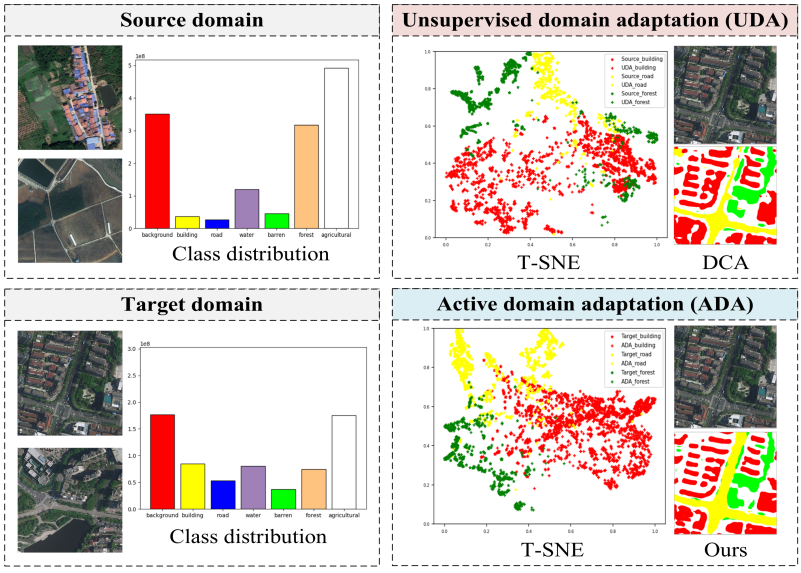

Figure 1. A domain shift exists between urban and rural scenarios and is manifested by differences in target characteristics and imbalances in the class distribution. The t-SNE [25] feature visualization for the UDA method using the DCA [22] and our ADA method is shown on the right.

To strike a balance between model performance and labeling costs, recent proposed ADA for classification methods jointly consider diversity and uncertainty criteria to select representative samples, e.g., AADA [26], CLUE [27]. Recently, Ning et al. put forward a multi-anchor method [28] for the ADA segmentation task, which chooses active images by the clustering of source image-level center features with K-means. However, this approach could result in an inefficient and wasteful use of the annotation budget, owing to the redundant labeling of objects. Additionally, the RIPU method [29] proposed by Xie et al. employs prediction region impurity and uncertainty as selection criteria, but lacks consideration of inter-domain relationships.

Existing AL strategies for cross-domain segmentation in RSIs have the following limitations: (1) The complex characteristics of remote-sensing scenes pose challenges in accurately constructing source-domain features using existing methods. In the current approach [28], it is presupposed that data from the source domain adhere to a distribution by uniform variance, and the Euclidean distance is employed to estimate inter-domain similarity. However, this approach may be inadequate for pixel-level features in remote-sensing segmentation. For example, variations in forest features due to geographic location and illumination, as well as differences in features such as building height, result in a heterogeneity in intra-class features. (2) Due to the imbalanced category distribution, certain challenging and disrupted samples within the source domain might not aid the target-domain's performance, and they could even interfere with its performance. For instance, in a rural scenario, forest and farmland areas may be more prevalent compared to an urban scenario, and the features of buildings may differ significantly between domains. Therefore, it is crucial to filter the training data for targets with substantial feature gaps.

We introduce the active bidirectional self-training network (ABSNet) for the cross-domain segmentation of RSIs. The proposed ABSNet is structured around two training phases: the multi-prototype active region selection (MARS) phase and the source-weighted class-balanced self-training (SCBS) phase. In the MARS phase, we annotate active samples based on superpixel regions, significantly reducing the cost of pixel-level fine annotation and eliminating labeling redundancy. Starting from a pre-trained UDA model, we conduct clustering of source-domain features using Gaussian mixture models (GMMs) [30], ensuring precise construction of source-domain features based on multiple prototypes and covariance matrices. As illustrated in Figure 2, which depicts the domain offsets assuming a Gaussian distribution, we leverage the probability values of the target data under the source GMMs to assess the similarity from the target domain to the source domain. Samples with less similarity are labeled as active regions. By labeling these samples that have the domain offset property and incorporating them into self-training, we enhance the model's knowledge of the target domain. In the SCBS phase, the adaptability of the source-domain samples in terms of domain adaptation is measured based on the feature center distribution of target data. Then, the assessment is used to degrade the self-trained source samples that may be excessively challenging or disturbed. Meanwhile, class-wise average entropy is employed to alleviate the class imbalance.

In summary, the proposed ABSNet conducts active region selection by assessing the similarity from the target data to the source domain and measures the distance from the source data to the target domain, thereby achieving sample denoising and class-balance training. The main contributions of this paper are as follows:

1. We propose a novel active bidirectional self-training network for cross-domain semantic segmentation in RSIs. Different from previous UDA methods, our approach aims to learn the realistic distribution of the target-domain data whenever possible. It selectively trains advantageous samples from the source-domain data.
2. The multi-prototype active region selection (MARS) module is introduced, relying on multiple prototypes and covariances to more precisely characterize the feature distribution of the source domain. This enables the selection of representative samples from the target domain. Additionally, the region labeling based on superpixels is more convenient and involves less labeling redundancy.

3. Source-weighted class-balanced self-training (SCBS) is proposed for the fine-tuning of semi-supervised domain adaptation. This approach measures the domain adaptation capability of the source-domain samples and combines it with class average entropy to denoise the source-domain samples and alleviate class imbalance.

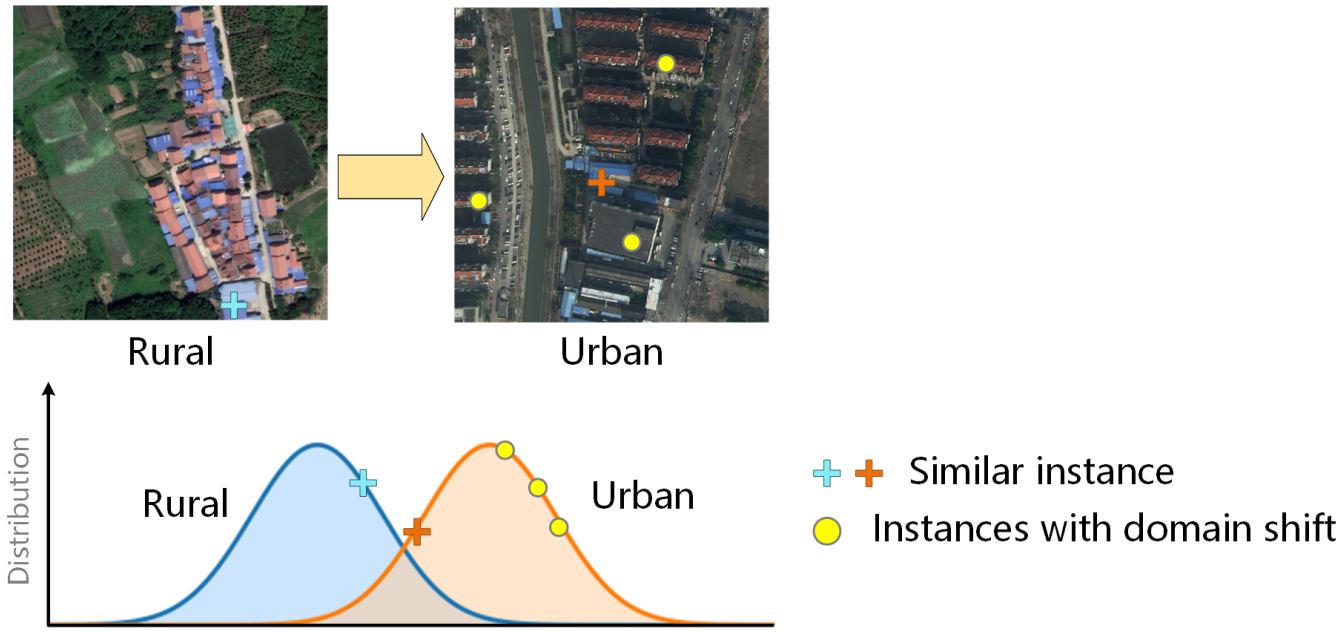

Figure 2. Domain offsets in the instances of the Urban domain from the Rural domain based on Gaussian distribution.

2. Related Works

2.1. Domain-Adaptive Semantic Segmentation

The study of semantic segmentation is dedicated to overcoming inter-domain differences to enhance the segmentation accuracy in target scenarios to avoid laborious fine-grained annotation. Previously, UDA in semantic segmentation has been extensively studied, employing two main pipelines: adversarial training-based [12–16] and self-training-based [18–24]. Within the adversarial training (AT) framework, a min–max adversarial optimization game is typically employed, where feature extractors are trained to deceive a domain discriminator, aligning the feature distributions across domains. Tsai et al. [12] developed a multi-level adversarial network that employs discriminators at different feature levels to distinguish whether the input originates from the source domain or from the target domain. Luo et al. [13] designed a category-level adversarial network to enhance local semantic consistency within the global alignment trend. Wang et al. [14] proposed a fine-grained adversarial learning framework adding category information to the discriminator, whose output contains the category attributes. In the self-training (ST) approach, reliable pseudo-labels are generated for target-domain data, and research in this area often focuses on refining and correcting these pseudo-labels. Zou et al. [18] attempted to find a suitable threshold for generating pseudo-labels for rare classes under the class-balanced principle. Mei et al. [19] proposed instance adaptive self-training, which selectively learns pseudo-labels for self-training guided by image-level entropy. Zhang et al. [20] introduced ProDA, relying on representative prototypes to reduce the noise of target pseudo-labels. It aligns prototypes using relative feature distances to achieve a more compact target feature space. Hoyer et al. [21] observed limitations in existing CNN-based UDA methods in terms of adaptation performance and proposed DAFormer. DAFormer combines a transformer encoder and a multiscale decoder, along with an improved training strategy to avoid overfitting to the source domain.

While UDA techniques are efficient, they unconditionally align the target data's distribution with that of the source domain, which may distort the underlying structural information of the target data and lead to stagnant performance. Semi-supervised domain adaptation has been explored as an alternative, incorporating a small amount of labeling to promote domain adaptation performance. Wang et al. [31] proposed a semi-supervised DA method, aiming to improve the consistency of the distribution to realize the migration of labeling information from synthetic source-domain images to real street scene images. Alonso et al. [32] designed a high-quality feature memory bank from labels and employed pixel-level contrastive learning with same-class features. These impressive semi-supervised domain adaptation works have significantly advanced the model's adaptation capability using a few labeled target samples.

2.2. Domain-Adaptive Semantic Segmentation for RSIs

To address the disparities in color texture and spatial layout among RSIs arising from geographic location and sensor variations, numerous UDA methods for semantic segmentation in RSIs have been proposed. Zheng et al. [15] introduced an entropy-guided adversarial learning algorithm that measures inter-domain differences by learning weights from predictions. This method also incorporates graph convolution to mine structural information among semantic regions. Wu et al. [22] developed a deep covariance alignment strategy for features, explicitly aligning category features based on the self-training method. Li et al. [23] successfully implemented the Transformer network in the UDA task for RSIs, aiming to address class imbalance and enhance pseudo-labels. Gao et al. [24] proposed the prototype and context-enhanced learning (PCEL) method, which assesses complex class relationships and fuses feature information from different RSIs.

In addition, Gao et al. [33] further explored the semi-supervised domain adaptation on RSIs and proposed cross-domain multi-prototype constraints to handle large inter- and intra-domain differences. However, their approach failed to take into account annotating the most informative data in the target domain and made insufficient utilization of the target-domain annotation data.

2.3. Active Domain Adaptation

The goal of AL is to improve model performance with less labeling effort by intelligently selecting and labeling the most informative samples. Common sampling strategies include uncertainty-based and diversity-based approaches. Based on uncertainty, active learning selects instances with high classification uncertainty for annotation, helping the model focus on challenging samples. The works [34,35] utilize the entropy, confidence, or margin derived from prediction outputs as criteria for selection. On the other hand, diversity-based active learning aims to choose samples that are representative and diverse for the current model, enriching the model's understanding of the overall data distribution and enhancing generalization performance. Clustering [36] or greedy selection [37,38] is used to realize the diversity in the data being processed by the query.

Unlike semi-supervised DA that trains labels for a randomly chosen subset of target samples, active domain adaptation focuses on selecting a limited set of informative samples to supplement the structural information of the target domain. Su et al. [26] introduced an ADA approach that combines prediction uncertainty with discriminator-based diversity based on the adversarial training DA method. Prabhu et al. [27] proposed clustering uncertainty-weighted embeddings (CLUE) for the classification task, which combines uncertainty to cluster the target data and select representative target samples. For the dense semantic segmentation task, Ning et al. [28] introduced the MADA method, which chooses active images using K-means clustering of image-level center features. Xie et al. [29] proposed region impurity and prediction uncertainty (RIPU) as the selection criteria, promising high information content.

However, as previously analyzed, these approaches have challenges in cross-domain semantic segmentation in RSIs. To mitigate redundancy and comprehensively capture inter-

domain relationships, our approach focuses on modeling multi-prototype source-domain feature distributions for the selection of complementary and representative active samples. Learning from these representative samples allows the model to more accurately grasp the distribution of target data, assess the adaptive capacity of source-domain samples more effectively, and enhance the benefits of the self-training-based model.

3. Methods

3.1. Overview

In the setting of ADA for semantic segmentation, we have a set of labeled source domain data, $\mathcal{D}_s = \left\{\left(\mathbf{x}_i^s, \mathbf{y}_i^s\right)\right\}_{i=1}^{N_s}$, where $\mathbf{y}_i^s$ is corresponding pixel-wise annotation for image $\mathbf{x}_i^s$ and N_s means the number of images in the source domain of $\mathcal{D}_s$, and the target-domain dataset $\mathcal{D}_t = \left\{\mathbf{x}_i^t, \tilde{\mathbf{y}}_i^t\right\}_{i=1}^{N_t}$, where $\tilde{\mathbf{y}}_i^t$ is the target active label that is initialized as ϕ and N_t means the number of images in the target domain of $\mathcal{D}_t$. We aim to learn a segmentation network parameterized by Θ that performs well on the target domain using minimal annotations. Typically, the segmentation model, G, consists of a feature extractor, E, and a classifier, F, with the relationship $G = E \circ F$. Models trained exclusively with labeled data from the source domain often produce suboptimal results when applied to the target domain. For effective knowledge transfer, the latest self-training paradigm creates pseudo-labels, $\hat{\mathbf{y}}_i^t$, for target input, $\mathbf{x}_i^t$, subsequently optimizing the cross-entropy loss. However, the performance is still limited to be competitive with the model under supervised learning. This arises because pseudo-labels tend to be imprecise, and only those pixels exceeding a specific confidence level are selected for further training. Moreover, the existence of stubborn inter-domain differences causes the reliability of the pixels, convinced by the model itself, to remain doubtful. To tackle this issue, we introduce a straightforward but effective AL method to assist domain adaptation via identifying image regions with sharp inter-domain differences. Furthermore, an enhanced self-training strategy is proposed for filtering unfavorable samples from the source domain.

The comprehensive structure of ABSNet comprises three stages: (1) Initially, train the model using the UDA technique as a preliminary step. (2) Execute multi-prototype active region selection (MARS) to determine the representative region of each image based on the inter-domain similarity. (3) Re-train the ADA segmentation model through source-weighted class-balanced self-training (SCBS) with the target-domain specific knowledge. The overview is shown in Figure 3, in which a ResNet50 is used as the feature extractor of the segmentation model. The MARS module is a forward inference procedure where the weight parameters of the segmentation network are frozen. Each image in the source-domain data, Ds, is fed into the model to obtain the source features, and those pixel-level feature vectors are clustered to obtain the source prototype. The MARS module then involves estimating the source-domain data distribution and picking a few active samples with the labeling budge, B, for the target domain guided by prototypes from the source domain. Conversely, the SCBS module is a backward training procedure with the back propagation algorithm. The source and target-domain images are simultaneously fed into the segmentation network, where the source and target-domain features are derived from the output of the feature extractor, and the source and target-domain predictions are derived from the decoder of the segmentation network (e.g., the Atrous Spatial Pyramid Pooling decoder [39]). We retrain the segmentation model by optimizing two segmentation losses (e.g., cross-entropy loss). The SCBS module screens out hard instances or sensitive samples in the source domain by considering the relationship between the source domain and target-domain centroids.

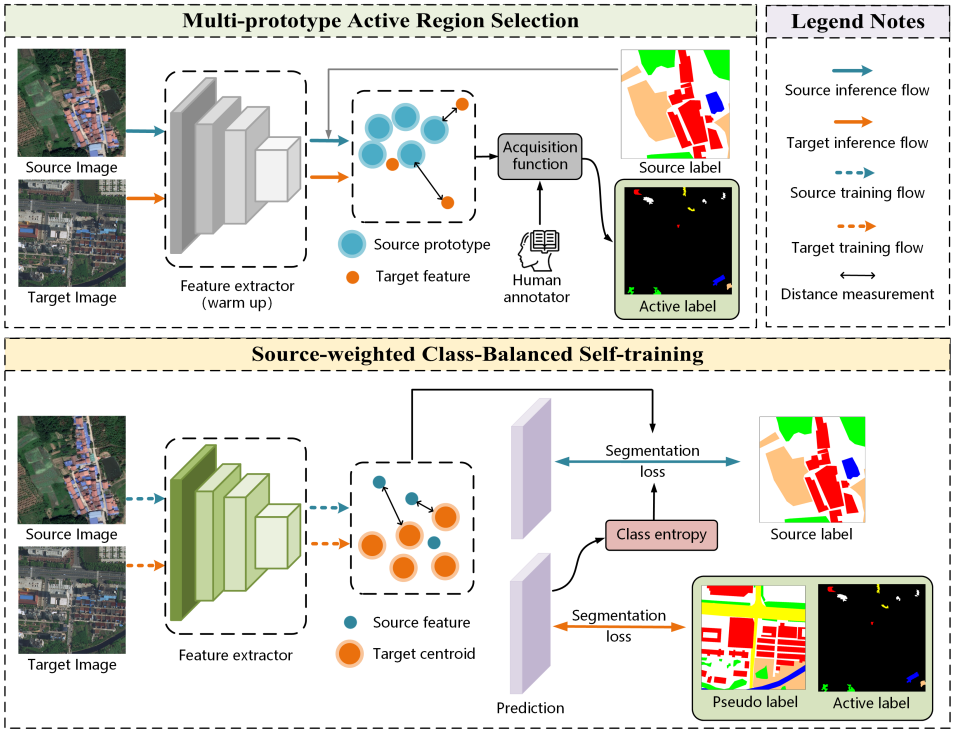

Figure 3. Architecture of the proposed ABSNet. The upper left part is the multi-prototype active region selection module, responsible for selecting and labeling the target-domain active samples that are the most informative for domain adaptation. The lower part represents the source-weighted class-balanced self-training process, in which the distance measurement from the source samples to the target distribution and the class entropy are incorporated for self-training.

3.2. Multi-Prototype Active Region Selection

Region Generation. To optimize labeling efforts for unlabeled target images, our initial step involves decomposing each image into superpixels. Superpixels serve as image primitives, automatically extracting object boundaries to some extent by grouping similar pixels in the image. Subsequently, we employ an AL strategy to select informative superpixel regions. Previous research [40] has shown that, unlike traditional polygon-based labeling, superpixel annotation offers a more advantageous labeling scheme. In line with this, each superpixel region in our experiments is assigned only one class label. To simulate annotation by an oracle, our work utilizes the ground truth of target-domain dataset.

We adopt the off-the-shelf SEEDS algorithm [41], which is a clustering-based algorithm used for superpixel generation. For the training image, $\mathbf{x}_i^t$, the set of obtained superpixels is denoted as:

$$S_i = \left\{ s_i^1, s_i^2, \ldots, s_i^k \right\}, \tag{1}$$

where an image is divided into k superpixels. For N_t images in the target domain, the set of obtained superpixels is denoted as:

$$S_{total} = \{ S_1, S_2, \ldots, S_{N_t} \}. \tag{2}$$

Therefore, the total achievable superpixels amount to $k * N_t$. A labeling budget, B, is established, which is much less than the potential labeling volume within the target-domain data, which is quantifiable as follows:

$$B = r\% * k * N_t, \tag{3}$$

where $r\%$ denotes the proportion of active labels.

Multi-prototype Domain Density Estimation. Previous ADA techniques typically rely on labeling with predictive uncertainty or measuring the similarity between their data

pairs based on Euclidean distances. However, they may be sub-optimal for RSIs that have strong intra-class variation. For effective labeling of the most informative target-domain data, we employ a soft clustering approach to estimate the density of the source-domain features, i.e., a Gaussian mixture model (GMM) [30], and evaluate the likelihood of target-domain samples being associated with a specific class. Leveraging the GMMs, we enhance the measurement of the domain gap and augment specific knowledge of the target domain to tailor the model more effectively to target-domain data, focusing on selecting samples that exhibit the largest domain gap.

We utilize a GMM to match source feature distributions due to its capability to predict complex distributions and estimate probability densities. GMMs measure the probability of a sample belonging to a cluster using probability densities. The use of a multiple-weighted Gaussian distribution enables generalization to non-Gaussian cases of training data. To acquire multiple prototypes of each class and the density distribution of the source domain, we utilize the feature extractor, E, to extract feature map $f^s = E(\mathbf{x}^s) \in \mathrm{R}^{N \times h \times w}$ from each source image, $\mathbf{x}^s$, where N is the channel number of the feature map and there are $h \times w$ pixel-level feature vectors in each feature map. We consider the pixel features that are correctly classified:

$$\Gamma_c = \left\{ f_j^s \middle| \arg\max G(\mathbf{x}^s)_j = c, \mathbf{y}_j^s = c, (\mathbf{x}^s, \mathbf{y}^s) \in \mathcal{D}_s \right\}, c = \{1, 2, \ldots, C\}, \tag{4}$$

where $G(\mathbf{x}^s)$ represents the prediction made by the initial warm-up model and $\mathbf{y}^s$ is the corresponding ground truth of $\mathbf{x}^s$. To align with the feature map's dimensions, $G(\mathbf{x}^s)$ is the intermediate output with Softmax function before the decoder of the model performing up-sampling, sized $h \times w$, whereas $\mathbf{y}^s$ is derived by down-sampling the ground truth to $h \times w$.) Thus, f_j^s denotes the pixel feature with index j, and Γ_c denotes the feature set of class c in the source domain.

We employ GMMs to characterize class-specific data distributions for each feature set within the Γ_c. Formally, the GMMs for class c, represented as the weighted sum of K Gaussian distributions, can be expressed as follows:

$$p_c(f^s) = \sum_{k=1}^{K} \pi_{c,k} \cdot \mathcal{N}(f^s | \mu_{c,k}, \Sigma_{c,k}) \tag{5}$$

$$\mathcal{N}(f^s | \mu_{c,k}, \Sigma_{c,k}) = \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_{c,k}|^{1/2}} \exp\left(-\frac{1}{2} (f^s - \mu_{c,k})^T \Sigma_{c,k}^{-1} (f^s - \mu_{c,k}) \right), \tag{6}$$

where $f^s \in \Gamma_c$ is the feature vector of class c in the source domain, μ and Σ represent the mean vector and covariance matrix of the Gaussian distribution, respectively, and π denotes the mixture weight. d is the dimension of feature vector f^s. In the GMM expression, K prototypes are obtained, where each prototype is characterized by the mean vector that describes the center of its distribution in the feature space. In addition, the covariance matrix provides information about whether the distribution expands or compresses in different directions, thus revealing the direction and strength of the shape of the distribution. We employ the Expectation–Maximization algorithm [30] to solve the GMMs equations iteratively, and random sampling is applied to restrict the number of features to no more than 300,000 per class to avoid excessive memory usage.

Active Target Region Selection Against Source Density. In the context of cross-domain active learning, we suggest that the segmentation network benefits more from target samples that exhibit greater dissimilarity to the source domain. To quantify dissimilarity, we calculate the probability value of target-domain samples under the source-domain density distribution. This computation serves as a measure of the significance of unlabeled target-domain samples for the cross-domain adaptive model. The complete process of our MARS is depicted in Figure 4.

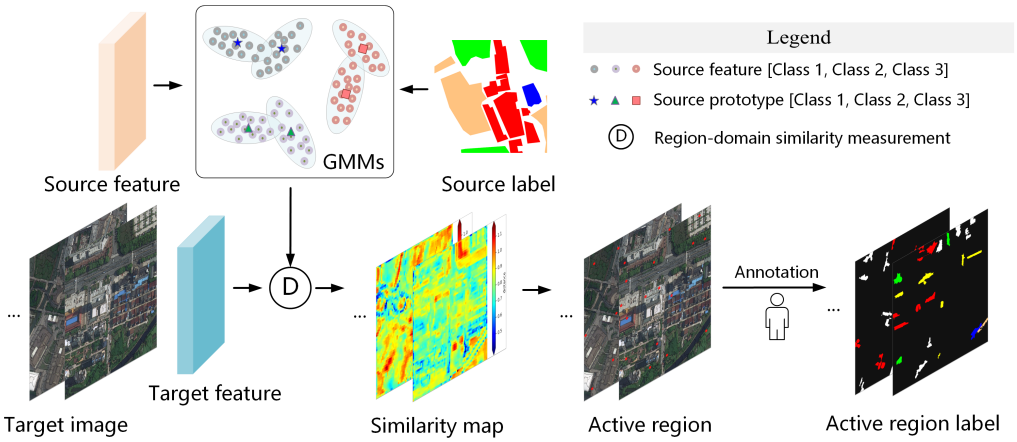

Figure 4. Illustration of multi-prototype active region selection module.

Given a target image, $\mathbf{x}^t$, we extract the feature map $f^t = E(\mathbf{x}^t) \in \mathrm{R}^{N \times h \times w}$ using the model's encoder. Then, for a target pixel feature with index j, the maximum probability value under the GMMs of the source domain with C categories is defined as the inter-domain similarity from the target pixel feature to the source domain:

$$D(f_j^t) = \arg\max_c p_c(f_j^t), c = \{1, 2, \ldots, C\}. \tag{7}$$

Next, since we consider a superpixel as the smallest labeling unit of the active sample, we obtain the inter-domain similarity of a superpixel region by computing the mean of $D(f_j^t)$ within that superpixel region. For the target image, $\mathbf{x}_i^t$, its superpixel region is represented as in Equation (1), and the inter-domain similarity of the k superpixels of this image is expressed as:

$$D^{S_i} = \left\{ D^{s_i^1}, D^{s_i^2}, \ldots, D^{s_i^k} \right\}. \tag{8}$$

Eventually, for all N_t target-domain images, there are a total of $k * N_t$ superpixels undergoing active region selection based on their inter-domain similarities, and the formula is as follows:

$$D^{total} = \left\{ D^{S_1}, D^{S_2}, \ldots, D^{S_{N_t-1}}, D^{S_{N_t}} \right\}. \tag{9}$$

With the labeling budget, B, we rank the similarity of these $k * N_t$ regions by $Sort_{\uparrow}(D^{total})$, and we select the top r% regions with the lowest similarity as active samples.

Intuitively, the similarity definition in Equation (7) assigns target-domain samples (i.e., superpixel regions) to the closest category in the source domain. Utilizing inter-domain similarity enables us to identify target-domain samples that significantly differ from the entire source domain. By labeling these samples with their true categories, we acquire target-domain-specific information that is challenging to learn solely from pseudo-labels. Note that, for these active samples, we uniformly assign a class label to image pixels within the same superpixel region, disregarding the outcome that superpixels may not be strict in segmenting edges. This approach significantly reduces the labeling cost of active samples.

3.3. Source-Weighted Class-Balanced Self-Training

In our study, the segmentation model, guided by the proposed AL strategy, is equipped to grasp specific knowledge from the target domain. However, unlike the conventional self-training learning in UDA, which learns consistent category features across both domains, our model may encounter disruption if the source data is trained without any filtering. This interference may arise from source-domain samples that significantly deviate from the target-domain characteristics. Hence, we present a method to degrade source-domain samples by utilizing the clustering centers of the target domain. This approach mitigates the impact on domain adaptation by adjusting the adaptive weighting of the source-domain

training loss. The detailed procedure is illustrated in Figure 5, where source- and target-domain images are trained simultaneously by optimizing two cross-entropy (CE) losses. We incorporate the distance measurement from the source-domain pixels to the target domain and the class-wise average entropy to obtain the weighting map, which is used in the optimization process of the source-domain CE loss.

Target Clustering Center Generation. To maximize the utilization of a priori knowledge in the target-domain data, we employ the fine-tuned model guided by active labels to compute the pixel features of unlabeled target-domain samples by $f^t = E(\mathbf{x}^t) \in \mathrm{R}^{N\times h\times w}$, and we obtain the set of target-domain features $\Omega_{\text{target}} = \left\{ f_j^t | \mathbf{x}^t \in D_t \right\}$. As the target-domain data is unlabeled, we perform unsupervised clustering algorithm K-means on this feature set to derive center anchor features that effectively represent the distribution of the target data. Specifically, we organize them into V clusters, aiming to minimize the error:

$$\sum_{v=1}^{V} \sum_{f_j^t \in \Omega_{\text{target}}} \left\| f_j^t - A_v^t \right\|_2^2, \tag{10}$$

where $\|\cdot\|_2^2$ represents the squared Euclidean distance and A_v^t is the centroid of cluster $\mathcal{C}_v$:

$$A_v^t = \frac{1}{|\mathcal{C}_v|} \sum_{f_j^t \in \mathcal{C}_v} f_j^t, \tag{11}$$

where $|\mathcal{C}_v|$ indicates the quantity of features assigned to cluster $\mathcal{C}_v$. The centroids $\left\{ A_v^t \right\}_{v=1}^{V}$ are then employed to assess the domain adaptability of the source samples during the self-training process. Notably, due to the potential bias of the initial fine-tuned model's clustering centers towards the actual target distribution, we dynamically update the centroids during self-training. After a specific number of iterations, we conduct inference on the target-domain training images, and new pixel features are used to update the centers $\left\{ A_v^t \right\}_{v=1}^{V}$.

Source Weighting Class-Balanced Factor. To filter out challenging source samples or those with insufficient contribution to domain adaptation, we assess each sample's level of contribution by measuring the feature distance to the target centroids. The initial step involves computing the distance of the source feature to the target domain as follows:

$$D(\mathbf{x}^s)_j = \min_v \left\| f_j^s - A_v^t \right\|_2, v = \{1, 2, \ldots, V\}, \tag{12}$$

where f^s is the feature map of source image, $\mathbf{x}^s$, extracted by the encoder, E, j is the index on feature map, f^s, and $D(\mathbf{x}^s)_j$ represents the nearest distance of the source feature from the target centers.

Based on the calculation of this distance, we define the source weighting factor of the source image, $\mathbf{x}^s$, as:

$$w_i^s = \exp\left(-D(\mathbf{x}^s)_i^2 \cdot \left(d_{\text{mean}}^2 \right)^{-1} \right), \tag{13}$$

where d_{mean} denotes the average distance of $D(\mathbf{x}^s)$ over all source images and is used for normalization in the equation. During training, d_{mean} is computed via the exponential moving average as follows:

$$d_{\text{mean}}^{t+1} = \alpha d_{\text{mean}}^t + (1-\alpha) D(\mathbf{x}^s)_i, \tag{14}$$

where α denotes the smoothing parameter to avoid large fluctuations.

Additionally, as illustrated in Figure 1, there is a noticeable difference in category distribution, impacting the model's training results on the target domain, especially on categories that are dominated by the source domain and have fewer samples in the target domain. To address this issue, we utilize the prediction entropy of the segmentation model

to measure the training difficulty of the target-domain data on each category, which is used to avoid repeated training on samples with a larger proportion of the source domain, while paying more attention to hard-to-learn classes in the target domain.

Hence, we employ class-wise average entropy, Ent, to achieve class-balanced self-training. As depicted in Figure 5, we acquire the Softmax output of the fine-tuned model on the target image, $\mathcal{D}_t$, to compute the entropy for each image. This entropy is then combined with the pseudo-label to calculate the average entropy under each class. Specifically, We use the pretrained segmentation model to output per-pixel classification predictions for the target-domain training images. Let the classification probability prediction result for a certain pixel be $[p_1, p_2, \ldots, p_C]$, where C is the number of categories and p_i represents the probability that the pixel belongs to category i, and then use the entropy calculation formula $E = -\sum_{i=1}^{C} p_i \log(p_i)$ to obtain the prediction entropy for each pixel. Finally, we calculate the average entropy, Ent, for each category based on the classification categories of the pseudo-labels using the pixel-level prediction entropy.

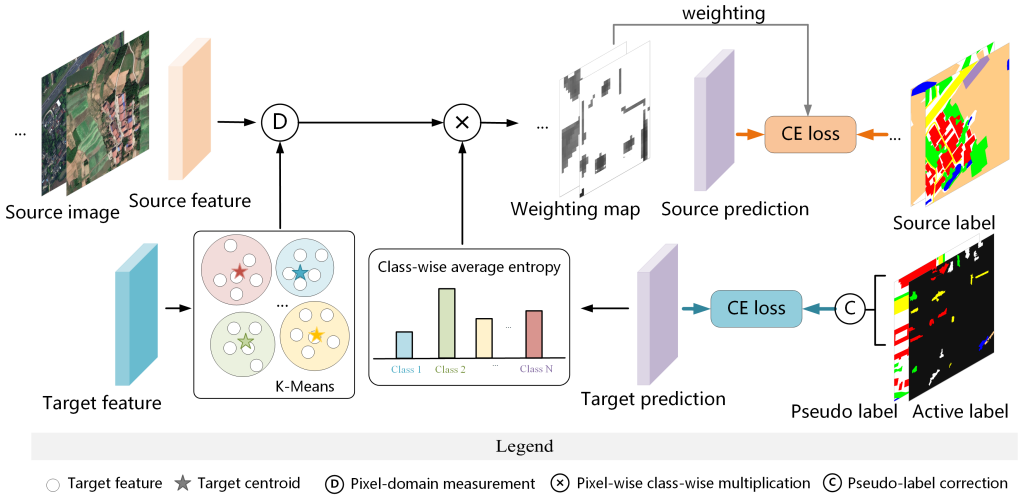

Figure 5. Illustration of the source-weighted class-balanced self-training process.

Finally, we adjust the source weighting factor by multiplying the inter-domain nearest distance and the class-wise average entropy, Ent, as follows:

$$w_i^s = \exp\left(-D(\mathbf{x}^s)_i^2 \cdot \left(d_{\text{mean}}^2\right)^{-1}\right) \times Ent_c, \mathbf{y}_i^s = c, \tag{15}$$

where Ent_c is the normalized value of class-wise average entropy corresponding to the category of $\mathbf{y}_i^s$. In this way, the class-wise average entropy limits the degree of degradation of the source-domain samples on that category, thus mitigating the effects of inter-domain category imbalance. Then, we adjust the loss function for source data by applying a class-balanced weighting factor, and the cross-entropy loss for the source image $\mathbf{x}^s$ is expressed as:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{ce}(G(\mathbf{x}^s), \mathbf{y}^s) = \frac{1}{HW} \sum_{i=1}^{H \times W} \sum_{c=1}^{C} -w_i^s \cdot y_{i,c}^s \log(p_{i,c}), \tag{16}$$

where w_i^s, y_i^s, and p_i denote the source weighting factor, the ground truth, and the probability predicted on pixel ith, respectively.

3.4. Optimization Process

Leveraging the proposed MARS and SCBS methods, the segmentation model aims to achieve improved performance in the target-domain RSIs. With the selection of active region samples, we train these target active regions by:

$$\mathcal{L}_{\text{al}} = \mathcal{L}_{ce}\left(G(\mathbf{x}^t), \tilde{\mathbf{y}}^t\right) = \frac{1}{HW} \sum_{i=1}^{H\times W} \sum_{c=1}^{C} -\tilde{y}_{i,c}^t \log(p_{i,c}), \tag{17}$$

where $\tilde{y}^t$ denotes the active label. Meanwhile, we use pseudo-labels on the target-domain unlabeled samples for self-training:

$$\mathcal{L}_{\text{pseudo}} = \mathcal{L}_{ce}\left(G(\mathbf{x}^t), \hat{\mathbf{y}}^t\right) = \frac{1}{HW} \sum_{i=1}^{H\times W} \sum_{c=1}^{C} -\hat{y}_{i,c}^t \log(p_{i,c}), \tag{18}$$

where $\hat{y}^t$ denotes the pseudo-label. The overall objective function is defined as

$$loss = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{al}} + \mathcal{L}_{\text{pseudo}}. \tag{19}$$

The complete training pipeline is outlined in Algorithm 1.

Algorithm 1 The Optimization Process of the ABSNet.

Require: Labeled source-domain data $\mathcal{D}_s = \left\{\left(\mathbf{x}_i^s, \mathbf{y}_i^s\right)\right\}_{i=1}^{N_s}$, unlabeled target-domain data $\mathcal{D}_t = \left\{\mathbf{x}_i^t\right\}_{i=1}^{N_t}$. Annotation budget B. The segmentation network G, parameterized by Θ. Number of iterations N.
Define: Target active label $\left\{\tilde{\mathbf{y}}_i^t\right\}_{i=1}^{N_t} = \phi$.

Stage 0:
Pre-train the model Θ^0 on $\mathcal{D}_s$ and $\mathcal{D}_t$ with the UDA method [12].
Stage 1:
Regionalize the target training image using SEEDS algorithm to obtain S_{total}.
Apply the GMMs density estimation on the source feature vector set $\left\{f_j^s\right\}$ to model the source-domain distribution.
Calculate similarity $D(f_j^t)$, and rank the target superpixel regions with similarity $Sort_{\uparrow}(D^{total})$.
Select the top r% regions with the lowest similarity serving as active samples, obtaining set $\left\{\left(\mathbf{x}_i^t, \tilde{\mathbf{y}}_i^t\right)\right\}_{i=1}^{N_t}$.
Stage 2:
Fine-tune the model with $\left\{\left(\mathbf{x}_i^s, \mathbf{y}_i^s\right)\right\}_{i=1}^{N_s}$ and $\left\{\left(\mathbf{x}_i^t, \tilde{\mathbf{y}}_i^t\right)\right\}_{i=1}^{N_t}$ by minimizing cross-entropy loss, and obtain Θ^1.
Initialize centroids $\left\{A_v^t\right\}_{v=1}^{V}$ with K-means clustering on the target feature vector set $\left\{f_j^t\right\}$.
Compute the class-wise average entropy set $\{Ent_c\}_{c=1}^{C}$.
for $iter = 1, \ldots, N$ **do**
    Calculate the source weighting class-balanced factor w^s.
    Calculate $\mathcal{L}_{\text{seg}}$ according to Equation (16) with $(\mathbf{x}^s, \mathbf{y}^s)$.
    Calculate $\mathcal{L}_{\text{al}}$ according to Equation (17) with $\left(\mathbf{x}^t, \tilde{\mathbf{y}}^t\right)$.
    Calculate $\mathcal{L}_{\text{pseudo}}$ according to Equation (18) with $\left(\mathbf{x}^t, \hat{\mathbf{y}}^t\right)$.
    Update Θ^2 with the overall objective function *loss* according to Equation (19).
    if $iter$%1000 == 0 **do**
        Update centroids $\left\{A_v^t\right\}_{v=1}^{V}$ with $\left\{\mathbf{x}_i^t\right\}_{i=1}^{N_t}$.
    end if
end for
Return: Final segmentation network G with parameters Θ^2.

4. Experimental Results

4.1. Dataset Description

To verify the effectiveness of the proposed approach in the cross-scene domain adaptation of semantic segmentation, we conducted extensive experiments on the LoveDA dataset [42] and the ISPRS semantic labeling benchmark. The LoveDA dataset comprises two subsets, namely Rural and Urban, while the ISPRS dataset consists of two subsets: Vaihingen (VH) and Potsdam (PD).

LoveDA presents a challenging spaceborne RSIs dataset, consisting of 1366 training images and 992 validation images from rural areas, as well as 1156 training images and 677 validation images from urban areas. Each image in the dataset has a resolution of 1024 × 1024 pixels. The dataset annotations encompass seven classes: background, building, road, water, wasteland, forest, and agriculture. The VH dataset comprises 33 images, including 16 training images and 17 test images. These images vary in size from 1996 × 1995 pixels to 3816 × 2550 pixels and consist of near-infrared, red, and green bands (IRRG). Meanwhile, the PD dataset consists of 38 images, with 24 training images and 14 test images. The images in the PD dataset have a fixed size of 6000 × 6000 pixels. Notably, the IRRG and RGB images of the PD dataset are officially provided separately, and our experiments are conducted on the PD data using RGB images, introducing more significant domain differences. Both the VH and PD datasets share the same 9 cm spatial resolution and are annotated with six categories: impervious surface, building, low vegetation, tree, car, and others.

We conduct three validation experiments for cross-domain semantic segmentation: rural-to-urban, urban-to-rural, and VH-to-PD, respectively. For the rural-to-urban task, there are 1366 labeled samples from rural and 677 unlabeled samples from urban used in ADA training. The segmentation model is evaluated using 1156 labeled samples from the urban area. In the urban-to-rural task, there are 1156 labeled samples from urban and 992 unlabeled samples from rural used for ADA training, while 1366 labeled samples from rural are used for evaluation. In the VH-to-PD task, there are 16 labeled samples from the training set of VH and 14 unlabeled samples from the test set of PD used in ADA training, and evaluation is performed using 24 samples from the training set of PD.

4.2. Implementation Details

Preprocessing. For unlabeled target-domain images, we apply the SEEDS algorithm to generate superpixel regions using the toolkit in OpenCV. The SEEDS algorithm, derived from [41], is designed to segment an image into uniform, compact blocks of superpixels. In our work, the following hyperparameter settings are employed: $priority = 3$, $num\_levels = 5$, $num\_histogram\_bins = 10$, and *double_step* is set to *True* to enhance the quality of the superpixels. The number of superpixels (*num_superpixels*) is 500 for the images in LoveDA. The original images in ISPRS are cropped to 512 × 512 pixel slices, thus their parameter *num_superpixels* is set to 125. It should be highlighted that our suggested framework is generic with any superpixel algorithm. We select 5% of the target region samples as active samples to minimize the annotation effort while maximizing performance improvement.

Training Setting. To benchmark against existing domain adaptation methods, we employ the Deeplabv2 [39] model with the ResNet50 [43] backbone for comparison experiments on the LoveDA dataset. Additionally, we utilize the DAFormer [21] model with the Mix Transformer-B5 [44] encoder for experiments on the ISPRS dataset. In the experiments, the pretrained results on ImageNet [45] are used as initialization parameters for the backbone networks. For Deeplabv2, we employ the stochastic gradient descent (SGD) optimizer during training, setting the momentum to 0.9 and the weight decay to 0.0005. The "poly" learning rate strategy [39] is applied with the initial value set to 0.01 and a power of 0.9. For DAFormer, the model is trained using the AdamW [46] optimizer with betas = (0.9, 0.999). The learning rate is set to 0.0001 for the encoder and 0.001 for the decoder, with a weight decay of 0.01. The number of iterations, N, is set to 10,000. A batch size of 8 images

is used for training. For training images of LoveDA with 1024 × 1024 pixels, we randomly crop them to obtain 512 × 512 pixels as inputs of the model. The training images undergo data augmentation methods, including random flips in horizontal and vertical directions, and random rotations.

Evaluation Metrics. We quantitatively assess the segmentation results of the model using standard metrics for accuracy evaluation, including mean intersection-over-union (mIoU) and pixel accuracy (PA). To analyze the impact on different categories, we also present the intersection-over-union for each category in comparison results.

4.3. Comparisons with the State-of-the-Art

We begin by comparing the proposed ABSNet with state-of-the-art (SOTA) DA methods of semantic segmentation from quantitative and qualitative perspectives. The methods for comparison encompass UDA methods as well as semi-supervised domain adaptation (semi-DA) methods, where the target image is partially labeled.

4.3.1. Quantitative Results

Tables 1 and 2 present the results of accuracy evaluation for different DA methods on Rural-to-Urban and Urban-to-Rural tasks, respectively. UDA methods include AdaptNet [12] and CLAN [13], based on adversarial learning, and CBST [18], IAST [19], and DCA [22], based on self-training. Semi-DA methods include the pixel-level contrastive learning method with class memory bank proposed by Alonso et al. [32], the MADA [28] and RIPU [29] methods. Following them, Alonso's method randomly selects 5% of the target-labeled images for training, MADA obtains 5% of the target-labeled images through active learning, and RIPU selects 5% of the target regions for labeling, which are regularly shaped squares centered on each pixel of the image.

For the Rural-to-Urban task in Table 1, we present the PA, the IoU per class, and the mIoU of the different methods on the test images. Firstly, the UDA approach suffers from domain biases, leading to performance limitations, such as AdaptNet and CLAN. The Semi-DA approaches can facilitate domain adaptation with the assistance of target-labeled data. Among them, Alonso's method, which randomly selects target-domain samples, pays insufficient attention to highly informative samples and may repeatedly select samples that are well migrated across domains. Moreover, the MADA method labels the whole training image, resulting in information redundancy and less efficient model learning from the labeled samples. RIPU focuses on selecting pixel regions with high prediction uncertainty, ignoring the semantic consistency of objects. In contrast, it can be observed that our ABSNet achieves the optimal PA and mIoU evaluation results. Our PA achieves 65.66%, and mIoU is 45.54%, representing an improvement of 2.77% mIoU compared to the current advanced RIPU method. In addition, for per-class IoU, our method significantly outperforms existing DA methods for the segmentation of Building and Road in the urban scene, with IoU accuracies of 52.04% and 49.89% on Building and Road, respectively.

Table 1. Comparison with different DA of semantic segmentation methods in Rural-to-Urban task.

Setting	Method	Type	PA(%)	BG *	Building	Road	Water	Barren	Forest	Agricultural	mIoU(%)
UDA	AdaptNet [12]	AT	54.81	50.90	28.43	17.73	46.86	13.10	17.83	11.21	26.58
	CLAN [13]	AT	53.61	47.71	35.33	27.04	40.93	22.93	22.52	8.88	29.34
	CBST [18]	ST	63.07	**53.09**	46.90	40.23	72.14	18.13	14.14	21.66	38.04
	IAST [19]	ST	61.50	52.54	42.98	40.40	70.86	15.67	4.85	19.03	35.19
	DCA [22]	ST	63.15	52.21	49.65	37.59	61.73	26.60	20.32	24.04	38.88
Semi-DA	Alonso's [32]	Random	60.83	45.58	49.11	46.51	59.77	**32.69**	22.95	23.81	40.06
	MADA [28]	AL	62.30	48.35	50.13	44.51	70.53	31.25	23.74	22.79	41.62
	RIPU [29]	AL	62.57	49.96	49.31	47.03	71.03	22.62	22.99	**36.47**	42.77
	ABSNet	AL	**65.66**	52.83	**52.04**	**49.89**	**72.93**	28.13	**27.36**	35.61	**45.54**

* BG: Background.

The results of PA, IoU per class, and mIoU evaluation for different approaches regarding the Urban-to-Rural task are presented in Table 2. Semi-DA methods provide a greater advantage in overall evaluation compared to UDA methods. The proposed ABSNet achieves 69.02% PA and 50.05% mIoU, which is significantly better than the state-of-the-art DA methods. For Building and Forest, our method is able to overcome objective differences between domains, obtaining a better adaptation for these two categories in the rural scene. Therefore, improvement of the classes with far inter-domain distances is vital to enhance the overall performance of cross-domain segmentation models. Overall, compared to existing SOTA methods, our method is more advantageous in selecting target-domain samples conducive to domain migration and better utilizes the annotation information to make the segmentation model adapt well to the target-domain data.

Table 2. Comparison with different DA of semantic segmentation methods in Urban-to-Rural task.

Setting	Method	Type	PA(%)	BG *	Building	Road	Water	Barren	Forest	Agricultural	mIoU(%)
UDA	AdaptNet [12]	AT	59.51	38.24	34.08	27.54	53.94	18.69	52.26	39.22	37.71
	CLAN [13]	AT	63.78	40.29	41.28	27.53	51.58	14.27	55.25	50.98	40.17
	CBST [18]	ST	62.41	29.30	39.33	36.01	55.70	16.75	53.89	53.88	40.69
	IAST [19]	ST	62.83	32.34	50.43	34.78	46.43	23.94	54.21	46.04	41.17
	DCA [22]	ST	63.13	28.89	51.58	33.42	52.99	29.54	51.42	53.38	43.03
Semi-DA	Alonso's [32]	Random	66.56	41.53	51.50	33.32	59.86	20.33	53.86	54.60	45.00
	MADA [28]	AL	67.53	**41.82**	53.58	35.07	60.48	16.20	56.15	**56.40**	45.67
	RIPU [29]	AL	66.30	41.10	52.19	32.57	57.48	**34.37**	54.86	52.99	46.51
	ABSNet	AL	**69.02**	41.77	**57.94**	**39.94**	**61.03**	34.33	**59.27**	56.07	**50.05**

* BG: Background.

We also present comparative experiments on the VH-to-PD task, as shown in Table 3. UDA methods include the Advent [16] and Zheng's [15] method based on adversarial training, and ProDA [20], Li's method [23], DAFormer [21], and PCEL [24] based on self-training. Semi-DA methods also consist of Alonso's [32], the MADA [28], and RIPU [29]. Comparison experiments on the VH-to-PD task are performed on the DAFormer model, which is a Transformer with context-aware multi-level feature fusion. This allows for the convenient comparison with existing SOTA methods and verifies the generality of our proposed pipeline for segmentation models. It can be seen that ABSNet's PA and mIoU results are 86.59% and 77.32%, respectively, which is an improvement of 2.73% mIoU compared to the RIPU method. Our method yields a significant improvement in IoU accuracy over existing methods in the Building and Low Vegetation classes. Therefore, the Semi-DA method proposed in this paper is remarkable in promoting the accuracy of the cross-domain segmentation model, especially compared to the current mainstream UDA domain adaptation methods for RSIs, e.g., Advent, Zheng's, Li's, and PCEL methods. Compared to the methods proposed for natural scene images, e.g., Alonso's, the MADA, and RIPU methods, our Semi-DA incorporating the AL algorithm effectively utilizes the statistical prior knowledge of remote-sensing data and is more advantageous in the performance of the cross-domain model.

Table 3. Comparison with different DA of semantic segmentation methods in VH-to-PD task.

Setting	Method	Type	PA(%)	Impervious Surface	Building	Low Vegetation	Tree	Car	mIoU(%)
UDA	Advent [16]	AT	60.03	49.80	54.85	40.19	26.94	46.71	43.70
	Zheng's [15]	AT	60.89	47.63	48.77	34.92	41.17	51.58	44.81
	ProDA [20]	ST	74.59	67.67	78.59	47.01	45.02	72.20	62.10
	Li's [23]	ST	77.98	72.64	82.39	54.12	48.69	60.51	63.67
	DAFormer [21]	ST	79.25	66.09	78.23	63.06	56.83	77.57	68.36
	PCEL [24]	ST	81.32	65.04	82.64	63.09	70.96	**76.95**	71.74
Semi-DA	Alonso's [32]	Random	84.23	77.93	86.43	65.38	68.48	66.95	73.03
	MADA [28]	AL	85.03	78.14	86.59	67.38	68.70	71.17	74.40
	RIPU [29]	AL	85.18	77.65	87.07	67.31	69.61	71.30	74.59
	ABSNet	AL	**86.59**	**79.47**	**89.21**	**70.46**	**71.24**	76.24	**77.32**

4.3.2. Qualitative Results

The visualization results of different methods for segmentation of test images in three tasks are shown in Figures 6–8.

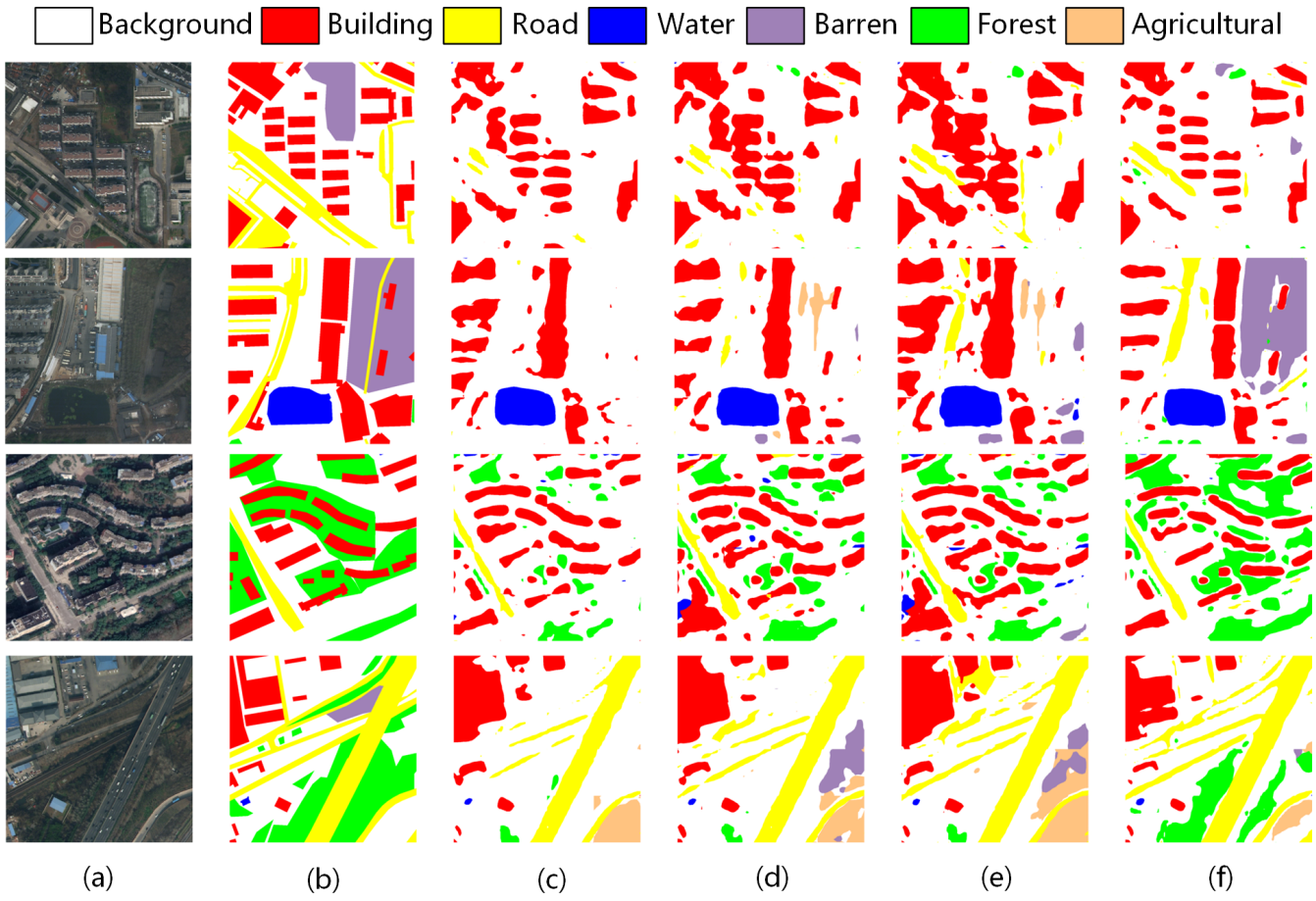

Figure 6. Segmentation visualization results with different DA of semantic segmentation methods in Rural-to-Urban task. (**a**) Image. (**b**) Ground truth. (**c**) DCA. (**d**) MADA. (**e**) RIPU. (**f**) ABSNet.

For Rural-to-Urban and Urban-to-Rural, we show segmentation visualizations for DCA, MADA, RIPU, and our ABSNet. From Figure 6, in urban scenes, buildings are arranged in a dispersed manner, making it challenging to distinguish vegetation or backgrounds located near the buildings. Additionally, roads in the target urban scene become wider and exhibit large intra-class variations compared to the source rural scene, making

them difficult to recognize as they become confused with the background. As depicted in Figure 6f, our ABSNet recognizes buildings and roads more accurately compared to the DCA, MADA, and RIPU methods. This improvement can be attributed to our method's capture of representative target samples and the learning of target-domain-specific knowledge. From Figure 7, in rural scenes, there is more agricultural land and water, and the roads become narrower. It is difficult to distinguish between agricultural land, water, and forest areas. Compared to other methods, our segmentation map greatly improves the recognition of water, and thin and narrow roads are better segmented.

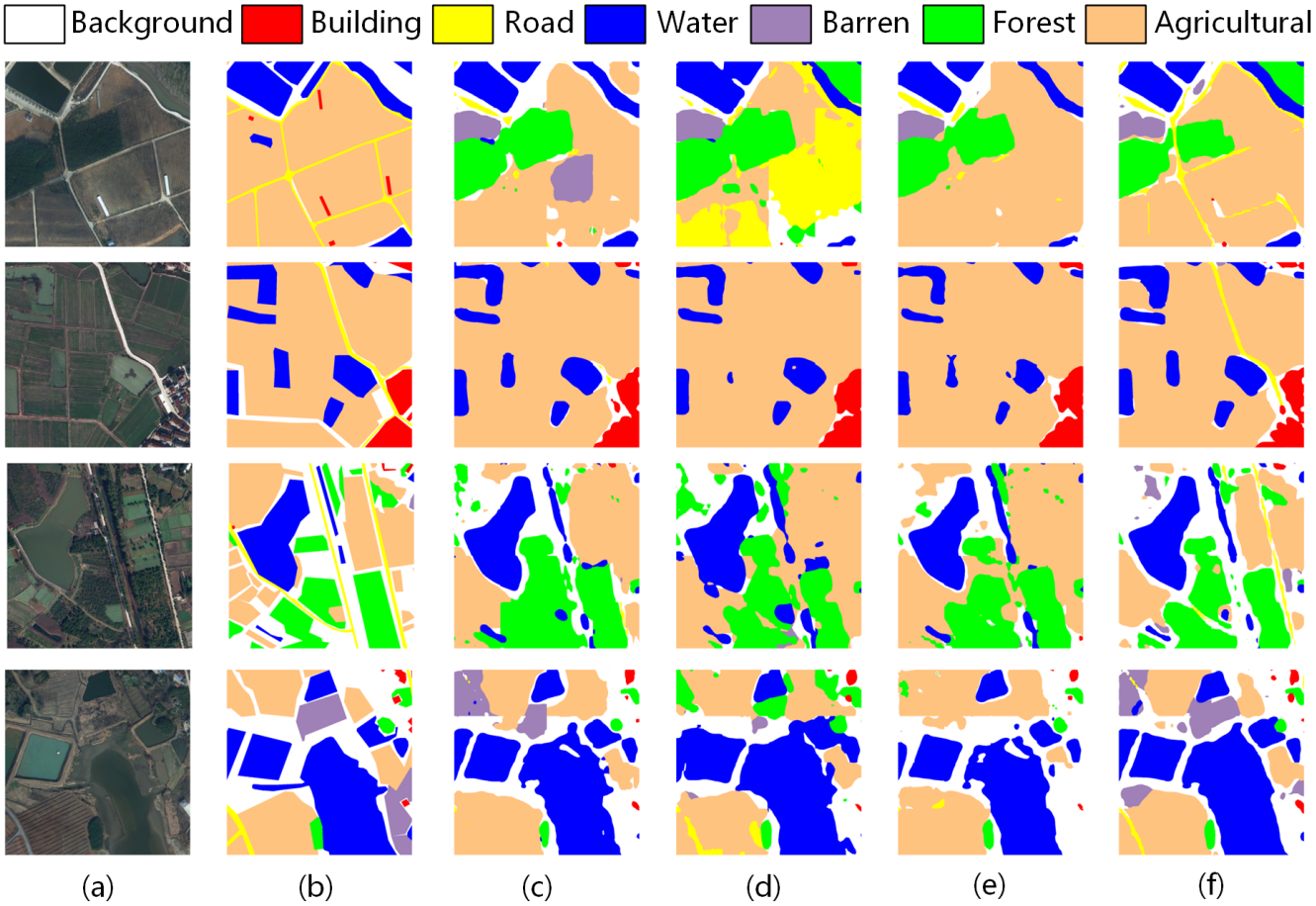

Figure 7. Segmentation visualization results with different DA of semantic segmentation methods in Urban-to-Rural task. (**a**) Image. (**b**) Ground truth. (**c**) DCA. (**d**) MADA. (**e**) RIPU. (**f**) ABSNet.

For the VH-to-PD task, we show segmentation visualizations of PD test images for Alonso's, MADA, RIPU, and our ABSNet in Figure 8. It can be observed that the buildings and trees in the PD dataset exhibit blurred boundaries, and existing methods yield coarse segmentation results at the edges of the objects. Our approach in Figure 8f enhances the segmentation of edges and more effectively mitigates inter-category confusion.

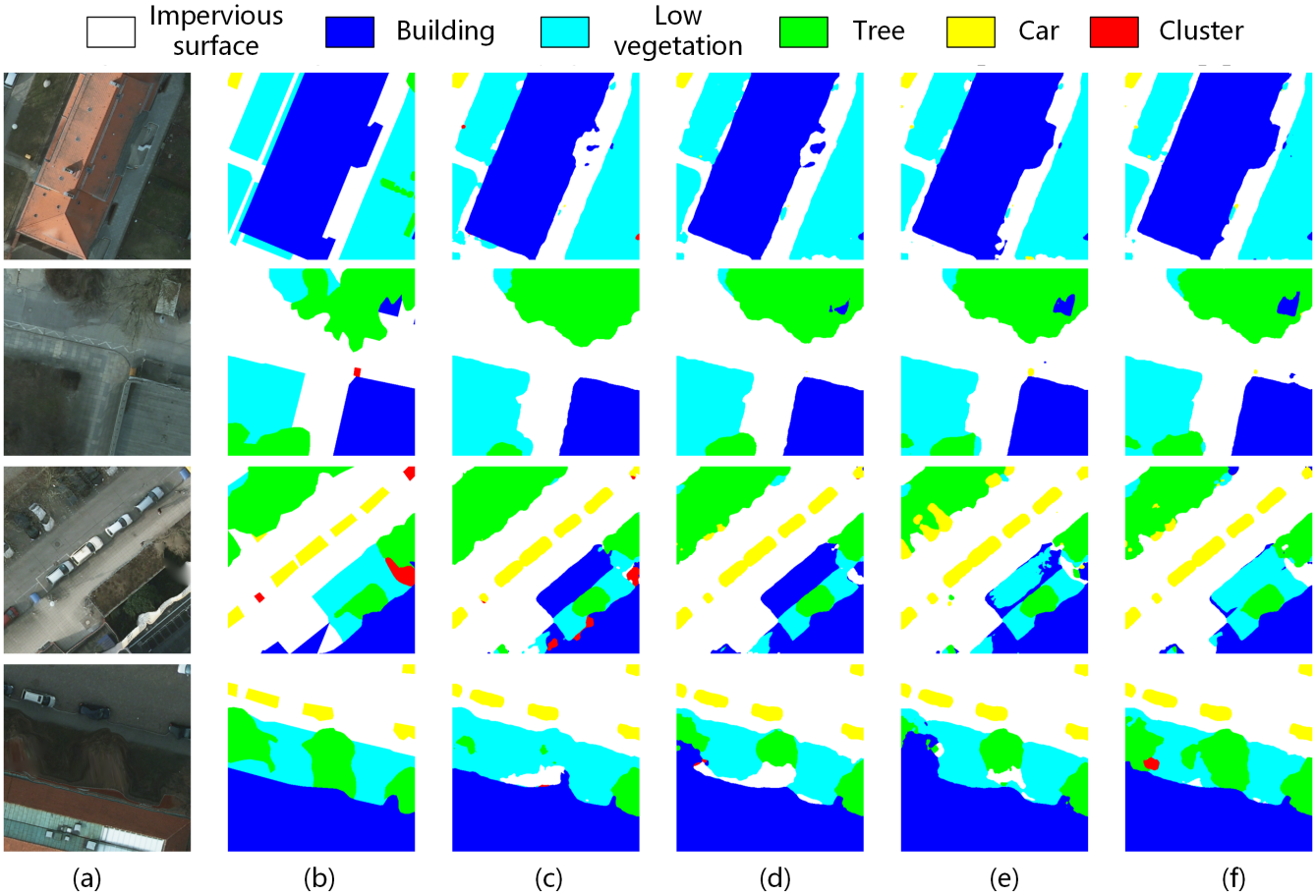

Figure 8. Segmentation visualization results with different DA of semantic segmentation methods in VH-to-PD task. (**a**) Image. (**b**) Ground truth. (**c**) Alonso's. (**d**) MADA. (**e**) RIPU. (**f**) ABSNet.

4.4. Analysis and Discussion

In this section, we analyze and evaluate in detail the effectiveness of the two proposed phases of our study. We first conduct an ablation study to recognize the detailed contribution of each module to cross-domain semantic segmentation. Next, we compare with other active learning methods. Finally, we report the optimal values of the hyper-parameters in the proposed approach, as well as the effect of the number of annotations.

4.4.1. Ablation Study

In order to validate the effectiveness of the proposed AL method, MARS, and the source-domain degradation method, SCBS, we sequentially implement these two modules to improve the baseline method. We take the pre-trained UDA model [12] learning pseudo-labels of the target domain for self-training as the baseline method. The experimental results of the Rural-to-Urban and Urban-to-Rural tasks are displayed in Table 4.

First, fine-tuning by selecting the target informative samples with the MARS module brings 3.54% and 7.84% mIoU improvement in Rural-to-Urban and Urban-to-Rural tasks, respectively. Next, we analyze the effects of the source weighting factor (Equation (13)) and the class-wise average entropy (Equation (15)), respectively. When the baseline method is trained with the SCBS module, the mIoU of the baseline method is boosted by 2.93% and 7.76% on Rural-to-Urban and Urban-to-Rural tasks, which implies that our adaptive weighting of the source samples leads to the model adapting more closely to the target domain. Together with the proposed two phases, our method achieves 45.54% and 50.05% in mIoU evaluation, which is an improvement of 6.5% and 10.01%, respectively, compared to the baseline. This validates the promotive effects of the proposed modules on each other,

and it can be appreciated that our MARS first learns representative knowledge in the target data to provide more accurate target-domain modeling for clustering in the SCBS phase. Finally, the cross-domain adaptation performance is further enhanced by source-domain degradation and class-balanced self-training.

Table 4. Ablation study results of the proposed different modules on Rural-to-Urban and Urban-to-Rural tasks.

Tasks	Baseline	MARS	SCBS[1]*	SCBS[2]*	PA(%)	mIoU(%)
Rural-to-Urban	✓				61.86	39.04
	✓	✓			63.09	42.58
	✓		✓		63.53	41.31
	✓		✓	✓	64.44	41.97
	✓	✓	✓	✓	**65.66**	**45.54**
Urban-to-Rural	✓				63.96	40.04
	✓	✓			66.92	47.88
	✓		✓		67.31	47.20
	✓		✓	✓	68.57	47.80
	✓	✓	✓	✓	**69.02**	**50.05**

* SCBS[1] : The source image is weighted according to Equation (13). * SCBS[2] : The source image is weighted according to Equation (15).

In addition, we use radar charts to visualize the per-class performance gains of our approach with respect to the baseline approach in Figure 9. Significant improvements can be seen in those hard categories, with major inter-domain differences, e.g., Road, Building, and Barren.

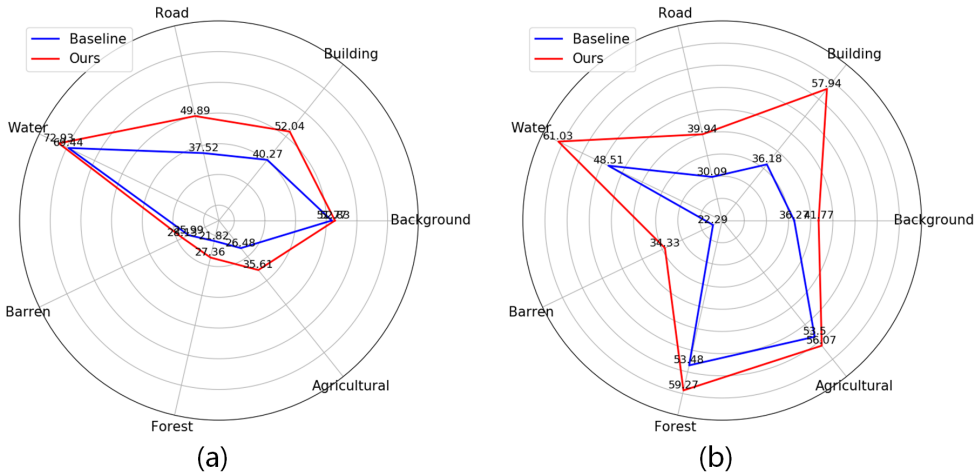

Figure 9. Radar charts of mIoU(%) on the 7 categories of the LoveDA dataset with the baseline method and our ABSNet. (**a**) Rural-to-Urban. (**b**) Urban-to-Rural.

4.4.2. Comparison of Active Sample Selection Methods

We compare the proposed multi-prototype density estimation-based active sample selection strategy with other AL methods, including RAND, ENT [34], CONF [34], and CLUE [27]. These AL methods obtain active samples and, as in our MARS experiments, they are used to fine-tune pre-trained UDA model and then test the segmentation performance of their models.

RAND. Five percent of the samples are randomly selected from all superpixel regions in the target data.

ENT [34]. The entropy of the prediction is utilized to measure the uncertainty of the data, and the method calculates the entropy of the pixel, i, of the target image, $\mathbf{x}^t$, using Equation (20). To be compatible with our region-based annotation method, we calculate

the average entropy of pixels within each superpixel region as the region entropy. Finally, the 5% regions with the largest entropy among the target samples are selected as the active samples.

$$D_{\text{ent}}\left(\mathbf{x}^t\right)_i = \frac{-1}{\log(C)} \sum_{c=1}^{C} p_{i,c}^t \log\left(p_{i,c}^t\right) \tag{20}$$

CONF [34]. The method takes the least confidence as the measure of AL. The prediction confidence is the maximum probability of the Softmax layer, illustrated in Equation (21). We also calculate the average confidence in each superpixel region and select the 5% regions with the lowest confidence in the target sample for labeling.

$$D_{\text{conf}}\left(\mathbf{x}^t\right)_i = \arg\max_c p_{i,c}^t, c = \{1, 2, \ldots, C\} \tag{21}$$

CLUE [27]. The clustering uncertainty-weighted embeddings method proposed by Prabhu et al. combines the predicted entropy values to perform weighted K-means clustering of target embeddings. The number of clustering centers is set based on the budget, B, and the samples closest to each center are selected. In the same way, we set the budget, B, to 5% of the total target samples.

Comparison results with other active learning methods on Rural-to-Urban and Urban-to-Rural tasks are reported in Table 5. ENT and CONF pick samples with uncertain predictions that might give little reward due to redundant sample learning or outlier instances. CLUE combines sample diversity with uncertainty but lacks consideration of inter-domain distance. Our proposed scheme based on multi-prototype density estimation yields the outperforming mIoU metric. This indicates that our AL strategy can alleviate sample redundancy and capture more representative and informative target samples.

Table 5. Comparison results of the proposed MARS with other active learning methods.

Tasks	Method	mIoU(%)
Rural-to-Urban	RAND	40.07
	ENT [34]	41.66
	CONF [34]	41.19
	CLUE [27]	42.01
	MARS	**42.58**
Urban-to-Rural	RAND	43.49
	ENT [34]	46.02
	CONF [34]	44.31
	CLUE [27]	46.07
	MARS	**47.88**

4.4.3. Impact of the Number of Prototypes

We next analyze the effect of the number of prototypes, K, of GMMs in MARS. Considering the diversity of remote-sensing scenarios, we propose multi-prototype source-domain clustering based on density estimation for accurate modeling of source data distribution. The validity of the multi-prototype density estimation and the effect of different K are demonstrated in Table 6. The results show that, for Rural-to-Urban and Urban-to-Rural tasks, the selected active samples give the optimal fine-tuning performance for the model when the number of prototype, K, is 4 and 6, respectively. Meanwhile, compared to the single prototype estimation ($K = 1$), the mIoU is improved by 1.74% and 2.25%, respectively. This may be related to the richness of scenarios in the data, and the urban as source data with richer scenario types requires a greater number of prototypes than the rural as source data.

Table 6. Impact of the number of prototypes in the MARS on Rural-to-Urban and Urban-to-Rural tasks.

Tasks	K = 1	K = 2	K = 4	K = 6	K = 8
Rural-to-Urban	40.84	41.85	**42.58**	42.19	41.46
Urban-to-Rural	45.63	46.96	46.46	**47.88**	47.65

4.4.4. Impact of the Number of Centroids

In the SCBS phase, we dynamically cluster unlabeled target data by K-means, and we analyze the effect of the number of cluster centroids, V, on this phase. We seek the optimal value of V based on the model that achieves the best performance in the MARS phase with different values of V selected, where V is set to 50, 100, 150, 200, 250. As shown in Table 7, our source-weighted self-training method improves on the fine-tuned model for every alternative value of the cluster centroids. The model acquires the highest effectiveness when V is set to 200 for the Rural-to-Urban task and 150 for the Urban-to-Rural task.

Table 7. Impact of the number of centroids in the SCBS on Rural-to-Urban and Urban-to-Rural tasks.

Tasks	V = 50	V = 100	V = 150	V = 200	V = 250
Rural-to-Urban	44.99	45.06	45.34	**45.54**	45.17
Urban-to-Rural	49.45	49.54	**50.05**	49.73	49.85

4.4.5. Impact of the Smoothing Parameter

We conduct experiments by adjusting the value of α in Equation (14) in order to analyze its effect on the SCBS module. Table 8 reports the effect of α in the SCBS on Rural-to-Urban and Urban-to-Rural tasks. As shown in Table 8, the SCBS module acquires the highest effectiveness when α is set to 0.999. We attribute this to the fact that the updating of d_{mean} ensures that the distance measurement is not influenced by outliers and more accurately describes the similarity between the source-domain data and the target-domain data.

Table 8. The effect of α in the SCBS on Rural-to-Urban and Urban-to-Rural tasks.

Tasks	α = 0.5	α = 0.9	α = 0.99	α = 0.999	α = 0.9995
Rural-to-Urban	42.82	44.50	45.19	**45.54**	45.39
Urban-to-Rural	47.49	49.67	49.83	**50.05**	49.87

4.4.6. Impact of the Number of Active Samples

To verify the robustness of our approach, we vary the value of the labeling budget, B, to analyze the effect of the number of active samples on cross-domain semantic segmentation. As shown in Table 9, we report the results of model training with the proposed MARS and SCBS modules under the number of active samples of 0, 5%, 10%, 50%, and 100%, respectively. As the number of labeled samples increases, the model performance gradually improves, and our SCBS can further improve the domain adaptation ability of the model based on the active sample's fine-tuning the target domain.

To more clearly visualize the superiority of our approach, in Figure 10 we present the line plot showing the evolution of the model's mIoU accuracy according to the annotation budget. As can be seen, the model performance is significantly improved after providing only 5% of the annotation. For the Rural-to-Urban task, our ABSNet achieves a 45.54% mIoU after labeling 5% of the samples, which is close to the result of labeling all the target data ($B = 100\%$). This implies that if the model can adequately exploit the training data using our method, focusing more on those informative and representative samples, whether from the source or target domain, we can enable the model to cope with variations in data from different scenarios with a powerful adaptation capability. In addition, the proposed SCBS brings 1.63% and 1.94% mIoU improvement under labeled 100% samples, respectively, which indicates that our cross-domain-trained model outperforms that of traditional supervised learning by learning the source data adaptively.

Table 9. Impact of the selection of the number of active samples for cross-domain segmentation on Rural-to-Urban and Urban-to-Rural tasks.

B	Method	Rural-to-Urban		Urban-to-Rural	
		PA(%)	mIoU(%)	PA(%)	mIoU(%)
0	-	61.86	39.04	63.96	40.04
	w/ SCBS	**64.44**	**41.97**	**68.57**	**47.80**
5%	MARS	63.09	42.58	66.92	47.88
	w/ SCBS	**65.66**	**45.54**	**69.02**	**50.05**
10%	MARS	63.48	43.00	67.56	48.35
	w/ SCBS	**67.40**	**45.65**	**69.67**	**51.75**
50%	MARS	65.61	44.79	68.83	50.16
	w/ SCBS	**67.36**	**46.94**	**70.36**	**52.90**
100%	-	67.08	45.89	70.10	51.30
	w/ SCBS	**68.05**	**47.52**	**70.94**	**53.24**

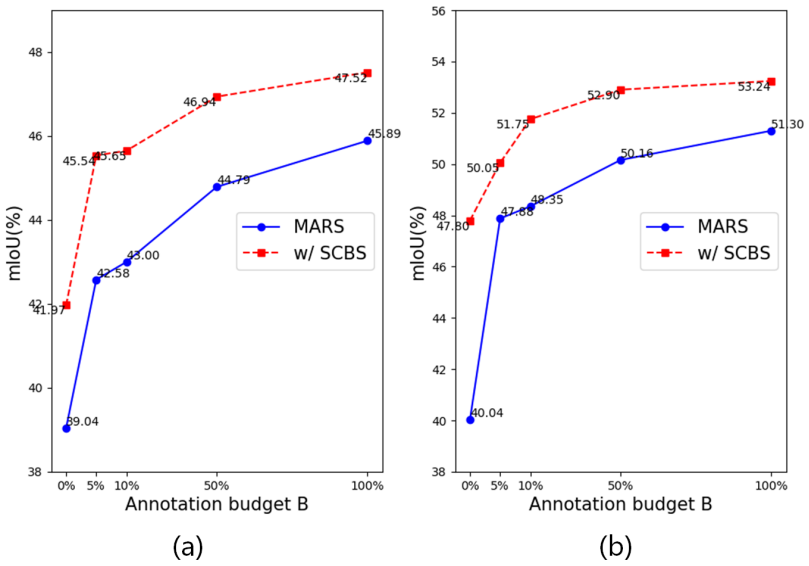

Figure 10. Performance improvement of our method for different number of active samples. (**a**) Rural-to-Urban. (**b**) Urban-to-Rural.

4.4.7. Evaluation of Inference Speed

In this section, we evaluate the inference speed of the proposed cross-domain segmentation model Deeplabv2 for the Rural-to-Urban and Urban-to-Rural tasks and DAFormer for the VH-to-PD task. We used frames per second (FPS) to evaluate the inference speed of Deeplabv2 and DAFormer. Additionally, we evaluate the number of parameters of the model (denoted as Params) and the number of float-point operations (denoted as FLOPs) for the models to compute a sliced image with 512 × 512 pixels. The detailed configuration of the experimental platform is as follows: the operating system is Ubuntu 18.04, with 256 GB of memory, an Intel(R) Xeon(R) CPU E5-2640 v4 2.40 GHz, and an NVIDIA Tesla P100 GPU with 16 GB of VRAM.

As shown in Table 10, the FPS metric for the Deeplabv2 segmentation model is 14.7, while the FPS metric for the DAFormer segmentation model is 5.3. We note that the larger number of parameters and FLOPs of DAFormer lead to slower inference speed. Nevertheless, the DAFormer model based on the Transformer structure is better in segmentation performance, as shown by comparing the segmentation results in Figures 6–8. Therefore, the segmentation model can be selected according to the needs of segmentation accuracy and inference speed in practical applications.

Table 10. The inference speeds for different tasks.

Tasks	Model	Params(M)	FLOPs(G)	FPS
Rural-to-Urban	Deeplabv2	39.1	47.9	14.7
Urban-to-Rural	Deeplabv2	39.1	47.9	14.7
VH-to-PD	DAFormer	85.2	183.3	5.3

4.5. Limitations and Future Works

In this section, we further discuss the limits of our method and potential improvement of our cross-domain segmentation network.

In this paper, we observe that the segmentation model Deeplabv2, which is commonly used in the current research of semantic segmentation [12–16,18,19,22], fails to perform well for challenging remote-sensing objects in our experimental results. For example, there are scale variations among the objects in the high-resolution RSIs in the LoveDA dataset, and the complex background samples in the RSIs. These factors pose challenges for segmentation models to effectively capture object features and achieve accurate segmentation. In addition, the clustering methods (GMMs and K-means) used in this paper may be less effective in describing the source- and target-domain data distributions when handling large-scale remote-sensing data or when the remote-sensing scenarios are more complex.

In this paper, the proposed method achieves superior cross-domain segmentation results on the datasets of optical remote-sensing images. We consider that cross-domain segmentation can be further extended on the cross-modal remote-sensing data. we plan to further explore domain adaptation issues between different modalities, such as from optical to infrared and from optical to synthetic aperture radar (SAR). These studies will help improve the generalization ability of models on data from different sensors and broaden their application scope. Furthermore, we expect to study the effect of the quality of remote-sensing images on the generalization of cross-domain segmentation models, for instance, where there are different meteorological conditions, or where the cloud and cloud shadows influence the quality of the training images.

5. Conclusions

In this paper, we propose a novel and effective active domain adaptation method for cross-domain segmentation in remote-sensing images called the active bidirectional self-training network (ABSNet). In particular, the proposed multi-prototype active region selection (MARS) and source weighted class-balanced self-training (SCBS) are two substages, which measure the similarity from the target samples to the source domain for active learning, and measure the distance from the source samples to the target domain for filtering the favorable samples, respectively. MARS models the source-domain feature distributions by using the GMM algorithm to obtain the multiple source prototypes and the covariance matrices. We use the density estimates of the target features as the inter-domain similarity and select target-labeled samples that are far away from the source domains based on the principle of least similarity. The SCBS clusters the target features based on the fine-tuned model and measures the inter-domain distance of source samples based on the target centers. This distance is combined with the class average entropy to achieve weighted training on the source data and mitigate the negative impact of interference-prone samples. Furthermore, extensive experiments on the LoveDA and ISPRS datasets are performed, which validate the performance superiority of the proposed ABSNet over state-of-the-art domain adaptation methods.

Author Contributions: Conceptualization, Z.Y. (Zhujun Yang); methodology, Z.Y. (Zhujun Yang); software, Z.Y. (Zhujun Yang); validation, Z.Y. (Zhujun Yang), Z.Y. (Zhiyuan Yan), W.D. and Y.M.; formal analysis, Z.Y. (Zhujun Yang) and Z.Y. (Zhiyuan Yan); writing—original draft preparation, Z.Y. (Zhujun Yang); writing—review and editing, Z.Y. (Zhujun Yang), Z.Y. (Zhiyuan Yan), W.D., Y.M., X.L. and X.S.; visualization, Z.Y. (Zhujun Yang); supervision Z.Y. (Zhiyuan Yan) and X.S.; project

administration, Z.Y. (Zhiyuan Yan) and X.S.; funding acquisition, Z.Y. (Zhiyuan Yan) and X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China under Grant No. 2021YFB3900504.

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

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

References

1. Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying urban land use by integrating remote sensing and social media data. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [CrossRef]
2. Marcos, D.; Volpi, M.; Kellenberger, B.; Tuia, D. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 96–107. [CrossRef]
3. Maboudi, M.; Amini, J.; Malihi, S.; Hahn, M. Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remotely sensed images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 151–163. [CrossRef]
4. Hamuda, E.; Glavin, M.; Jones, E. A survey of image processing techniques for plant extraction and segmentation in the field. *Comput. Electron. Agric.* **2016**, *125*, 184–199. [CrossRef]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
6. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS4Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5398–5413. [CrossRef]
7. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Chen, K.; Fu, K. Hybrid multiple attention network for semantic segmentation in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603018. [CrossRef]
8. Li, X.; He, H.; Li, X.; Li, D.; Cheng, G.; Shi, J.; Weng, L.; Tong, Y.; Lin, Z. PointFlow: Flowing semantics through points for aerial image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4217–4226.
9. Mou, L.; Hua, Y.; Zhu, X.X. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7557–7569. [CrossRef]
10. Niu, R.; Sun, X.; Tian, Y.; Diao, W.; Feng, Y.; Fu, K. Improving semantic segmentation in aerial imagery via graph reasoning and disentangled learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5611918. [CrossRef]
11. Yang, Z.; Yan, Z.; Sun, X.; Diao, W.; Yang, Y.; Li, X. Category correlation and adaptive knowledge distillation for compact cloud detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623318. [CrossRef]
12. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7472–7481.
13. Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; Yang, Y. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2507–2516.
14. Wang, H.; Shen, T.; Zhang, W.; Duan, L.Y.; Mei, T. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 642–659.
15. Zheng, A.; Wang, M.; Li, C.; Tang, J.; Luo, B. Entropy guided adversarial domain adaptation for aerial image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5405614. [CrossRef]
16. Vu, T.H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2517–2526.
17. Liu, W.; Zhang, W.; Sun, X.; Guo, Z. Unsupervised Cross-Scene Aerial Image Segmentation via Spectral Space Transferring and Pseudo-Label Revising. *Remote Sens.* **2023**, *15*, 1207. [CrossRef]
18. Zou, Y.; Yu, Z.; Kumar, B.; Wang, J. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 289–305.
19. Mei, K.; Zhu, C.; Zou, J.; Zhang, S. Instance adaptive self-training for unsupervised domain adaptation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVI 16; Springer: Berlin/Heidelberg, Germany, 2020, pp. 415–430.
20. Zhang, P.; Zhang, B.; Zhang, T.; Chen, D.; Wang, Y.; Wen, F. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12414–12424.

21. Hoyer, L.; Dai, D.; Van Gool, L. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9924–9935.
22. Wu, L.; Lu, M.; Fang, L. Deep covariance alignment for domain adaptive remote sensing image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620811. [CrossRef]
23. Li, W.; Gao, H.; Su, Y.; Momanyi, B.M. Unsupervised domain adaptation for remote sensing semantic segmentation with transformer. *Remote Sens.* **2022**, *14*, 4942. [CrossRef]
24. Gao, K.; Yu, A.; You, X.; Qiu, C.; Liu, B. Prototype and Context Enhanced Learning for Unsupervised Domain Adaptation Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5608316. [CrossRef]
25. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
26. Su, J.C.; Tsai, Y.H.; Sohn, K.; Liu, B.; Maji, S.; Chandraker, M. Active adversarial domain adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 739–748.
27. Prabhu, V.; Chandrasekaran, A.; Saenko, K.; Hoffman, J. Active domain adaptation via clustering uncertainty-weighted embeddings. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 8505–8514.
28. Ning, M.; Lu, D.; Wei, D.; Bian, C.; Yuan, C.; Yu, S.; Ma, K.; Zheng, Y. Multi-anchor active domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9112–9122.
29. Xie, B.; Yuan, L.; Li, S.; Liu, C.H.; Cheng, X. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8068–8078.
30. Reynolds, D.A.; Rose, R.C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 72–83. [CrossRef]
31. Wang, Z.; Wei, Y.; Feris, R.; Xiong, J.; Hwu, W.M.; Huang, T.S.; Shi, H. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 936–937.
32. Alonso, I.; Sabater, A.; Ferstl, D.; Montesano, L.; Murillo, A.C. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 8219–8228.
33. Gao, K.; Yu, A.; You, X.; Qiu, C.; Liu, B.; Zhang, F. Cross-Domain Multi-Prototypes with Contradictory Structure Learning for Semi-Supervised Domain Adaptation Segmentation of Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3398. [CrossRef]
34. Wang, D.; Shang, Y. A new active labeling method for deep learning. In Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, China, 6–11 July 2014; pp. 112–119.
35. Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; Lin, L. Cost-effective active learning for deep image classification. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 2591–2600. [CrossRef]
36. Ash, J.T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv* **2019**, arXiv:1906.03671.
37. Kirsch, A.; Van Amersfoort, J.; Gal, Y. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Adv. Neural Inf. Process. Syst.* **2019**, pp. 7026–7037.
38. Wu, T.H.; Liu, Y.C.; Huang, Y.K.; Lee, H.Y.; Su, H.T.; Huang, P.C.; Hsu, W.H. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15510–15519.
39. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
40. Cai, L.; Xu, X.; Liew, J.H.; Foo, C.S. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10988–10997.
41. Van den Bergh, M.; Boix, X.; Roig, G.; De Capitani, B.; Van Gool, L. Seeds: Superpixels extracted via energy-driven sampling. In Proceedings of the Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part VII 12; Springer: Berlin/Heidelberg, Germany, 2012; pp. 13–26.
42. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
44. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.

45. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
46. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.