DDF: A Novel Dual-Domain Image Fusion Strategy for Remote Sensing Image Semantic Segmentation with Unsupervised Domain Adaptation

Lingyan Ran, Lushuang Wang, Tao Zhuo, Yinghui Xing, *Member, IEEE*, Houjun He, Yanning Zhang, *Senior Member, IEEE*

***Abstract*—Semantic segmentation of remote sensing images is a challenging and hot issue due to the large amount of unlabeled data and domain variation. Unsupervised domain adaptation (UDA) has proven to be advantageous in leveraging unlabeled information from the target domain. However, traditional approaches of independently fine-tuning UDA models in the source and target domains have a limited effect on the result. In this paper, we propose a hybrid training strategy that boosts self-training methods with domain fusion images. First, we introduce a novel dual-domain fusion (DDF) strategy to effectively utilize the original image, the style-transferred image, and the intermediate-domain information. Second, to further refine the precision of pseudo-labels, we present a region-specific re-weighting strategy which assigns different weights to pseudo-label regions based on their spatial context. Finally, we conduct a series of extensive benchmark experiments and ablation studies on the ISPRS Vaihingen and Potsdam datasets. These results show the efficiency of our approach and establish a practical basis for implementing semantic segmentation in remote sensors.**

***Index Terms*—Semantic segmentation, domain adaptation, feature fusion**

I. Introduction

RECENTLY, the application of semantic segmentation to remote sensing (RS) images has become prevalent for tasks such as land use analysis, road network extraction, and building inspection [1]–[4], which requires accurate pixel-level labeling. Advances in deep learning have notably improved the effectiveness of these applications. However, challenges persist, mainly due to two factors: insufficient labels and a significant domain gap. The efficacy of deep learning techniques is highly dependent on the availability of extensive labeling, with performance significantly declining when this is not met. Meanwhile, the process of annotating RS images requires considerable time and effort. For example, annotating a single Cityscapes image consumes approximately 1.5 hours [5], and RS images typically present more complexity in terms of both content and dimensions. What's more, differences in geographical areas, variable timing, and the use of diverse sensors for RS imaging, as depicted in Fig. 1, further contribute to the complexity of this task.

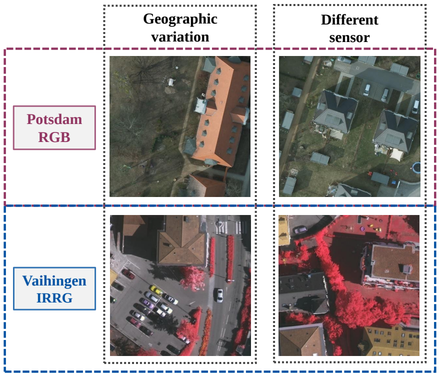

Fig. 1. Semantic segmentation of RS images faces challenges due to geographic variations and the use of different sensors. These fundamental differences between datasets pose a challenge for models trained on one dataset to generalize effectively to others.

To tackle these issues, researchers have proposed the use of unsupervised domain adaptation (UDA) [6]–[9] in the context of semantic segmentation. The primary objective of this strategy is to train a model using labeled data from a source domain and achieve accurate predictions on unlabeled data from a target domain. The focus of UDA is primarily on mitigating the disparities between the source and target domains to enhance the model's performance on the target domain. Consequently, various techniques have been developed to reduce the domain discrepancy. Key techniques include generative training (GT) methods (CycleGAN [10], ColorMapGAN [11]), adversarial methods (AdaptSeg [12], ADVENT [13], CCGDA [14]), and self-training methods (CBST [15], DAFormer [16]).

The fundamental aim of generative training [10], [11] is to modify the visual characteristics of an image to minimize color and texture disparities between the images of the source and target domains. This approach addresses the domain-shift issue at the input level. However, the experimental results are heavily dependent on the quality of the produced images. Distortions in these images might result in detrimental migration, degrading segmentation outcomes.

This work was supported in part by the Natural Science Basic Research Program of Shaanxi under Grant 2024JC-YBQN-0719, the Key R & D Program of Shaanxi under Grant 2023-YBGY-012, the Natural Science Foundation of NingBo under Grant 2023J262, the Shandong Excellent Young Scientists Fund Program (Overseas) 2023HWYQ-114. *(Corresponding author: Tao Zhuo.)*

Lingyan Ran, Lushuang Wang, Yinghui Xing, and Yanning Zhang are with the Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, and the National Engineering Laboratory for Integrated Aerospace-GroundOcean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. Tao Zhuo is in the College of Information Engineering, Northwest A&F University, Yangling, 712100, China. Houjun He is in the Information Center of YRCC, Zhengzhou, 450003, China.

Manuscript received April 19, 2021; revised August 16, 2021.

Another popular approach is adversarial training, focusing on adaptations at the feature level [12] [17] or the pixel level [13]. For feature level adaptation, the objective is to align the feature distribution of different domains, whereas pixel-level domain adaptation seeks to modify the output to reduce prediction discrepancies between the source and target domains. However, these techniques often fail to perform well when used directly in domain adaptation for remote semantic segmentation. Currently, the most prevalent method is self-training [15], [16], which involves generating pseudo-labels for unlabeled target images to enhance the model. Despite self-training's proven higher efficacy over adversarial approaches, it faces challenges in producing high-confidence pseudo-labels and utilizing them effectively in the target domain.

We propose a hybrid training approach that combines self-training with generative training methods. This combination reduces the negative effects of noise that may arise from generative training, enhancing the accuracy of pseudo-labels. By combining these methods, we were able to mitigate the limitations of each method and achieve higher performance. Specifically, the generative training method is used to adjust the image style while preserving the original semantic information. The self-training method is augmented with a dual-domain image fusion (DDF) module and a pseudo-label regional reweighting (PRR) strategy to improve the generalization ability of the model. Compared with existing fusion techniques, the DDF module introduces an innovative way to integrating images across different domains. The DDF generates an intermediate domain containing information from both source and target domains, which helps to achieve better generalization by mitigating the domain gap and enhancing feature representation. Additionally, we introduce a PRR strategy that assigns different weights to pseudo-label regions based on their spatial context and difficulty level. This approach allows for more efficient utilization of the labels and focuses on improving the segmentation accuracy for challenging categories.

To summarize, the main contributions are as follows:

- We propose a novel DDF module for image fusion to mitigate the domain gap issue. This module combines images from both the source and target domains to produce fusion images. Fusion images containing dual-domain information perform alignment at the input level, thus reducing the domain gap. The network learns from these fused images, thereby enhancing its ability to generalize to the target domain.
- We utilize a hybrid training strategy that focuses on a self-training framework, supplemented by a generative training method. By combining these two approaches strategically, we are able to reduce the negative effects of noise that may arise from the generative training method. As a result, the accuracy of the pseudo-labels is enhanced. Furthermore, the PRR module provides additional support for pseudo-labels, and focuses on improving segmentation accuracy for challenging categories.
- Our method surpasses existing techniques by achieving a mIoU of 66.81% and an F1-score of 79.61% when carrying out the segmentation task from Potsdam R-G-B to Vaihingen. These findings indicate a substantial enhancement of 5.40% and 4.69% compared to the current state-of-the-art method, emphasizing the efficacy of our approach. And ablation experiments demonstrated a notable improvement in the above modules.

II. Related Work

A. Semantic Segmentation

In recent years, there has been significant progress in the field of semantic segmentation. This progress can be attributed to the introduction of advanced architectures, innovative techniques, and improved performance levels. [18] proposed a groundbreaking fully convolutional neural network (FCN) architecture, which differed from the traditional approach of using fully connected layers. This marked a turning point in the field. Subsequently, convolutional neural networks (CNNs) have become the dominant paradigm. [19] introduced dilated convolutions as a means to efficiently capture contextual information at multiple scales. The UNet [20] architecture and its various variants, such as UNet++ [21], have been developed to enhance the segmentation capabilities of the original architecture. Most networks now follow an encoder-decoder architecture, which enables the capture of hierarchical features and the refinement of segmentation maps.

Although CNNs have been successful in extracting local features, they have limitations in capturing long-range dependencies and global contextual information, which are crucial for visual tasks. To address this, researchers have looked to natural language processing (NLP) for inspiration and adapted the Transformer architecture [22] for computer vision tasks like semantic segmentation. The self-attention mechanism in Transformers allows the model to assign different weights to different positions in the input sequence, effectively capturing both local and global dependencies. As a result, there is a growing trend in computer vision to explore hybrid architectures that combine the strengths of both CNNs and Transformers.

B. Unsupervised Domain Adaptation for Semantic Segmentation

Unsupervised domain adaptation deals with the problem of adjusting a model that has been trained in a source domain, where the labeled data are accessible, to perform effectively on a target domain, where labeled data is not accessible. In UDA, the source and target domains typically exhibit dissimilar distributions, leading to a domain shift. UDA techniques can be categorized into generative training methods, adversarial-based methods, and self-training methods.

The objective of adversarial training methods is to align the distributions of the source and target domains [23], [24]. An example is domain-invariant representation, which involves a game of least-maximum adversarial optimization [25]. In this game, feature extractors try to deceive domain discriminators to achieve aligned feature distributions. However, these adversarial training methods often have inconsistent performance.

Self-training is a technique that involves utilizing a model trained on labeled data from a specific domain to generate

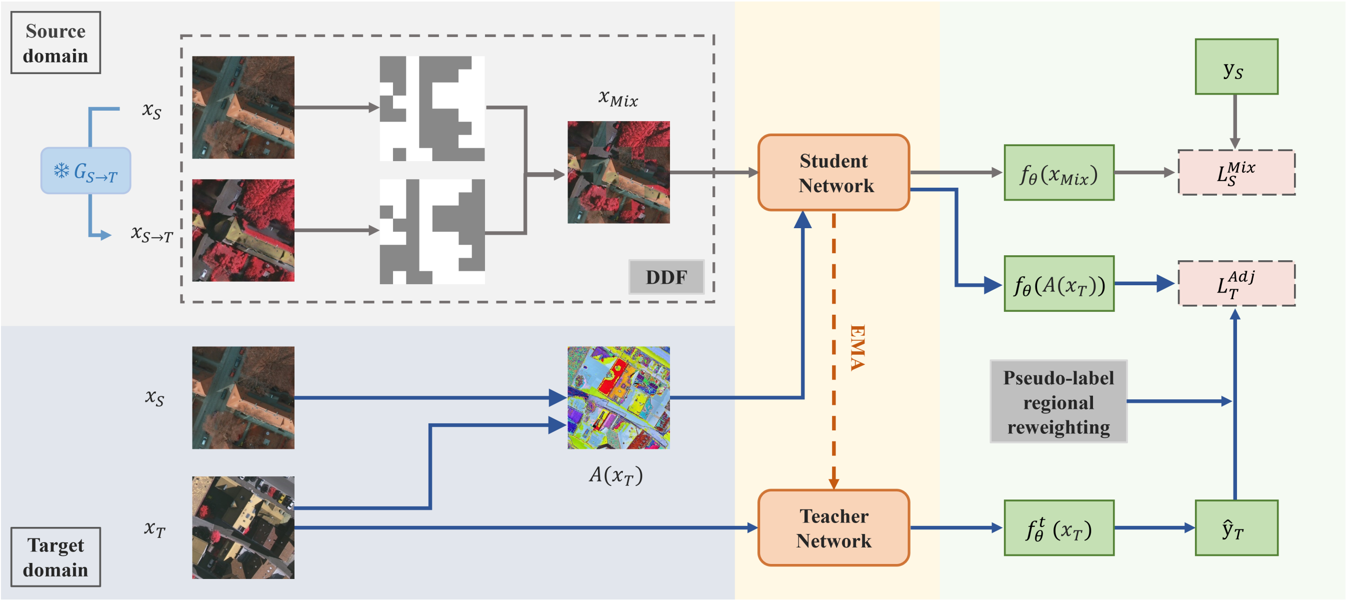

Fig. 2. The structure we propose comprises three primary elements: a self-training model that utilizes GT augmented images, a dual-domain fusion module (DDF), and a strategy for assigning weights to pseudo-label regions. The DDF integrates the original image from the source domain with its corresponding transferred image into the student model. Data augmentation is performed on the target domain image to form $A(x_T)$. The teacher model assigns pseudo-labels $\hat{y}_T$ to the target domain image, and the student model generates predictions of $A(x_T)$. We also perform regional adjustments to the weights assigned to pseudo-labels, thereby obtaining weighted pseudo-labels.

pseudo-labels for data from a different domain [15], [26]. These pseudo-labels are then used to retrain the model. Most methods for unsupervised domain adaptation (UDA) follow one of two approaches. The first approach involves pre-computing pseudo-labels offline, training the model, and repeating this process iteratively [27]. The second approach involves computing pseudo-labels online during training [16]. However, since there are inherent differences in the distribution of data between the domains, the pseudo-labels generated are likely to contain some level of noise. To mitigate the impact of incorrect labeling, pseudo-labels with a high level of confidence are often utilized.

Moreover, the application of domain mix-ups is also incorporated [28], [29]. This approach involves merging different characteristics from both the source and target domains during the training phase. By doing so, it facilitates a more seamless transition between domains, reducing the model's susceptibility to domain shifts. This methodology is influenced by the concept of mix-up regularization.

C. Unsupervised Domain Adaptation for Remote Sensing

Due to varying factors, including task attributes, operational hours, and environmental meteorological conditions, there are significant differences in data distribution in remote sensing scenes. These differences are characterized by large inter-domain variation and inter-class similarity. If a model pretrained on an experimental dataset is naively applied to another dataset, it may fail to yield satisfactory segmentation outcomes due to the aforementioned factors. For the semantic segmentation of remote sensing images, Benjdira et al. [30] first addressed the domain adaptation challenge, employing a Generative Adversarial Network (GAN)-based framework and achieving promising results. Following their pioneering work, Li et al. [31] introduced DualGAN, which further extended the basic GAN to two coupled GANs. This innovation facilitates the learning of image inversion tasks and introduces image reconstruction losses, which significantly improves the quality of the generated images. Continuing in this vein, Chen et al. [32] proposed a region and category adaptive domain discriminator, which aims to reduce region and class differences during domain alignment, surpassing traditional adversarial techniques in terms of results. Zhao et al. [33] designed a resi-dual GAN to solve the scale variation within remote sensing datasets, thereby bolstering the capacity for style transfer across images. Considering the characteristics of remote sensing data, Zhang et al. [34] proposed a novel domain adaptive algorithm termed OSDA-ETD including the transferability and discriminability strategy. It is designed to reduce the global and local distribution differences between domains, and enhance the distribution differences of different categories in different domains. With the advancement of large models, Hong et al. [35] introduced the SpectralGPT, a universal RS foundational model, which is purposefully built to handle spectral RS images leveraging an innovative 3D generative pretrained transformer (GPT).

III. Methodology

This section begins by presenting a broad definition and notation of self-training for UDA, along with an explanation of its training process. Following that, we offer a comprehensive explanation of the hybrid training framework, as well as the DDF and PRR modules that have been proposed.

A. Self-training for UDA

In the conventional UDA approach, a neural network is trained using a set of images in the source domain $X_S = \{x_S^{(i)}\}_{i=1}^{N_S}$ along with their corresponding labels $Y_S = \{y_S^{(i)}\}_{i=1}^{N_S}$. Then, fine-tune the network in order to achieve satisfactory performance on a different set of target domain images $X_T = \{x_T^{(i)}\}_{i=1}^{N_T}$, where the labels Y_T are unknown. The accuracy of the model drops noticeably when directly applied to the target domain as a result of the domain gap. Therefore, our aim is to make accurate predictions on the target domain, without having access to the target domain labels. To accomplish this, this study utilizes a self-training approach to generate pseudo-labels for the target domain. These pseudo-labels are subsequently employed to guide the training of target domain images.

The self-training method comprises a semantic segmentation network consists of a student network f_θ and a teacher network f_θ^t. We first train the student network f_θ with backward propagation to minimize cross-entropy loss in the source domain.

$$L_S = -\sum y_S log(f_\theta(x_S)) \tag{1}$$

The teacher network f_θ^t is then initialized as a copy of f_θ. To avoid an increase in misclassification probability due to erroneous results of the student model, it is recommended not to share weights with the teacher model during the training procedure. Instead, the exponential moving average (EMA) method [36] can be used to aggregate information from each step, resulting in smoother output and improved pseudo-labeling quality. Generally, the weights of f_θ^t are set as the exponential moving average of the weights of f_θ at training step τ. The teacher model parameters are updated as follows:

$$\theta_\tau^t = \alpha\theta_{\tau-1}^t + (1-\alpha)\theta_\tau. \tag{2}$$

To assist the model in learning from the unlabeled target domain data, we request the teacher model f_θ^t to generate pseudo-labels $\hat{y}_T = f_\theta^t(x_T)$ for the target domain data. And then they are used to additionally train the network f_θ on the target domain.

$$L_T = -\sum \hat{y}_T log(f_\theta(x_T)) \tag{3}$$

The overall loss function for the student network f_θ can be expressed as:

$$L_{total} = L_S + \lambda L_T, \tag{4}$$

with λ being a hyper parameter that leverages both parts.

B. The proposed hybrid training framework

Pseudo-label quality is crucial for the self-training approach to UDA. Incorrect pseudo-labels primarily arise from a substantial mismatch between the two domains. As the distributional gap between the domains narrows, the model's efficacy on the target domain enhances. Style transfer commonly serves to make the source domain visually similar to the target domain. Nevertheless, this technique has a limitation: It cannot ensure that the images after transformation are devoid of noise. Should these transformed images yield erroneous semantic details, it could detrimentally affect the training of the model.

To address this problem, we introduce a hybrid training framework that utilizes images with transferred styles while preserving semantic content to enhance the self-training process. During the style transfer phase, a generative network consists of a generator $G_{S\to T}$ and a discriminator D_S is utilized. $G_{S\to T}$ modifies X_S into the style of X_T, producing $X_{S\to T}$, which alters only the style but retains the semantic content. D_S assesses if the image is synthetically produced, and $G_{S\to T}$ aims to create an image that can mislead D_S. The cooperative effort of these networks results in the production of style-transferred images.

During training, we further improves the use of style transfered images with DDF, as shown in Fig. 2. x_S and $x_{S\to T}$ are fused through DDF to obtain the fusion image x_{Mix}. The fused image retains the semantic information, so its corresponding label remains y_S. The fused image not only reduces the domain gap, but also mitigates the impact of noise. We train the student network f_θ with a cross-entropy loss in the source domain where x_{Mix} belongs.

$$L_S^{Mix} = -\sum y_S log(f_\theta(x_{Mix})) \tag{5}$$

In our approach, the accuracy of pseudo-labels is of paramount importance, as they serve as a guide for training the model on the target domain data. To enhance the reliability of these pseudo-labels, we propose a novel regional reweighting strategy. This method evaluates the quality of pseudo-labels and assigns weights to different regions based on the difficulty of the categories, with a focus on challenging areas such as object boundaries. As the training progresses, we dynamically adjust the weights in w to prioritize regions that are harder to detect. With the regional reweighting strategy in place, we can now define the adjusted loss for the unlabeled target domain images, which incorporates the weighted pseudo-labels:

$$L_T^{Adj} = -\sum w * \hat{y}_T log(f_\theta(x_T)) \tag{6}$$

Ultimately, the loss calculations for our hybrid training model are defined as follows.

$$L_{total} = \lambda_1 * L_S^{Mix} + \lambda_2 * L_T^{Adj}, \tag{7}$$

and λ_1, λ_2 are hyperparameters.

C. Dual-domain image fusion

In the conventional training approaches for cross-domain segmentation networks, the strategy to reduce domain differences involves exclusively using style-transferred images during the training process. However, this method might lead to errors that could impair the model's performance. To tackle these challenges and further diminish the domain gap, a novel dual-domain image fusion module is developed. This module merges the original and style-transferred images to create a composite image. The fused image enhances the information from the target domain without increasing the size of the dataset. Consequently, the network derives advantages from learning within both domains, thereby acquiring uniform features.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3433564

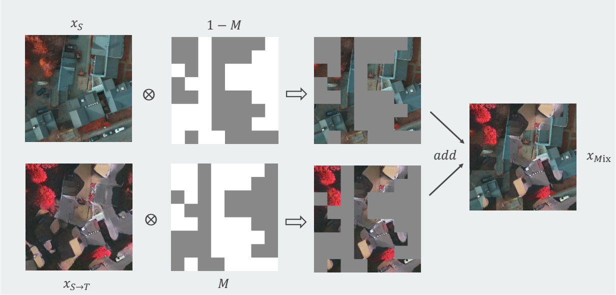

Fig. 3. An illustration of the proposed Naïve Fusion module.

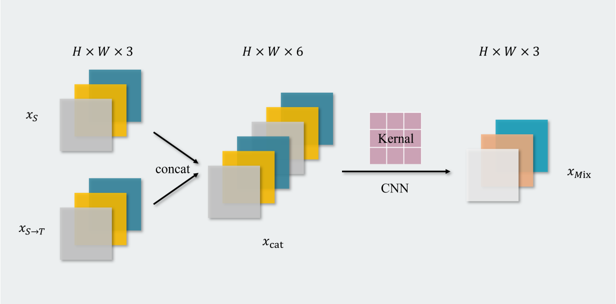

Fig. 4. An illustration of the proposed CNN Fusion module.

1) Naïve Fusion: The process begins with the creation of a fusion mask through region-wise entropy filtering. Initially, a transferred image $x_{S\to T}$ is entered into the student model to determine the probability of the output logits $p_{S\to T}$. The image is then segmented into N patches of equal size, each $k\times k$. Following this, we calculate the total entropy for each patch.

$$E[x_{S\to T}]=-\sum_{i=1}^{N}(p^i_{S\to T}*log(p^i_{S\to T})) \tag{8}$$

Patches exhibiting lower entropy demonstrate reduced noise in the transferred image, whereas those with higher entropy are associated with increased noise, potentially harming network training. Therefore, we select the part of the transferred image that has a lower entropy and eliminate the part with a higher entropy. By sorting the entropy values of N patches from lowest to highest, we can determine the smallest value τ_E among the blocks that have the lowest entropy of $c\%$.

$$\tau_E=percentile(E[x_{S\to T}],c) \tag{9}$$

Next, the patch mask M is calculated using the following equation:

$$M=\mathbb{1}[E[X_{S\to T}]<\tau_E], \tag{10}$$

where $\mathbb{1}$ is an indicator function.

Ultimately, we select segments from the transferred image using the mask M and complete the remaining areas with parts of the original image to form a composite image x_{Mix}, as illustrated in Fig. 3.

$$x_{Mix}=M\otimes x_{S\to T}+(1-M)\otimes x_S \tag{11}$$

$\otimes$ repesents dot multiplication.

2) CNN Fusion: Naïve Fusion involves combining sections of the original and transferred images. When there is a considerable gap between the two domains, the style of the combined image tends to become disordered. Additionally, the method of fusing by regions often leads to mismatches at the boundaries. Consequently, we suggest the adoption of CNN Fusion when the domains exhibit significant divergence. This approach incorporates a convolutional network with little effect on the efficiency of training. The images produced through CNN Fusion are more aligned with the intermediate domain, thereby reducing the domain disparity.

Firstly, the image x_S and its corresponding transferred image $x_{S\to T}$ should be concatenated. Both images have a size of $H\times W\times C$.

$$x_{cat}=concat(x_S,x_{S\to T})\in\mathbb{R}^{H\times W\times 2C} \tag{12}$$

Next, the input x_{cat} undergoes convolution using a convolutional network, where the convolutional kernel size is set to 3 and the number of convolutional kernels is set to 3.

$$x_{Mix}=fusion\_conv(x_{cat})\in\mathbb{R}^{H\times W\times C} \tag{13}$$

Fig. 4 shows the processing details, the resulting x_{Mix} is physically closer to the intermediate domain and preserves the semantic information of the initial image x_S. Subsequently, this image is utilized as an enhanced source domain image for network training.

D. Pseudo-label regional reweighting

The accuracy of pseudo-labels is crucial for the training process, and inevitably, some are less reliable. Assessing the quality of pseudo-labels is vital. We suggest a dynamic assessment of pseudo-label quality to reduce the impact of inaccurate pseudo-labels during training. This involves calculating a quality matrix w for the pseudo-labels in each training batch. A greater value of w signifies increased reliability.

The initial setting is with w_{init}, where all pseudo-labels are treated uniformly. By establishing a threshold δ, the pixels exceeding this threshold can be considered trusted. Calculate the ratio of trustworthy pixels to the total number of pixels.

$$w_{init}=A*\frac{[f_\theta(x_T)>\delta]}{H\times W} \tag{14}$$

where A is an all-ones matrix.

Throughout several iterations of training, there exists a tendency for the network to favor simpler categories, which might lead to neglecting the more intricate ones. Additionally, classifying object boundaries poses a significant challenge, prompting the allocation of greater weights to these regions. This method thus prioritizes the more difficult categories, especially boundaries, while less attention is given to easily recognizable objects. To identify the object's boundary area, superpixel clustering [37] is employed, which aggregates pixels based on shared characteristics. Subsequently, a binary mask M_b is created to depict the boundary.

$$M_b = \begin{cases} 0 & x \notin boundary \quad region \\ 1 & x \in boundary \quad region \end{cases} \tag{15}$$

To increase the significance of positions on the region boundaries, the weights of the pseudo-labels are adjusted. The parameters in the initial matrix are kept constant for the non-edge positions that lie outside the mask area.

$$w = w_{init}[M_b == 1] + \beta, \beta \in (0, 1) \tag{16}$$

The parameter β in Eq. 16 is introduced as a means to adjust the weights of the pseudo-labels, particularly focusing on the regions of interest within the image. By increasing the weights for positions that are on the boundaries ($M_b = 1$), our model becomes more sensitive to these regions, which are typically harder to classify correctly. The adjusted weights help in improving the overall segmentation accuracy, especially for categories which have complex boundaries and are often confused with the background. β allows for a dynamic adjustment of the weights during training. As the network learns, the value of β can be fine-tuned to better capture the difficult-to-detect features in images.

IV. Experiments

A. Implementation Details

1) Datasets: Potsdam and Vaihingen are two datasets with images captured using different sensors and from different locations. They are subsets of the ISPRS 2D open-source RS semantic segmentation benchmark dataset[1]. The Potsdam dataset consists of three band modes: IR-R-G, R-G-B, and IR-R-GB. The experiments use Potsdam IR-R-G and Potsdam R-G-B, each containing 38 very high-resolution top-of-atmosphere reflectance products (VHR TOPs) with fixed dimensions of 6000 x 6000 pixels and a spatial resolution of 5 cm. The Vaihingen dataset has one band mode: IR-R-G, with 33 TOPs, each having dimensions of 2000 x 2000 pixels and a spatial resolution of 9 cm. The images are cropped to 896 x 896 pixels for Potsdam and 512 x 512 pixels for Vaihingen, resulting in a total of 1764 images for Potsdam IR-R-G and Potsdam R-G-B and 1,696 images for Vaihingen. The Potsdam dataset is divided into training and testing sets, with 1323 images in the training set and 441 images in the testing set. The Vaihingen dataset is also divided into training and testing sets, with 1256 images in the training set and 440 images in the testing set.

We propose four cross-domain RS semantic segmentation tasks, which are described as follows:

- *Task 1:* Potsdam IR-R-G to Vaihingen IR-R-G.
- *Task 2:* Vaihingen IR-R-G to Potsdam IR-R-G.
- *Task 3:* Potsdam R-G-B to Vaihingen IR-R-G.
- *Task 4:* Vaihingen IR-R-G to Potsdam R-G-B.

2) Network architecture and training: We adopt SegFormer [38] as the base architecture, which is pre-trained in ImageNet-1k. To train the network, we employ AdamW [39] optimizer with a learning rate of 6×10^{-5} for the encoder and 6×10^{-4} for the decoder. A weight decay of 0.01 is applied, along with a linear learning rate warm-up for 1.5k steps, followed by linear decay. In Equation 7, both λ_1 and λ_2 are set to 1 followed previous work. In Task 1 and 3, the parameter c in Equation 9 is set to 50. In Task 2 and 4, c is set to 25. Within the PRR module, the threshold δ is established at 0.9. The data augmentation method employs DACS [29]. Concurrently, optical perturbations are used to enhance the model's generalization capacity, addressing differences in environmental conditions (e.g., temperature and humidity, atmospheric effects) and instrument configurations (e.g., sensor noise) [40]. All models are trained utilizing a single GPU from the NVIDIA Tesla V100.

3) Evaluation metric: For straightforward comparison with alternative approaches, this study uses the commonly adopted evaluation metrics, mIoU and F1-score, for semantic segmentation. The intersection of union (IoU) for each category is determined using the equation $IoU = A \cap B / A \cup B$. The mIoU denotes the mean IoU score in various classes. The F1-score is calculated as F1-score= $(2 \cdot Precision \cdot Recall)/(Precision + Recall)$. And we also use the mF-score to represent the mean F1-score value of all classes.

B. Comparison with SOTA methods

Typically, methods are categorized into two distinct groups based on their backbones. The initial group utilizes DeepLabV3, including methods such as AdaptSegNet [12], ProDA [41], Li's [31], Zhang's [43], Wang's [44], and CIA-UDA [45]. The second group employs Segformer, featuring methods such as DAFormer [16] and ST-DASegNet [46]. For our approach, we selected the latter as the baseline. In addition, we evaluated the enhancements of our proposed modules on various other backbones in a separate ablation study.

1) Comparison experiments on Task 1 from Potsdam IR-R-G to Vaihingen IR-R-G: In this study, Potsdam IR-R-G images serve as the source domain, while Vaihingen IR-R-G images are used as the target domain. The training involves 1764 annotated images from Potsdam and 1296 unannotated

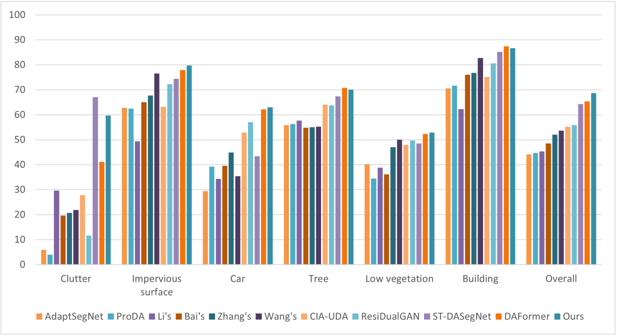

Fig. 5. The histogram results for cross-domain semantic segmentation from Potsdam IR-R-G to Vaihingen IR-R-G.

[1] https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3433564

TABLE I
THE QUANTITATIVE RESULTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM POTSDAM IR-R-G TO VAIHINGEN IR-R-G.

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
AdaptSegNet [12]	5.84	9.01	62.81	76.88	29.43	44.83	55.84	71.45	40.16	56.87	70.64	82.66	44.12	56.95
ProDA [41]	3.99	8.21	62.51	76.85	39.20	56.52	56.26	72.09	34.49	51.65	71.61	82.95	44.68	58.05
Li's [31]	29.66	45.65	49.41	66.13	34.34	51.09	57.66	73.14	38.87	55.97	62.30	76.77	45.38	61.43
Bai's [42]	19.60	32.80	65.00	78.80	39.60	56.70	54.80	70.80	36.20	53.20	76.00	86.40	48.50	63.10
Zhang's [43]	20.71	31.34	67.74	80.13	44.90	61.94	55.03	71.90	47.02	64.16	76.75	86.65	52.03	66.02
Wang's [44]	21.85	35.87	76.58	86.73	35.44	52.33	55.22	71.15	49.97	66.64	82.74	90.56	53.63	67.21
CIA-UDA [45]	27.80	43.51	63.28	77.51	52.91	69.21	64.11	78.13	48.03	64.90	75.13	85.80	55.21	69.84
ResiDualGAN [33]	11.64	18.42	72.29	83.89	57.01	72.51	63.81	77.88	49.69	66.29	80.57	89.23	55.83	68.04
ST-DASegNet [46]	**67.03**	**80.28**	74.43	85.36	43.38	60.49	67.36	80.49	48.57	65.37	85.23	92.03	64.33	77.34
DAFormer* [16]	41.21	58.37	77.95	87.61	62.21	76.70	**70.80**	**82.90**	52.38	68.75	**87.44**	**93.30**	65.33	77.94
Ours	59.69	74.76	**79.72**	**88.72**	**63.04**	**77.33**	70.03	82.37	**52.94**	**69.23**	86.70	92.88	**68.69**	**80.88**

"*" are our re-implemented version for RS images.

TABLE II
THE QUANTITATIVE RESULTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM VAIHINGEN IR-R-G TO POTSDAM IR-R-G.

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
AdaptSegNet [12]	8.36	15.33	49.55	64.64	40.95	58.11	22.59	36.79	34.43	61.50	48.01	63.41	33.98	49.96
ProDA [41]	10.63	19.21	44.70	61.72	46.78	63.74	31.59	48.02	40.55	57.71	56.85	72.49	38.51	53.82
Li's [31]	11.48	20.56	51.01	67.53	48.49	65.31	34.98	51.82	36.5	53.48	53.37	69.59	39.30	54.71
Zhang's [43]	12.31	24.59	64.39	78.59	59.35	75.08	37.55	54.60	47.17	63.27	66.44	79.84	47.87	62.66
Wang's [44]	11.65	19.47	73.43	84.55	63.86	77.85	32.68	47.36	47.69	63.45	76.32	87.43	50.94	63.31
CIA-UDA [45]	10.87	19.61	62.74	77.11	65.35	79.04	47.74	64.63	54.40	70.47	72.31	83.93	52.23	65.80
ST-DASegNet [46]	0.18	0.35	76.45	86.65	**73.54**	**84.76**	**62.89**	**77.22**	61.04	75.80	83.81	91.19	59.65	69.33
DAFormer* [16]	12.97	22.96	69.65	82.11	71.42	83.33	58.34	73.69	57.79	73.25	**90.00**	**94.74**	60.03	71.68
Ours	**16.34**	**28.09**	**78.51**	**87.96**	72.90	84.32	60.57	75.45	**66.11**	**79.60**	89.16	94.27	**63.93**	**74.95**

"*" are our re-implemented version for RS images.

TABLE III
THE QUANTITATIVE RESULTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM POTSDAM R-G-B TO VAIHINGEN IR-R-G.

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
AdaptSegNet [12]	6.49	9.82	55.70	71.24	33.85	50.05	47.72	64.31	22.86	36.75	65.70	79.15	38.72	51.89
ProDA [41]	2.39	5.09	49.04	66.11	31.56	48.16	49.11	65.86	32.44	49.06	68.94	81.89	38.91	52.70
Li's [31]	3.94	13.88	46.19	61.33	40.31	57.88	55.82	70.66	27.85	42.17	65.44	83.00	39.93	54.82
Bai's [42]	10.80	19.40	62.40	76.90	38.90	56.00	53.90	70.00	35.10	51.90	74.80	85.60	46.00	60.00
ResiDualGAN [33]	9.76	16.08	55.54	71.36	48.49	65.19	57.79	73.21	29.15	44.97	78.97	88.23	46.62	59.84
Zhang's [43]	12.38	21.55	64.47	77.76	43.43	60.05	52.83	69.62	38.37	55.94	76.87	86.95	48.06	61.98
Wang's [44]	12.61	22.39	**73.80**	**84.92**	43.24	60.38	44.41	61.50	43.27	60.40	83.76	91.16	50.18	63.46
CIA-UDA [45]	13.50	23.78	62.63	77.02	52.28	68.66	63.43	77.62	33.31	49.97	79.71	88.71	50.81	64.29
ST-DASegNet [46]	36.03	50.64	68.36	81.28	43.15	60.28	64.65	78.31	34.69	47.08	84.09	91.33	55.16	68.15
DAFormer* [16]	39.66	56.79	69.98	82.34	**58.01**	**73.43**	69.21	81.81	45.76	62.79	**85.82**	**92.37**	61.41	74.92
Ours	**56.82**	**72.47**	72.19	83.85	57.73	73.20	**73.12**	**84.47**	**55.85**	**71.67**	85.18	91.99	**66.81**	**79.61**

"*" are our re-implemented version for RS images.

TABLE IV
THE QUANTITATIVE RESULTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM VAIHINGEN IR-R-G TO POTSDAM R-G-B.

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
AdaptSegNet [12]	6.11	11.50	37.66	59.55	42.31	55.95	30.71	45.51	15.10	25.81	54.25	70.31	31.02	44.75
ProDA [41]	11.13	20.51	44.77	62.03	41.21	59.27	30.56	46.91	35.84	52.75	46.37	63.06	34.98	50.76
Li's [31]	**13.56**	**23.84**	45.96	62.97	39.71	56.84	25.80	40.97	41.73	58.87	59.01	74.22	37.63	52.95
Zhang's [43]	13.27	23.43	57.65	73.14	56.99	72.27	35.87	52.80	29.77	45.88	65.44	79.11	43.17	57.77
Wang's [44]	10.84	17.49	66.11	79.75	65.45	80.17	28.64	43.51	35.47	51.85	68.63	81.32	45.86	59.74
CIA-UDA [45]	9.20	16.86	53.39	69.61	63.36	77.57	44.90	61.97	43.96	61.07	70.48	82.68	47.55	61.63
DAFormer* [16]	5.85	11.05	66.75	80.06	69.08	81.71	52.41	68.77	**65.99**	**79.51**	78.09	87.70	56.36	68.13
ST-DASegNet [46]	3.70	7.38	69.83	83.12	**75.99**	**87.89**	57.41	73.47	50.76	67.64	83.46	90.67	56.86	68.37
Ours	4.41	8.46	**75.82**	**86.24**	66.17	79.64	**62.58**	**76.99**	64.45	78.38	**87.67**	**93.43**	**60.18**	**70.52**

"*" are our re-implemented version for RS images.

images from Vaihingen. Assessments are performed on 440 test images from Vaihingen. The results are presented in Table I and a more visual comparison is available in Fig. 5. Our approach surpasses the DAFormer, which is based on SegFormer, showing a 3.36% increase in mIoU and a 2.94% increase in mF-score. In particular, we record superior performance in the categories of 'Impervious surface', 'Car', and 'Low vegetation', with only slight differences in the categories of 'Tree' and 'Building' compared to the best results.

2) Comparison experiments on Task 2 from Vaihingen IR-R-G to Potsdam IR-R-G: In this study, Vaihingen IR-R-G images serve as the source domain, while Potsdam IR-R-G images are used as the target domain. The training involves 1696 labeled images from Vaihingen and 1223 unlabeled images from Potsdam. The assessment is performed on 441 test images from Potsdam, with the results comparison shown in Table II. Compared to the former SOTA method DAFormer, this approach shows a 3.9% enhancement in the mIoU score and a 3.27% increase in the mF-score. Specifically, gains of 3.37% in 'Clutter' and 8.32% in 'Low vegetation' were noted. In other categories with less impressive scores, the variations were minor.

3) Comparison experiments on Task 3 from Potsdam R-G-B to Vaihingen IR-R-G: In this study, Potsdam R-G-B images are utilized as the source domain, while Vaihingen IR-R-G images serve as the target domain. The training involves 1764 labeled images from Potsdam and 1296 unlabeled images from Vaihingen. Assessments are performed using 440 test images from Vaihingen. This challenge is greater than the *Task 1* challenge due to sensor differences, leading to a wider disparity between the domains. Nonetheless, our findings indicate substantial improvements in this area. These improvements are detailed in Table III. Notably, our method sets a new benchmark in performance, leading in five categories and showing remarkable results in the 'Clutter' category by significantly outperforming the runner-up(Specifically, our technique achieves a 17.16% increase in IoU value and an 15.68% rise in F1-score.).

4) Comparison experiments on Task 4 from Vaihingen IR-R-G to Potsdam R-G-B: In this study, Vaihingen IR-R-G images are utilized as the source domain while Potsdam R-G-B images serve as the target domain. The training involves 1696 labeled images from Vaihingen and 1223 unlabeled images from Potsdam. Assessments are performed using 441 test images from Potsdam. Our findings demonstrate excellent overall results as indicated in Table IV. These results are competitive in various categories without any noticeable flaws. In intricate situations, the categories 'Impervious surface' and 'Tree' show high effectiveness. Additionally, the network's ability to recognize large-scale objects remains impressively stable.

C. Ablation Study

In our ablation study, we compared various configurations to evaluate the influence of individual components on overall performance. The findings are presented in Table V, Table VI, Table VII, and Table VIII. We investigated the DDF module and the PRR strategy. Throughout training, the network architecture was maintained uniformly in all tasks. The 'Base' task was similar to the others, with the only difference being the lack of the mentioned modules.

The DDF module demonstrated superior results in nearly all categories, whether combined with Na"ive Fusion or CNN Fusion. This improvement can be linked to the DDF module's ability to effectively bridge the domain gap by merging data from different sources, enhancing performance. Moreover, the use of various sensors accentuated the disparities between the source and target domains, making the tasks of "Potsdam R-G-B to Vaihingen IR-R-G" and "Vaihingen IR-R-G to Potsdam R-G-B" particularly difficult. However, our module introduced demonstrated considerable improvements in these areas. The integration of these modules successfully reduces the domain gap, particularly in regions with marked differences. Significant improvements in overall outcomes were observed, marked by uniform performance in all categories.

The DDF ablation studies reveal varying results with the introduction of different fusion techniques. According to Table V and Table VI, Naïve Fusion outperforms CNN Fusion. However, the findings in Table VII and Table VIII show the opposite effect. This variation is attributed to differences in the data distribution. The experiments involved two distinct image

TABLE V
ABLATION EXPERIMENTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM POTSDAM IR-R-G TO VAIHINGEN IR-R-G

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
Base	30.22	46.42	77.42	87.27	60.47	75.37	70.63	82.79	51.76	68.21	87.06	93.08	62.93	75.52
Base +DDF (Naïve)	54.00	70.13	79.16	88.37	63.14	77.40	70.64	82.80	53.89	70.04	86.53	92.78	67.89	80.25
Base +DDF (CNN)	55.95	71.75	77.53	87.35	56.45	72.16	71.01	83.05	52.60	68.93	85.57	92.22	66.52	79.24
Base +PRR	52.62	68.96	79.27	88.44	62.29	76.77	71.39	83.31	55.29	71.21	86.23	92.61	67.85	80.21
Base +DDF(Naïve) +PRR	59.69	74.76	79.72	88.72	63.04	77.33	70.03	82.37	52.94	69.23	86.70	92.88	68.69	80.88

TABLE VI
ABLATION EXPERIMENTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM VAIHINGEN IR-R-G TO POTSDAM IR-R-G

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
Base	12.33	23.15	77.42	86.42	70.49	82.67	58.44	74.16	65.02	77.02	86.82	92.52	61.77	72.66
Base +DDF (Naïve)	14.21	24.88	73.52	84.74	76.8	86.88	61.27	75.98	64.94	78.75	87.98	93.59	63.11	74.13
Base +DDF (CNN)	13.25	23.39	77.37	87.24	68.28	81.85	59.91	74.93	66.58	79.94	88.63	93.97	62.50	73.55
Base +PRR	12.61	22.40	75.72	86.18	75.64	86.13	57.09	72.69	66.84	80.13	87.40	93.27	62.55	73.47
Base +DDF(Naïve) +PRR	16.34	28.09	78.51	87.96	72.9	84.32	60.57	75.45	66.11	79.60	89.16	94.27	63.93	74.95

TABLE VII
ABLATION EXPERIMENTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM POTSDAM R-G-B TO VAIHINGEN IR-R-G

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
Base	14.37	25.14	69.20	81.79	61.90	76.47	69.51	82.02	40.17	57.32	85.95	91.86	56.68	69.10
Base +DDF (Naïve)	47.89	64.76	75.45	86.01	59.94	74.95	69.48	81.99	47.93	64.80	85.30	92.07	64.33	77.43
Base +DDF (CNN)	46.34	63.34	75.25	85.88	58.93	74.16	72.17	83.83	55.29	71.21	85.60	92.24	65.60	78.44
Base +PRR	40.91	58.06	76.22	86.51	60.68	75.53	69.29	81.86	51.14	67.68	85.44	92.15	63.95	76.96
Base +DDF(CNN) +PRR	56.82	72.47	72.19	83.85	57.73	73.20	73.12	84.47	55.85	71.67	85.18	91.99	66.81	79.61

formats: IR-R-G and R-G-B. For the ’IR-R-G to IR-R-G’ task, Naïve Fusion, using entropy-based selection, achieved superior results due to the compatibility of image formats across domains. This approach successfully captured the nuances in the intermediate domains. On the other hand, in the ’R-G-B to IR-R-G’ or ’IR-R-G to R-G-B’ tasks, where there is a significant difference in visual representation, employing a trainable CNN network yielded better outcomes. Our research, which spans both uniform and varied imaging conditions, consistently validates these observations. Thus, we argue that the selection between Naïve and CNN fusion techniques should be based on the characteristics of the data set. In cases where datasets have slight differences, Naïve Fusion offers quicker and more effective results. In contrast, for datasets with greater discrepancies, a trainable CNN is more advantageous. Regarding computational expenses, the Naïve approach necessitates the computation of a mask matrix and its subsequent pixel-by-pixel multiplication with the image, leading to a computational cost of 0.263 MFLOPs. Conversely, the CNN approach employs a compact convolutional network with a weight parameter of 162, which incurs 84.935 MFLOPs.

In contrast, the PRR strategy concentrates on challenging categories. By employing its region weights, the PRR method adeptly identifies hard-to-detect features in images, thus considerably improving the classification success for the ’Clutter’ and ’Low Vegetation’ categories. In the ’Base’ setting, these categories typically present classification challenges due to their complex visual features and their resemblance to the background, often leading to sub-par results. However, the adoption of the PRR technique has resulted in a significant improvement in classification accuracy for these categories. This enhancement is largely due to the PRR method’s effectiveness in handling unclear boundaries.

D. Style transfer

Style transfer plays a crucial role in our research as it is directly related to the DDF module. By exploring various methods, we opted to employ a bi-directional GAN [33] as our style transfer network. Regardless of the style transfer network, incorporating our fusion module will benefit the final result. Specifically, in ablation experiments, the results of “Base” solely utilize the original images without any image migration stage or fusion module. We also conducted experiments using only the transferred images on “Vaihingen IR-R-G to Potsdam IR-R-G” task and “Potsdam R-G-B to Vaihingen IR-R-G” task shown in Table IX.

TABLE VIII
ABLATION EXPERIMENTS OF THE CROSS-DOMAIN SEMANTIC SEGMENTATION FROM VAIHINGEN IR-R-G TO POTSDAM R-G-B

Methods	Clutter		Impervious surface		Car		Tree		Low vegetation		Building		Overall	
	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	IoU	F1-score	mIoU	mF-score
Base	8.97	16.47	66.04	79.55	56.97	72.59	37.46	54.50	60.38	75.30	82.60	90.47	52.07	64.81
Base +DDF (Naïve)	10.92	19.69	75.33	85.93	62.17	76.67	58.64	73.93	62.55	76.96	83.97	91.29	58.93	70.74
Base +DDF (CNN)	9.30	17.01	75.55	86.07	66.30	79.74	56.08	71.86	65.31	79.01	86.79	92.93	59.89	71.10
Base +PRR	12.65	22.46	66.71	86.07	70.34	79.74	34.52	71.86	60.29	79.01	88.50	92.93	55.50	71.10
Base +DDF(CNN) +PRR	4.41	8.46	75.82	86.24	66.17	79.64	62.58	76.99	64.45	78.38	87.67	93.43	60.18	70.52

TABLE IX
EXPERIMENTS WITH DIFFERENT INPUTS ON TWO TASKS

Methods	Clutter	Impervious Surface	Car	Tree	Low Vegetation	Building	Overall
Vaihingen IR-R-G to Potsdam IR-R-G:							
original images	12.33	77.42	70.49	58.44	65.02	86.82	61.77
transferred images	15.75	74.44	69.65	64.06	62.8	81.66	61.39
DDF images	16.34	78.51	72.9	60.57	66.11	89.16	63.93
Potsdam R-G-B to Vaihingen IR-R-G:							
original images	14.37	69.20	61.90	69.51	40.17	85.95	56.68
transferred images	36.51	67.96	59.73	70.86	44.68	84.85	60.77
DDF images	56.82	72.19	57.73	73.12	55.85	85.18	66.81

TABLE X
EXPERIMENTS OF DIFFERENT NETWORK FROM VAIHINGEN IR-R-G TO POTSDAM IR-R-G

Methods	Clutter	Impervious Surface	Car	Tree	Low Vegetation	Building	Overall
ResNet-Base	4.61	60.82	48.49	58.62	38.99	73.61	47.52
ResNet-Base+DDF(Naïve)	6.05	68.44	41.63	66.25	32.56	80.85	49.30
UNetFormer-Base	3.73	74.75	51.37	60.35	43.56	86.37	53.35
UNetFormer-Base+DDF(Naïve)	7.50	76.20	57.59	70.80	55.12	86.84	59.01

As observed in the results, experiments solely employing transferred images often yield sub-optimal performance. We hypothesize that this phenomenon is closely tied to the quality of the images generated during the style transfer process. However, by introducing the DDF module, we achieved consistent improvements in the results. The DDF module effectively minimizes noise in the transferred images, leading to improved outcomes.

E. General utility of the DDF module

Our methodology is applicable not only to Segformer but also to other networks. We have previously endeavored to extend its applicability to various networks such as ResNet [47] and UNetFormer [48]. Table X presents the IoU results for the ”Potsdam IR-R-G to Vaihingen IR-R-G” task using DDF. Owing to the distinct network architectures of ResNet and UNetFormer, where the former is a convolutional network and the latter is a Transformer network. ResNet relies on convolutional layers for local information, which might not align as effectively with the DDF module's strategy for Naïve fusing. Because Naïve fusion is block-by-block fusion, the network needs to obtain the relationship between blocks. ResNet with small receptive fields may have a limited capacity to adjust to the intermediate domain created by the DDF module. This could result in a loss of discriminative power for certain classes that are already challenging to segment, especially for classes with intricate details such as 'Car' and 'Low Vegetation'. The self-attention mechanism in UNetFormer is beneficial for capturing remote dependencies and global context. Therefore, the architecture of UNetFormer may be more suitable for the DDF module, which can make better use of the intermediate domain information to improve the segmentation accuracy.

F. Visualization

This subsection is centered on the visualization results of our experimental study. Our technique outperforms traditional methods in several dimensions. The illustrations in Fig. 6 show that our model precisely segments and identifies complex mixed objects, thus decreasing the frequency of misclassifications into the 'Clutter' category. Moreover, our approach markedly improves performance in the 'Impervious surface' category. Compared to competing techniques, our method more effectively captures the details and textures of the surface, leading to more defined segmentation edges with less blurring and misalignment. Regarding the segmentation of the 'Car' category, our technique excels at clearly differentiating cars from their environments.

By evaluating our approach across various scene images, we not only enhance accuracy but also improve boundary clarity. Our approach's heightened sensitivity to detecting edges diminishes the chances of misclassification.

The visualization results confirm the outstanding performance of our proposed technique in the three main categories: 'Clutter', 'Impervious surface', and 'Car'. These results show the efficiency of our approach and establish a practical basis for implementing semantic segmentation in remote sensing images. These benefits will contribute to improved precision and dependability in the analysis of remote sensing images for future practical applications.

V. CONCLUSION

In this paper, we present a novel approach for cross domain semantic segmentation of remote sensing images. Our method consists of three main elements: a hybrid training approach, dual-domain image fusion, and regional weight pseudo-labeling. The hybrid training strategy improves the performance of self-training by using images augmented in the source domain. The dual domain image fusion strategy generates intermediate domain information and reduces the discrepancy between different domains. The regional weighting of

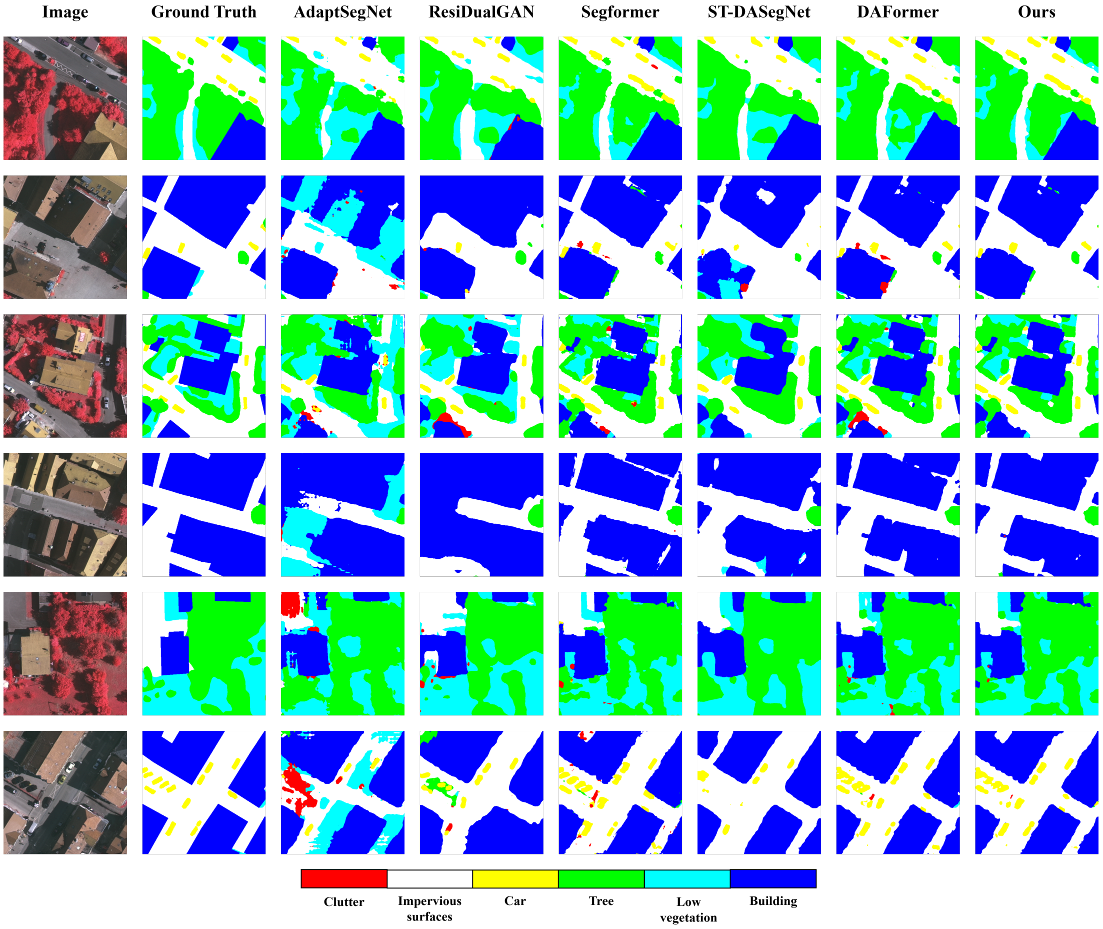

Fig. 6. The visualization of the cross-domain semantic segmentation results from Potsdam IR-R-G to Vaihingen IR-R-G.

pseudolabels assigns higher weights to categories that are more difficult to identify, leading to significant improvements in segmentation accuracy for those categories. To demonstrate the effectiveness of the proposed approach, extensive benchmark experiments and ablation studies are conducted on the ISPRS Vaihingen and Potsdam datasets. Moving forward, we intend to take a deeper look at the utilization of intermediate domain information. This encompasses devising innovative methods to generate and leverage intermediate representations that can further minimize the domain gap between the source and target domains. Based on the existing DDF strategy, our objective is to explore a more general image fusion strategy capable of achieving robust performance across diverse networks and datasets.

REFERENCES

[1] I. Kotaridis and M. Lazaridou, “Remote sensing image segmentation advances: A meta-analysis,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 309–322, 2021.

[2] X. Zhang, Z. Zhao, L. Ran, Y. Xing, W. Wang, Z. Lan, H. Yin, H. He, Q. Liu, B. Zhang, and Y. Zhang, “Fasticenet: A real-time and accurate semantic segmentation model for aerial remote sensing river ice image,” *Signal Processing*, vol. 212, p. 109150, Nov. 2023.

[3] X. Yuan, J. Shi, and L. Gu, “A review of deep learning methods for semantic segmentation of remote sensing imagery,” *Expert Systems with Applications*, vol. 169, p. 114417, 2021.

[4] X. Wu, J. Feng, R. Shang, J. Wu, X. Zhang, L. Jiao, and P. Gamba, “Multi-task multi-objective evolutionary network for hyperspectral image classification and pansharpening,” *Information Fusion*, vol. 108, p. 102383, 2024.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[6] D. Tuia, C. Persello, and L. Bruzzone, “Domain adaptation for the classification of remote sensing data: An overview of recent advances,” *IEEE geoscience and remote sensing magazine*, vol. 4, no. 2, pp. 41–57, 2016.

[7] J. Chen, J. Zhu, Y. Guo, G. Sun, Y. Zhang, and M. Deng, “Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[8] L. Ran, C. Ji, S. Zhang, X. Zhang, and Y. Zhang, “An unsupervised domain adaption framework for aerial image semantic segmentation based on curriculum learning,” in *2022 7th International Conference on Image, Vision and Computing (ICIVC)*. Xi’an, China: IEEE, Jul. 2022, pp. 354–359.

[9] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, “Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[11] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, “Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks,” *IEEE Transactions*

on Geoscience and Remote Sensing, p. 7178–7193, 2020. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2020.2980417

[12] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472–7481.

[13] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2512–2521.

[14] J. Feng, Z. Zhou, R. Shang, J. Wu, T. Zhang, X. Zhang, and L. Jiao, "Class-aligned and class-balancing generative domain adaptation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[15] Y. Zou, Z. Yu, B. V. K. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 297–313.

[16] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9914–9925.

[17] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. PP, no. 99, pp. 1–11, 2020.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[19] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[21] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, \. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37. PMLR, 2015, pp. 1180–1189. [Online]. Available: https://proceedings.mlr.press/v37/ganin15.html

[24] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[25] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2502–2511.

[26] K. Mei, C. Zhu, J. Zou, and S. Zhang, "Instance adaptive self-training for unsupervised domain adaptation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 415–430.

[27] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, "Confidence regularized self-training," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5981–5990.

[28] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1368–1377.

[29] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1378–1388.

[30] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sensing*, p. 1369, Jun 2019. [Online]. Available: http://dx.doi.org/10.3390/rs11111369

[31] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "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*, p. 20–33, May 2021. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2021.02.009

[32] X. Chen, S. Pan, and Y. Chong, "Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator," *IEEE Transactions on Geoscience and Remote Sensing*, p. 1–13, Jan 2022. [Online]. Available: http://dx.doi.org/10.1109/tgrs.2022.3200246

[33] Y. Zhao, P. Guo, Z. Sun, X. Chen, and H. Gao, "Residualgan: Resize-residual dualgan for cross-domain remote sensing images semantic segmentation," *Remote Sensing*, vol. 15, no. 5, p. 1428, 2023.

[34] J. Zhang, J. Liu, B. Pan, Z. Chen, X. Xu, and Z. Shi, "An open set domain adaptation algorithm via exploring transferability and discriminability for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[35] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia *et al.*, "Spectralgpt: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[37] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 2274–2282, 2012. [Online]. Available: http://dx.doi.org/10.1109/tpami.2012.120

[38] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018.

[40] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1923–1938, 2018.

[41] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 414–12 424.

[42] L. Bai, S. Du, X. Zhang, H. Wang, B. Liu, and S. Ouyang, "Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive learning and adversarial learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[43] B. Zhang, T. Chen, and B. Wang, "Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[44] L. Wang, P. Xiao, X. Zhang, and X. Chen, "A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 4109–4121, 2023.

[45] H. Ni, Q. Liu, H. Guan, H. Tang, and J. Chanussot, "Category-level assignment for cross-domain semantic segmentation in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[46] Q. Zhao, S. Lyu, H. Zhao, B. Liu, L. Chen, and G. Cheng, "Self-training guided disentangled adaptation for cross-domain remote sensing image semantic segmentation," *International Journal of Applied Earth Observation and Geoinformation*, vol. 127, p. 103646, 2024.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[48] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.