Self-Training Guided Disentangled Adaptation for Cross-Domain Remote Sensing Image Semantic Segmentation

Qi Zhao[a], Shuchang Lyu[a], Hongbo Zhao[a,*], Binghao Liu[a], Lijiang Chen[a] and Guangliang Cheng[b,**]

[a] *Beihang University, Xueyuan Road No.37, Haidian district, Beijing, 100191, , China*
[b] *University of Liverpool, Foundation Building, Brownlow Hill,, Liverpool, L693BX, , UK*

ARTICLE INFO

Keywords:
Remote sensing image semantic segmentation
Unsupervised domain adaptation
Self-training
Domain disentangling
Adversarial learning

ABSTRACT

Remote sensing (RS) image semantic segmentation using deep convolutional neural networks (DCNNs) has shown great success in various applications. However, the high dependence on annotated data makes it challenging for DCNNs to adapt to different RS scenes. To address this challenge, we propose a cross-domain RS image semantic segmentation task that considers ground sampling distance, remote sensing sensor variation, and different geographical landscapes as the main factors causing domain shifts between source and target images. To mitigate the negative impact of domain shift, we propose a self-training guided disentangled adaptation network (ST-DASegNet) that consists of source and target student backbones to extract source-style and target-style features. To align cross-domain single-style features, we adopt feature-level adversarial learning. We also propose a domain disentangled module (DDM) to extract universal and distinct features from single-domain cross-style features. Finally, we fuse these features and generate predictions using source and target student decoders. Moreover, we employ an exponential moving average (EMA) based cross-domain separated self-training mechanism to ease the instability and disadvantageous effect during adversarial optimization. Our experiments on several prominent RS datasets (Potsdam, Vaihingen, and LoveDA) demonstrate that ST-DASegNet outperforms previous methods and achieves new state-of-the-art results. Visualization and analysis also confirm the interpretability of ST-DASegNet. The code is publicly available at `https://github.com/cv516Buaa/ST-DASegNet`.

1. Introduction

Remote sensing (RS) technology has been broadly applied in various real-world vision tasks, such as remote sensing scene classification Xia et al. (2017); Cheng et al. (2017); Zhao et al. (2021); Wang et al. (2022), object detection Xia et al. (2018); Ding et al. (2019); Lin et al. (2022), and semantic segmentation Lyu et al. (2020); Hou et al. (2022b); Zhao et al. (2022a); Mao et al. (2023). Among these applications, remote sensing scene image semantic segmentation has garnered significant research interest, aiming to predict accurate categories for each pixel of RS scene images. Deep Convolutional Neural Networks (DCNNs) Shelhamer et al. (2017); Zhao et al. (2017); Chen et al. (2018); Xie et al. (2021); Cai et al. (2022) have recently boosted the performance of RS image semantic segmentation tasks. However, their effectiveness heavily relies on large amounts of annotated training samples with similar data distribution to the testing samples. In practical applications, the scene of testing (target) images drastically differ from the scene of training (source) images. To address the domain shift and bridge the domain gap between the source and target images, cross-domain RS image semantic segmentation has become a hot research topic.

Cross-domain semantic segmentation task mainly involves training a model with source images and using it to generate pixel-wise predictions on target images, which may have significant domain shifts due to differences between the source and target domains. While many notable methods Tsai et al. (2018a); Hoffman et al. (2018); Zhu et al. (2017); Zhou et al. (2022a); Hoyer et al. (2022) have been proposed to address this problem on natural scene images, the domain shift problem for RS images is primarily caused by differences in ground sampling distance, remote sensing sensor variation, and geographical landscapes, as illustrated in Fig. 1. Specifically, different ground sampling distances can result in severe scale variation, remote sensing sensor variation can directly amplify the discrepancy of certain categories between source and target images (e.g., "Tree" on R-G-B and IR-R-G images exhibit different colors), and different geographical landscapes can cause elements of the same category to display different characteristics (e.g., "rural building" may have different patterns than "urban building").

To improve representation generalization and address the domain shift in cross-domain RS image semantic segmentation, we propose four principles as theoretical instruction for our method. First, RS images from different domains have large visual differences, but human beings can still easily distinguish between geographical elements, suggesting that images from different domains contain both unique and invariant features. Second, since RS images cover large geographical areas, image-level alignment using generation techniques Zhu et al. (2017); Isola et al. (2017a); Benjdira et al. (2019); Li et al. (2021b); Zhao et al. (2023) (e.g., image translation) may be challenging, while

[*] Primary Corresponding author
[**] Secondary Corresponding author

`zhaoqi@buaa.edu.cn` (Q. Zhao); `lyushuchang@buaa.edu.cn` (S. Lyu); `bhzhb@buaa.edu.cn` (H. Zhao); `liubinghao@buaa.edu.cn` (B. Liu); `chenlijiang@buaa.edu.cn` (L. Chen); `Guangliang.Cheng@liverpool.ac.uk` (G. Cheng)

ORCID(s): `0000-0001-9769-7083` (S. Lyu)

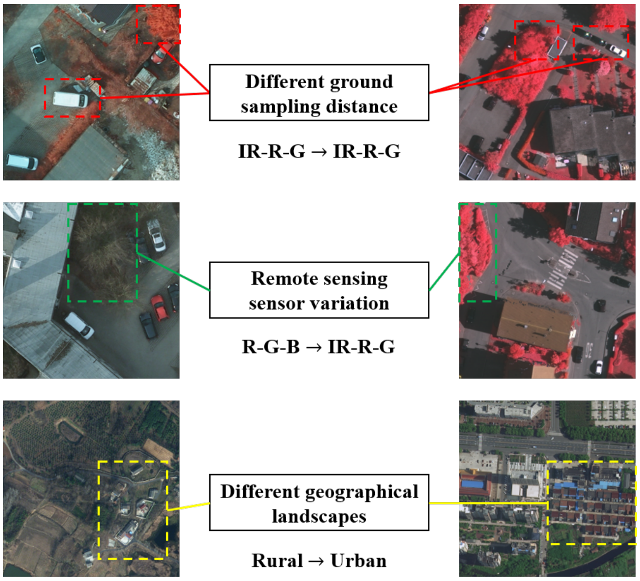

Different ground sampling distance
IR-R-G → IR-R-G
Remote sensing sensor variation
R-G-B → IR-R-G
Different geographical landscapes
Rural → Urban

Figure 1: Intuitive cases to explain the main problems in cross-domain remote sensing image semantic segmentation.

feature-level alignment may be more effective. Thirdly, RS images contain complex geographical information, making the direct adversarial learning process suffer from instability. Finally, while adversarial learning Tsai et al. (2018a); Bai et al. (2022); Ma et al. (2023); Wang et al. (2023a) can fill the domain gap, the lack of target annotations can lead to a representation tendency on source images. Based on these principles, we propose a novel Self-Training Guided Disentangled Adaptation Network (ST-DASegNet) to tackle the cross-domain RS image semantic segmentation task.

Following the first two principles, we design a Disentangled Adaptation Network (DASegNet). Specifically, we first design two student backbones to represent source and target images in two different styles. The source student backbone is designed to represent source and target images as “source-domain source-style feature” and “target-domain source-style feature”, respectively. On the other hand, the target student backbone is designed to represent source and target images as “source-domain target-style feature” and “target-domain target-style feature”, respectively. With this design, we expect to obtain a target-sensitive backbone that is more suitable for target images. Then, towards cross-domain single-style features (source and target output features from the same student backbone), we adopt adversarial learning for feature-level alignment. This can explore the potential of student backbones to adapt source annotations to the target domain. Next, towards single-domain cross-style features (source or target output features from different student backbones), we propose Domain Disentangled Module (DDM) to extract the universal feature and purify distinct features by fusion and disentangling operations. Finally, we design the source student decoder and target student decoder to respectively map source-style and target-style features into final predictions. During optimization with only source-labeled samples, the target student backbone and decoder tend to be more sensitive to target-style features. It means the target student can find the instinct target-style features in both source and target images. Intuitively, the target image contains more instinct target-style features than the source image, so the target student can represent the target image better. Similarly, source student is more suitable for source images.

Following the last two principles, we propose an EMA-based cross-domain separated self-training paradigm to further address the representation tendency and optimize the performance of our model. In this paper, we propose two self-training paradigms: “Decoder-only” and “Single-target”. In the “Decoder-only” paradigm, we use only two teacher decoders (source teacher decoder and target teacher decoder) instead of the entire teacher network. This efficient paradigm saves a significant amount of training memory and computation cost since the decoder of a segmentation network is much smaller than the backbone. Additionally, we propose the "Single-target" paradigm, which utilizes the target teacher backbone and target teacher decoder for pseudo label generation. This paradigm provides another perspective on integrating self-training into DASegNet. The EMA-based cross-domain separated self-training paradigm can be seamlessly integrated into DASegNet to form the unified ST-DASegNet.

To validate the effectiveness of ST-DASegNet, we conducted extensive experiments on several predominant benchmarks, including ISPRS (Potsdam/Vaihingen) Gerke (2014) and LoveDA Wang et al. (2021a). Comparison experiments demonstrate that ST-DASegNet outperforms previous SOTA methods and achieves new SOTA performance. Visualization and analysis further illustrate the interpretability of ST-DASegNet.

In summary, the main contributions are listed as follows:

- We propose a self-training guided disentangled adaptation network (ST-DASegNet) to address the cross-domain RS image semantic segmentation task. Extensive experiments on several predominant datasets (Potsdam, Vaihingen, and LoveDA) demonstrate that ST-DASegNet outperforms previous state-of-the-art methods.

- We propose a domain disentangled module (DDM) in ST-DASegNet that extracts cross-domain universal features and purifies single-domain distinct features. This provides insight into the use of feature disentangling to bridge the domain gap and improve the adaptability of deep learning models to different RS scenes.

- We adopt feature-level adversarial learning in ST-DASegNet to align cross-domain single-style features, which enhances the feature consistency of features from cross-domain images.

- We propose an efficient EMA-based cross-domain separated self-training paradigm in ST-DASegNet and integrate it with adversarial learning. This design rectifies the representation tendency and stabilizes optimization during adversarial training, leading to improved segmentation performance.

2. Related Work

2.1. Semantic Segmentation on RS Images

Semantic segmentation is a classical computer vision task, which plays an important role in many real-world applications. Fully convolutional networks (FCN) Shelhamer et al. (2017) first proposes an end-to-end deep learning architecture for semantic segmentation. From then, many notable generalized segmentation networks Badrinarayanan et al. (2017); Chen et al. (2018); Zhao et al. (2017); Fu et al. (2019); Huang et al. (2019); Wu et al. (2021); Yu et al. (2021); Xie et al. (2021) are directly applied and fitted in well on RS images.

Compared to natural scene images, remote sensing images contain more specific detailed information like element boundary and corner, irregular shape, etc. Moreover, large geographical coverage and confusing geographical elements on RS images cause large intra-class variance and inter-class similarity. On these issues, many novel and strong RS segmentation networks are proposed. Li et al. (2019) propose a novel adaptive multi-scale deep fusion (AMDF) ResNet to fuse multiple hierarchy features in an adaptive manner. With abundant and discriminative feature extraction, AMDF-ResNet achieves high-level performance. Nogueira et al. (2019) propose an efficient dilated network, which improves the network by exploring the multi-context features. Besides exploiting the effective information on multiple features, some works focus on integrating attention mechanisms into RS segmentation networks. Mou et al. (2020); Li et al. (2021a); Zhao et al. (2022b) utilize spatial-wise and channel-wise attention modules, which can overcome misunderstanding by capturing long-range spatial relationships and finding important channels. Liu et al. (2018); Yang et al. (2020); Li et al. (2022a); Hou et al. (2022a) all pay attention to boundary parsing. These edge-sensitive architectures offer an inspiring perspective on enhancing the understanding of complex geographical elements in pixel-wise.

2.2. Unsupervised Cross-Domain Adaptation for Semantic Segmentation

With domain shift alleviating mechanism, unsupervised domain adaptation (UDA) techniques can make source-trained (trained only with source samples) models adapt to target samples. Recently, many remarkable works have made huge progress in applying UDA to cross-domain semantic segmentation tasks.

Adversarial learning is frequently adopted in many excellent methods. Some methods utilize image generalization techniques like image translation Zhu et al. (2017); Isola et al. (2017a) to align image appearance between source and target images. Chen et al. (2019); Yang and Soatto (2020); Guo et al. (2021); Chen et al. (2020); Zou et al. (2020) employ image-level adaption in the first step and then train the segmentation networks with cross-domain synthetic data. Other methods explore the domain-invariant features between source-style and target-style features, which takes advantage of the feature-level alignment strategy to solve the domain shift problem. Tsai et al. (2018b); Du et al. (2019); Wang et al. (2020a); Zeng et al. (2020) insert discriminators into networks for consistency alignment on intermediate feature maps or output entropy maps.

As another typical non-adversarial UDA paradigm, self-training has attracted much attention in cross-domain semantic segmentation tasks. Zou et al. (2019); Zhou et al. (2022b,a); Hoyer et al. (2022); Chen et al. (2022a) promote the adaption ability by generating reliable, consistent, and class-balanced pseudo labels. Supervised by target pseudo ground-truth, models can quickly adapt to target images.

2.3. Cross-Domain Semantic Segmentation on RS Images

Although the cross-domain RS image semantic segmentation task has not been fully studied and the cross-domain adaptive potential has not been fully exploited, there are still many awesome works proposed in recent years. Benjdira et al. (2019) first address the domain adaptation issue for the RS image semantic segmentation task. They use generative adversarial networks (GANs) based architecture to tackle this task and achieve convincing results. Following their pioneer work, Li et al. (2021b) introduce DualGAN Yi et al. (2017) and demonstrate its strong adaptation ability on the RS image semantic segmentation task. Similarly, Chen et al. (2022c,b); Zhao et al. (2023) all design GANs-based networks and explore the adaptive potential of GANs on cross-domain RS image semantic segmentation tasks. Bai et al. (2022); Ma et al. (2023); Wang et al. (2023a) propose a novel network integrating adversarial learning and contrastive learning. Combined with adversarial loss and pixel-wise contrastive loss, the segmentation network can learn rich domain-invariant features. Chen et al. (2022d) propose DNT to jointly align the distribution in image-level and feature-level, which makes effect on reducing the domain shift. Wu et al. (2022) also focus on extracting domain-invariant features. Instructed by this principle, they propose a deep covariance alignment (DCA) module, which achieves competitive results on a popular benchmark, LoveDA Wang et al. (2021a). Li et al. (2022b) propose a step-wise RS segmentation network with covariate shift alleviation to close the gap between source and target domains. Zhang et al. (2022b) propose DA-MSCDNet to align feature-level inconsistency between optical and SAR images. Zhang et al. (2022a) propose a local-to-global RS segmentation framework that follows a curriculum-style approach. They design a two-stage cross-domain adaptation: “source domain to Easy-to-adapt” and “Easy-to-adapt to Hard-to-adapt”. All above-mentioned brilliant works advance the cross-domain RS image semantic segmentation task.

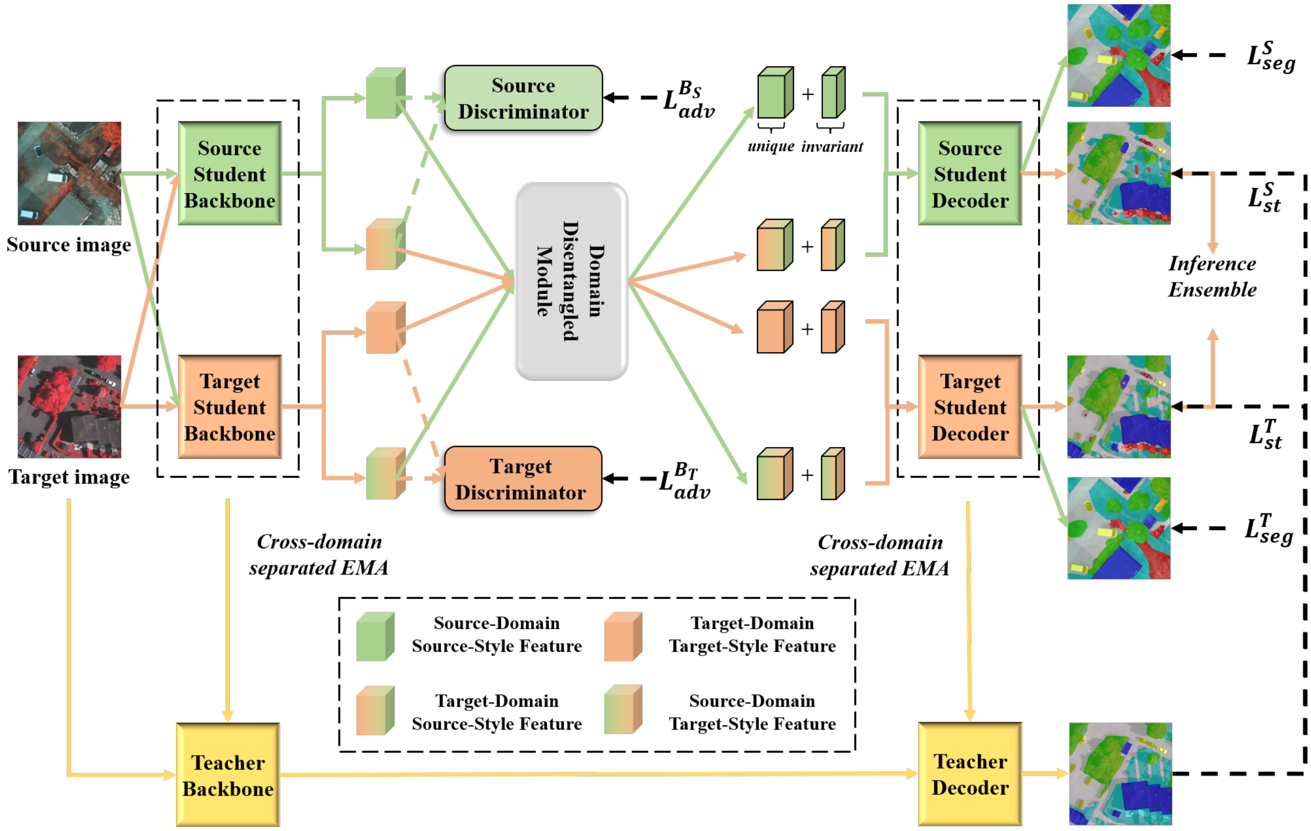


Figure 2: The overview of ST-DASegNet. In this architecture, source, and target student backbones are designed for feature extraction on two-style features. Domain disentangled module is designed to extract the cross-domain universal features and single-domain unique features. Source and target student decoders are designed to respectively map source and target features into predictions. Teacher backbone and decoder are respectively updated by student backbones and student decoders with the EMA technique. D_S and D_T are two discriminators to align cross-domain single-style features.

3. Proposed Method

To decrease the negative influence of domain shift between source and target RS images, we propose ST-DASegNet. Fig. 2 shows the overview of ST-DASegNet. In this paper, we follow the unsupervised domain adaptation constraint, which means annotations of source samples are given while no annotation of target samples is available.

3.1. Source and Target Student Backbones

From our proposed first principle in Sec.1, we believe that unique features lead to large visual discrepancy while invariant features lead to easy identification. Naturally, if the model can precisely extract two-style features from two-domain images, it will benefit the unique and invariant feature representation. Motivated by this reasonable assumption, we design two student backbones rather than common-applied weight-sharing backbone. Here, the source student backbone (B_S) is expected to extract source-style features from two-domain images. Similarly, the target student backbone (B_T) is expected to extract target-style features from two-domain images.

As shown in Fig. 2, source and target images ($\boldsymbol{x}_s$ and $\boldsymbol{x}_t$) will be respectively fed into the source and target student backbones. Eq. 1 and Eq. 2 show this process.

$$\boldsymbol{F}_{S-s} = B_S(\boldsymbol{x}_s), \quad \boldsymbol{F}_{S-t} = B_S(\boldsymbol{x}_t) \tag{1}$$

$$\boldsymbol{F}_{T-s} = B_T(\boldsymbol{x}_s), \quad \boldsymbol{F}_{T-t} = B_T(\boldsymbol{x}_t) \tag{2}$$

where $\boldsymbol{F}_{S-s}$ and $\boldsymbol{F}_{S-t}$ are two output features from source student backbone, which respectively denote source-domain source-style feature and target-domain source-style feature. $\boldsymbol{F}_{T-s}$ and $\boldsymbol{F}_{T-t}$ are two output features from target student backbone, which respectively denote source-domain target-style feature and target-domain target-style feature.

3.2. Adversarial Learning on Cross-Domain Single-Style Features

From our proposed second principle in Sec.1, we believe that feature-level alignment may be more suitable than image-level alignment on cross-domain RS image semantic segmentation task. Feature alignment can guarantee the basic adaptation ability of student backbones and enhance the representation consistency between cross-domain images. In this paper, we design two discriminators to apply adversarial learning on cross-domain single-style features. Specifically, source discriminator (D_S) is proposed to

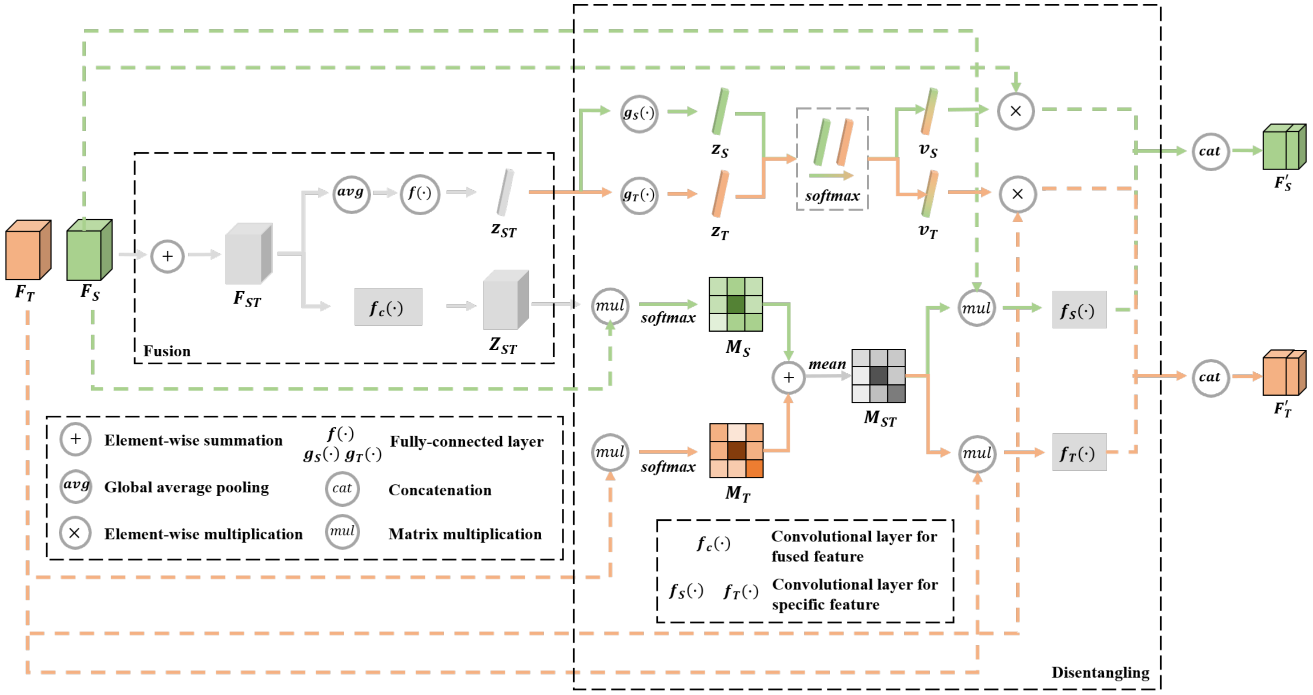

F_T
F_S
+
F_{ST}
avg
$f(\cdot)$
z_{ST}
$f_c(\cdot)$
Z_{ST}
Fusion
$g_S(\cdot)$
$g_T(\cdot)$
z_S
z_T
softmax
v_S
v_T
×
mul
softmax
M_S
mean
M_{ST}
mul
softmax
M_T
$f_S(\cdot)$
$f_T(\cdot)$
cat
F'_S
F'_T
+ Element-wise summation
avg Global average pooling
× Element-wise multiplication
$f(\cdot)$ $g_S(\cdot)$ $g_T(\cdot)$ Fully-connected layer
cat Concatenation
mul Matrix multiplication
$f_c(\cdot)$ Convolutional layer for fused feature
$f_S(\cdot)$ $f_T(\cdot)$ Convolutional layer for specific feature
Disentangling

Figure 3: The module structure of domain disentangled module. In this module, $\boldsymbol{F_S}$ and $\boldsymbol{F_T}$ respectively denote source-style and target-style features. Particularly, two features are extracted from the single-domain image (e.g., $\boldsymbol{F_{S-s}}$ and $\boldsymbol{F_{T-s}}$). This module consists of fusion and disentangling blocks. A fusion block is used to generate a fused prototype and feature map. In the disentangling block, a fused prototype is used to generate unique features while a fused feature map is used to generate invariant features. Finally, these two features are concatenated together.

align source-domain source-style feature and target-domain source-style feature. target discriminator (D_T) is proposed to align source-domain target-style feature and target-domain target-style feature. The mathematical operations are shown in Eq. 3 and Eq. 4.

$$\begin{aligned}\mathcal{L}_{adv}^{B_S}(B_S, D_S) = &\ \mathbb{E}_{x^s\sim X^s}[log(D_S(\boldsymbol{F_{S-s}}))] \\ &+ \mathbb{E}_{x^t\sim X^t}[log(1 - D_S(\boldsymbol{F_{S-t}}))]\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{adv}^{B_T}(B_T, D_T) = &\ \mathbb{E}_{x^t\sim X^t}[log(D_T(\boldsymbol{F_{T-t}}))] \\ &+ \mathbb{E}_{x^s\sim X^s}[log(1 - D_T(\boldsymbol{F_{T-s}}))]\end{aligned} \tag{4}$$

where $\mathcal{L}_{adv}^{B_S}(B_S, D_S)$ and $\mathcal{L}_{adv}^{B_T}(B_T, D_T)$ are adversarial losses. To optimize B_S and B_T. We adopt "min-max" criterion, which can be expressed in Eq. 5.

$$\begin{cases} min_{B_S} max_{D_S} \mathcal{L}_{adv}^{B_S}(B_S, D_S) \\ \\ min_{B_T} max_{D_T} \mathcal{L}_{adv}^{B_T}(B_T, D_T) \end{cases} \tag{5}$$

3.3. Domain Disentangled Module on Single-Domain Cross-Style Features

From source and target student backbones, we can obtain two-style features for source and target images. With feature-level adversarial learning on these features, the domain shift problem is alleviated. Naturally, we prepare to extract universal and unique features from single-domain cross-style features. In this paper, we propose a domain disentangled module (DDM). The module structure is shown in Fig. 3.

In DDM, two input features are extracted by two student backbones from the single domain image, so two features have different styles. Here, we propose a fusion block to generate a fused prototype ($\boldsymbol{z_{ST}}$) and feature map ($\boldsymbol{Z_{ST}}$). The forward process of the fusion block is shown in Eq. 6.

$$\boldsymbol{z_{ST}} = f(avg(\boldsymbol{F_{ST}})), \quad \boldsymbol{Z_{ST}} = f_c(\boldsymbol{F_{ST}}) \tag{6}$$

As shown in Fig. 3, $f(\cdot)$ denotes a fully-connected layer (dimensionality-reduction). $f_c(\cdot)$ denotes a convolutional layer. $avg(\cdot)$ denotes channel-wise global average pooling operation. $\boldsymbol{F_{ST}} = \boldsymbol{F_S} + \boldsymbol{F_T}$.

We then propose a disentangling block to decouple unique and invariant features. This block can be separated into two parts. The first part is designed to extract unique features. Even though source and target student backbones can represent single-domain images as source and target styles, the output features are still mixtures partially containing other style features (E.g., source-domain source-style feature contains some target-style information). Therefore, unique feature extraction can be regarded as purifying the style-specific feature from the fused feature. As shown in Fig.3, $\boldsymbol{z_{ST}}$ is served as a guided prototype to generate two complementary vectors ($\boldsymbol{v_S}$ and $\boldsymbol{v_T}$). The generation process is shown in Eq. 7 and Eq. 8.

$$\boldsymbol{v_S} = \frac{exp(\boldsymbol{z_S})}{exp(\boldsymbol{z_S}) + exp(\boldsymbol{z_T})}, \ \boldsymbol{v_T} = \frac{exp(\boldsymbol{z_T})}{exp(\boldsymbol{z_S}) + exp(\boldsymbol{z_T})} \quad (7)$$

$$\boldsymbol{z_S} = g_S(\boldsymbol{z_{ST}}), \quad \boldsymbol{z_T} = g_T(\boldsymbol{z_{ST}}) \quad (8)$$

where $g_S(\cdot)$ and $g_T(\cdot)$ are two style-specific fully-connected layers (dimensionality-increase). $\boldsymbol{v_S} + \boldsymbol{v_T} = \boldsymbol{1}$. Eq. 7 shows $softmax$ operation.

With $\boldsymbol{v_S}$ and $\boldsymbol{v_T}$, we conduct channel-wise multiplication on input features ($\boldsymbol{F_S}$ and $\boldsymbol{F_T}$). Here, $\boldsymbol{v_S} = [v_S^1, v_S^2, \cdots, v_S^C]$. $\boldsymbol{v_T} = [v_T^1, v_T^2, \cdots, v_T^C]$. $\boldsymbol{F_S} = [\boldsymbol{F_S^1}, \boldsymbol{F_S^2}, \cdots, \boldsymbol{F_S^C}]$. $\boldsymbol{F_T} = [\boldsymbol{F_T^1}, \boldsymbol{F_T^2}, \cdots, \boldsymbol{F_T^C}]$. Eq. 9 shows the channel-wise multiplication operation.

$$\boldsymbol{F_{U-S}^i} = \boldsymbol{F_S^i} \times v_S^i, \quad \boldsymbol{F_{U-T}^i} = \boldsymbol{F_T^i} \times v_T^i \quad (9)$$

where v_S^i and v_T^i are scalars of $\boldsymbol{v_S}$ and $\boldsymbol{v_T}$ ($\mathbb{R}^C$). $\boldsymbol{F_S^i}$ and $\boldsymbol{F_T^i}$ ($\mathbb{R}^{H \times W}$) are channels of $\boldsymbol{F_S}$ and $\boldsymbol{F_T}$ ($\mathbb{R}^{C \times H \times W}$). $\boldsymbol{F_{U-S}}$ and $\boldsymbol{F_{U-T}}$ are respectively source-style and target-style unique features.

The second part of DDM is designed to extract invariant features. The key idea of this part is to find the invariant channel-wise relation mask between input features ($\boldsymbol{F_S}$ and $\boldsymbol{F_T}$) and fused feature map ($\boldsymbol{Z_{ST}} = [\boldsymbol{Z_{ST}^1}, \boldsymbol{Z_{ST}^2}, \cdots, \boldsymbol{Z_{ST}^C}]$). Here, we reshape each element ($\boldsymbol{F_S^i}$ and $\boldsymbol{F_T^i}$) of $\boldsymbol{F_S}$ and $\boldsymbol{F_T}$ to $\mathbb{R}^N$, where $N = H \times W$. Similarly, we reshape $\boldsymbol{Z_{ST}^i}$ to $\mathbb{R}^N$. After reshaping, we apply dot product with $softmax$ operation, which is shown in Eq. 10 ∼ Eq. 12.

$$M_S^{j,i} = \frac{exp(\boldsymbol{F_S^i} \cdot \boldsymbol{Z_{ST}^j})}{\sum_{i=1}^{C}(\boldsymbol{F_S^i} \cdot \boldsymbol{Z_{ST}^j})}, \ M_T^{j,i} = \frac{exp(\boldsymbol{F_T^i} \cdot \boldsymbol{Z_{ST}^j})}{\sum_{i=1}^{C}(\boldsymbol{F_T^i} \cdot \boldsymbol{Z_{ST}^j})} \quad (10)$$

$$\boldsymbol{M_S} = \begin{bmatrix} M_S^{1,1} & \dots & M_S^{1,C} \\ \vdots & \ddots & \vdots \\ M_S^{C,1} & \dots & M_S^{C,C} \end{bmatrix} \quad (11)$$

$$\boldsymbol{M_T} = \begin{bmatrix} M_T^{1,1} & \dots & M_T^{1,C} \\ \vdots & \ddots & \vdots \\ M_T^{C,1} & \dots & M_T^{C,C} \end{bmatrix} \quad (12)$$

where $\boldsymbol{M_S}$ ($\mathbb{R}^{C \times C}$) represents the channel-wise relation mask between $\boldsymbol{F_S}$ and $\boldsymbol{Z_{ST}}$ while $\boldsymbol{M_T}$ ($\mathbb{R}^{C \times C}$) represents the channel-wise relation mask between $\boldsymbol{F_T}$ and $\boldsymbol{Z_{ST}}$. The invariant relation mask is denoted as $\boldsymbol{M_{ST}} = (\boldsymbol{M_S} + \boldsymbol{M_T})/2$. In $\boldsymbol{M_{ST}}$, if the value ($\boldsymbol{M_{ST}^{j,i}}$) is larger, it means that the i^{th} channel of $\boldsymbol{F_S}$ and $\boldsymbol{F_T}$ probably both have high impact on the j^{th} channel of $\boldsymbol{M_{ST}}$.

Since $\boldsymbol{M_{ST}}$ reflects the universal relation, we perform a matrix multiplication between $\boldsymbol{M_{ST}}$ and $\boldsymbol{F_S}, \boldsymbol{F_T}$ to generate features with channel-wise attention. Eq. 13 shows this process.

$$\boldsymbol{F_{I-S}^j} = \sum_{i=1}^{C}(\boldsymbol{F_S^i} \times M_{ST}^{j,i}), \quad \boldsymbol{F_{I-T}^j} = \sum_{i=1}^{C}(\boldsymbol{F_T^i} \times M_{ST}^{j,i}) \quad (13)$$

where $\boldsymbol{F_{I-S}} = [\boldsymbol{F_{I-S}^1}, \boldsymbol{F_{I-S}^2}, \cdots, \boldsymbol{F_{I-S}^C}]$ and $\boldsymbol{F_{I-T}} = [\boldsymbol{F_{I-T}^1}, \boldsymbol{F_{I-T}^2}, \cdots, \boldsymbol{F_{I-T}^C}]$ are respectively source-style and target-style invariant features.

Finally, we fuse the unique and invariant feature through concatenation. The outputs of DDM are $\boldsymbol{F_S'}$ and $\boldsymbol{F_T'}$. The pipeline of DDM is shown in Alg.1.

Algorithm 1 Domain disentangled module

```
Input: Source-style feature map $\boldsymbol{F_S}$ and target-style feature map $\boldsymbol{F_T}$ extracted from single-domain image. $\boldsymbol{F_S}, \boldsymbol{F_T} \in \mathbb{R}^{C \times H \times W}$. Trainable parameters in fully-connected layers ($f(\cdot)$, $g_S(\cdot)$, $g_T(\cdot)$) and convolutional layers ($f_c(\cdot)$, $f_S(\cdot)$, $f_T(\cdot)$)
Output: New source-style feature map $\boldsymbol{F_S'}$ and target-style feature map $\boldsymbol{F_T'}$ containing unique and invariant information. $\boldsymbol{F_S'}, \boldsymbol{F_T'} \in \mathbb{R}^{C \times H \times W}$.
  Feature Fusion:
  Get fused prototype $\boldsymbol{z_{ST}}$ using Eq. 6.
  Get fused feature map $\boldsymbol{Z_{ST}}$ using Eq. 6.
  Feature Disentangling:
    Unique feature disentangling:
    Get complementary channel-wise weighted vectors $\boldsymbol{v_S}$ and $\boldsymbol{v_T}$ using Eq. 7 and Eq. 8.
    Get source-style and target-style unique features $\boldsymbol{F_{U-S}}$ and $\boldsymbol{F_{U-T}}$ using Eq. 9.
    Invariant feature disentangling:
    Get channel-wise relation masks $\boldsymbol{M_S}$ and $\boldsymbol{M_T}$ using Eq. 10 ∼ Eq. 12.
    Get source-style and target-style invariant features $\boldsymbol{F_{I-S}}$ and $\boldsymbol{F_{I-T}}$ using Eq. 13.
    Feature concatenation:
    Get source-style new features $\boldsymbol{F_S'} = cat(\boldsymbol{F_{U-S}}, \boldsymbol{F_{I-S}})$.
    Get target-style new features $\boldsymbol{F_T'} = cat(\boldsymbol{F_{U-T}}, \boldsymbol{F_{I-T}})$.
```

3.4. EMA-based Cross-Domain Separated Self-Training Mechanism

From our proposed third and fourth principles in Sec.1, we believe that directly applying an adversarial learning mechanism can ease the domain shift, but it will still cause instability and representation tendency on source images because of no target annotations. Motivated by further improving the representation capability on target images, we propose an EMA-based cross-domain separated self-training mechanism. As shown in Fig. 4, the traditional EMA-based self-training paradigm generates a teacher network with EMA updating operation on the student network. Compared to traditional paradigm, our self-training paradigm will not generate a whole teacher network, but only generate some components. In this paper, we propose two self-training paradigms, which are “Decoder-only” and “Single-target”.

3.4.1. Decoder-only

As shown in Fig. 4, “Decoder-only” paradigm introduces two teacher decoders rather than two teacher networks. In

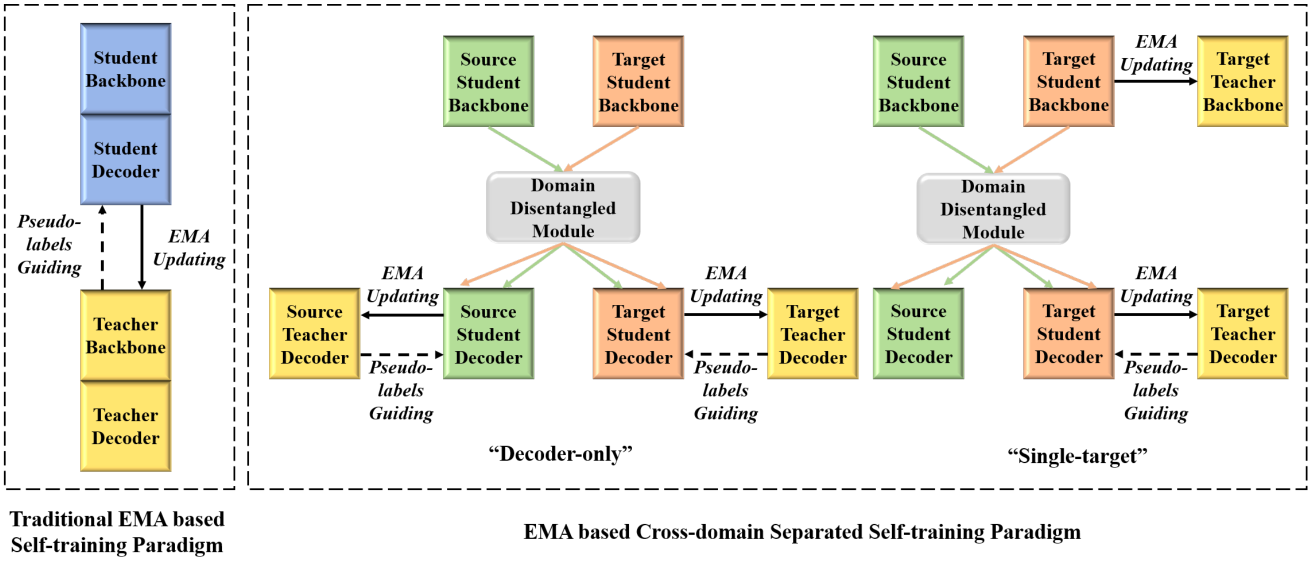


Figure 4: Comparison between traditional EMA-based self-training paradigm and our proposed EMA-based cross-domain separated self-training paradigm.

this paradigm, student networks get credible and mature knowledge from teacher decoders. Specifically, we first construct two teacher decoders, which are the source teacher decoder (H_S^{te}) and the target teacher decoder (H_T^{te}). Then, at training step t, the teacher decoders' weights are updated by student decoders' weights by EMA operation, which can be formulated in Eq. 14.

$$\begin{cases} \phi_t^{H_S^{te}} = \alpha\phi_{t-1}^{H_S^{te}} + (1-\alpha)\theta_t^{H_S} \\ \\ \phi_t^{H_T^{te}} = \alpha\phi_{t-1}^{H_T^{te}} + (1-\alpha)\theta_t^{H_T} \end{cases} \tag{14}$$

where α denotes the EMA decay factors controlling the updating rate. $\phi_t^{H_S^{te}}, \phi_t^{H_T^{te}}$ refer to weights of two teacher decoders at t^{th} step. $\theta_t^{H_S}, \theta_t^{H_T}$ refer to weights of two student decoders (H_S, H_T) at t^{th} step.

After updating the weights of teacher decoders, we transfer the reliable knowledge to student by pseudo-label guidance. With two teacher decoders, the generation of pseudo-label is formulated in Eq. 15.

$$\hat{\boldsymbol{P}}_t^{te} = \arg\max_c \frac{(H_S^{te}(\boldsymbol{F}'_{S-t}) + H_T^{te}(\boldsymbol{F}'_{T-t}))}{2} \tag{15}$$

where $\boldsymbol{F}'_{S-t} \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{F}'_{T-t} \in \mathbb{R}^{C\times H\times W}$ are respectively target-domain source-style feature and target-domain target-style feature after DDM (Alg.1). $\hat{\boldsymbol{P}}_t^{te} \in \mathbb{R}^{H\times W}$ is pseudo-label (index map) after channel-wise *argmax* operation. Here, we apply a soft-voting ensemble strategy to integrate the predictions after two teacher decoders, which makes the pseudo-label more credible.

"Decoder-only" paradigm has two main advantages. (1) It can efficiently transfer reliable knowledge to both source and target students (including backbone and decoder) for target-style feature extraction, which maximally alleviates the representation tendency on source images. (2) It can efficiently save a lot of training memory and computation cost because the decoder of a segmentation network is always much smaller than the backbone.

3.4.2. Single-target

In this paper, we propose "Single-target" paradigm, which is another self-training paradigm on our proposed DASegNet. This paradigm mainly aims to enhance the target-style feature representation of the target student backbone and decoder. As shown in Fig. 4, we construct a whole target teacher network including target teacher backbone (B_T^{te}) and target teacher decoder (H_T^{te}). Similar to Eq. 14, we apply EMA updating on the target teacher network (Eq. 16).

$$\begin{cases} \phi_t^{B_T^{te}} = \alpha\phi_{t-1}^{B_T^{te}} + (1-\alpha)\theta_t^{B_T} \\ \\ \phi_t^{H_T^{te}} = \alpha\phi_{t-1}^{H_T^{te}} + (1-\alpha)\theta_t^{H_T} \end{cases} \tag{16}$$

where $\phi_t^{B_T^{te}}$ refers to weights of target teacher backbone at t^{th} step. $\theta_t^{B_T}$ refers to weights of target student backbone at t^{th} step.

With a whole teacher network, the pseudo-label generation of "Single-target" paradigm can be expressed in Eq. 17.

$$\hat{\boldsymbol{P}}_t^{te} = \arg\max_c H_T^{te}(\boldsymbol{F}'_{T-t}) \tag{17}$$

Compared to "Decoder-only" paradigm, "Single-target" paradigm only employs pseudo-label guiding on target students. Theoretically, "Single-target" paradigm can enhance the target-style feature extraction power on the target student, which leads to a larger representation gap between the source and student networks. In practical experiments, we find that "Decoder-only" paradigm can also enhance the target-style feature extraction power on target students with less memory and training computation cost. Basically, both these two paradigms improve the target-domain adaptation ability of DASegNet by credible pseudo-label guiding.

In summary, we show our proposed cross-domain separated self-training mechanism in Alg.2. In this paper, we mainly adopt "Decoder-only" paradigm. "Single-target" paradigm provides another perspective on integrating the self-training mechanism into DASegNet. In experiments, we will make further comparisons between these two paradigms.

3.5. Optimization with Combined Loss

In this paper, we apply segmentation loss, adversarial loss, and self-training loss to jointly optimize the ST-DASegNet.

As for segmentation loss ($\mathcal{L}_{seg}^{S}$ and $\mathcal{L}_{seg}^{T}$) shown in Fig. 2, we apply conventional cross-entropy loss using annotations of source images. We formulate the segmentation loss function in Eq. 18 and Eq. 19.

$$\mathcal{L}_{seg}^{S} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}(\boldsymbol{y_s}(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})log(H_S(\boldsymbol{F}'_{\boldsymbol{S-s}}))) \tag{18}$$

$$\mathcal{L}_{seg}^{T} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}(\boldsymbol{y_s}(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})log(H_T(\boldsymbol{F}'_{\boldsymbol{T-s}}))) \tag{19}$$

where, $\boldsymbol{y_s}$ is the ground truth of source images. $\boldsymbol{F}'_{\boldsymbol{S-s}}$ and $\boldsymbol{F}'_{\boldsymbol{T-s}}$ are respectively source-domain source-style and source-domain target-style disentangled features after DDM.

As for self-training loss ($\mathcal{L}_{st}^{S}$ and $\mathcal{L}_{st}^{T}$) shown in Fig. 2, we also adopt cross-entropy loss using pseudo-labels of target images. We formulate self-training loss function in Eq. 20 and Eq. 21.

$$\mathcal{L}_{st}^{S} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}(\hat{\boldsymbol{P}}_{\boldsymbol{t}}^{\boldsymbol{te}}(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})log(H_S(\boldsymbol{F}'_{\boldsymbol{S-t}}))) \tag{20}$$

$$\mathcal{L}_{st}^{T} = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}(\hat{\boldsymbol{P}}_{\boldsymbol{t}}^{\boldsymbol{te}}(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})log(H_T(\boldsymbol{F}'_{\boldsymbol{T-t}}))) \tag{21}$$

where, $\hat{\boldsymbol{P}}_{\boldsymbol{t}}^{\boldsymbol{te}}$ denotes the pseudo-labels of target images, which are transformed from index maps ($\mathbb{R}^{H\times W}$) to one-hot labels ($\mathbb{R}^{C\times H\times W}$). The calculation operation of $\hat{\boldsymbol{P}}_{\boldsymbol{t}}^{\boldsymbol{te}}$ is shown in Eq. 15 or Eq. 17 respectively for "Decoder-only" and "Single-target" paradigms.

To optimize ST-DASegNet, we combine three types of losses together. The final combined loss function ($\mathcal{L}$) is formulated in Eq. 22.

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda\mathcal{L}_{st} + \beta\mathcal{L}_{adv} \tag{22}$$

where $\mathcal{L}_{seg} = \mathcal{L}_{seg}^{S} + \mathcal{L}_{seg}^{T}$ (Eq. 18 and Eq. 19). $\mathcal{L}_{st} = \mathcal{L}_{st}^{S} + \mathcal{L}_{st}^{T}$ (Eq. 20 and Eq. 21). $\mathcal{L}_{adv} = \mathcal{L}_{adv}^{B_S} + \mathcal{L}_{adv}^{B_T}$ (Eq. 3 and Eq. 4). λ and β are two factors to adjust the proportion of self-training loss and adversarial loss. In this paper, we set $\lambda = 0.25$ and $\beta = 0.005$.

Algorithm 2 Cross-domain separated self-training mechanism

```
Input: Target images $\boldsymbol{x_t}$. Source student backbone $B_S$, target student backbone $B_T$, source student decoder $H_S$ and target student decoder $H_T$.
Output: Trainable parameters of teacher network components at $t^{th}$ step. Pseudo-label $\hat{\boldsymbol{P}}_{\boldsymbol{t}}^{\boldsymbol{te}}$.
  if Self-training paradigm is "Decoder-only" then
    EMA Updataing:
    Update source teacher decoder $H_S^{te}$ and target teacher decoder $H_T^{te}$ at $t^{th}$ step using Eq. 14.
    Pseudo-label Generation:
    Extract features ($\boldsymbol{F_{S-t}}$ , $\boldsymbol{F_{T-t}}$) on $\boldsymbol{x_t}$ respectively from $B_S$ and $B_T$ using Eq. 1 and Eq.2
    Get disentangle features ($\boldsymbol{F}'_{\boldsymbol{S-t}}$ , $\boldsymbol{F}'_{\boldsymbol{T-t}}$) from DDM with $\boldsymbol{F_{S-t}}$ and $\boldsymbol{F_{T-t}}$ as input using Alg.1.
    Get pseudo-label $\hat{\boldsymbol{P}}_{\boldsymbol{t}}^{\boldsymbol{te}}$ with $\boldsymbol{F}'_{\boldsymbol{S-t}}$ and $\boldsymbol{F}'_{\boldsymbol{T-t}}$ as input using Eq. 15.
  end if
  if Self-training paradigm is "Single-target" then
    EMA Updataing:
    Update target teacher backbone $B_T^{te}$ and target teacher decoder $H_T^{te}$ at $t^{th}$ step using Eq. 16.
    Pseudo-label Generation:
    Extract features ($\boldsymbol{F_{S-t}}$ , $\boldsymbol{F}_{\boldsymbol{T-t}}^{\boldsymbol{te}}$) on $\boldsymbol{x_t}$ respectively from $B_S$ and $B_T^{te}$ using Eq. 1 and Eq. 2.
    Get disentangle features ($\boldsymbol{F}'_{\boldsymbol{S-t}}$ , $\boldsymbol{F}'_{\boldsymbol{T-t}}$) from DDM with $\boldsymbol{F_{S-t}}$ and $\boldsymbol{F}_{\boldsymbol{T-t}}^{\boldsymbol{te}}$ as input using Alg.1.
    Get pseudo-label $\hat{\boldsymbol{P}}_{\boldsymbol{t}}^{\boldsymbol{te}}$ with $\boldsymbol{F}'_{\boldsymbol{T-t}}$ as input using Eq. 17.
  end if
```

4. Experiments and Analysis

4.1. Datasets and Evaluation Metric

To evaluate our method on cross-domain RS image semantic segmentation task, we use three benchmark datasets, which are Potsdam, Vaihingen, and LoveDA.

Potsdam and Vaihingen. These two datasets belong to ISPRS 2D semantic segmentation benchmark dataset Gerke (2014). The Potsdam dataset contains 38 VHR TOP (very-high-resolution True Orthophotos) with the size of 6000 × 6000 (fixed size). The Potsdam dataset has three different imaging modes, which are IR-R-G, R-G-B, and R-G-B-IR. The first two modes are 3-channel while the last mode is 4-channel. In this paper, we choose to use the first two modes. The Vaihingen dataset contains 33 VHR TOP with the size of 2000 × 2000 (approximate size). The Vaihingen dataset only has one imaging mode, which is IR-R-G. For more efficient computation cost, we crop the images of these two datasets into smaller patches with the size of 512 × 512. Specifically, we respectively select 512 and 256 as cropping strides for Potsdam and Vaihingen, generating 4598 and 1696 patches. Moreover, we split the Potsdam and Vaihingen datasets into training and testing sets. For Potsdam, the training and testing set contains 2904 and 1694 images, respectively. For Vaihingen, the training and testing set contain 1296 and 440 images, respectively. It is worth noting that, all the data preprocessing methods followed previous works Li et al. (2021b); Bai et al. (2022); Chen et al. (2022c); Zhang et al. (2022a); Zhao et al. (2023).

On Potsdam/Vaihingen datasets, we design four cross-domain RS semantic segmentation tasks, which are listed as follows.

- Potsdam IR-R-G to Vaihingen IR-R-G (Potsdam IR-R-G → Vaihingen IR-R-G).
- Vaihingen IR-R-G to Potsdam IR-R-G (Vaihingen IR-R-G → Potsdam IR-R-G).
- Potsdam R-G-B to Vaihingen IR-R-G (Potsdam R-G-B → Vaihingen IR-R-G).
- Vaihingen IR-R-G to Potsdam R-G-B (Vaihingen IR-R-G → Potsdam R-G-B).

LoveDA. This dataset is recently proposed to advance both RS semantic segmentation and domain adaptation tasks. It consists of 5987 high spatial resolution (1024 × 1024) RS images from three cities including Nanjing, Changzhou, and Wuhan. LoveDA dataset contains images from two domains (urban and rural), which focuses on challenging the model's generalized representation capacity on different geographical elements of urban and rural scenes. This dataset has 1833 urban images, which are split into 1156 training images and 677 validation images. For rural images, there are 2358 images in total, where 1366 images are used for training and the rest 992 images are used for validation. In addition, LoveDA also contains 1796 testing images (976 for rural and 820 for urban), which can be evaluated on online server[1]. On LoveDA dataset, we design two cross-domain RS semantic segmentation tasks, which are urban-to-rural (urban → rural) and rural-to-urban (rural → urban) tasks. All our experiments setting is followed Wang et al. (2021a); Wu et al. (2022).

Evaluation Metric. In this paper, we select common-used IoU (intersection of union) and $F1$-score as evaluation metrics. Specifically, for a specific class i, IoU is formulated as $IoU_i = tp_i/(tp_i + fp_i + fn_i)$, where tp_i, fp_i, fn_i denote true positive, false positive and false negative, respectively. The mIoU is the mean value of all categories' IoU_i. Additionally, $F1$-score is defined as $F1\text{-score} = (2\times Precision\times Recall)/(Precision + Recall)$.

4.2. Implementation Details

Followed previous works, we select DeeplabV3 Chen et al. (2018) as baseline segmentation network. We also select a recent outstanding segmentation network, SegFormer Xie et al. (2021) as another baseline. If applying DeeplabV3, the student backbone will be ResNet-50 He et al. (2016) and the student decoder will be ASPP block (a combination of ASPP module and several convolutional blocks). To optimize DeeplabV3 based ST-DASegNet, we use SGD (Stochastic Gradient Descent) as an optimizer, where the initial learning rate is 0.001, the momentum value is 0.9 and the weight decay value is 0.0001. If applying SegFormer, the student backbone will be mit-b5 Xie et al. (2021) and the student decoder will be "All MLP" module. To optimize SegFormer based ST-DASegNet, we use Adam as optimizer and the initial learning rate is set as 0.0001.

For DeeplabV3 and SegFormer based ST-DASegNet, we both apply Adam as an optimizer in which the initial learning rate is 0.00025. The structure of our proposed discriminators (D_S and D_T) is followed PatchGAN Isola et al. (2017b). Specifically, the discriminator consists of 4 convolutional blocks with kernels as size of 4 × 4. The stride of the first two and the last two blocks is respectively set as 2 and 1. The output channels of each block are 64, 128, 256, and 1.

All experiments are implemented on mmsegmentation[2] semantic segmentation framework and all models are trained on two NVIDIA RTX 3090. Our code is available at `https://github.com/cv516Buaa/ST-DASegNet`.

4.3. Experimental Results

4.3.1. *Cross-domain RS image semantic segmentation on Potsdam and Vaihingen*

As mentioned above, we design four cross-domain tasks between Potsdam and Vaihingen. On these four tasks, we conduct abundant experiments to show the effectiveness of our proposed ST-DASegNet. Previous methods always select DeeplabV3 as the baseline model and hardly apply the transformer as the baseline model. Therefore, we reimplement DAFormer Hoyer et al. (2022) to fairly compare with our SegFormer based ST-DASegNet. It is worth noting that DAFormer Hoyer et al. (2022) is the first method applying transformer (SegFormer) on cross-domain natural scene image semantic segmentation task. Similar to our method, DAFormer also uses mmsegmentation to implement their method, so we can easily apply DAFormer on cross-domain RS image semantic segmentation tasks.

Comparison experiments from Potsdam IR-R-G to Vaihingen IR-R-G. In this task, Postsdam IR-R-G and Vaihingen IR-R-G images are respectively served as source-domain and target-domain. The 2904 annotated training images from Potsdam and 1296 no-annotation training images from Vaihingen are used to train the model. The 440 Vaihingen testing images are used for evaluation. The comparison results are shown in Tab.1. Compared to DeeplabV3 based methods, ST-DASegNet (DeeplabV3) surpasses the current SOTA method Zhao et al. (2023). Compared to SegFormer based method (DAFormer Hoyer et al. (2022)), ST-DASegNet (SegFormer) achieves a 2.03% improvement on $mIoU$ value and 3.29% improvement on mF-score.

[1] https://github.com/Junjue-Wang/LoveDA

[2] https://github.com/open-mmlab/mmsegmentation

Table 1
Cross-domain RS image semantic segmentation comparison results (%) from Potsdam IR-R-G to Vaihingen IR-R-G. Methods with "*" are our reimplemented version.

Methods	Clutter		Impervious surfaces		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
DeeplabV3 Chen et al. (2018) (Baseline)	2.33	4.56	43.90	60.78	24.26	39.07	52.25	64.19	25.76	40.87	60.35	71.81	34.81	46.88
SegFormer Xie et al. (2021) (Baseline)	4.22	9.47	61.03	76.07	31.13	47.89	66.31	78.87	44.47	60.38	75.50	87.95	46.11	60.11
AdaptSegNet Tsai et al. (2018a)	4.60	8.76	54.39	70.39	6.40	11.99	52.65	68.96	28.98	44.91	63.14	77.40	35.02	47.05
FSDAN Ji et al. (2021)	10.00	-	57.40	-	37.00	-	58.40	-	41.70	-	57.80	-	43.70	-
ProDA Zhang et al. (2021)	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
DualGAN Li et al. (2021b)	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 *et al.* Bai et al. (2022)	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 *et al.* Zhang et al. (2022a)	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 *et al.* Wang et al. (2023b)	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
DNT Chen et al. (2022d)	14.77	25.74	69.74	82.18	53.88	70.03	59.19	74.37	47.51	64.42	80.04	88.91	54.19	67.61
CIA-UDA Ni et al. (2023)	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 Zhao et al. (2023)	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
DAFormer Hoyer et al. (2022)*	48.26	60.17	74.09	84.12	38.96	56.41	**70.88**	**81.36**	**57.53**	**71.48**	84.07	90.75	62.30	74.05
ST-DASegNet (DeeplabV3)	21.17	32.64	70.88	82.20	51.81	67.63	68.01	80.10	41.97	57.97	82.57	89.24	56.07	68.30
ST-DASegNet (SegFormer)	**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**

Table 2
Cross-domain RS image semantic segmentation comparison results (%) from Vaihingen IR-R-G to Potsdam IR-R-G. Methods with "*" are our reimplemented version.

Methods	Clutter		Impervious surfaces		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
DeeplabV3 Chen et al. (2018) (Baseline)	5.28	11.58	55.57	71.07	50.02	66.97	14.0	27.69	43.67	60.70	60.62	75.02	38.19	52.17
SegFormer Xie et al. (2021) (Baseline)	1.08	2.65	60.63	76.47	58.99	73.14	30.07	46.24	51.91	68.82	74.85	87.18	46.29	59.08
AdaptSegNet Tsai et al. (2018a)	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 Zhang et al. (2021)	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
DualGAN Li et al. (2021b)	11.48	20.56	51.01	67.53	48.49	65.31	34.98	51.82	36.50	53.48	53.37	69.59	39.30	54.71
DNT Chen et al. (2022d)	11.51	20.65	61.91	76.48	49.50	66.22	35.46	52.36	37.61	54.67	66.41	79.81	43.74	58.36
Zhang *et al.* Zhang et al. (2022a)	**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 *et al.* Wang et al. (2023b)	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 Ni et al. (2023)	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
DAFormer Hoyer et al. (2022)*	2.56	5.02	68.42	79.07	65.20	79.31	**70.65**	**82.13**	56.39	72.48	78.94	87.64	57.03	67.61
ST-DASegNet (DeeplabV3)	5.21	10.21	74.19	85.26	**76.76**	**86.90**	43.33	60.44	51.56	68.62	82.15	90.28	55.53	66.95
ST-DASegNet (SegFormer)	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**

From Tab.1, we also find that even though the baseline model performs strong, ST-DASegNet (SegFormer) can still gain 20.22% improvement on $mIoU$ value and 17.23% improvement on mF-score. Particularly, on "Clutter" category, ST-DASegNet (SegFormer) shows outstanding performance over previous methods. On this single category, ST-DASegNet surpasses the second best method (DAFormer) by 18.77% on IoU value and 20.11% on $F1$-score.

Comparison experiments from Vaihingen IR-R-G to Potsdam IR-R-G. In this task, Vaihingen IR-R-G and Potsdam IR-R-G images are respectively served as source-domain and target-domain. The 1296 annotated training images from Vaihingen and 2904 no-annotation training images from Potsdam are used to train the model. The 1694 Potsdam testing images are used for evaluation. As shown in Tab.2, DeeplabV3 based ST-DASegNet performs much stronger than previous methods. Compared to previous SOTA method Ni et al. (2023), it achieves 3.30% improvement on $mIoU$ value and 1.15% on mF-score. SegFormer based ST-DASegNet also shows obvious superiority over DAFormer and achieves the SOTA performance.

Comparison experiments from Potsdam R-G-B to Vaihingen IR-R-G. In this task, Postsdam R-G-B and Vaihingen IR-R-G images are respectively served as source-domain and target-domain. The 2904 annotated training images from Potsdam and 1296 no-annotation training images from Vaihingen are used to train the model. The 440 Vaihingen testing images are used for evaluation. As shown in Fig. 1, "Potsdam R-G-B to Vaihingen IR-R-G" adaptation is more complex than "Potsdam IR-R-G to Vaihingen IR-R-G" adaptation. Besides suffering from different ground sampling distances, this task also suffers from remote sensing sensor variation. The comparison results on this task are shown in Tab.3. Among DeeplabV3 based methods, DNT Chen et al. (2022d) shows best performance, which may partially benefit from larger encoder (ResNet-101). Our DeeplabV3 based ST-DASegNet achieves comparable results with Ni et al. (2023); Wang et al. (2023a). Obviously, SegFormer based ST-DASegNet outperforms DAFormer Hoyer et al. (2022) and achieves SOTA results. In addition, Similar to the results on "Potsdam IR-R-G to Vaihingen IR-R-G" task (Tab.1), ST-DASegNet also outperforms the second best method by a large margin in "Clutter" category.

Comparison experiments from Vaihingen IR-R-G to Potsdam R-G-B. In this task, Vaihingen IR-R-G and Potsdam R-G-B images are respectively served as source-domain and target-domain. The 1296 annotated training images from Vaihingen and 2904 no-annotation training images from Potsdam are used to train the model. The 1694 Potsdam

Table 3
Cross-domain RS image semantic segmentation comparison results (%) from Potsdam R-G-B to Vaihingen IR-R-G. Methods with "*" are our reimplemented version.

Methods	Clutter		Impervious surfaces		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
DeeplabV3 Chen et al. (2018) (Baseline)	0.58	1.16	40.42	57.57	12.52	22.25	30.88	47.19	12.12	21.62	54.23	70.33	25.12	36.68
SegFormer Xie et al. (2021) (Baseline)	1.43	2.81	51.34	67.85	37.97	55.04	52.62	68.96	5.18	9.85	73.18	84.51	36.95	48.17
AdaptSegNet Tsai et al. (2018a)	2.99	5.81	51.26	67.77	10.25	18.54	51.51	68.02	12.75	22.61	60.72	75.55	31.58	43.05
ProDA Zhang et al. (2021)	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
DualGAN Li et al. (2021b)	3.94	13.88	49.16	61.33	40.31	57.88	55.82	70.66	27.85	42.17	65.44	83.00	39.93	54.82
Bai *et al.* Bai et al. (2022)	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
Zhang *et al.* Zhang et al. (2022a)	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
ResiDualGAN Zhao et al. (2023)	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
Wang *et al.* Wang et al. (2023b)	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 Ni et al. (2023)	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
DNT Chen et al. (2022d)	11.55	20.71	67.94	80.91	**52.64**	**68.97**	58.43	73.76	**43.63**	**61.05**	81.09	89.56	52.60	65.83
DAFormer Hoyer et al. (2022)*	22.57	33.72	67.44	79.65	45.60	60.13	**66.27**	**80.41**	40.49	54.93	81.34	90.07	53.95	66.49
ST-DASegNet (DeeplabV3)	20.53	33.74	62.60	76.39	47.32	64.30	61.71	74.89	29.72	44.43	75.58	86.13	49.58	63.31
ST-DASegNet (SegFormer)	**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**

Table 4
Cross-domain RS image semantic segmentation comparison results (%) from Vaihingen IR-R-G to Potsdam R-G-B. Methods with "*" are our reimplemented version.

Methods	Clutter		Impervious surfaces		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
DeeplabV3 Chen et al. (2018) (Baseline)	4.61	8.82	46.02	63.03	59.71	74.77	1.63	3.53	7.1	13.25	37.34	54.37	26.07	36.30
SegFormer Xie et al. (2021) (Baseline)	2.36	4.61	62.45	76.89	72.16	83.83	5.38	10.21	31.52	48.65	72.61	84.13	41.08	51.39
AdaptSegNet Tsai et al. (2018a)	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 Zhang et al. (2021)	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
DualGAN Li et al. (2021b)	**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
DNT Chen et al. (2022d)	8.43	15.55	56.41	72.13	46.78	63.74	36.56	53.55	30.59	46.85	69.95	82.32	41.45	55.69
Zhang *et al.* Zhang et al. (2022a)	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 *et al.* Wang et al. (2023b)	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 Ni et al. (2023)	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 Hoyer et al. (2022)*	1.07	1.88	65.12	78.16	70.40	84.28	**61.25**	**76.59**	49.02	65.51	82.44	89.70	54.88	66.02
ST-DASegNet (DeeplabV3)	2.66	4.29	65.48	79.27	75.15	85.86	34.46	47.95	45.59	63.13	78.06	87.66	50.23	61.36
ST-DASegNet (SegFormer)	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**

testing images are used for evaluation. From Tab.4, we also obtain encouraging results. On this task, DeeplabV3 based ST-DASegNet shows surprising improvement. Compared to baseline results, it gains 24.16% on $mIoU$ value and 25.06% and mF-score. Compared to the recent best DeeplabV3 based method Ni et al. (2023), it leads by 2.68% on $mIoU$ value. Moreover, SegFormer based ST-DASegNet achieves the best performance on 4 categories (Impervious surfaces, Cars, Low vegetation, and Building) and achieves the current SOTA results on $mIoU$ value and mF-score.

4.3.2. Cross-domain RS image semantic segmentation on LoveDA

Besides cross-domain adaptation between Potsdam and Vaihingen datasets, we further conduct comparison experiments on LoveDA dataset. Tab.5 and Tab.6 show the rural-to-urban and urban-to-rural results, respectively. From the results, we analyze the following aspects. (1) Compared to previous methods, our method provides insight into integrating self-training and adversarial training to tackle cross-domain remote sensing image semantic segmentation tasks. (2) Compared to baseline models, ST-DASegNet makes huge progress. On rural-to-urban adaptation task, ST-DASegNet (DeepLabV3) and ST-DASegNet (SegFormer) respectively surpass the baseline models by 8.94% and 9.43% on $mIoU$ value. On urban-to-rural adaptation task, ST-DASegNet (DeepLabV3) and ST-DASegNet (SegFormer) respectively surpass the baseline models by 11.90% and 10.03% on $mIoU$ value. (3) Compared to previous published SOTA methods, DCA Wu et al. (2022), our proposed DeepLabV3 based ST-DASegNet achieves comparable results. It has 1.98% inferiority on rural-to-urban adaptation task and 0.52% superiority on urban-to-rural adaptation task. Our proposed SegFormer based ST-DASegNet outperforms DCA by a large margin on both these two tasks. (4) For a fair comparison with ST-DASegNet (SegFormer), we reimplement DAFormer. It is clear that ST-DASegNet (SegFormer) achieves better results. (5) These two tasks belong to online competitions. The track is "LoveDA Unsupervised Domain Adaptation Challenge". Among all the submitted results on rural-to-urban and urban-to-rural adaptation tasks, our method respectively ranks 6^{th} and 3^{rd} place. Among published methods, our method encouragingly achieves new SOTA results.

4.4. Ablation Study

To separately show the performance of each component, we conduct ablation experiments on "Potsdam IR-R-G to Vaihingen IR-R-G" and "Potsdam R-G-B to Vaihingen IR-R-G" adaptation tasks shown in Tab.7 and Tab.8.

Table 5
Cross-domain RS image semantic segmentation comparison results (%) from Rural to Urban of LoveDA test dataset. Methods with "*" are our reimplemented version. "AT" indicates adversarial training and "ST" indicates self-training.

Methods	Type	IoU							$mIoU$
		Background	Building	Road	Water	Barren	Forest	Agriculture	
DeeplabV3 Chen et al. (2018) (Baseline)	-	32.81	43.41	27.59	79.23	14.84	29.24	20.97	35.44
SegFormer Xie et al. (2021) (Baseline)	-	44.98	43.93	27.46	85.82	16.24	37.02	30.48	40.85
DDC Tzeng et al. (2014)	-	43.60	15.37	11.98	79.07	14.13	33.08	23.47	31.53
AdaptSegNet Tsai et al. (2018a)	AT	42.35	23.73	15.61	81.95	13.62	28.70	22.05	32.68
FADA Wang et al. (2020b)	AT	43.89	12.62	12.76	80.37	12.70	32.76	24.79	31.41
CLAN Luo et al. (2019)	AT	43.41	25.42	13.75	79.25	13.71	30.44	25.80	33.11
TransNorm Wang et al. (2021b)	AT	38.37	5.04	3.75	80.83	14.19	33.99	17.91	27.73
PyCDA Lian et al. (2019)	ST	38.04	35.86	45.51	74.87	7.71	40.39	11.39	36.25
CBST Zou et al. (2018)	ST	48.37	46.10	35.79	80.05	19.18	29.69	30.05	41.32
IAST Mei et al. (2020)	ST	48.57	31.51	28.73	86.01	**20.29**	31.77	36.50	40.48
DCA Wu et al. (2022)	ST	45.82	49.60	51.65	80.88	16.70	42.93	**36.92**	46.36
DAFormer Hoyer et al. (2022)*	ST	50.94	**56.66**	**62.83**	**89.41**	11.99	45.81	25.26	48.99
ST-DASegNet (DeeplabV3)	AT+ST	49.68	51.62	52.41	74.76	10.69	35.67	35.79	44.38
ST-DASegNet (SegFormer)	AT+ST	**51.01**	54.23	60.52	87.31	15.18	**47.43**	36.26	**50.28**

Table 6
Cross-domain RS image semantic segmentation comparison results (%) from Urban to Rural of LoveDA test dataset. Methods with "*" are our reimplemented version. "AT" indicates adversarial training and "ST" indicates self-training.

Methods	Type	IoU							$mIoU$
		Background	Building	Road	Water	Barren	Forest	Agriculture	
DeeplabV3 Chen et al. (2018) (Baseline)	-	26.64	48.14	23.37	43.97	11.35	32.47	50.61	33.79
SegFormer Xie et al. (2021) (Baseline)	-	26.60	55.80	35.62	65.44	17.13	40.13	39.65	40.05
DDC Tzeng et al. (2014)	-	25.61	44.27	31.28	44.78	13.74	33.83	25.98	31.36
AdaptSegNet Tsai et al. (2018a)	AT	26.89	40.53	30.65	50.09	16.97	32.51	28.25	32.27
FADA Wang et al. (2020b)	AT	24.39	32.97	25.61	47.59	15.34	34.35	20.29	28.65
CLAN Luo et al. (2019)	AT	22.93	44.78	25.99	46.81	10.54	37.21	24.45	30.39
TransNorm Wang et al. (2021b)	AT	19.39	36.30	22.04	36.68	14.00	40.62	3.30	24.62
PyCDA Lian et al. (2019)	ST	12.36	38.11	20.45	57.16	18.32	36.71	41.90	32.14
CBST Zou et al. (2018)	ST	25.06	44.02	23.79	50.48	8.33	39.16	49.65	34.36
IAST Mei et al. (2020)	ST	29.97	49.48	28.29	64.49	2.13	33.36	61.37	38.44
DCA Wu et al. (2022)	ST	36.38	55.89	40.46	62.03	**22.01**	38.92	60.52	45.17
DAFormer Hoyer et al. (2022)*	ST	**37.39**	52.84	41.99	72.05	11.46	46.79	61.27	46.25
ST-DASegNet (DeeplabV3)	AT+ST	33.79	55.95	39.69	69.28	14.19	44.79	62.16	45.69
ST-DASegNet (SegFormer)	AT+ST	36.78	**59.83**	**43.77**	**73.83**	19.38	**49.96**	**67.01**	**50.08**

4.4.1. Effectiveness of DDM

Before evaluating the performance of DDM, we first construct a dual path baseline model. This baseline model contains two student backbones and two student decoders. To form it, we simply remove DDM, discriminators and teacher network from ST-DASegNet. Compared to baseline model, it simply adds one more student network. During inference, two target predictions from the source and target decoders will be integrated with the soft voting strategy. Tab.7 and Tab.8 show that the dual path baseline model has limited improvement compared to the baseline model. This little improvement may come from the model ensemble strategy in the inference phase.

Based on the dual path baseline model, we can easily insert DDM and evaluate its performance. As shown in Tab.7 and Tab.8, when adding DDM, new models make huge progress. It means that the feature fusion and disentangling mechanism enhances the capability of source and target student backbones to extract different style features on different domain images. With DDM, the source and target student backbones can respectively represent the target images in source and target style with only source annotations.

4.4.2. Effectiveness of feature-level adversarial learning

After inserting DDM into the model, we further apply feature-level adversarial learning. From Tab.7 and Tab.8, it is obvious that the performance of models is skyrocketing. DeepLabv3 based models improve by an average of 10.54% on $mIoU$ value and 10.98% on mF-score on "Potsdam IR-R-G to Vaihingen IR-R-G" and "Potsdam R-R-B to Vaihingen IR-R-G" adaptation tasks. SegFormer based models

Table 7
Ablation comparison experiments on "Potsdam IR-R-G to Vaihingen IR-R-G" adaptation task (%). "Adv" indicates adversarial learning. "ST" indicates self-training.

Methods	Clutter		Impervious surfaces		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
DeeplabV3 Chen et al. (2018) (Baseline)	2.33	4.56	43.90	60.78	24.26	39.07	52.25	64.19	25.76	40.87	60.35	71.81	34.81	46.88
Dual path baseline model	4.14	7.83	44.18	60.09	28.72	42.97	50.60	63.26	27.80	41.35	56.03	68.52	35.26	47.34
+ DDM	6.02	13.72	58.10	72.67	18.62	33.28	54.90	70.04	38.80	56.08	66.05	77.70	40.42	53.92
+ DDM + Adv	16.86	32.77	62.70	76.44	31.51	47.29	74.48	85.32	50.19	67.02	73.56	84.81	51.55	65.61
+ DDM + Adv + ST	21.17	32.64	70.88	82.20	51.81	67.63	68.01	80.10	41.97	57.97	82.57	89.24	56.07	68.30
with Single-target	25.91	38.16	68.57	81.47	49.11	64.17	67.93	80.91	42.08	58.50	80.47	88.71	55.68	68.65
SegFormer Xie et al. (2021) (Baseline)	4.22	9.47	61.03	76.07	31.13	47.89	66.31	78.87	44.47	60.38	75.50	87.95	46.11	60.11
Dual path baseline model	5.91	12.16	62.32	76.03	26.39	40.83	67.31	81.23	46.69	63.49	72.88	86.98	46.92	60.55
+ DDM	27.53	44.01	61.61	73.46	36.98	54.05	63.46	74.70	47.17	62.75	79.74	88.73	52.75	66.28
+ DDM + Adv	46.49	63.49	73.1	84.46	41.6	58.77	61.14	75.88	37.45	54.46	84.94	91.86	57.45	71.49
+ DDM + Adv + ST	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
with Single-target	75.24	85.84	73.28	84.57	49.35	66.10	66.54	79.90	50.20	66.79	86.72	92.89	66.89	79.35

Table 8
Ablation comparison experiments on "Potsdam R-G-B to Vaihingen IR-R-G" adaptation task (%). "Adv" indicates adversarial learning. "ST" indicates self-training.

Methods	Clutter		Impervious surfaces		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
DeeplabV3 Chen et al. (2018) (Baseline)	0.58	1.16	40.42	57.57	12.52	22.25	30.88	47.19	12.12	21.62	54.23	70.33	25.12	36.68
Dual path baseline model	0.32	0.91	43.27	58.99	18.36	34.58	25.42	43.30	14.75	24.83	51.04	66.16	25.53	38.13
+ DDM	2.80	5.44	50.72	67.30	18.42	31.11	54.34	70.42	21.34	35.18	50.35	66.98	32.99	46.07
+ DDM + Adv	11.92	23.76	53.95	68.00	42.20	59.33	63.37	77.57	20.99	31.88	65.18	77.50	42.94	56.34
+ DDM + Adv + ST	20.53	33.74	62.60	76.39	47.32	64.30	61.71	74.89	29.72	44.43	75.58	86.13	49.58	63.31
with Single-target	24.71	36.53	61.10	76.33	46.53	63.81	57.25	72.70	30.16	45.04	73.25	85.83	48.83	63.37
SegFormer Xie et al. (2021) (Baseline)	1.43	2.81	51.34	67.85	37.97	55.04	52.62	68.96	5.18	9.85	73.18	84.51	36.95	48.17
Dual path baseline model	2.92	5.17	53.84	69.20	37.15	55.28	53.91	69.48	5.64	9.33	71.89	83.10	37.56	48.59
+ DDM	13.42	27.26	55.22	71.13	37.33	54.37	53.71	69.86	10.65	19.20	78.86	88.18	41.53	55.00
+ DDM + Adv	21.14	35.27	65.62	79.25	40.53	57.16	60.82	72.91	34.39	51.70	78.35	87.86	50.14	64.03
+ DDM + Adv + ST	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
with Single-target	35.91	51.16	64.57	78.47	45.11	62.17	67.93	80.91	32.08	48.58	80.47	89.31	54.35	68.43

improve by 6.66% on $mIoU$ value and 7.12% on mF-score on the two tasks. These results demonstrate that feature-level adversarial learning can ease the domain shift problem by aligning single-style features from different domain images. Moreover, these results also prove that applying feature-level adversarial learning can enhance the power of the DDM.

4.4.3. *Effectiveness of EMA-based cross-domain separated self-training mechanism*

Based on DDM and adversarial learning, we add EMA-based cross-domain separated self-training mechanism into the model and form the final ST-DASegNet. As shown in Tab.7 and Tab. 8, with our proposed self-training mechanism, the performance gets further improved. From these results, it is clear that the models benefit from high-quality pseudo-labels, which suppress the representation tendency on source annotated images. The results also show that our proposed self-training mechanism can efficiently incorporate DDM and adversarial learning.

4.4.4. *"Decoder-only" Versus "Single-target"*

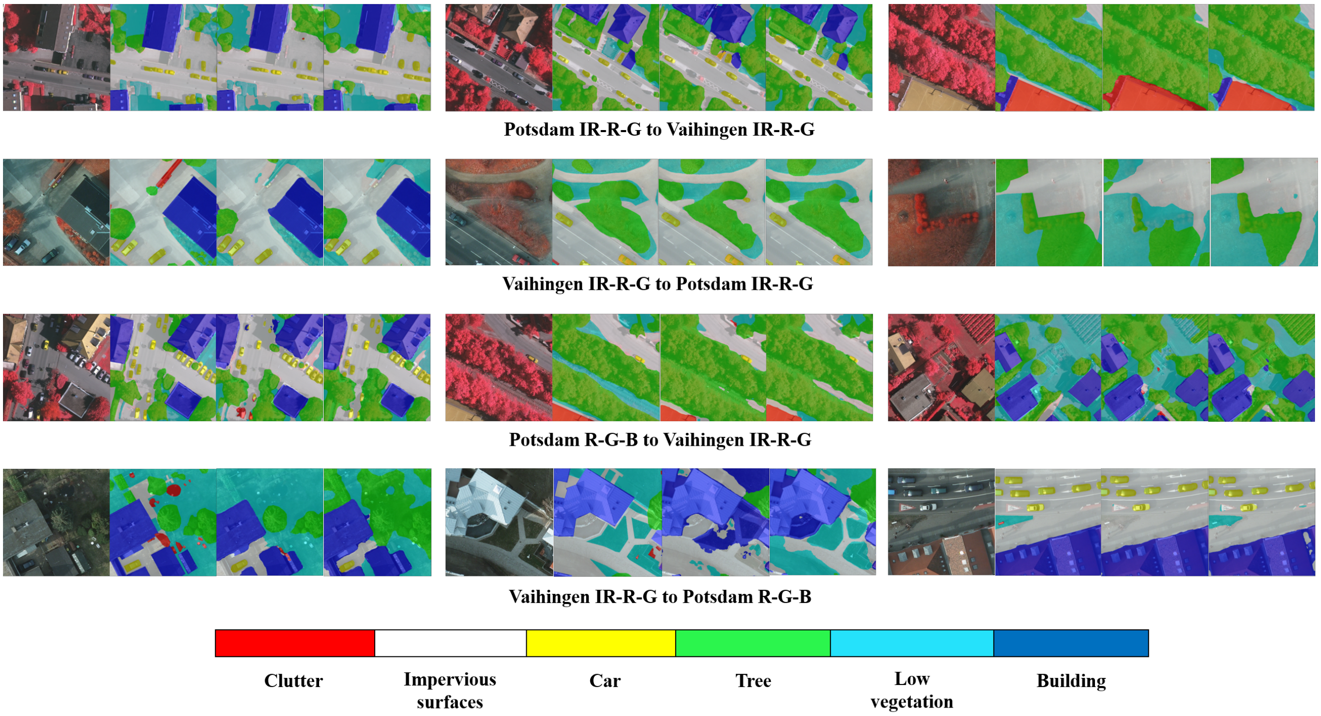
In this paper, we propose two self-training paradigms. To compare their performance, we conduct experiments in Tab.7 and Tab.8. On "Potsdam IR-R-G to Vaihingen IR-R-G" task, ST-DASegNet (SegFormer) with "Single-target" paradigm outperforms ST-DASegNet (SegFormer) with "Decoder-only" paradigm by 2.56% on $mIoU$ value and 2.01% on mF-score. ST-DASegNet (DeepLabV3) achieves comparable results with these two paradigms. From the results of "Potsdam R-G-B to Vaihingen IR-R-G" task, "Single-target" paradigm has little inferiority compared to "Decoder-only" paradigm. Generally, "Single-target" and "Decoder-only" has close performance. Due to less training computation cost, we mainly adopt "Decoder-only" in this paper. Despite that "Single-target" can not beat "Decoder-only", it still provides another inspiring paradigm for EMA-based cross-domain separated self-training mechanism.

As a whole, DDM is designed to improve the model by enhancing feature representation capacity. Feature-level adversarial learning is employed to ease the domain-shift problem. EMA-based cross-domain separated self-training mechanism makes an effect on balancing the representation tendency between source and target domain images. These three key components of ST-DASegNet show great compatibility, which improves the model from three different perspectives.

4.5. Visualization and Analysis

4.5.1. *Qualitative visualization of segmentation results*

From Tab.1 to Tab.6, we report the experimental results of our proposed ST-DASegNet on 6 cross-domain RS image semantic segmentation tasks. Here, we visualize the predictions using above-mentioned trained models to intuitively show the improvement of ST-DASegNet.

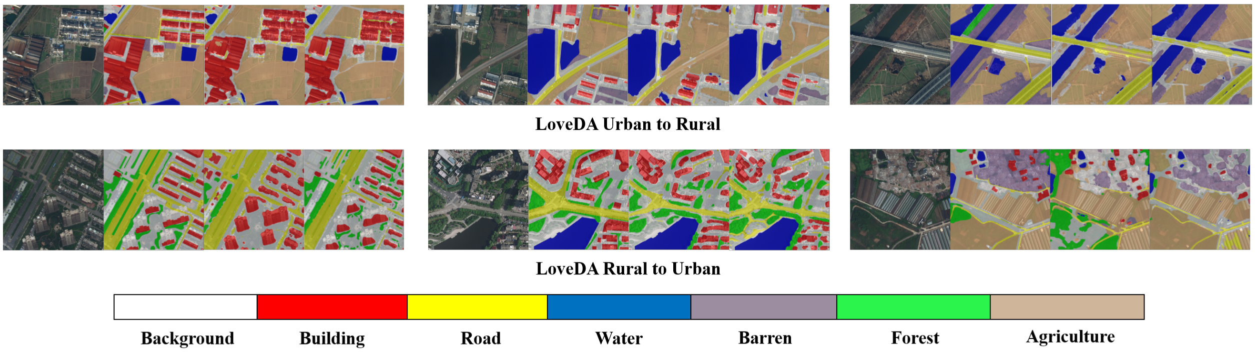

Potsdam IR-R-G to Vaihingen IR-R-G

Vaihingen IR-R-G to Potsdam IR-R-G

Potsdam R-G-B to Vaihingen IR-R-G

Vaihingen IR-R-G to Potsdam R-G-B

Figure 5: Qualitative visualization of results on Potsdam and Vaihingen datasets. For each task, we provide 3 cases containing 4 images. From left to right, the 4 images are respectively target images, ground truth, prediction of ST-DASegNet (DeeplabV3) and prediction of ST-DASegNet (SegFormer).


LoveDA Urban to Rural

LoveDA Rural to Urban

Figure 6: Qualitative visualization of results on LoveDA datasets. For each task, we provide 3 cases containing 4 images. From left to right, the 4 images are respectively target images, ground truth, prediction of ST-DASegNet (DeeplabV3), and prediction of ST-DASegNet (SegFormer).

As shown in Fig. 5, we can intuitively find that ST-DASegNet has strong performance on cross-domain RS image semantic segmentation tasks between Potsdam and Vaihingen datasets. Specifically, from the most right case on the "Potsdam IR-R-G to Vaihingen IR-R-G" task, we find ST-DASegNet has excellent perceptual ability on the "Clutter" category, which coincides with the results shown in Tab.1. From the middle and the most right cases on the "Vaihingen IR-R-G to Potsdam IR-R-G" task, ST-DASegNet shows a strong ability on distinguishing "Tree" and "Low vegetation". Similar results appear in Tab.2, where results of the "Tree" and "Low vegetation" categories improve by a large margin compared to baseline results. On the "Potsdam R-G-B to Vaihingen IR-R-G" task, the 3 visualization cases further confirm the results in Tab.3, where ST-DASegNet makes huge progress on segmenting "Clutter", "Impervious surfaces" and "Building" categories. On the "Potsdam R-G-B to Vaihingen IR-R-G" task, the most right case shows outstanding performance on recognizing "Car". As shown in Tab.4, both ST-DASegNet (DeeplabV3) and ST-DASegNet (SegFormer) surpass previous methods by a large margin on the results of the "Car" category.

As shown in Fig. 6, we provide the visualization results on LoveDA dataset. Since images in the testing dataset do

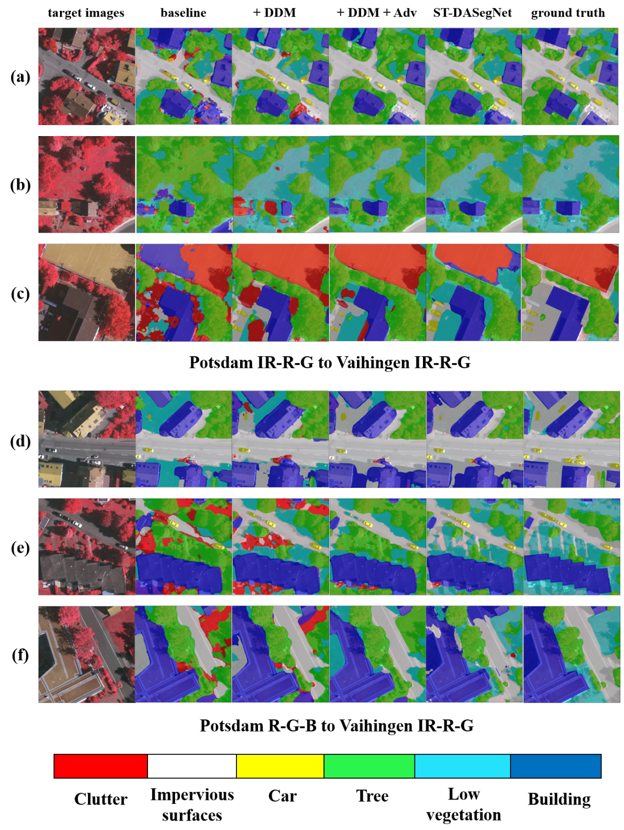


Figure 7: Ablation visualization on "Potsdam IR-R-G to Vaihingen IR-R-G" task (Tab.7) and "Potsdam R-G-B to Vaihingen IR-R-G" task (Tab.8). For each task, we provide 3 cases containing 6 images. From left to right, the first image is the target image. The rest 5 images are respectively predictions of the baseline model, model with DDM, model with DDM and adversarial learning mechanism, model with DDM, adversarial learning mechanism and self-training mechanism (ST-DASegNet), and ground truth. Here, we select SegFormer as the baseline model. Particularly, we do not provide results of the dual path baseline model, because it does not contain key components and improves little compared to baseline model.

not have annotations, we display the results of images in the validation dataset. It can be seen that ST-DASegNet can produce more accurate and reasonable predictions. Especially on some hard categories like "Barren" and "Forest", ST-DASegNet can generate reasonable predictions with clear boundary. The visualization results also coincide with the results in Tab.5 and Tab.6.

4.5.2. Visual ablation study

To separately show the effectiveness of key components of ST-DASegNet in an intuitive manner, we provide ablation visualization on "Potsdam IR-R-G to Vaihingen IR-R-G" and "Potsdam R-G-B to Vaihingen IR-R-G" tasks shown in Fig. 7. When adding DDM, many large-region mistakes are avoided. In the baseline result of case (b), large regions of "Low vegetation" are classified into "Tree". In the baseline result of case (c), large regions of "Clutter" are classified into "Building". After adding DDM, these large-region mistakes are correctly classified. When adding adversarial learning, the edge of each category is clearer and the whole region of each category becomes smooth. In all 6 cases, we find that many small regions will be incorrectly classified into "Clutter". After adding adversarial learning, this problem is almost completely solved. When further applying EMA-based cross-domain separated self-training mechanism, many detailed mistakes for almost every category are corrected, which makes the results seem much better. As shown in cases (a), (b), and (f), some little rectifications on "Building" make predictions seem closer to ground truth. Similarly in cases (b) and (e), detailed corrections on "Impervious surfaces" really help improve the quality of predictions.

Theoretically, with DDM, source, and target student backbones can respectively represent the target images into source style and target style with only source annotations. In other words, besides domain universal features, some target-specific features can be extracted. However, feature inconsistency caused by domain shift will harm the effectiveness of DDM. When further adding feature-level adversarial learning, the domain shift problem is eased and cross-domain single-style features are aligned to become more consistent. Therefore, the feature representation capability of backbones and DDM are both improved. When integrating our proposed self-training mechanism, it can fundamentally improve this task by providing high-quality target annotations (pseudo-labels). In summary, besides ablation comparison experiments (Tab.7 and Tab.8), ablation visualization further proves the effectiveness and great compatibility of the three key components of ST-DASegNet.

4.5.3. Visualization and analysis on feature maps

To intuitively show the interpretability of ST-DASegNet, we visualize the multi-level feature maps extracted from backbones and make an analysis. We select "Potsdam IR-R-G to Vaihingen IR-R-G" and "Postdam R-G-B to Vaihingen IR-R-G" tasks to conduct experiments. We select SegFormer as the baseline model.

Feature map visualization comparison between baseline model and ST-DASegNet. As shown in Fig. 8, we will analyze the visualization results with 4 cases from the following 2 points. (1) Baseline model and ST-DASegNet both have well-represented low-level feature maps, which have abundant detailed information and clear edges among regions of each category. (2) On high-level feature maps, it is obvious that two backbones of ST-DASegNet have better representation capability than baseline backbones. As shown in the "yellow box" regions of the case (a), ST-DASegNet's high-level feature maps have clearer boundaries between regions of "Tree" and "Low vegetation". As shown in case (a) and case (b), we find that the edge of "Building" on ST-DASegNet's high-level feature maps is also in better shape. Case (c) and case (d) reveal that ST-DASegNet has better performance on localizing "Low vegetation". From Fig. 8, we find that the high-level feature maps of the baseline backbone will miss some important information. Even though

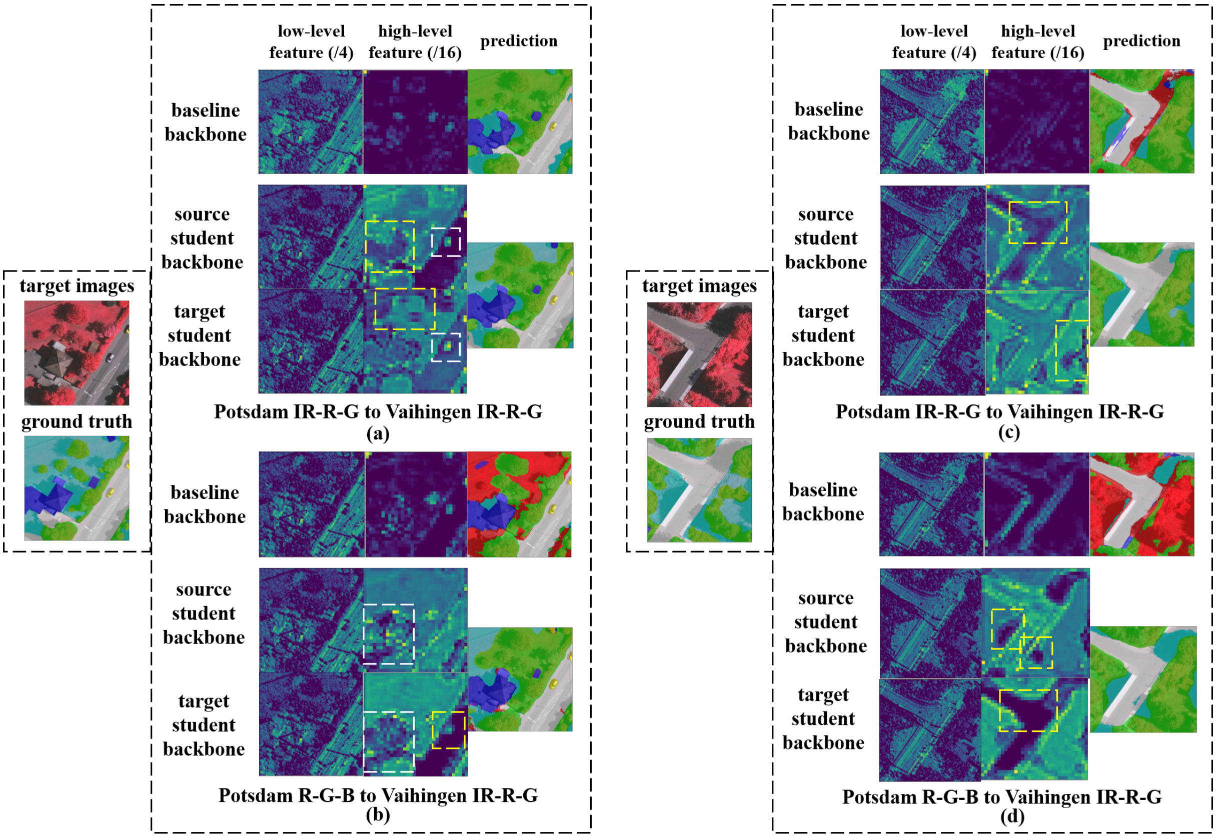


Figure 8: Feature map visualization comparison between baseline model and ST-DASegNet. We select the feature maps extracted from backbones for visualization. Feature maps listed here are channel-wise average feature maps. "/4" and "/16" respectively indicate the downsampling rate compared to the original image. "Prediction" indicates the final segmentation results of the baseline model and ST-DASegNet.

the baseline backbone can represent well on some categories (e.g. Impervious surfaces in case (d)), the generalized representation ability is far behind ST-DASegNet.

Feature map visualization comparison between source and target student backbones of ST-DASegNet. Cases in Fig. 8 intuitively prove the effectiveness of our dual backbones structure. First, the two backbones of ST-DASegNet are complementary in representing target images. As shown in the "yellow box" regions of the case (a), the source student backbone can better find the "Building" edges while the target student backbone can better distinguish the boundaries between regions of "Tree" and "Low vegetation". In case (b), when the source student backbone fails to localize "Car", target student backbone successfully finds it. In cases (c) and (d), one backbone has advantages in dividing the edge of "Impervious surfaces" and the other has better performance on recognizing "Low vegetation". Second, two backbones can extract universal features. As shown in the "white box" regions of cases (a) and (b), the two backbones both have strong perceptions of "Car" and "Building". Third, source and target student backbones in ST-DASegNet can respectively represent target images in source style and target style. In case (a), we find that the source student backbone tends to focus on "Building" and "Impervious surfaces", which is similar to the source-only trained baseline model. Target student backbone tends to have a better focus on "Low vegetation" because intuitively distinguishing "Low vegetation" and "Tree" in target images (Vaihingen IR-R-G) is easier than source images (Potsdam IR-R-G). From another perspective, if comparing case (c) with case (d), we find the representation style (preference) of the source style backbone completely changes, because source-domain images change from Potsdam IR-R-G to Potsdam R-G-B.

Visualization on domain disentangling. Achieving high-quality domain disentangling with DDM is important for ST-DASegNet. With DDM, we expect to extract the invariant feature and purify the unique feature of source-style and target-style features. If DDM works well, more distinct target-style feature extraction can make an effect on better-representing target images. Moreover, differences between source-style and target-style features can improve the representation diversity. In Fig. 9, we adopt feature map visualization comparison to intuitively analyze domain disentangling. First, low-level features have minor changes because low-level features are always invariant, similar, and "no-style". Second, after DDM, high-level source-style and target-style features both become more distinct. In case (a) and (c), we observe that high-level source-style feature tends

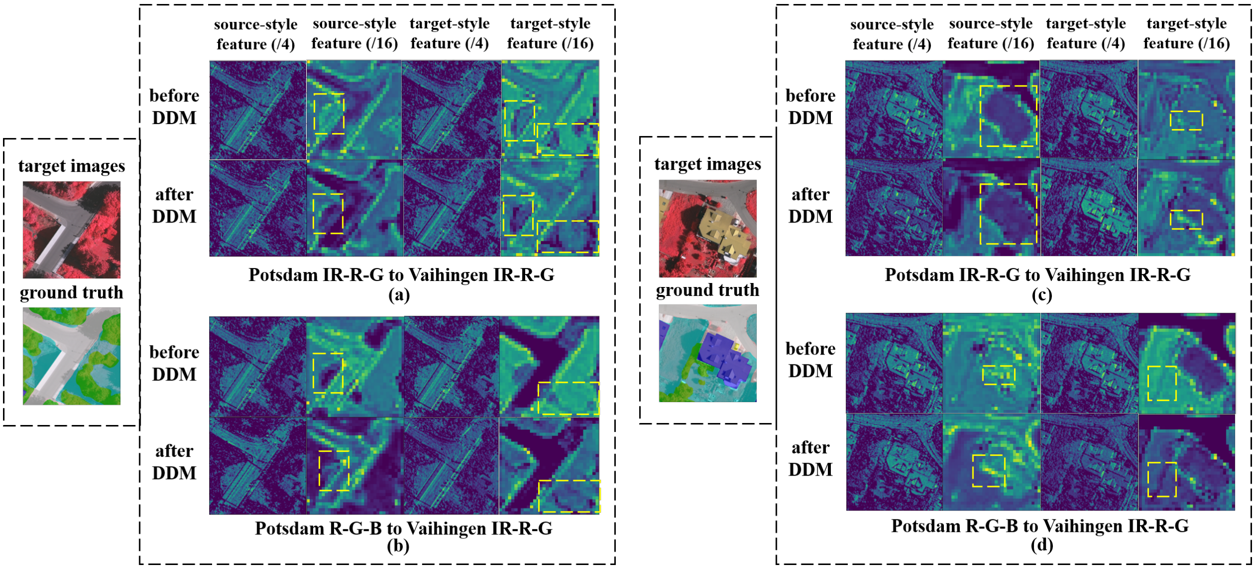

target images
ground truth
source-style feature (/4)
source-style feature (/16)
target-style feature (/4)
target-style feature (/16)
before DDM
after DDM
Potsdam IR-R-G to Vaihingen IR-R-G
(a)
before DDM
after DDM
Potsdam R-G-B to Vaihingen IR-R-G
(b)
target images
ground truth
source-style feature (/4)
source-style feature (/16)
target-style feature (/4)
target-style feature (/16)
before DDM
after DDM
Potsdam IR-R-G to Vaihingen IR-R-G
(c)
before DDM
after DDM
Potsdam R-G-B to Vaihingen IR-R-G
(d)

Figure 9: Visualization on domain disentangling. We select the feature maps extracted from backbones for visualization. Feature maps listed here are channel-wise average feature maps. "/4" and "/16" respectively indicate the downsampling rate compared to the original image.

to focus on "Impervious surfaces" and "Building", which is also proved in Fig. 8. After DDM, the boundary of "Impervious surfaces" becomes clearer and expanded (case (a)). The expanded part is shadow and source student backbone tends to recognize it as "Impervious surfaces". In addition, after DDM, the boundary of "Building" is in better shape (case (c)). As a complementary, target-style feature tends to focus on "Low vegetation" and "Tree". The features after DDM obviously have more tendency in these two categories (case (a) and (c)). Particularly on the shadow area in case (a), the target student backbone tends to recognize it as "Low vegetation". Case (b) and case (d) also show that high-level source-style and target-style features become more distinct and different from each other, which proves the effectiveness of our proposed domain disentangling mechanism.

5. Conclusion

In this paper, we propose a self-training guided disentangled adaptation method (ST-DASegNet) to tackle cross-domain remote sensing image semantic segmentation. In ST-DASegNet, we propose a dual path structure with two student backbones and two student decoders, which aims to represent source and target images into source style and target style with only source annotations. Based on this structure, we propose feature-level adversarial learning on cross-domain single-style features to enhance the representation consistency by feature alignment mechanism. Furthermore, we propose a domain disentangled module (DDM) on single-domain cross-style features to make source-style and target-style features more distinct and diverse. Besides adversarial learning and DDM, we propose and integrate an EMA-based cross-domain separated self-training paradigm ("Decoder-only" and "Single-target"). Our proposed self-training mechanism can balance the representation tendency between source and target domains by generating pseudo-labels of target images. The three key components show great compatibility, which can improve the performance of ST-DASegNet from three different perspectives. Extensive experiments on different benchmark datasets show that ST-DASegNet outperforms the previous SOTA methods on cross-domain RS image semantic segmentation tasks. Abundant visualization analysis intuitively proves that our proposed ST-DASegNet is reasonable and interpretable.

CRediT Author Statement

Qi Zhao: Conceptualization, Supervision, Project administration. **Shuchang Lyu**: Conceptualization, Methodology, Software, Writing - Original Draft. **Hongbo Zhao**: Validation, Formal analysis, Writing - Review & Editing, Funding acquisition. **Binghao Liu**: Validation, Visualization. **Lijiang Chen**: Investigation, Writing - Review & Editing. **Guangliang Cheng**: Writing - Review & Editing, Supervision.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (grant numbers 62072021 and 61901015).

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 39, 2481–2495.

Bai, L., Du, S., Zhang, X., et al., 2022. Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive learning and adversarial learning. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–13.

Benjdira, B., Bazi, Y., Koubâa, A., et al., 2019. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. Remote Sensing (RS) 11, 1369.

Cai, Y., Dai, L., Wang, H., et al., 2022. Dlnet with training task conversion stream for precise semantic segmentation in actual traffic scene. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 33, 6443–6457.

Chen, C., Dou, Q., Chen, H., et al., 2020. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Transactions Medical Imaging (TMI) 39, 2494–2505.

Chen, J., Sun, B., Wang, L., et al., 2022a. Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas. International Journal of Applied Earth Observation and Geoinformation (JAG) 112, 102881.

Chen, J., Zhu, J., Guo, Y., et al., 2022b. Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–15.

Chen, L., Zhu, Y., Papandreou, G., et al., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: European Conference on Computer Vision (ECCV), pp. 833–851.

Chen, X., Pan, S., Chong, Y., 2022c. Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–13.

Chen, Y., Li, W., Chen, X., et al., 2019. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1841–1850.

Chen, Z., Yang, B., Ma, A., et al., 2022d. Joint alignment of the distribution in input and feature space for cross-domain aerial image semantic segmentation. International Journal of Applied Earth Observation and Geoinformation (JAG) 115, 103107.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE 105, 1865–1883.

Ding, J., Xue, N., Long, Y., et al., 2019. Learning roi transformer for detecting oriented objects in aerial images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2844–2853.

Du, L., Tan, J., Yang, H., et al., 2019. SSF-DAN: separated semantic feature based domain adaptation network for semantic segmentation, in: IEEE International Conference on Computer Vision (ICCV), pp. 982–991.

Fu, J., Liu, J., Tian, H., et al., 2019. Dual attention network for scene segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3146–3154.

Gerke, M., 2014. Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen). Technical Report. ITC, Univ. Twente, Enschede, The Netherlands. Doi: 10.13140/2.1.5015.9683.

Guo, S., Zhou, Q., et al., 2021. Label-free regional consistency for image-to-image translation, in: IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Hoffman, J., Tzeng, E., Park, T., et al., 2018. Cycada: Cycle-consistent adversarial domain adaptation, in: International Conference on Machine Learning (ICML), pp. 1994–2003.

Hou, J., Guo, Z., Wu, Y., Diao, W., Xu, T., 2022a. Bsnet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–22.

Hou, J., Guo, Z., Wu, Y., et al., 2022b. Bsnet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–22.

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

Huang, Z., Wang, X., Huang, L., et al., 2019. Ccnet: Criss-cross attention for semantic segmentation, in: IEEE International Conference on Computer Vision (ICCV), IEEE. pp. 603–612.

Isola, P., Zhu, J., Zhou, T., et al., 2017a. Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976.

Isola, P., Zhu, J., Zhou, T., et al., 2017b. Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976.

Ji, S., Wang, D., Luo, M., 2021. Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 59, 3816–3828.

Li, A., Jiao, L., Zhu, H., et al., 2022a. Multitask semantic boundary awareness network for remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–14.

Li, G., Li, L., Zhu, H., Liu, X., et al., 2019. Adaptive multiscale deep fusion residual network for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 57, 8506–8521.

Li, H., Qiu, K., Chen, L., Mei, X., Hong, L., Tao, C., 2021a. Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters (GRSL) 18, 905–909.

Li, J., Zi, S., Song, R., et al., 2022b. A stepwise domain adaptive segmentation network with covariate shift alleviation for remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–15.

Li, Y., Shi, T., Zhang, Y., et al., 2021b. 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 (ISPRS) 175, 20–33.

Lian, Q., Duan, L., Lv, F., et al., 2019. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach, in: IEEE International Conference on Computer Vision (ICCV), pp. 6757–6766.

Lin, Q., Zhao, J., Fu, G., et al., 2022. Crpn-sfnet: A high-performance object detector on large-scale remote sensing images. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 33, 416–429.

Liu, S., Ding, W., Liu, C., et al., 2018. Ern: Edge loss reinforced semantic segmentation network for remote sensing images. Remote Sensing (RS) 10, 1339.

Luo, Y., Zheng, L., Guan, T., et al., 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2507–2516.

Lyu, Y., Vosselman, G., Xia, G.S., et al., 2020. Uavid: A semantic segmentation dataset for uav imagery. ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS) 165, 108–119.

Ma, X., Zhang, X., Wang, Z., et al., 2023. Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 61, 1–15.

Mao, Z., Huang, X., Niu, W., et al., 2023. Improved instance segmentation for slender urban road facility extraction using oblique aerial images. International Journal of Applied Earth Observation and Geoinformation (JAG) 121, 103362.

Mei, K., Zhu, C., Zou, J., et al., 2020. Instance adaptive self-training for unsupervised domain adaptation, in: European Conference on Computer Vision (ECCV), pp. 415–430.

Mou, L., Hua, Y., Zhu, X., 2020. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 58, 7557–7569.

Ni, H., Liu, Q., Guan, H., et al., 2023. Category-level assignment for cross-domain semantic segmentation in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 61, 1–16.

Nogueira, K., Mura, M.D., Chanussot, J., et al., 2019. Dynamic multi-context segmentation of remote sensing images based on convolutional networks. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 57, 7503–7520.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 39, 640–651.

Tsai, Y., Hung, W., Schulter, S., et al., 2018a. Learning to adapt structured output space for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7472–7481.

Tsai, Y., Hung, W., Schulter, S., et al., 2018b. Learning to adapt structured output space for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7472–7481.

Tzeng, E., Hoffman, J., Zhang, N., et al., 2014. Deep domain confusion: Maximizing for domain invariance. CoRR abs/1412.3474.

Wang, H., Shen, T., Zhang, W., Duan, L.Y., Mei, T., 2020a. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation, in: European Conference on Computer Vision (ECCV).

Wang, H., Shen, T., Zhang, W., et al., 2020b. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation, in: European Conference on Computer Vision (ECCV), pp. 642–659.

Wang, J., Zheng, Z., Ma, A., et al., 2021a. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation, in: Neural Information Processing Systems (NeurIPS).

Wang, J., Zhong, Y., Zheng, Z., et al., 2021b. Rsnet: The search for remote sensing deep neural networks in recognition tasks. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 59, 2520–2534.

Wang, L., Xiao, P., Zhang, X., et al., 2023a. 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 (J-STARS) 16, 4109–4121.

Wang, L., Xiao, P., Zhang, X., et al., 2023b. 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 (J-STARS) 16, 4109–4121.

Wang, Q., Huang, W., Xiong, Z., et al., 2022. Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) 33, 1414–1428.

Wu, L., Lu, M., Fang, L., 2022. Deep covariance alignment for domain adaptive remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–11.

Wu, T., Tang, S., Zhang, R., et al., 2021. Cgnet: A light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing (TIP) 30, 1169–1179.

Xia, G., Bai, X., Ding, J., et al., 2018. DOTA: A large-scale dataset for object detection in aerial images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3974–3983.

Xia, G., Hu, J., Hu, F., et al., 2017. Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 55, 3965–3981.

Xie, E., Wang, W., Yu, Z., et al., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 12077–12090.

Yang, G., Zhang, Q., Zhang, G., 2020. Eanet: Edge-aware network for the extraction of buildings from aerial images. Remote Sensing (RS) 12, 2161.

Yang, Y., Soatto, S., 2020. FDA: fourier domain adaptation for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4084–4094.

Yi, Z., Zhang, H., Tan, P., et al., 2017. Dualgan: Unsupervised dual learning for image-to-image translation. IEEE International Conference on Computer Vision (ICCV) , 2868–2876.

Yu, C., Gao, C., Wang, J., et al., 2021. Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation. International Journal of Computer Vision (IJCV) 129, 3051–3068.

Zeng, G., Schmaranzer, F., Lerch, T.D., et al., 2020. Entropy guided unsupervised domain adaptation for cross-center hip cartilage segmentation from MRI, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 447–456.

Zhang, B., Chen, T., Wang, B., 2022a. Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–12.

Zhang, C., Feng, Y., Hu, L., et al., 2022b. A domain adaptation neural network for change detection with heterogeneous optical and SAR remote sensing images. International Journal of Applied Earth Observation and Geoinformation (JAG) 109, 102769.

Zhang, P., Zhang, B., Zhang, T., et al., 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12414–12424.

Zhao, H., Shi, J., Qi, X., et al., 2017. Pyramid scene parsing network, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239.

Zhao, Q., Liu, J., Li, Y., et al., 2022a. Semantic segmentation with attention mechanism for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–13.

Zhao, Q., Liu, J., Li, Y., et al., 2022b. Semantic segmentation with attention mechanism for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (TGRS) 60, 1–13.

Zhao, Q., Lyu, S., Li, Y., et al., 2021. Mgml: Multigranularity multilevel feature ensemble network for remote sensing scene classification. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) .

Zhao, Y., Guo, P., Sun, Z., et al., 2023. Residualgan: Resize-residual dualgan for cross-domain remote sensing images semantic segmentation. Remote Sensing (RS) 15, 1428.

Zhou, Q., Feng, Z., Gu, Q., et al., 2022a. Uncertainty-aware consistency regularization for cross-domain semantic segmentation. Computer Vision and Image Understanding (CVIU) 221, 103448.

Zhou, Q., Zhuang, C., Lu, X., et al., 2022b. Domain adaptive semantic segmentation via regional contrastive consistency regularization. IEEE International Conference on Multimedia and Expo (ICME) , 01–06.

Zhu, J., Park, T., Isola, P., et al., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251.

Zou, D., Zhu, Q., Yan, P., 2020. Unsupervised domain adaptation with dual-scheme fusion network for medical image segmentation, in: International Joint Conference on Artificial Intelligence (IJCAI), pp. 3291–3298.

Zou, Y., Yu, Z., Kumar, B.V.K.V., et al., 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training, in: European Conference on Computer Vision (ECCV), pp. 297–313.

Zou, Y., Yu, Z., Liu, X., et al., 2019. Confidence regularized self-training. IEEE International Conference on Computer Vision (ICCV) , 5981–5990.