CMT: Cross Mean Teacher Unsupervised Domain Adaptation for VHR Image Semantic Segmentation

Liang Yan, Bin Fan, *Senior Member, IEEE*, Shiming Xiang, *Member, IEEE*, and Chunhong Pan, *Member, IEEE*

***Abstract*— Semantic segmentation of remote sensing images has achieved superior results with the supervised deep learning models. However, their performance to unseen data domains could be very bad due to the domain shift between different domains. Recently, a series of unsupervised domain adaptation (UDA) methods has been developed to solve the domain shift problem in semantic segmentation. Most of them use adversarial learning to achieve global cross-domain alignment and use a self-training (ST) strategy to generate pseudo-labels for classwise alignment. However, these methods ignore the pixels that are not assigned pseudo-labels. Those pixels are mostly at the boundaries, which are vital to the final segmentation results. To solve this problem, this letter proposes a cross mean teacher (CMT) UDA method. The whole framework consists of two parts. On the one hand, the global cross-domain distribution alignment is performed, and then, reliable pseudo-labels are assigned to the target data. On the other hand, a cross teacher–student network (CTSN) is developed to effectively use those pixels with and without pseudo-labels. This network contains two student networks (S_1 and S_2) and two teacher networks (T_1 and T_2) for cross-consistency constraints that supervises S_2 (or S_1) by the prediction results of T_1 (or T_2). The cross supervision by CTSN is helpful to prevent performance bottlenecks caused by the high coupling of teacher–student network in existing methods. Extensive experiments on three different remote sensing adaptation scenes verify the effectiveness and superiority of the proposed method.**

***Index Terms*— Cross mean teacher (CMT), self-training (ST), semantic segmentation, unsupervised domain adaptation (UDA), very-high-resolution (VHR) image.**

I. Introduction

SEMANTIC segmentation, which aims at assigning label to each pixel in an image, has been widely used in the remote sensing community. Recently, deep learning-based methods have made great achievements for this task when there are a large number of annotated samples [1]–[5]. However, it is time-consuming and laborious to provide pixelwise annotations for each image, especially in the remote sensing community with diverse scenes. The suboptimal solution is to train a model with an existing labeled data set (source domain) and deploy it directly to the unlabeled target domain. Unfortunately, the distribution discrepancy between different domains degrades the performance of the learned model on the target domain, especially for very-high-resolution (VHR) remote sensing images. Such discrepancy is mainly caused by the diversity of data acquisition conditions, such as color saturation, different spectral bands, regions, and ground sampling distances (GSDs).

The distribution discrepancy between domains is usually named domain shift [6], which is expected to be eliminated by unsupervised domain adaptation (UDA) [7]. For semantic segmentation, a maximum classifier discrepancy model [8] is proposed by applying two task-specific classifiers as discriminators to perform distribution alignment. Yan *et al.* [9] developed an adversarial domain similarity discriminator to jointly consider dissimilar and similar information between domains on feature space for VHR image semantic segmentation. Tsai *et al.* [10] first performed distribution alignment through adversarial learning [11] in the output space of the segmentation network. Since the target domain has no annotations, these methods can only align the distribution of different domains globally and cannot guarantee the classwise alignment [12].

To perform cross-domain classwise alignment, many self-training (ST)-based UDA methods [12]–[14] have been developed. Those approaches generate pseudo-labels for the target domain to retrain the segmentation network for classwise alignment. Li *et al.* [13] proposed a bidirectional learning domain adaptation model (BDL) that alternatively trains the image translation model and the self-supervised segmentation adaptation model. In the process of self-supervised, BDL assigns pseudo-labels to pixels in the target domain whose predicted probability is greater than a threshold for retraining. Yan *et al.* [12] proposed a class-aware ST (CAST) method that considers both the outputs of the discriminator and the segmentation network for pseudo-labels generation. However, these methods leave many without any pseudo-labels pixels, which are mostly at the boundaries and have a crucial influence on the final segmentation results.

Manuscript received December 24, 2020; revised February 22, 2021; accepted March 10, 2021. This work was supported in part by the Major Project for New Generation of Artificial Intelligence under Grant 2018AAA0100400, in part by the National Natural Science Foundation of China under Grant U2013202, Grant 91646207, Grant 62076242, Grant 62071466, and Grant 61976208, and in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2018QNRC001. *(Corresponding author: Bin Fan.)*

Liang Yan and Shiming Xiang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: liang.yan@nlpr.ia.ac.cn; smxiang@nlpr.ia.ac.cn).

Bin Fan is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100080, China (e-mail: bin.fan@ieee.org).

Chunhong Pan is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: chpan@nlpr.ia.ac.cn).

Color versions of one or more figures in this letter are available at https://doi.org/10.1109/LGRS.2021.3065982.

Digital Object Identifier 10.1109/LGRS.2021.3065982

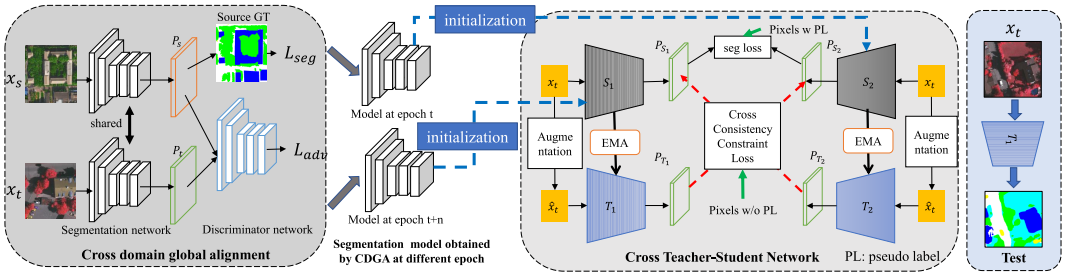

Fig. 1. Overall architecture of the proposed method. It mainly consists of two parts: the left part is the CDGA and the right part is the CTSN for classwise alignment. CTSN is applied after the CDGA, and the weights of two student networks (S_1 and S_2) in CTSN are initialized by the segmentation network obtained by CDGA at two different epochs. T_1 and T_2 use the same initialization way as S_1 and S_2. EMA represents the EMA algorithm used to update the teacher network. The T_1 model is used as the test model.

In the semisupervised learning, the teacher–student structure is usually adopted to use unlabeled data. The mean teacher (MT) [15] obtains an ensemble teacher by applying an exponential moving average (EMA) to the student. It obeys the smoothness assumption that if two samples are generated by a sample through different perturbations, then the teacher and student should have consistent prediction for them. However, under the setting of UDA, the student has strong convergence, leading to the EMA teacher that is coupled with the student. This phenomenon leads to the teacher, which cannot provide sufficient meaningful knowledge for the student [16].

To solve the above issues, this letter proposes a cross mean teacher (CMT) ST UDA method for VHR image semantic segmentation. First, CMT performs global distribution alignment by global alignment domain adaptation (DA) method, e.g., AdasegNet [12] and TriADA [10], and generates pseudo-labels of the target domain. Second, a cross teacher–student network (CTSN) is proposed to perform classwise alignment. CTSN contains two student networks (S_1 and S_2) and two teacher networks (T_1 and T_2). The pixels with pseudo-labels are trained directly through the standard cross-entropy loss. A cross-consistency constraint is developed to utilize pixels without pseudo-labels. As shown in the right part of Fig. 1, CTSN uses the predicted results of T_1 (or T_2) as the supervisory information of S_2 (or S_1) for cross-consistency constraints. In addition, we initialize S_1 and S_2 with different weights and optimize them independently. In this way, CTSN not only inherits the stability of the MT but also avoids the excessive coupling of the teacher–student network. The main contributions of this letter are summarized as follows.

1) A CMT ST UDA method is proposed for cross-domain semantic segmentation in VHR remote sensing images, which can perform both global and classwise distribution alignment.
2) A CTSN is developed in CMT to effectively utilize the pixels that are not assigned pseudo-labels during the classwise alignment.
3) Experiments on three different remote sensing adaptation situations demonstrate that the proposed method outperforms the state-of-the-art methods.

II. Proposed Method

In the setting of UDA, the labeled source domain is denoted as $D_S = \{(\mathbf{x}_s, \mathbf{y}_s)|\mathbf{x}_s \in \mathbb{R}^{H\times W\times 3}, \mathbf{y}_s \in \mathbb{R}^{H\times W\times C}\}$, and the unlabeled target domain is denoted as $D_T = \{\mathbf{x}_t|\mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}\}$, where W and H are the width and height of the image, respectively, and C is the number of classes. The proposed network consists of two parts, the cross-domain global alignment (CDGA) module and the CTSN, as shown in Fig. 1.

A. Cross-Domain Global Alignment (CDGA)

Note that the proposed CTSN model requires pseudo-labels of the target domain and different weights to initialize the two teacher–student networks. Therefore, CDGA is used to generate pseudo-labels and provide different initializations and will not be updated when training CTSN. It can be any global alignment DA methods, e.g., AdasegNet [10] and TriADA [12]. In this letter, we adopt AdasegNet as CDGA for its popularity. The results of using TriADA will be discussed in Section III. Here, we briefly introduce the principle of AdasegNet (please refer to [10] for details).

To achieve global alignment, AdasegNet plays a min-max game between the segmentation network **F** and the discriminator **D**. In this process, **D** is treated as a binary classifier to distinguish whether the input features are generated from the source domain or the target domain. **F** is updated to trick **D** into distinguishing features that are originally generated from the target domain as if they are generated from the source domain. Thus, **F** can generate domain-invariant features and achieve global alignment. This process is formulated as follows:

$$L_{\text{adv},D}(\mathbf{x}_s, \mathbf{x}_t; \theta_{\mathbf{D}}) = -\mathop{\mathbb{E}}_{\mathbf{x}_s\sim D_S}[\log(D(F(\mathbf{x}_s)))] - \mathop{\mathbb{E}}_{\mathbf{x}_t\sim D_T}[\log(1 - D(F(\mathbf{x}_t)))] \quad (1)$$

$$L_{\text{adv},F}(\mathbf{x}_t; \theta_{\mathbf{F}}) = -\mathop{\mathbb{E}}_{\mathbf{x}_t\sim D_T}[\log(D(F(\mathbf{x}_t)))] \quad (2)$$

where $\boldsymbol{\theta}_D$ and $\boldsymbol{\theta}_F$ denote the weights of **D** and **F**, respectively, and $L_{\text{adv},D}$ and $L_{\text{adv},F}$ are alternately optimized to update **D** and **F**, respectively.

Meanwhile, to ensure the performance of **F** on source domain, AdasegNet optimizes the standard supervised segmentation loss on the source domain. The learning objective is as follows:

$$L_{\text{seg},S}(\mathbf{x}_s, \mathbf{y}_s; \theta_{\mathbf{F}}) = -\frac{1}{\text{HW}}\sum_{n=1}^{\text{HW}}\sum_{c=1}^{C} y_s^{n,c}\log(p_n(c|\theta_{\mathbf{F}}, \mathbf{x}_s)) \quad (3)$$

in which $p_n(c|\theta_{\mathbf{F}}, \mathbf{x}_s)$ denotes the probability that the nth position in source sample $\mathbf{x}_s$ is predicted as class c. Here, $\mathbf{y}_s^n$ is the ground truth that is expressed as a one-hot vector.

B. Cross Teacher–Student Network (CTSN)

CTSN can effectively utilize all pixels in the target domain, including those pixels without pseudo-labels. Specifically, after global alignment, we first use $\mathbf{F}$ to predict the target data and assign pseudo-labels to pixels whose predicted probabilities are higher than a threshold λ. The pseudo-labels generation process is as follows:

$$c^* = \arg\max_c p_n(c|\theta_{\mathbf{F}}, \mathbf{x}_t) \tag{4}$$

$$\hat{y}_t^{n,c^*} = \begin{cases} 1, & \text{if } \ p_n(c^*|\theta_{\mathbf{F}}, \mathbf{x}_t) > \lambda \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where c^* is the predicted class at the nth position for $\mathbf{x}_t$. If $\hat{\mathbf{y}}_t^n$ taht equals zero, vector indicates that the nth position does not assign a pseudo-label, and vice versa. M_t^p and M_t^{up} denote masks with and without pseudo-labels in $\mathbf{x}_t$, respectively

$$M_t^{\text{up},n} = \begin{cases} 1, & \hat{\mathbf{y}}_t^n == \vec{0} \\ 0, & \hat{\mathbf{y}}_t^n \neq \vec{0}, \end{cases} \quad M_t^{p,n} = \begin{cases} 1, & \hat{\mathbf{y}}_t^n \neq \vec{0} \\ 0, & \hat{\mathbf{y}}_t^n == \vec{0}. \end{cases} \tag{6}$$

Second, to utilize the pixels that without pseudo-labels in the target domain, CTSN is designed, as shown in the right of Fig. 1. CTSN consists of two student networks (S_1 and S_2) and two teacher networks (T_1 and T_2). The weights of T_1 and T_2 are updated by EMA of S_1 and S_2, as shown in the following:

$$\theta_T^t = \alpha\theta_T^{t-1} + (1-\alpha)\theta_S^t \tag{7}$$

in which θ_T^t and θ_S^t represent the weights of the teacher network and the student network in the training step of t, respectively, and $\alpha \in [0, 1]$ is a smoothing coefficient.

Different from MT [15] that directly uses the output of the teacher network as the supervision information of the student network for consistency constraint, this may lead to a performance bottleneck [16]. CTSN utilizes the predicted probability of T_1 (or T_2) as the supervision information of S_2 (or S_1) to perform cross-consistency constraint, which can be formulated as follows:

$$\begin{aligned} L_{3c}(\mathbf{x_t}) = &\left\| M_t^{\text{up}} \odot T_2(\hat{\mathbf{x}}_t) - M_t^{\text{up}} \odot S_1(\mathbf{x}_t) \right\|^2 \\ &+ \left\| M_t^{\text{up}} \odot T_1(\hat{\mathbf{x}}_t) - M_t^{\text{up}} \odot S_2(\mathbf{x}_t) \right\|^2 \end{aligned} \tag{8}$$

where $\hat{\mathbf{x}}_t$ is an augmented image of sample $\mathbf{x}_t$ after adding Gaussian noise and random color jitter, as MT did [15]. Here, $\odot$ is an elementwise dot product. Since S_1 and S_2 are initialized by the segmentation model $\mathbf{F}$ obtained at different iterations in CDGA and independently optimized, T_2 and T_1 are decoupled from S_1 and S_2, respectively.

For target data with pseudo-labels, the cross-entropy loss is applied to S_1 and S_2. The object function is as follows:

$$\begin{aligned} &L_{\text{seg},T}(\mathbf{x}_t, \hat{\mathbf{y}}_t) \\ &= -\frac{1}{\text{HW}} \sum_{n=1}^{\text{HW}} \sum_{c=1}^{C} M_t^{p,n} \cdot \hat{y}_t^{n,c} \log(p_n(c|\theta_{\mathbf{S_1}}, \mathbf{x}_t)) \\ &\quad - \frac{1}{\text{HW}} \sum_{n=1}^{\text{HW}} \sum_{c=1}^{C} M_t^{p,n} \cdot \hat{y}_t^{n,c} \log(p_n(c|\theta_{\mathbf{S_2}}, \mathbf{x}_t)). \end{aligned} \tag{9}$$

Algorithm 1 Overall Steps of the Proposed Method

Input: Source domain D_S, target domain D_T, segmentation network $\mathbf{F}$, discriminator $\mathbf{D}$. The training iterations of CDGA N, the training epochs of CTSN K. The hyperparameters of adversarial loss λ_{adv} and target segmentation loss β.
Output: The segmentation network $\mathbf{F}$.

```
for i = 1 to N do
    Updating F by minimizing L_{seg,S}(θ_F) + λ_adv L_{adv,F}(θ_F);
    Updating D by minimizing L_{adv,D}(θ_D);
end for
Generating target pseudo-labels (Ŷ_T) and masks (M_T^p and M_T^up) through Eq ((4), (5), (6));
Initializing CTSN (S_1, S_2 and T_1, T_2) by the model of F obtained at different iterations;
for k = 1 to K do
    Updating CTSN by minimizing L_3c + β L_{seg,T}.
end for
```

C. Overall Steps

The main optimization process consists of two parts. First, CDGA is optimized to perform global distribution alignment and obtain reliable target pseudo-labels. Then, we use the weights of the fifth epoch from the end and the tenth epoch from the end of the segmentation network in CDGA to initialize to the weights of S_1 and S_2, respectively. T_1 and T_2 use the same initialization way as S_1 and S_2. Finally, CTSN is optimized by utilizing the target data with and without pseudo-labels. The whole training steps are shown in Algorithm 1.

III. Experiments and Analysis

A. Data Sets Description

The International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen data set (VAI) [19] is a 2-D semantic labeling benchmark data set. It consists of 16 annotated three-band infrared, red, green (IRRG) VHR images with five categories taken by the airborne sensor from the city of Vaihingen, Germany. The size of each image is about 2500×2000, with a GSD of 9 cm. We randomly sample ten images as the training set and the rest as the testing set. We cropped the training set to a number of 512×512 patches with an overlap of 200 pixels (no overlap in the testing set).

ISPRS Postdam data set [19] has two different color bands, i.e., three-band Postdam IRRG (POTIRRG) and three-band Postdam RGB (POTRGB). It has 24 annotated VHR images with the same five categories as VAI. It is obtained by airborne sensor from the city of Postdam, Germany. The resolution is about 6000×6000 with a GSD of 5 cm. There are 15 images sampled as the training set and the rest as the testing set. We process these data in the same way as VAI.

BeiJing City data set (BEJ) [12] is collected on the Baidu map with 30-cm GSD. There are 202 three-band RGB annotated images with the size of about 1800×800. All images are cropped to 512×512 patches without overlap and flip them horizontally and vertically. Finally, we sample 800 images as

TABLE I
PERFORMANCE (%) OF DIFFERENT ADAPTATION SCENES

Scene	BEJ to VAI					POTIRRG to VAI						POTIRRG to POTRGB					
Method	car	building	tree	road	mIoU	car	building	tree	low veg	road	mIoU	car	building	tree	low veg	road	mIoU
Source-only	2.3	30.8	0	44.0	19.3	6.0	46.6	42.0	23.9	27.5	29.2	94.7	94.5	75.5	65.8	82.2	82.5
DANN [17]	19.3	27.1	1.8	49.6	24.4	33.1	68.3	55.5	26.9	64.0	49.5	90.8	88.2	68.0	68.3	82.4	79.5
ADDA [18]	5.9	61.7	26.8	42.3	31.1	30.4	67.6	43.1	29.7	62.3	46.7	91.4	93.2	73.0	74.1	87.7	83.9
MCD [8]	9.7	47.5	19.2	38.2	28.6	8.3	52.9	32.0	25.3	55.0	34.7	88.0	93.7	60.3	64.8	85.2	78.4
AdasegNet [10]	10.3	66.5	9.6	34.3	30.2	35.6	68.7	54.8	33.7	65.0	51.6	91.8	93.3	74.2	72.8	81.2	83.9
ADDS [9]	18.3	61.9	12.0	38.3	32.6	38.0	70.8	53.3	29.6	65.6	51.5	95.1	94.7	77.6	72.9	87.4	85.5
TriADA-CAST [12]	13.0	74.6	**68.1**	54.6	52.6	46.7	75.7	58.0	34.5	69.1	56.8	94.4	96.1	79.0	79.2	91.2	88.0
CMT(Ours)	**21.4**	**79.4**	67.4	**59.6**	**57.0**	**48.1**	**77.5**	**61.9**	**41.1**	**71.9**	**60.1**	**94.7**	**96.5**	**80.9**	**82.3**	**92.7**	**89.4**

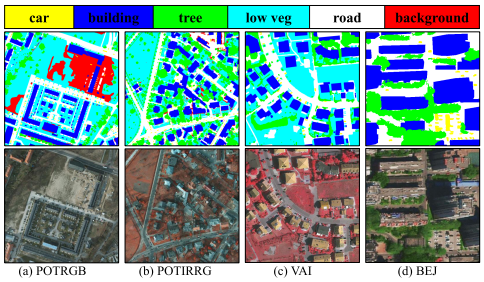

Fig. 2. Images and the corresponding ground truths sampled from different remote sensing data sets.

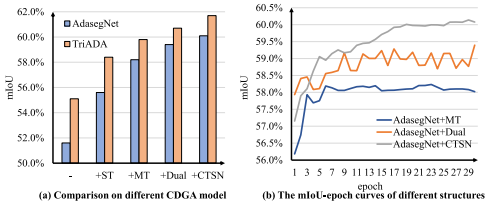

Fig. 3. Ablation studies on POTIRRG to VAI adaptation scenario. (a) Ablation experiments with different ST methods and different CDGA models. (b) mIoU curves of three different structures.

the testing set and the rest of 3916 images are regarded as the training set.

To demonstrate the effectiveness and generalization of CMT, we construct three cross-domain scenarios based on those data sets. Fig. 2 shows the images and the corresponding ground truths in different data sets.

B. Implementation Details

In this letter, all methods are implemented by the Pytorch framework on a Titan XP GPU with 12-GB RAM. For a fair comparison, we use DeeplabV3 [20] as the segmentation network in all experiments. The segmentation network is optimized by the stochastic gradient descent optimizer whose momentum is set as 0.9 and the weight decay is 5×10^{-4}. The learning rate is 2.5×10^{-4} with the polynomial decrease strategy according to [20]. The structure of discriminator and the optimizer is the same as AdasegNet [10]. The learning rate of discriminator is 1×10^{-4}. The hyperparameters of λ_{adv}, λ, α, and β are set as 0.01, 0.9, 0.999, and 1.0 in all experiments. The intersection over union (IoU) is taken as the evaluation criterion: $IoU(P_{\text{pre}}, P_{gt}) = (|P_{\text{pre}} \cap P_{gt}|/|P_{\text{pre}} \cup P_{gt}|)$, in which P_{pre} and P_{gt} are the set of predicted pixels and ground truth, respectively.

C. Comparison With the State of the Arts

Table I presents the performance comparison of CMT and other competitive methods under three different adaptation scenarios. Note that "Source-only" is a model that only uses the source domain for training and is directly tested on the target domain. The results show that in the adaptation scene between images taken from different areas by different sensors with different spectral bands (BEJ to VAI), CMT is superior to other adversarial and ST-based UDA methods, e.g., AdasegNet [10] and TriADA-CAST [12]. Meanwhile, CMT achieves the best performance in both the POTIRRG to VAI (the adaptation scene that images are acquired by the same sensor at different areas) and POTIRRG to POTRGB (the adaptation scene between images of different spectral bands taken from the same region by the same sensor). Overall, CMT is superior to other methods in most categories in all three scenarios, which proves its effectiveness and generalization.

D. Ablation Analysis

To investigate the effectiveness of CTSN, ablation studies on the adaptation of POTIRRG to VAI are conducted, as shown in Fig. 3. "−" denotes the result of only using the CDGA module (AdasegNet [10] or TriADA [12]). "+ST" represents directly using the generated pseudo-labels to fine-tune the segmentation network after CDGA. "+MT" denotes that using MT to process all target data, some of which are even without pseudo-labels. "+Dual" denotes that using the outputs of two individual network (e.g., S_1 and S_2) to perform consistency constraints. "+CTSN" is the proposed CTSN. Note that when the initial weights of S_1 and S_2 in CTSN are the same, CTSN degenerates into two identical MT structures, and the decoupling of CTSN is lost. This is why S_1 and S_2 have to be initialized with different weights in our work.

Fig. 3(a) reports the mIoUs of applying different ST methods with different CDGA models. The comparison between "+ST" and "+MT" proves that the performance can be further improved after using the target data without pseudo-labels. Meanwhile, "+CTSN" achieves the best performance, indicating that the proposed CTSN is more suitable for processing the target data without pseudo-labels. In addition, better performance can be obtained by applying better global alignment methods, as shown by the results of TriADA.

Fig. 3(b) reports the mIoU curves of three different structures (MT, Dual, and CTSN). It shows that although MT can produce stable results, the improvement is limited. Although the Dual structure can improve the performance as well, the training is unstable. CTSN not only improves the performance of MT, but also the training process is more stable than the Dual structure.

Fig. 4 shows the segmentation results of different methods. The segmentation result suffers from domain shift seriously

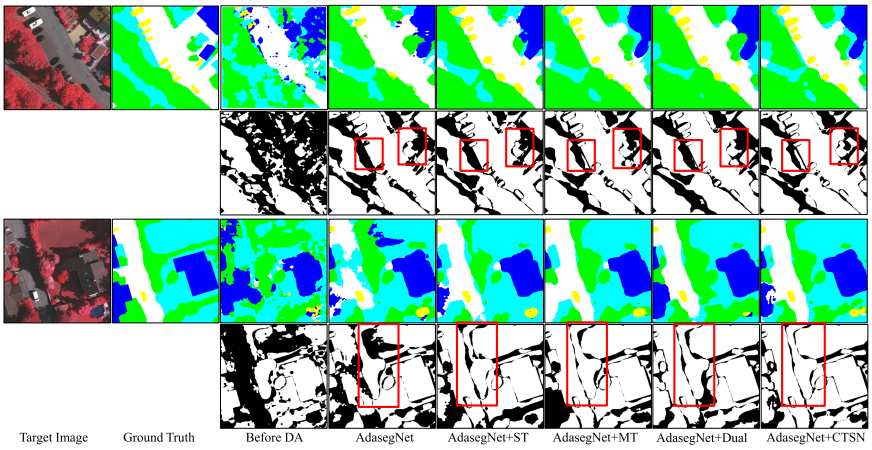

Fig. 4. Visualization results of POTIRRG to VAI adaptation scene on target images. The first and third rows represent the segmentation results. The second and fourth rows represent the discrepancy between the segmentation result and the ground truth. Best viewed in zoom-up.

before DA. After DA, the segmentation results of all methods are improved. Since the performance improvement of AdasegNet+{ST, MT, Dual, CTSN} is mainly reflected at the edges, no significant improvement can be observed from the segmentation results alone. To better observe the improvement, we visualize the discrepancy between the segmentation results and the ground truth in Fig. 4 as well. As shown by the red rectangle, compared to not using data without pseudo-labels (AdasegNet+ST), the models using data without pseudo-labels (AdasegNet + {MT, Dual, CTSN}) have better performance at edges. This shows the effectiveness of using data without pseudo-labels. Meanwhile, AdasegNet + CTSN can obtain more refined results at edges than AdasegNet + {ST, MT, Dual}.

IV. Conclusion

In this letter, a CMT UDA method is proposed for VHR images semantic segmentation. CMT mainly consists of two parts, CDGA and CTSN. CDGA is able to achieve global cross-domain alignment and generate reliable pseudo-labels for target domain. CTSN can effectively use both pixels in the target domain with and without pseudo-labels by cross-consistency constraint. Comprehensive experiments on three different adaptation scenes demonstrate the effectiveness and generalization of the proposed method.

References

[1] F. Hu, G.-S. Xia, W. Yang, and L. Zhang, "Recent advances and opportunities in scene classification of aerial images with deep models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 4371–4374.

[2] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018.

[3] L. Lv, Y. Guo, T. Bao, C. Fu, H. Huo, and T. Fang, "MFALNet: A multi-scale feature aggregation lightweight network for semantic segmentation of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, early access, Aug. 6, 2021, doi: 10.1109/LGRS.2020.3012705.

[4] X. Zhang, G. Wang, P. Zhu, T. Zhang, C. Li, and L. Jiao, "GRS-det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Sep. 4, 2020, doi: 10.1109/TGRS.2020.3018106.

[5] X. Zhang, W. Ma, C. Li, J. Wu, X. Tang, and L. Jiao, "Fully convolutional network-based ensemble method for road extraction from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1777–1781, Oct. 2020.

[6] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[7] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.

[8] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

[9] L. Yan, B. Fan, S. Xiang, and C. Pan, "Adversarial domain adaptation with a domain similarity discriminator for semantic segmentation of urban areas," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1583–1587.

[10] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[11] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[12] L. Yan, B. Fan, H. Liu, C. Huo, S. Xiang, and C. Pan, "Triplet adversarial domain adaptation for pixel-level classification of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3558–3573, May 2020.

[13] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.

[14] Z. Zhang, K. Doi, A. Iwasaki, and G. Xu, "Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training," *IEEE Geosci. Remote Sens. Lett.*, early access, Aug. 4, 2020, doi: 10.1109/LGRS.2020.2982783.

[15] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[16] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6727–6735.

[17] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 2, pp. 59:1–59:35, 2016.

[18] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.

[19] *International Society for Photogrammetry and Remote Sensing 2D Semantic Labeling Challenge*. Accessed: Mar. 2021. [Online]. Available: http://www2.isprs.org/commissions/comm2/ wg4/semantic-labeling

[20] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, pp. 1–14, Jun. 2017.