Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level context memory

Jingru Zhu, Ya Guo, Geng Sun, Libo Yang, Min Deng and Jie Chen, Member, IEEE

Abstract—Semantic segmentation is a key technique involved in automatic interpretation of high-resolution remote sensing (HRS) imagery and has drawn much attention in the remote sensing community. Deep convolutional neural networks (DCNNs) have been successfully applied to the HRS imagery semantic segmentation task due to their hierarchical representation ability. However, the heavy dependency on a large number of training data with dense annotation and the sensitiveness to the variation of data distribution severely restrict the potential application of DCNNs for the semantic segmentation of HRS imagery. This study proposes a novel unsupervised domain adaptation semantic segmentation network (MemoryAdaptNet) for the semantic segmentation of HRS imagery. MemoryAdaptNet constructs an output space adversarial learning scheme to bridge the domain distribution discrepancy between source domain and target domain and to narrow the influence of domain shift. Specifically, we embed an invariant feature memory module to store invariant domain-level context information because the features obtained from adversarial learning only tend to represent the variant feature of current limited inputs. This module is integrated by a category attention-driven invariant domain-level context aggregation module to current pseudo invariant feature for further augmenting the pixel representations. An entropy-based pseudo label filtering strategy is used to update the memory module with high-confident pseudo invariant feature of current target images. Extensive experiments under three cross-domain tasks indicate that our proposed MemoryAdaptNet is remarkably superior to the state-of-the-art methods.

Index Terms—Unsupervised domain adaptation, high-resolution remote sensing (HRS) imagery, semantic segmentation, invariant domain-level context, memory module, category attention, pseudo label filtering strategy.

I. INTRODUCTION

RIVEN by the rapid growth of Earth observation technology, large amounts of remote sensing imagery with high spatial resolution are increasingly available, which makes Earth observation possible. Automatic interpretation of high-resolution remote sensing (HRS) imagery plays a vital role in the field of remote sensing analysis, such as urban planning, intelligent transportation, agricultural production, and natural disaster monitoring. Specifically, semantic segmentation model is an important tool for the automatic interpretation of HRS imagery. Compared with remote sensing scene classification task that assigns a subject category to an image scene, it aims to assign a land cover category (such as building, tree, etc.) to each pixel in the image. Methods based on feature description have been proposed to classify remote sensing images pixel by pixel by investigating the spatial-spectral features [1-3]. However, these traditional methods depend heavily on hand-crafted feature and experts’ experience, which fail to fully represent the detailed and structural information driven by the gradually increased spatial resolution. In the past 10 years, deep convolutional neural networks (DCNNs) have achieved great success in remote sensing image automatic interpretation tasks, such as scene classification [4, 5], image caption [6, 7], object detection [8, 9], and semantic segmentation [10, 11], due to their excellent capability in exhibiting representations and high-level features. A fully convolutional network (FCN) [12] is a classic semantic segmentation tool based on the DCNN structure. Some studies, including [13, 14], have applied FCNs to remote sensing imagery semantic segmentation task and drawn much attention. More complex encoder-decoder frameworks with skip connection, such as SegNet [15], U-Net [16], PSPNet [17], and Deeplab [18], have been used to HRS imagery semantic segmentation.

However, some major problems are found in applying DCNNs to the semantic segmentation of HRS imagery. (1) The superior performance of DCNN-based semantic segmentation models relies heavily on a massive number of high-quality training samples with dense annotation. Although many HRS semantic segmentation datasets with dense annotation are available in the community, they are often limited to a certain area or an application. Manually annotating the dense semantic label is labor intensive and time consuming. Overall, the insufficient training data limit the availability of DCNN-based semantic segmentation models for HRS imagery. (2) The DCNN-based semantic segmentation methods prefer to be particularly sensitive to the distribution variation. The appearance and structural characteristics of HRS imagery vary because of the diverse imaging conditions, including difference in geolocations, imaging sensors, and observational illumination, as shown in Fig. 1. This condition makes the deep

Manuscript received ; revised ; accepted ; date of publication ; date of current version. This work was supported by the National Key Research and Development Program of China (Grant 2020YFA0713503) and the National Natural Science Foundation of China (Grant 42071427). (Corresponding author: Jie Chen.)

J. R. Zhu, Y. Guo, G. Sun, L. B. Yang, M. Deng, and J. Chen are with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China (e-mail: cj2011@csu.edu.cn mailto:dengmin@csu.edu.cn).
models trained on a certain HRS dataset with annotated labels invalid when dealing with images obtained under different imaging conditions. Therefore, finding an accurate and efficient semantic segmentation method for HRS imagery with the ability to deal with domain gap is crucial.

![Vahingen (IR-RG)Potsdam (IR-RG) Potsdam (R-G-B)](image)

**Fig. 1.** Cross-domain remote sensing imagery.

Unsupervised domain adaptation (UDA) is developed to bridge the domain discrepancy between the source and target domains. UDA can transfer knowledge learned from source domain with dense annotation to unlabeled target domain by reducing the shift of domain distributions. Traditional UDA methods in semantic segmentation learn domain-invariant representations by minimizing the domain distribution discrepancy between the source and target domains. Maximum mean discrepancy (MMD) [19], central moment discrepancy (CMD) [20], correlation alignment [21], and Wasserstein distance [22] are the common discrepancy measures to reduce the cross-domain distribution discrepancy. Recently, adversarial learning [23] has raised great attention in the fields of computer vision and has been applied to the UDA semantic segmentation. In accordance with the levels of data that generate adversarial loss, adversarial adaptation methods can be divided into pixel-level adversarial adaptation methods, feature-level adversarial adaptation methods, and output-level adversarial adaptation methods. The pixel-level adversarial adaptation methods usually implement generative adversarial network-based image translation model to transfer the source domain images into target domain-styled images and then realize the semantic segmentation model on the transferred images [24-26]. They can bridge the domain discrepancy between the source and target domains in terms of image style, including illumination, color, and texture. However, they cannot narrow the domain differences in the content structure. Their excellent performance relies on a successful image translation model. The feature-level adversarial adaptation methods focus on the alignment of source and target domains at the feature space through adversarial learning based on the feature representations of input [27, 28]. However, adversarial learning at high-dimension feature space may be unstable and unreliable due to the heterogeneity and noninterpretability of high-dimension feature. The output-level adaptation methods adopt adversarial learning in the output space [29-31]. The features from the source and target domains can maintain the semantic consistency through a minimax game between segmentation network and discriminator.

Although the above UDA methods based on adversarial learning have achieved impressive performance in semantic segmentation, they are mainly designed for natural images [24, 25, 29, 30]. Given that some differences are found between the remote sensing and natural images in terms of imaging angle, observation subject, scene complexity, and spatial layout, the above methods cannot be directly applied to the cross-domain semantic segmentation of HRS imagery. The domain adaptation semantic segmentation method based on adversarial learning needs to cut the large-size source domain and target domain images into small-size patches, which are inputted into the model in the form of batch (1 ≤ batch size ≪ dataset size) for training. However, the distribution of these batch sizes of source or target images cannot represent the data distribution of the entire source or target domain. In other words, the network propagates the gradient based on the adversarial loss generated by these batch images, which causes the extracted features only represent the invariant features of current inputs (batch images of source or target domain) rather than the domain invariant features, which we call pseudo invariant features. Theoretically, a certain bias and variance is observed between pseudo invariant features and domain invariant features, which makes the knowledge learned in the source domain cannot be transferred to the target domain to the greatest extent. We aim that domain adaptation models based on adversarial learning can represent domain invariant features based on all data in the two domains rather than the pseudo invariant features. Therefore, how to obtain domain invariant features from pseudo invariant features is extremely important for cross-domain semantic segmentation of HRS imagery.

Considering these issues, we develop a novel UDA semantic segmentation method for HRS imagery by integrating invariant domain-level context information. Given that the output-level adaptation methods can ensure the semantic consistency of features by conducting adversarial learning in the output space, we use an output-level-based adversarial strategy to reduce the distribution difference in the source domain and target domain images, and to obtain the pseudo invariant feature with semantic consistency. As illustrated in Fig. 2, we use a feature memory module to store the pseudo invariant features of historical source images and target images and dynamically update them to obtain the domain invariant feature in category wise form because a certain bias and variance is found between the pseudo invariant feature and the domain invariant feature. We use an invariant domain-level context aggregation module to adaptively integrate the domain invariant feature in feature memory module and the current pseudo invariant feature. Current pseudo invariant feature representations are augmented by integrating the domain invariant feature representations, which enables the model to maximize the transferability of knowledge learned from the source domain to the target domain, thereby enhancing the performance of semantic segmentation model on target HRS imagery. Our main contributions can be summarized as follows:
1) This study develops an invariant domain feature memory module (IFMM) to integrate invariant domain-level context information for the UDA semantic segmentation of HRS imagery, where the invariant domain feature memory module can be dynamically updated by the pseudo invariant features of historical source images and target images to obtain the invariant domain-level presentation. Concretely, invariant domain-level presentation is stored classwise for the alignment of category-level joint distribution. To the best of our knowledge, this study is the first to apply a memory module to the UDA semantic segmentation of HRS imagery.

2) This study introduces a category attention-driven invariant domain-level context aggregation module to better integrate the invariant domain-level context information to current pseudo invariant feature for enhancing the feature representations of current input images.

3) This study proposes an entropy-based pseudo label filtering strategy to better utilize the pseudo invariant feature of current target images for dynamically updating the IFMM. Specifically, it reserves the prediction in which its entropy is less than or equal to the threshold.

The remainder of this study is organized as follows. Section II summarizes the related work to our study. Section III introduces the proposed method in detail. Section IV discusses the experiment settings, comparative analysis of experimental results, ablation experiments, parameter sensitivity analysis of entropy threshold, and the effects of different data augmentation strategies. Section V provides the conclusion and future research directions.

II. RELATED WORK

A. UDA Semantic Segmentation in the Computer Vision Field

UDA can narrow the distribution shift between different data domains, in which we denote the labeled dataset as source domain, and the unlabeled dataset with different distributions from source domain as target domain. Traditional UDA semantic segmentation methods reduced the domain shift by minimizing the domain distribution discrepancy between the source and target domains. Deep adaptation network [19] used the multiple kernel variant of MMD (MK-MMD) to jointly maximize the difference of source and target domains. As the extension of MMD, [20] explored a CMD to learn the domain-invariant representation by matching the higher order central moments of probability distribution. [22] designed a sliced Wasserstein discrepancy to reduce the marginal distributions between the output of task-specific classifier. As another choice of UDA methods, adversarial learning bridges the domain shift by forcing the generator to produce features or predictions that confuse the discriminator. CyCADA [24] transferred source training data into the target domain by adversarial learning in pixel-level to bridge the low-level discrepancy between the source and target domains. DCAN [25] leveraged channel wise alignment to bridge the domain discrepancy at the pixel-level and feature-level. [26] utilized a bidirectional learning framework to promote the image translation model and segmentation adaptation model simultaneously, which can gradually reduce the domain gap. These UDA methods pursue the alignment of source and target domains at the pixel-level, which can bridge the domain discrepancy between the source and target domains in image style, including color illumination and texture. Some adversarial-based UDA methods pursued the adaptation between the source and target domains at the feature space so that the aligned features can generalize to the two domains. [27] used a convolutional domain adversarial training technique to align the distribution of source and target domains at the global-level and category-level. [28] proposed a domain adversarial learning framework joining global and class-specific measures for performing cross-city semantic segmentation task. Considering the heterogeneity and noninterpretability of adversarial-based domain adaptation
methods at feature space, the output-level adversarial adaptation methods employed adversarial learning in output space via a minimax game between semantic segmentation network and discriminator. Considering that the outputs generated by semantic segmentation model share a massive number of similarities in spatial layout and local context between the source and target domains, [29] employed output space-based adversarial learning at different feature levels for domain adaptation. [30] used a category-level adversarial learning framework on the output space to enforce the local semantic consistency and global alignment. [31] introduced an adversarial training method based on the entropy of semantic predictions to address the UDA semantic segmentation from the source domain to target domain. The output-level adversarial adaptation methods can ensure the semantic similarity and consistency of source domain and target domain features by adversarial learning in the output space.

Although the adversarial learning-based UDA methods have achieved good performance in the semantic segmentation of natural images, they cannot be directly applied to the UDA semantic segmentation task of HRS imagery because they ignore the special properties of HRS imagery in terms of content complexity, image resolution, and spatial layout.

B. UDA Semantic Segmentation in Remote Sensing

UDA has been proposed to remote sensing for bridging the data shift due to the differences in imaging sensor, geographic location, and atmospheric conditions. Previous UDA semantic segmentation studies of remote sensing mainly focused on scene classification task. [32] considered the denoising autoencoders and domain-adversarial neural networks to learn domain-invariant feature representations and applied them to the UDA scene classification task of hyperspectral and multispectral images. [33] used a subspace alignment method based on DCNN to deal with the UDA task in remote sensing image scene classification. [34] presented a correlation subspace dynamic distribution alignment framework to narrow the distribution difference between the source and target domains for the cross-domain scene classification of remote sensing image. In recent years, increasing researchers have focused on the UDA semantic segmentation task of remote sensing images. [35] designed an appearance adaptation approach with semantic consistency for the UDA semantic segmentation of remote sensing images. To deal with the domain shift issue in road segmentation, [36] presented a stagewise domain adaptation model to align the feature of source and target domains via the interdomain adaptation based on the generative adversarial networks. To learn an excellent semantic segmentation network for UDA remote sensing image semantic segmentation task, [37] exploited an objective function under multiple weakly supervised constraints to minimize the disadvantageous effects of data distribution shift between source and target domains. [38] designed a category-certainty attention to adaptively deal with the unadapted regions and categories for the UDA semantic segmentation of HRS imagery. The input size of the existing UDA methods is usually limited to the batch size (1 ≤ batch size ≤ dataset size) level, which makes the features extracted from the network pseudo invariant features rather than the true domain invariant features, so that the knowledge learned in the source domain cannot be transferred to the target domain to the maximum extent. Therefore, how to obtain the domain invariant features from pseudo invariant features needs to be further studied on the cross-domain HRS imagery semantic segmentation task.

C. Memory Module

Memory module enables the DCNNs to store variables and data beyond a small local range. It has been widely used in various fields of computer vision, such as image captioning [39], video object detection [40], video object segmentation [41], and semantic segmentation [42, 43]. [39] exploited the memory as a context repository of prior knowledge and attached previously generated words to memory for seizing the long-term dependency for personalized image captioning. For video object detection task, MEGA [40] utilized a long range memory module to effectively integrate global and local information, which is vital for recognizing the object in a video. For the semantic segmentation task we focus on in this work, [42] employed a feature memory module to store the dataset-level representations for mining the contextual information more than the current input images. [43] maintained a class wise memory module to yield similar pixel-level feature representations with relevant features from labels data for semi supervised semantic segmentation. In this study, we utilize an effective feature memory module to retain the invariant domain-level representations in category wise form to enable the segmentation network in seizing the contextual information more than the current input images and enhance the performance on the target domain image. To the best of our knowledge, this work is the first to employ the feature memory module to the UDA semantic segmentation of HRS imagery.

III. METHODOLOGY

We concentrate on the UDA semantic segmentation task of HRS imagery. Given source domain data \( X_s \) with \( C_K \)-class pixel-level labels \( Y_s \in \{0, 1\}^{H \times W} \), where \( H \) and \( W \) denote the height and width of image data, and \( C_K \) represents the number of categories. Our goal is to learn a semantic segmentation model that well performs on unlabeled target domain data \( X_T \) by transferring the knowledge learned from the labeled source domain data \( X_s \) and \( Y_s \).

As shown in Fig. 3, our proposed MemoryAdaptNet makes up a feature domain adaptation (FDA) branch and an invariant domain-level memory aggregation (IDMA) branch. It uses the FDA branch to adapt the features between the source and target domain images through adversarial learning in the output space, which is presented in Section III-A. In IDMA, we use a memory module to store each pseudo invariant feature of the current input generated by FDA and dynamically update them to obtain invariant domain-level representation, which is integrated to the current feature representations for further refining the prediction of the final class probability distribution, as introduced in Section III-B. When updating the invariant
domain-level presentation in the IDMA branch with pseudo invariant feature of current target images, we use the entropy-based pseudo label filtering strategy to provide a high-confidence pseudo label for the target domain images, which contribute to the calculation of each target category representation for invariant domain-level presentation. The details are provided in Section III-C.

A. FDA

Domain adaptation algorithms based on the adaptation in output space are one of the effective methods to solve the domain shift between two domains, which can preserve the semantic consistency of source domain and target domain features by performing adversarial learning in the output space. On this basis, we construct an adversarial learning scheme on the output space to align the features between the source and target domains in our proposed framework. Intuitively, the overview of FDA branch is visually shown in Fig. 4.

As shown in Fig. 4, the FDA contains a feature extractor $F$, a classifier $C_1$, and a discriminator $D$, where $F$ and $C_1$ make up the segmentation network $G_1$. $F$ extracts features from source and target images, and can be any CNN-based network. To realize a high-quality semantic segmentation performance, we utilize the ResNet-101 [44] pretrained on ImageNet [45] as our backbone for feature extractor $F$. It composes 33 residual blocks, each of which has 3 convolutional layers with a skip connection. We apply an atrous convolution layer with rate = 2 to replace the ordinary convolution layer with kernel $3 \times 3$ in the last residual blocks for extracting dense feature maps and capturing long range context. The detailed structure of the applied ResNet-101 is shown in Fig. 5 (a); $C_1$ classifies the features yielded from $F$ into predefined semantic categories pixel by pixel, such as building, tree, and car. Fig. 5 (b) shows the detailed structure of classifier $C_1$. It consists of an atrous spatial pyramid pooling layer with dilations of 6, 12, and 18 to effectively capture multi-scale information in HRS imagery and two convolution-BatchNorm-ReLU layers to obtain the pixel-level prediction with predefined semantic categories. We utilize an interpolation algorithm to fit the size of pixel-level prediction and input image; $D$ attempts to distinguish whether the input comes from the source domain or target domain, which encourages $F$ to generate feature distributions with semantic consistency between the source domain and the target domain. As shown in Fig. 5 (c), $D$ is constructed by five $4 \times 4$ convolution layers with stride of 2, and its channel number is [64, 128, 256, 512, 1].

B. IDMA

The domain adaptation semantic segmentation method based on adversarial learning normally propagating gradient in each iteration is only based on the adversarial loss generated by the current batch size images at each iteration update, causing the extracted features of feature extractor $F$ to represent the invariant features of the current inputs, which cannot represent the invariant features of the two domains. Therefore, the invariant features obtained through the FDA branch belong to pseudo invariant features, preventing the knowledge learned in the source domain to be transferred to the target domain to the greatest extent.

The memory module can enable the DCNNs to depict the valid information beyond the current input image. Contrary to previous memory modules [46, 47] that saved image-level features, classwise memory modules used in [42, 43] have
shown their advantage in semantic segmentation task by storing per-pixel features in a class wise form. In this study, the class-wise memory module is taken to convert the pseudo invariant feature obtained from the FDA branch to domain invariant feature. Specifically, an IDMA branch is embedded into the FDA branch. As shown in Fig. 3, it makes up an invariant feature memory module \( M \), an invariant domain-level context aggregation module \( A \), and a classifier \( C_2 \), where \( F \), \( M \) and \( A \) make up the segmentation network \( G_2 \). \( M \) is used to store the pseudo invariant feature of current input images and update them dynamically in real time. Given that \( M \) integrated the pseudo invariant feature of all source and target domain images during the dynamic update, the feature stored in \( M \) can approximate the invariant features of source and target domains. Specifically, \( M \) stored the feature in a class-wise format to enhance the intraclass consistency and interclass difference of HRS imagery. The invariant domain-level context aggregation module \( A \) is designed to integrate the invariant domain-level context information in \( M \) to the current feature for enhancing the feature representations of current input images. We use classifier \( C_2 \) to classify the enhanced features into a precise pixel-level prediction with predefined semantic classes. The details of \( M \) and \( A \) are presented in Sections III-B-1 and III-B-2, respectively.

1) **Invariant Feature Memory Module**

Invariant feature memory module \( M \) stores the invariant domain-level representations in category format to enable the network depict invariant-feature information more than the current input images. \( M \) is a data vector \( \mathcal{M} \) with size \( C' \times C_k \times 1 \), where \( C' \) is the vector size of each category. During training, same as [42], we first calculate the average value for each category from one randomly feature generated by \( F \) to initialize \( \mathcal{M} \). \( \mathcal{M} \) is then updated on every training iteration with a pseudo invariant feature subset of \( F_p = (f_p^i, f_p^j, f_p^k) \in R^{C'\times HW} \) generated by \( F \). Specifically, given the current pseudo invariant feature \( f_p \), we first permute it as size \( HW \times C \), that is, \( R^{HW \times C} \). Subsequently, we calculate the representation of each category \( c_k \) existing in \( f_p \)

\[
R_{c_k} = [R^{HW \times C}|GT = c_k] \in R^{N_{c_k} \times C},
\]

where \( R_{c_k} \) is the feature representations of each category \( c_k \), \( GT \in R^{HW} \) denotes the ground truth with category labels, and \( N_{c_k} \) is the number of pixels labeled as \( c_k \) in \( f_p \). We calculate the similarity \( S_{c_k} \) between \( R_{c_k} \) and memory of each category \( \mathcal{M}_{c_k} \) in two means (cosine similarity and mean)

\[
S_{c_k, \text{cosine similarity}} = \frac{\langle R_{c_k}, \mathcal{M}_{c_k} \rangle}{\| R_{c_k} \|_2 \| \mathcal{M}_{c_k} \|_2},
\]

\[
S_{c_k, \text{mean}} = \frac{1}{N_{c_k}} \sum_{i=1}^{N_{c_k}} R_{c_k}^i,
\]

where \( \mathcal{M}_{c_k} \in 1 \times C' \), and \( \| \|_2 \) represents the \( L_2 \)-norm. We update the representation of \( c_k \) as

\[
R_{c_k}' = \sum_{i=1}^{N_{c_k}} \frac{1-s^i_{c_k}}{\sum_{j=1}^{N_{c_k}} (1-s^j_{c_k})} R_{c_k}^i.
\]

We calculate the updated memory of each category by leveraging moving average

\[
\mathcal{M}_{c_k}' = (1 - m) \cdot \mathcal{M}_{c_k} + m \cdot R_{c_k}',
\]

where \( m \) is the momentum, and we employ polynomial annealing policy to update it

\[
m_t = (1 - \frac{t}{T})^p \cdot (m_0 - \frac{m_0}{100}) + \frac{m_0}{100}, t \in [0, T].
\]

where \( T \) and \( T' \) represent the current and total number of training iterations, respectively. \( p \) and \( m_0 \) are default parameters and set to 0.9.

| Feature extractor \( F \) | Classifier \( C_1 \) | Discriminator \( D \) |
|-------------------------|-------------------|-------------------|
| Conv1 \( \times 7, 64, \text{stride 2} \) | Conv1 \( \times 3, 256, \text{rate 12} \) | Conv1 \( \times 4 \times 4, 64, \text{stride 2} \) |
| Max pool \( 3 \times 3 \text{ max pool, stride 2} \) | Concat, Conv 1 \( \times 1, 256 \) | Conv2 \( \times 4 \times 4, 128, \text{stride 2} \) |
| Conv2 \( \times 3, 1 \times 1, 256 \) | ASPP \( \times 4 \times 4 \times 3 \) | Conv4 \( \times 4 \times 4, 256, \text{stride 2} \) |
| Conv3 \( \times 1 \times 1, 256 \) | Conv 1 \( \times 1 \), Batchnorm2D, 256, Red.U | Conv4 \( \times 4 \times 4, 512, \text{stride 2} \) |
| Conv4 \( \times 1 \times 1, 256 \) | Conv4 \( \times 1 \times 1, 1, \text{stride 2} \) | Classifier |
2) Category Attention-driven Invariant Domain-level Context Aggregation Module

Invariant domain-level context aggregation module aims to adaptively integrate invariant domain-level context stored in $M$ to current feature representations and obtain an enhanced feature. Given that the feature memory in $M$ is stored in the form of categories, we design a category attention mechanism to realize the feature aggregation, and Fig. 6 shows the detailed structure. Given the memory feature $M \in R^{C \times CK \times 1}$ stored in memory module and current pseudo invariant feature $F_p = \{f_p| f_{pi}\} \in R^{C \times HW}$, where $C$ denotes the number of channels in $F_p$. As shown in Fig. 6, three 1x1 convolutional layers are applied to them and generate three features $Q \in R^{C \times C_K \times 1}$, $K \in R^{C \times C_K \times 1}$, and $V \in R^{C \times HW}$, respectively. Subsequently, we matrix multiply the reshaped feature $K$ and feature $V$. A softmax function is utilized to compute the affinity attention map $S \in R^{HW \times CK}$ on each category between the current pseudo invariant feature and memory feature

$$s_{ij} = \frac{\exp(k_i v_j)}{\sum_{k=1}^{C} \exp(k_i v_j)},$$

where $s_{ij} \in S$ measures the correlation between the $i^{th}$ category memory of feature $K$ and $j^{th}$ channel of feature $V$, $i = \{1, 2, \ldots, C\}$, $j = \{1, 2, \ldots, C\}$ . We weighted the affinity attention map $S$ and feature $Q$ to obtain the memory feature $M'$s attention map $S'$ on the current pseudo invariant feature $F_p$ and reshape it to $R^{C \times HW}$. The current pseudo invariant feature $F_p$ and the attention map $S'$ are concatenated to acquire the enhanced feature $F_T$ as follows:

$$F_T = \theta(cat(\varphi(S'), F_p)),$$

where $\varphi(\cdot)$ and $\theta(\cdot)$ represent the mapping functions implemented by 1x1 convolution-BatchNorm-ReLU layer. The enhanced feature $F_T$ can adaptively enhance the presentation of current pseudo invariant features by calculating the correlation between the memory feature and original pseudo invariant features. The segmentation network can seize the contextual information more than the current input images, which can enhance the performance of semantic segmentation network on the target domain HRS imagery.

Fig. 6. Structure of category attention-driven invariant domain-level context aggregation module.

C. Entropy-based Pseudo Label Filtering

In accordance with Section III-B-1, the ground truth is required to calculate the feature representations of each category. However, no ground truth of target domain images is found in the UDA semantic segmentation task of HRS imagery. If the category feature representations of the target domain images are not reserved in the memory module, the memory feature is more inclined to represent the context information of the source domain images and fails to depict the feature distribution of the target domain images, which is detrimental to the semantic segmentation of target domain images. Therefore, the memory module with the pseudo invariant features of the target domain images must be updated. We can use the segmentation predictions of target images as pseudo labels to update the memory module. However, not all pixels in the segmentation predictions are high-confidence, so directly using the segmentation predictions as pseudo labels of target images is unreasonable. On the basis of the observation in Fig. 7, the predictions produced by the model trained only on the source domain have high confidence, low entropy on source images and low confidence, high entropy on target images. Therefore, we propose an entropy-based pseudo label filtering strategy to provide a high-confidence pseudo label for the target domain image. It can assist the updating of target image invariant features in the memory module. This strategy can also make the memory module better represent the invariant domain features, so as to realize superior segmentation performance on the target domain image.

Given the target image set $X_T = \{x_{ti}\}_{i=1}^{N_T}$, where $N_T$ is the size of target image set. We place each image $x_{ti}$ to MemoryAdaptNet and obtain a softmax prediction set $P_T = \{p_{ti} \in R^{H \times W \times CK}\}_{i=1}^{N_T}$ generated by classifier $C_2$, where

$$P_T^{(h,w)} = \{P_{T_{c1}}, P_{T_{c2}}, \ldots, P_{T_{cK}}\};$$

$P_T^{(h,w)}$ represents the probability that pixel $(h , w)$ in $p_{ti}$ belongs to class $c_k, c_k \in \{1, 2, \ldots, C\}$, and $C_K$ is the number of classes.

Fig. 7. Entropy map of pseudo labels.

We use the probability value to determine whether a pixel is associated with a label, so we calculate the highest probability value of $P_T^{(h,w)}$ by $\mu = \max(p_{T_{cK}}^{(h,w)})$ and assign the index of $\mu$ as the category of pixel $(h , w)$ of $p_{ti}$. The predicted classification label matrix $l_{i,t}$ can be obtained as

$$l_{i,t} \in R^{H \times W}, l_{i,t} = index(\mu).$$

Given that the pixels of $l_{i,t}$ are not all high confidence, we use the entropy map of $l_{i,t}$ to retain the high-confidence pixels and remove the low-confidence pixels. Given the predicted
classification probability matrix $p_{t,l}$. The entropy map $E_t \in [0,1]^{H \times W}$ is composed of independent pixelwise entropies

$$E_{t}^{(h,w)} = -\frac{1}{\log(C)} \sum_{k=1}^{C} p_{t,l}^{(h,w,k)} \log p_{t,l}^{(h,w,k)}.$$  \hspace{1cm} (11)

where $E_{t}^{(h,w)}$ represents the entropy at pixel $(h,w)$ of $E_t$. Specifically, high $E_{t}^{(h,w)}$ represents low confidence of $l_{t,l}$, whereas low $E_{t}^{(h,w)}$ represents high confidence of $l_{t,l}$, so we set a threshold $\sigma$ of $E_{t}^{(h,w)}$ to select the $l_{t,l}$ with high confidence $l_{t,l}^* = \{H \times W | E_t \leq \sigma \}$.  \hspace{1cm} (12)

In accordance with Eq. 12, the pseudo label with entropy value greater than the threshold $\sigma$ is discarded, whereas the pseudo label with entropy value less than or equal to the threshold $\sigma$ is reserved.

Owing to the entropy-based pseudo label filtering strategy, we can obtain high-confidence pseudo labels $l_{t,l}^*$ of $x_t$, which makes up the pseudo label set $GT_t = \{l_{t,l}^* \}_l \subseteq T$. $GT_t$ and $GT_s$ make up the ground truth set $GT$, which is used to update the invariant feature memory module.

D. Network Training

The proposed MemoryAdaptNet model is optimized by two training steps. In step 1, we only train the FDA branch in some epochs for reducing the domain shift between the source and target domains, and obtaining the pseudo invariant feature. Subsequently, we simultaneously optimize the FDA branch and IDMA branch in step 2 to enhance the performance of semantic segmentation network on the HRS imagery.

1) Step 1: FDA Branch Training

In the first $\tau$ iterations, we only train the FDA branch. Given a source domain image $x_s \in X_S$ and a target domain image $x_t \in X_T$, we forward them to $F$ and obtain the feature maps $f_s$ and $f_t$, which are inputted to classifiers $C_1$ to acquire the pixel-level predictions $p_s$ and $p_t$, respectively. $p_s$ is used to compute a segmentation loss under the supervision of ground truth $y_s \in Y_S$ for optimizing $G_1$. In this study, we adopt the cross-entropy loss function as the segmentation loss, which is shown as Eq. (13)

$$L_{seg_1}(X_S, Y_S) = -E_{(x_s,y_s)} \sum_{i=1}^{Y_S} y_s(i) \log G_1(x_s)(i),$$  \hspace{1cm} (13)

where $L_{seg_1}(X_S, Y_S)$ is the segmentation loss, $x_s \in X_S$ represents the images from source domain, and $y_s \in Y_S$ represents the ground truth corresponding to $x_s$.

In addition to the segmentation loss, we forward $p_t$ to $D$ to yield an adversarial loss for optimizing $G_1$. The adversarial loss can be defined as follows:

$$L_{adv}(X_T) = -E_{x_t \sim p_T(x)}[\log D(G_1(x_t))].$$  \hspace{1cm} (14)

where $L_{adv}(X_T)$ represents the adversarial loss generated by target images, and $P_T(x)$ represents the distributions of target domains. The network propagates gradients from $D$ to $G_1$, which encourages $F$ to extract the feature distributions with semantic consistency between the target domain and the source domain.

In summary, the training objective for the segmentation network $G_1$ can be extended from Eqs. (13) and (14) as

$$L(X_S, X_T) = L_{seg_1}(X_S, Y_S) + \lambda_{adv} L_{adv}(X_T).$$  \hspace{1cm} (15)

where $\lambda_{adv}$ is the weight used to balance the adversarial loss $L_{adv}(X_T)$.

We then optimize $D$ with the pixel-level prediction $p_s$ and $p_t$. Specifically, we forward $p_s$ and $p_t$ to $D$ to output a single scalar, which denotes the probability that the input came from the source domain rather than the target domain (i.e., label 0 for target training sample and 1 for source training sample). The single scalar is used to compute a cross-entropy loss $L_d$ for optimizing $D$. The loss can be written as

$$L_d(X_S, X_T) = -E_{x \sim p_s(x)}[\log D(G_1(X_S))] - E_{x \sim p_t(x)}[\log (1 - D(G_1(X_T))].$$  \hspace{1cm} (16)

2) Step 2: FDA Branch and IDMA Branch Training

After $\tau$ iterations, we simultaneously optimize the FDA and IDMA branches. The optimization process of FDA branch is the same as Section III-D-1. For the optimization of IDMA branch, we first use the source feature $f_s$ and the target feature $f_t$ obtained from feature extractor $F$ to update the invariant feature memory module $M$. We then forward $f_s$ and $M$ in $M$ to invariant domain-level context aggregation module $A$ for integrating the invariant domain-level context information in $M$ to the current pseudo invariant feature $f_t$, thereby obtaining the enhanced feature of source images. The enhanced feature is inputted to classifiers $C_2$ to generate the pixel-level prediction $p'_t$, which is used to compute the segmentation loss by cross-entropy loss function as follows:

$$L_{seg_2}(X_S, Y_S) = -E_{(x_s,y_s)} \sum_{i=1}^{Y_S} y_s(i) \log C_2(A(F(x_s), M)).$$  \hspace{1cm} (17)

In summary, the total semantic segmentation loss of MemoryAdaptNet can be expressed as

$$L_{seg}(X_S, Y_S) = L_{seg_1}(X_S, Y_S) + L_{seg_2}(X_S, Y_S).$$  \hspace{1cm} (18)

The whole optimization process of MemoryAdaptNet is summarized in Algorithm I for better understanding.

Algorithm I Optimization process of MemoryAdaptNet

Input: Source image $x_s \in X_S$ and target image $x_t \in X_T$;
Output: The predicted label $p'_t$ of source image $x_s$ and the predicted label $p'_t$ of target image $x_t$:

1: for $\tau = 0; \tau < \tau' ; \tau++$ do
2: Use source image $x_s$ and target image $x_t$ to optimize the FDA branch with Eqs. 15 and 16;
3: end for

4: for $\tau = \tau' ; \tau < \tau'' ; \tau++$ do
5: Simultaneously optimize the FDA branch and IDMA branch;
6: Using the source feature $f_s$ and the target feature $f_t$ from feature extractor $F$ to update the memory value in invariant feature memory module $M$;
7: Use source image $x_s$ and target image $x_t$ to optimize the FDA branch with Eqs. 15 and 16;
8: Use the source feature $f_s$ from feature extractor $F$ to optimize the IDMA branch with Eq. 17;
9: end for
IV. Experiments and Result Analysis

In this section, we first present the experimental settings in Section IV-A, including the description of datasets, task settings, implementation details, and evaluation metrics. We conduct comparison experiments with existing UDA semantic segmentation methods to verify the effectiveness of our proposed network in Section IV-B. In Section IV-C, we perform an ablation study to demonstrate the effectiveness of each module in our MemoryAdaptNet. We conduct parameter sensitivity analysis to select the appropriate \( \sigma \) value for entropy-based pseudo label filtering strategy in Section IV-D. We discuss the effects of different data augmentation strategies on the UDA semantic segmentation of HRS imagery in Section IV-E.

A. Experimental Settings

1) Datasets

To demonstrate the importance of MemoryAdaptNet on the UDA semantic segmentation of HRS imagery, we perform our experiments on two very high-resolution datasets: Vaihingen 2D dataset and Potsdam 2D dataset supplied by the International Society for Photogrammetry and Remote Sensing (ISPRS) WG II/4 [48]. All images in the two datasets are provided with semantic labels, which consist of six general land cover categories: Imp. surf. (impervious surfaces), building (buildings), low veg. (low vegetation), tree (trees), car (cars), and clutter (clutter/background). In the training process, we do not employ the semantic labels of the target domain images.

Vaihingen dataset: The Vaihingen dataset draws a small village with sparse layout pattern. It contains 33 high-resolution true orthophoto tiles (TOPs) with an average size of 2,494 \( \times \) 2,064 pixels and a spatial resolution of 9 cm. As shown in Fig. 8 (a), the near-infrared (IR), red (R), and green (G) channels are provided in this dataset. We employ ID: 2, 5, 7, 8, 13, 20, 22, 24 for testing, and the remaining 25 images for training and validation.

Potsdam dataset: The Potsdam dataset depicts a city scene with crowded residential pattern. It contains 38 TOPs with a fixed size of 6,000 \( \times \) 6,000 and a spatial resolution of 5 cm. The dataset provides near-IR, red (R), green (G), and blue (B) channels, which are combined into three channel compositions: [IR, R, G], [R, G, B], and [IR, R, G, B]. We employ ID: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, 7_13 for testing, and the remaining 24 images for training and validation. As shown in Figs. 8 (b) and (c), we employ the [IR, R, G] and [R, G, B] channel compositions in our experiments.

Fig. 9 represents the pixel percentage of each category in the Vaihingen and Potsdam training datasets. Concretely, the pixel proportions of imp. surf., building, low veg., tree, car, and clutter in the Potsdam training datasets are 28.46\%, 26.72\%, 23.54\%, 14.62\%, 1.69\%, and 4.96\%, respectively. The pixel proportions of imp. surf., building, low veg., tree, car, and clutter in the Vaihingen training datasets are 28.69\%, 26.70\%, 20.56\%, 21.91\%, 1.29\%, and 0.86\%, respectively. The two datasets have the characteristics of unbalanced category samples. For example, the proportion of car and cluster categories in the two datasets is remarkably lower than other categories.

2) Task Settings

We set up three UDA semantic segmentation tasks in our experiments: (1) P2V_S task, where the Potsdam dataset with [IR, R, G] channel composition serves as the source domain dataset, and the Vaihingen dataset with [IR, R, G] channel composition serves as the target domain dataset. (2) P2V_D task, where the Potsdam with [R, G, B] channel composition serves as the source domain dataset, and the Vaihingen dataset with [IR, R, G] channel composition serves as the target domain dataset. (3) V2P task, where the Vaihingen dataset with [IR, R, G] channel composition serves as the source domain dataset, and the Potsdam dataset with [IR, R, G] channel composition serves as the target domain dataset. In the training phase, we crop the Potsdam ([IR, R, G] and [R, G, B] channel compositions) datasets and their paired labels into 512 \( \times \) 512 size images with horizontal and vertical strides of 256 pixels, and yield a total of 13,310 images. With regard to the Vaihingen-[IR, R, G] dataset, we crop the images and their paired labels into 512 \( \times \) 512 size patches with horizontal and vertical strides of 256 pixels and yield approximately 1,700 images. In P2V_S and P2V_D tasks, we utilize the Potsdam training set for training and the Vaihingen testing and validation sets for testing and validation. In V2P task, we utilize the Vaihingen training set for training and the Potsdam testing and validation sets for testing and validation.

3) Implementation Details
We implement our network using the PyTorch toolbox on a single RTX 3090 GPU with 24 GB memory. To train the semantic segmentation networks $G_1$ and $G_2$, we employ the stochastic gradient descent optimizer [49] with the initial learning rate of $2.5 \times 10^{-4}$, momentum of 0.9, and weight decay of $10^{-4}$. For the discriminator training, we employ the Adam optimizer [50] with the initial learning rate of $10^{-4}$ and momentum of 0.9 and 0.99.

4) Evaluation Metric

We employ four general evaluation metrics, namely, $F_1$-score, OA, MA, and mIoU [12], to assess the performance of different UDA semantic segmentation methods. The higher the values of these metrics, the better the semantic segmentation performance.

$F_1$-score is a common evaluation metric applied to the semantic segmentation task. It can be expressed as follows:

$$ F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} $$

(19)

where the Precision and Recall can be computed as follows:

$$ \text{Precision} = \frac{TP}{TP + FP}, $$

(20)

$$ \text{Recall} = \frac{TP}{TP + FN}, $$

(21)

where $TP$ denotes the number of true positive pixels, $FP$ denotes the number of false positive pixels, and $FN$ denotes the number of false negatives.

OA represents the overall accuracy of semantic segmentation performance, MA represents the mean accuracy of semantic segmentation performance, and mIoU represents the intersection over union on each category of predicted labels. They can be expressed as follows:

$$ \text{OA} = \frac{\sum_{c} n_{c,i} / \sum_{c} \sum_{i} n_{c,i}}{\sum_{c} \sum_{i} n_{c,i}}, $$

(22)

$$ \text{MA} = (1/C) \sum_{c} \left( \frac{n_{c,i}}{\sum_{c} n_{c,i}} \right), $$

(23)

$$ \text{mIoU} = (1/C) \sum_{c} \left( \frac{\sum_{i} n_{c,i}}{\sum_{c} \sum_{i} n_{c,i} + \sum_{c} n_{c,i} - n_{c,i}} \right), $$

(24)

where $n_{c,i}$ denotes the number of pixels of category $c_i$ predicted as category $c_j$, $n_{c,i}$ denotes the number of pixels of category $c_j$ predicted as category $c_i$.

B. Comparative Experimental Result Analysis

To confirm the effectiveness of MemoryAdaptNet, we conduct the comparative experiment on the P2V_S task, where the Potsdam dataset with [IR, R, G] channel composition serves as the source domain and the Vaihingen dataset with [IR, R, G] channel composition as the target domain. The domain shift in this task is mainly found in geographic location, spatial resolution, and illumination. The quantitative evaluation results of the compared domain adaptation methods on this domain shift are presented in Table I.

From Table I, the source-only model achieves the worst performance with OA, MA, and mIoU values of 67.19%, 60.55%, and 43.58%, respectively. The performance of all UDA semantic segmentation methods is improved after some domain adaptations. In particular, the proposed MemoryAdaptNet achieves the highest OA, MA, and mIoU values of 77.87%, 77.05%, and 56.05%, respectively, compared with other state-of-the-art methods. Among all the comparison models, MCD_DA obtains the worst performance with OA, MA, and mIoU values of 69.86%, 51.70%, and 41.51%, respectively, which shows the performance reduction by 8.01% on OA, 25.35% on MA, and 14.54% on mIoU compared with our MemoryAdaptNet. AdvEnt obtains the second highest performance with OA, MA, and mIoU values of 74.43%, 65.29%, and 51.37%, respectively, which presents the performance reduction by 3.44%, 11.76%, and 4.68% on OA, MA, and mIoU compared with our MemoryAdaptNet. These performance differences demonstrate that our MemoryAdaptNet can better deal with the domain discrepancy on P2V_S semantic segmentation task.

In Table I, the $F_1$ scores of MemoryAdaptNet in all the categories, including imp. surf., building, low veg., tree, car, and clutter, realize the best accuracies of 79.65%, 86.45%, 68.22%, 78.52%, 61.09%, and 50.30%, respectively, compared with those of other models, which bring the improvement of 1.96%–3.97% on imp. surf., 0.46%–6.15% on building, 7.25%–15.29% on low veg., 4.12%–8.04% on tree, 6.81%–19.81% on car, and 19.79%–40.41% on clutter, respectively. Specifically, our method has a great $F_1$ score improvement on low veg., car, and clutter compared with other models that all have a negative transfer on clutter. This result shows that our method can relieve the effect of sample category imbalance on the UDA semantic segmentation of HRS imagery by using the effective and representative prototype features of each category stored in the memory module.

Fig. 10 shows the qualitative results on the P2V_S task, where the segmentation results of source-only model show great difference due to the serious domain shift problem. The segmentation performance is improved to varying degrees after adaptation. However, the semantic segmentation performance
Table I: Quantitative Evaluation Results (%) of Different UDA Models on the P2V_S Task

| Methods          | Imp. surf | Building | Low veg | Tree | Car | Clutter | OA  | MA  | mIoU |
|------------------|-----------|----------|---------|------|-----|---------|-----|-----|------|
| Source-only      | 72.85     | 74.40    | 56.42   | 67.96| 46.76| 28.49   | 67.19| 60.55| 43.58|
| MCD_DA           | 77.00     | 80.30    | 54.94   | 70.48| 54.28| nan     | nan | 57.66| 46.63|
| State-of-the-art |           |          |         |      |     |         |     |     |      |
| methods          | 75.68     | 83.88    | 60.76   | 74.40| 41.28| 9.89    | 71.62| 58.37| 44.02|
| AdaptSegNet      | 77.69     | 85.99    | 60.97   | 74.06| 52.01| 26.43   | 73.70| 64.00| 48.65|
| AdvEnt           | 77.55     | 84.42    | 52.93   | 74.23| 53.98| 30.51   | 74.43| 65.29| 51.37|
| MemoryAdaptSegNet| 79.65     | 86.45    | 68.22   | 78.52| 61.09| 50.30   | 77.87| 77.05| 56.05|

Fig. 10. Qualitative results on the P2V_S task.

The segmentation boundary of ground objects is rough. Compared with other domain adaptation methods, our MemoryAdaptNet achieves the best semantic segmentation performance, in which the segmentation performance for each category and the problems of semantic errors and unsmoothed boundary are improved.

1) Comparative Studies on P2V_D Task

In this section, we implement the comparative experiment on the P2V_D task, where the Potsdam dataset with [R, G, B] channel composition serves as the source domain and the Vaihingen dataset with [IR, R, G] channel composition as the target domain. In addition to the difference in geographical...
location, spatial resolution, and illumination, the domain shift in this task includes the difference in imaging sensors, which has a larger domain gap than the P2V_S task. The quantitative evaluation results of the compared domain adaptation methods on this domain shift are presented in Table II.

From Table II, the source-only model achieves the worst performance with OA, MA, and mIoU values of 57.38%, 48.07%, and 33.58%, respectively. The performance of semantic segmentation is improved after some domain adaptations. As expected, our method reaches the highest OA, MA, and mIoU values of 69.96%, 70.18%, and 46.22%, respectively, which obtains the improvement of 12.58%, 22.11%, and 12.64% on OA, MA, and mIoU compared with the source-only model. Our proposed method achieves the highest

---

**Table II**

Quantitative Evaluation Results (%) of different UDA models on the P2V_D task

| Methods              | Imp. surf. | Building | Low veg. | Tree | Car | Clutter | OA   | MA    | mIoU   |
|----------------------|------------|----------|----------|------|-----|---------|------|-------|--------|
| Source-only          | 64.91      | 63.58    | 43.54    | 63.97| 40.44| 7.94    | 57.38| 48.07 | 33.58  |
| MCD_DA               | 69.75      | 75.46    | 51.62    | 50.85| 35.05| 0.00    | 61.75| 48.30 | 36.52  |
| CLAN                 | 67.67      | 77.04    | 47.26    | 58.96| 26.76| 8.43    | 62.94| 50.35 | 35.96  |
| AdaptSegNet          | 68.31      | 78.23    | 46.27    | 63.73| 40.44| 8.09    | 63.13| 51.08 | 37.55  |
| AdvEnt               | 69.49      | 75.87    | 32.96    | 73.72| 49.79| 6.39    | 62.04| 52.80 | 37.64  |
| MemoryAdaptSegNet    | **77.21**  | **83.12**| **51.98**| **58.86**| **23.30**| **59.96**| **70.18**| **46.22** |

Fig. 11. Qualitative results on the P2V_D task.
$F_1$ score in the categories of imp. surf., building, low veg., car, and clutter with the values of 77.21%, 83.12%, 51.98%, 58.86%, and 23.30%, which provides an improvement of 12.3%, 19.54%, 8.44%, 18.42%, and 15.36%, respectively, compared with the source-only model. These results demonstrate that our MemoryAdaptNet can deal well with the domain shift in terms of geographic location, spatial resolution, and imaging sensor.

Fig. 11 represents the qualitative results of compared domain adaptation methods on the P2V_D task, where the source-only model achieves the worst semantic segmentation performance. Although the semantic segmentation performance of the four state-of-the-art methods is improved after domain adaptation, it still has some shortcomings in the aspects of unsmooth boundary and semantic confusion. The offset of the input limited data prevents the extracted features to represent the true domain invariant feature, which results in the knowledge learned in the source domain cannot be transferred to the target domain to the greatest extent. Our MemoryAdaptNet represents the domain variant features by updating the memory module with the historical pseudo invariant features of source domain and target domain images, enabling the classifier to classify by learning the invariant domain-level presentation and the feature presentation of the current input image, thereby enhancing the performance of semantic segmentation network on the target HRS images. Fig. 11 confirms the effectiveness of our MemoryAdaptNet on the P2V_D task.

2) Comparative Studies on V2P Task

We conduct the comparative experiment on the V2P task, where the Vaihingen dataset with [IR, R, G] channel composition serves as the source domain, and the Potsdam dataset with [IR, R, G] channel composition as the target domain. The domain shift in this task is the same as the P2V_S task. The quantitative and qualitative evaluation results of the compared domain adaptation methods on this task are presented in Table III and Fig. 12.

From Table III, the source-only model obtains the worst performance with OA, MA, and mIoU values of 64.36%, 54.05%, and 41.22%, respectively. The performance of the state-of-the-art methods and our method is improved to a certain extent after adaptation. Our MemoryAdaptNet reaches the highest OA, MA, and mIoU values of 72.25%, 61.87%, and 49.82%, respectively, which obtains the improvement of 7.89%, 7.82%, and 8.60% on OA, MA, and mIoU compared with the source-only model, respectively. In Fig. 12, our MemoryAdaptNet achieves the best semantic segmentation performance with more accurate semantic information and smoother boundaries compared with other domain adaptation methods. These results demonstrate that our proposed MemoryAdaptNet can well deal with the V2P task.

C. Ablation Experiment

Our proposed MemoryAdaptNet benefits from the FDA and IDMA modules. We conduct four different ablation experiments (source-only, FDA, FDA + IDMA_cs, FDA + IDMA_mn) to evaluate the contribution of the MemoryAdaptNet’s components to model performance. Source-only represents the segmentation network only trained on the source domain and is directly tested on the target domain. FDA represents the domain adaptation method that only uses the FDA branch to construct adversarial learning on the output space to align the feature between two domains. FDA + IDMA represents the model embedded with IDMA branch on the basis of FDA branch. In accordance with the calculation of the similarity between the feature representations and memory of each category when updating the value of invariant feature memory module, we divide the FDA + IDMA to FDA + IDMA_cs and FDA + IDMA_mn, where FDA + IDMA_cs represents the use of cosine similarity to calculate the similarity between the feature representation and memory of each category, FDA + IDMA_mn represents the use of the mean of feature representations of each category as the similarity between the feature representations and memory of each category. The experiments are conducted on the P2V_S domain adaptation task, and the corresponding results are shown in Table IV.

As shown in Table IV, employing the FDA branch yields a result of 70.51%, 65.27%, and 47.60% in OA, MA, and mIoU, respectively, compared with the source-only model, which brings 3.32%, 4.72%, and 4.02% improvement, respectively. This finding verifies the effect of adversarial learning in the output space for UDA semantic segmentation. The model embedded with IDMA branch on the basis of FDA branch outperforms the FDA model in terms of OA, MA, and mIoU. Specifically, FDA + IDMA_mn yields a result of 73.57%, 68.51%, and 49.90% in OA, MA, and mIoU, respectively, which brings 3.06%, 3.24%, and 2.30% improvement compared with the FDA model, respectively. FDA + IDMA_cs yields a result of 77.87%, 77.05%, and 56.05% in OA, MA, and mIoU.

### Table III

| Methods              | Imp. surf. | Building | Low veg. | Tree | Car | Clutter | OA   | MA   | mIoU |
|----------------------|------------|----------|----------|------|-----|---------|------|------|------|
| Source-only          | 71.32      | 69.45    | 60.97    | 54.67| 69.10| 5.47    | 64.36| 54.05| 41.22|
| MCD DA               | 75.67      | 76.72    | 63.53    | 39.97| 60.68| 0.00    | 67.04| 56.42| 43.71|
| CLAN                 | 75.94      | 79.08    | 64.44    | 49.17| 66.91| 6.77    | 67.87| 57.10| 44.12|
| AdaptSegNet          | 76.55      | 77.93    | 62.15    | 45.45| 70.89| 6.16    | 68.93| 57.33| 44.94|
| AdvEnt               | 76.65      | **82.64**| 63.73    | 32.94| 71.74| 2.76    | 69.59| 58.03| 45.43|
| MemoryAdaptSegNet    | **79.70**  | 77.84    | **67.58**| **64.06**| **76.54**| **16.15**| **72.25**| **61.87**| **49.82**|

As shown in Table IV, employing the FDA branch yields a result of 70.51%, 65.27%, and 47.60% in OA, MA, and mIoU, respectively, compared with the source-only model, which brings 3.32%, 4.72%, and 4.02% improvement, respectively. This finding verifies the effect of adversarial learning in the output space for UDA semantic segmentation. The model embedded with IDMA branch on the basis of FDA branch outperforms the FDA model in terms of OA, MA, and mIoU. Specifically, FDA + IDMA_mn yields a result of 73.57%, 68.51%, and 49.90% in OA, MA, and mIoU, respectively, which brings 3.06%, 3.24%, and 2.30% improvement compared with the FDA model, respectively. FDA + IDMA_cs yields a result of 77.87%, 77.05%, and 56.05% in OA, MA, and mIoU.
mIoU, respectively, which brings 7.36%, 11.78%, and 8.45% improvement compared with the FDA model, respectively. These improvements benefit from the invariant feature memory module in our method that can integrate the pseudo invariant features of historical images to obtain category-level invariant domain-level features. Compared with the pseudo invariant features extracted from FDA, this feature can better represent the domain invariant features and maximize the transfer of knowledge learned in the source domain to the target domain. At the same time, FDA + IDMA_cs shows 4.30%, 8.54%, and 6.15% improvements in OA, MA, and mIoU compared with the FDA + IDMA_mn model, respectively. The use of cosine similarity to calculate the correlation between the current pseudo invariant feature and the domain invariant feature in the invariant feature memory module can reduce the intraclass differences in the domain invariant feature, which is conducive to the formation of the category prototype feature in the invariant domain. Therefore, MemoryAdaptNet chooses the cosine similarity to measure the correlation between the current pseudo invariant feature and the domain invariant feature in the

### TABLE IV

**ABLAISON STUDY RESULTS (%) ON THE P2V_S TASK**

| Methods       | Imp. surf | Building | $F_1$ score | Low veg. | Tree | Car | Clutter | OA   | MA   | mIoU  |
|---------------|-----------|----------|-------------|----------|------|-----|---------|------|------|-------|
| Source-only   | 72.85     | 74.40    | 56.42       | 67.96    | 46.76| 28.49| 67.19   | 60.55| 43.58|
| FDA           | 74.77     | 78.29    | 61.18       | 64.98    | 53.66| 45.06| 70.51   | 65.27| 47.60|
| FDA + IDMA_mn | 79.27     | 82.04    | 63.53       | 68.25    | 51.27| 47.64| 73.57   | 68.51| 49.90|
| FDA + IDMA_cs | 79.65     | 86.45    | 68.22       | 78.52    | 61.09| 50.30| 77.87   | 77.05| 56.05|

**Fig. 12.** Qualitative results on the V2P task.
invariant feature memory module for the update of the memory value.

D. Analysis on the Entropy Threshold of the Target Pseudo Label

To investigate how the entropy threshold $\sigma$ of target pseudo label affect the proposed entropy-based pseudo label filtering strategy, we assess the performance of our proposed MemoryAdaptNet with different thresholds $\sigma$ on the P2V_S task, P2V_D task, and V2P task. The results of different thresholds $\sigma$ on the three tasks are shown in Tables V, VI, and VII.

For the threshold $\sigma$, $\sigma = 0$ represents that the pseudo label of target images does not participate in the update of invariant feature memory module; $\sigma = 1$ represents that the pseudo label of target images all are reserved for updating the invariant feature memory module without being filtered. When $0 < \sigma < 1$, the higher the $\sigma$ value, the more area of the pseudo label are reserved. $\sigma = 0.5$ represents that the pseudo label areas with entropy $\leq 0.5$ are reserved for the update of the invariant feature memory module, and the pseudo label areas with $0.5 < \sigma$ are filtered. From Table V, MemoryAdaptNet obtains the lowest score with OA of 75.07\%, MA of 76.04\%, and mIoU of 52.71\% when the threshold $\sigma = 0$. This finding is because when the pseudo label of target images does not participate in the update of invariant feature memory module, the invariant feature stored in the invariant feature memory module is more inclined to represent the source domain data distribution and cannot consider the target domain data distribution, resulting in poor segmentation performance of the model on the target domain. When the pseudo label of target domain participates in the update of invariant feature memory module ($\sigma > 0$), the segmentation performance of the model is improved. The simultaneous update of the pseudo invariant feature from the source and target domains enables the invariant feature stored in invariant feature memory module to represent the invariant data distribution of source and target domains. Similar experimental phenomena appear in the P2V_D task and V2P task, as shown in Tables VI and VII. When the threshold $\sigma > 0$, MemoryAdaptNet obtains the highest score with OA of 77.87\%, MA of 77.05\%, and mIoU of 56.05\% under the threshold $\sigma$ of 0.5 on the P2V_S task. For the other tasks, MemoryAdaptNet obtains the highest score with OA of 69.96\%, MA of 70.18\%, and mIoU of 46.22\% under the threshold $\sigma$ of 0.7 on the P2V_D task and obtains the highest score with OA of 72.25\%, MA of 61.87\%, and mIoU of 49.82\% under the threshold $\sigma$ of 0.5 on the V2P task.

E. Effects of Different Data Augmentation Strategies

In this section, we discuss the effects of different data augmentation strategies on the UDA semantic segmentation of HRS imagery. We conducted three experiments with different data augmentation strategies (w/o aug., affine aug., color-space aug.) on the domain adaptation task from P2V_S and the domain adaptation task from P2V_D. The corresponding results can be found in Table VIII. In Table VIII, the w/o aug. represents that no data augmentation strategy is utilized during the MemoryAdaptNet training. The affine aug. represents that the data augmentation strategy of affine transformation, such as horizontal flip, vertical flip, random rotate, and shift scale rotate, is used during the MemoryAdaptNet training. The color-space aug. represents the data augmentation strategy based on the color-space transformation, such as Gaussian noise, Gaussian blur, random brightness, and random contrast, is used during the MemoryAdaptNet training.

As shown in Table VIII, the w/o aug. model achieves the performance of 74.04\%, 68.78\%, and 50.47\% in OA, MA, and mIoU on the P2V_S task and the performance of 61.76\%, 61.98\%, and 39.78\% in OA, MA, and mIoU on the P2V_D task, respectively. Compared with the w/o aug. model, the performance of the color-space aug. model decreases to varying degrees on the two domain adaptation tasks. Specifically, the OA, MA, and mIoU on the P2V_S task decreased by 3.39\%, 0.99\%, and 3.17\%, respectively, and the OA, MA, and mIoU on the P2V_D task decreased by 2.48\%, 12.61\%, and 4.18\%, respectively, which indicates that the data augmentation strategy based on color-space transformation has a negative effect on the domain adaptation.

### Table V

| $\sigma$ | Imp. surf. | Building | $F_0$ score | Tree | Car | Clutter | OA | MA | mIoU |
|--------|-----------|----------|-------------|------|-----|---------|----|----|------|
| 0.0    | 77.76     | 84.75    | 62.42       | 77.23| 60.90| 36.67   | 75.07| 76.04| 52.71|
| 0.3    | 78.57     | 85.91    | 66.50       | 78.00| 61.55| 52.26   | 76.27| 75.92| 55.35|
| 0.5    | 79.65     | 86.45    | 68.22       | 78.52| 61.09| 50.30   | 77.87| 77.65| 56.05|
| 0.7    | 78.85     | 85.97    | 66.65       | 77.48| 61.48| 46.56   | 76.76| 76.87| 54.74|
| 1.0    | 78.20     | 86.07    | 65.99       | 77.42| 59.41| 49.57   | 76.81| 76.70| 54.56|

### Table VI

| $\sigma$ | Imp. surf. | Building | $F_0$ score | Tree | Car | Clutter | OA | MA | mIoU |
|--------|-----------|----------|-------------|------|-----|---------|----|----|------|
| 0.0    | 74.96     | 78.55    | 51.59       | 67.25| 57.51| 19.88   | 66.82| 67.74| 43.57|
| 0.3    | 72.07     | 81.67    | 42.05       | 71.02| 58.82| 28.00   | 67.84| 68.57| 44.44|
| 0.5    | 72.80     | 79.74    | 47.84       | 70.38| 58.12| 26.52   | 67.77| 67.39| 44.25|
| 0.7    | 77.21     | 83.12    | 51.98       | 71.53| 58.86| 23.30   | 69.96| 70.18| 46.22|
| 1.0    | 73.79     | 82.15    | 54.48       | 69.44| 56.37| 23.31   | 68.08| 68.76| 45.21|
TABLE VII
PARAMETER ANALYSIS OF THE TARGET PSEUDO LABEL ENTROPY THRESHOLD ON THE V2P TASK

| σ   | Imp. sur. | Building | F_1 score | Low veg. | Tree | Car | Clutter | OA  | MA  | mIoU |
|-----|-----------|----------|------------|----------|------|-----|---------|-----|-----|------|
| 0.0 | 76.96     | 77.19    | 64.78      | 56.35    | 75.87| 8.97| 69.41   | 59.38| 46.39|
| 0.3 | 80.42     | 82.13    | 66.87      | 63.43    | 72.55| 7.72| 71.44   | 60.83| 49.09|
| 0.5 | 79.70     | 77.84    | 67.58      | 64.06    | 76.54| 16.15| 72.25   | 61.87| 49.82|
| 0.7 | 79.69     | 77.67    | 61.47      | 61.90    | 77.97| 10.53| 71.45   | 61.20| 48.78|
| 1.0 | 78.60     | 78.94    | 67.10      | 63.55    | 75.02| 6.57| 72.01   | 60.89| 48.40|

TABLE VIII
QUANTITATIVE RESULTS (%) OF DIFFERENT DATA AUGMENTATION STRATEGIES ON THE P2V_S TASK AND P2V_D TASK

| Methods          | Imp. sur. | Building | F_1 score | Low veg. | Tree | Car | Clutter | OA  | MA  | mIoU |
|------------------|-----------|----------|------------|----------|------|-----|---------|-----|-----|------|
| P2V_S task       | w/o aug.  | 76.67    | 80.04      | 64.21    | 64.44| 48.04| 74.74   | 74.04| 68.78| 50.47|
|                  | color-space aug. | 75.68    | 80.19      | 58.70    | 60.48| 52.65| 36.86   | 70.65| 67.79| 47.30|
|                  | affine aug. | 79.65    | 86.45      | 68.22    | 78.52| 61.09| 50.30   | 77.87| 70.05| 56.05|
| P2V_D task       | w/o aug.  | 68.51    | 76.13      | 51.07    | 46.83| 51.87| 46.95   | 61.76| 61.98| 39.78|
|                  | color-space aug. | 69.49    | 76.29      | 48.53    | 49.34| 47.66| 5.05    | 59.28| 49.37| 35.60|
|                  | affine aug. | 77.21    | 83.12      | 51.98    | 71.53| 58.86| 23.50   | 69.96| 70.18| 46.22|

V. CONCLUSION AND FUTURE WORK

In this study, we propose a novel UDA method MemoryAdaptNet to deal with the cross-domain semantic segmentation of HRS imagery. Given that the memory module can capture context information beyond the current inputs, the MemoryAdaptNet employs an invariant feature memory module to store invariant domain-level context information by the update of current pseudo invariant feature. It is integrated by a category attention driven invariant domain-level context aggregation module to current pseudo invariant feature for further augmenting the pixel representations, which enables the model to classify by considering the invariant domain-level presentation and the current feature presentation, thereby enhancing the performance of semantic segmentation network on the target HRS imagery. The results of extensive experiments indicate the effectiveness of our MemoryAdaptNet, which favorably outperforms the baseline model and the state-of-the-art models.

In the future, we aim to further improve the MemoryAdaptNet from the following aspects: 1) performing adversarial learning at the output space with different feature levels by considering that different level features of DCNN excavate different semantic and other detailed information, and 2) exploring a pseudo label filtering strategy based on adaptive entropy threshold rather than manually setting the entropy threshold for reducing the manual intervention and workload.

REFERENCES

[1] Umaselvi, M., S.S. Kumar, and M. Athithya. Color based urban and agricultural land classification by GLCM texture features. in Proceedings of Iet Chennai International on Sustainable Energy & Intelligent Systems. 2012. IET. p. 192-195.
[2] Yang, Y. and S. Newsam. Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery. in 2008 15th IEEE international conference on image processing. 2008. IEEE. p. 1852-1855.
[3] Yang, Y. and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. in Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. 2010. p. 270-279.
[4] Cheng, G., J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE, 2017. 105(10): p. 1865-1883.
[5] Hu, F., et al., Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sensing, 2015. 7(11): p. 14680-14707.
[6] Chen, J., et al., Geospatial relation captioning for high-spatial-resolution images by using an attention-based neural network. International Journal of Remote Sensing, 2019. 40(16): p. 6492-6498.
[7] Zhao, R., Z. Shi, and Z. Zou, High-resolution remote sensing image captioning based on structured attention. IEEE Transactions on Geoscience and Remote Sensing, 2021. 60: p. 1-14.
[8] Chen, J., et al., Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery. IEEE Geoscience and Remote Sensing Letters, 2019. 17(4): p. 681-685.
> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CCLICK HERE TO EDIT) <

[9] Li, K., et al., Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 2020. 159: p. 296-307.

[10] Chen, J., et al., SMAF-Net: Sharing multiscale adversarial feature for high-resolution remote sensing imagery semantic segmentation. IEEE Geoscience and Remote Sensing Letters, 2020. 16(11): p. 1921-1925.

[11] Zhang, J., et al., Multi-scale context aggregation for semantic segmentation of remote sensing images. Remote Sensing, 2020. 12(4): p. 701.

[12] Long, J., E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3431-3440.

[13] Zhao, T., J. Feng, and X. Chen. HF-FCN: Hierarchically fused fully convolutional network for robust building extraction, in Asian Conference on Computer Vision. 2016. Springer. p. 291-302.

[14] Yang, Y., et al., M-FCN: Effective fully convolutional network-based airplane detection framework. IEEE Geoscience and Remote Sensing Letters, 2017. 14(8): p. 1293-1297.

[15] Badrinarayanan, V., A. Kendall, and R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 2017. 39(12): p. 2481-2495.

[16] Ronneberger, O., P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. in International Conference on Medical image computing and computer-assisted intervention. 2015. Springer. p. 234-241.

[17] Zhao, H., et al. Pyramid scene parsing network. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 2881-2890.

[18] Chen, L.-C., et al., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 2017. 40(4): p. 834-848.

[19] Geng, B., D. Tao, and C. Xu. DAML: Domain adaptation metric learning. IEEE Transactions on Image Processing, 2011. 20(10): p. 2980-2989.

[20] Zellinger, W., et al., Central moment discrepancy (cmd) for domain-invariant representation learning. arXiv preprint arXiv:1702.08811, 2017.

[21] Sun, B. and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. in European conference on computer vision. 2016. Springer. p. 443-450.

[22] Lee, C.-Y., et al., Sliced wasserstein discrepancy for unsupervised domain adaptation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 10285-10295.

[23] Goodfellow, I., et al., Generative adversarial nets. Advances in neural information processing systems, 2014. 27.

[24] Hoffman, J., et al., Cycada: Cycle-consistent adversarial domain adaptation. in International conference on machine learning. 2018. PMLR. p. 1989-1998.

[25] Wu, Z., et al. Dual channel-wise alignment networks for unsupervised scene adaptation. in Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 518-534.

[26] Li, Y., L. Yuan, and N. Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 6936-6945.

[27] Hoffman, J., et al., F Cov in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649, 2016.

[28] Chen, Y.-H., et al. No more discrimination: Cross city adaptation of road scene segmenters. in Proceedings of the IEEE International Conference on Computer Vision. 2017. p. 1992-2001.

[29] Tsai, Y.-H., et al. Learning to adapt structured output space for semantic segmentation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 7472-7481.

[30] Luo, Y., et al. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 2507-2516.

[31] Vu, T-H., et al. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 2517-2526.

[32] Elshamli, A., et al., Domain adaptation using representation learning for the classification of remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017. 10(9): p. 4198-4209.

[33] Song, S., et al., Domain adaptation for convolutional neural networks-based remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters, 2019. 16(8): p. 1324-1328.

[34] Zhang, J., et al., Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2020. 58(11): p. 7920-7930.

[35] Wittich, D. and F. Rottensteiner, Appearance based deep domain adaptation for the classification of aerial images. ISPRS Journal of Photogrammetry and Remote Sensing, 2021. 180: p. 82-102.

[36] Zhang, L., et al., Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2021. 60: p. 1-13.

[37] Li, Y., et al., 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, 2021. 175: p. 20-33.

[38] Chen, J., et al., Unsupervised Domain Adaptation for Semantic Segmentation of High-Resolution Remote Sensing Imagery Driven by Category-Certainty Attention. IEEE Transactions on Geoscience and Remote Sensing, 2022. 60: p. 1-15.

[39] Chunseong Park, C., B. Kim, and G. Kim. Attend to you: Personalized image captioning with context sequence memory networks. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 895-903.

[40] Chen, Y., et al., Memory enhanced global-local aggregation for video object detection. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. p. 10337-10346.

[41] Oh, S.W., et al. Video object segmentation using space-time memory networks. in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019. p. 9226-9235.

[42] Jin, Z., et al., Mining contextual information beyond image for semantic segmentation. in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 7231-7241.

[43] Alonso, I., et al. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 8219-8228.

[44] He, K., et al., Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

[45] Deng, J., et al., Imagenet: A large-scale hierarchical image database. in 2009 IEEE conference on computer vision and pattern recognition. 2009. Ieee. p. 248-255.

[46] Wu, Z., et al. Unsupervised feature learning via non-parametric instance discrimination. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 3733-3742.

[47] He, K., et al., Momentum contrast for unsupervised visual representation learning. in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. p. 9729-9738.

[48] Rottensteiner, F., et al., The ISPRS benchmark on urban object classification and 3D building reconstruction. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1-3 (2012), Nr. 1, 2012. 1(1): p. 293-298.

[49] Bottou, L., Large-scale machine learning with stochastic gradient descent, in Proceedings of COMPSTAT’2010. 2010. Springer. p. 177-186.

[50] Kingma, D.P. and J. Ba, Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[51] Saito, K., et al. Maximum classifier discrepancy for unsupervised domain adaptation. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 3723-3732.