Multitarget Domain Adaptation Building Instance Extraction of Remote Sensing Imagery With Domain-Common Approximation Learning

Fayong Zhang, Kejun Liu, Yuanyuan Liu, Chaofan Wang, Wujie Zhou, *Senior Member, IEEE*, Hongyan Zhang, *Senior Member, IEEE*, and Lizhe Wang, *Fellow, IEEE*

***Abstract*— Deep learning-based building instance extraction on remote sensing imagery (RSI) has achieved tremendous success under the large-scale labeled training data. However, multitarget domain adaptation building instance extraction (MD-BIE) is still a challenge task that involves transferring knowledge from a source domain to multiple unlabeled target domains, which poses various semantic gaps between and within multiple domains, e.g., style, illumination, resolution, density, and scale. Most current methods for single-target domain adaptation are not applicable to the more realistic MD-BIE task. To this end, we propose a novel domain-common approximation learning (DAL) for both modeling intradomain and interdomain adaptation, thus obtaining robust MD-BIE. DAL contains three main modules: multidomain style transfer (MST), multidomain feature approximation (MFA), and multidomain cascaded instance extraction (MCIE). To alleviate the semantic gaps between multiple domains for interdomain adaptation, we first employ the MST to learn multiple target-domain-like features that preserve both the styles of target domains and the content of the source domain, and then use the MFA to approximate these features toward a central domain-common space, thus producing domain-common semantic representations. Moreover, we develop the MCIE with hierarchical extraction losses for intradomain adaptation to extract precise building instance contours from the domain-common semantic representations, further eliminating the potential gaps within multiple domains. By colearning these three modules in an end-to-end manner, the DAL bridges the semantic gaps between and within multiple domains. Extensive experiments on different popular MD-BIS tasks (SAB → Crowd & WHU, Crowd → SAB & WHU, SAB → Crowd & SAB & WHU, and SAB → WHU) show that our DAL outperforms the current methods by a significant margin.**

Manuscript received 3 August 2023; revised 24 November 2023; accepted 8 March 2024. Date of publication 13 March 2024; date of current version 26 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62076227, in part by the Natural Science Foundation of Hubei Province Grant 2023AFB572, and in part by Hubei Key Laboratory of Intelligent Geo-Information Processing under Grant KLIGIP-2022-B10. *(Corresponding author: Yuanyuan Liu.)*

Fayong Zhang is with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: zhangfayong@cug.edu.cn).

Kejun Liu, Chaofan Wang, Hongyan Zhang, and Lizhe Wang are with the School of Computer Science, China University of Geosciences, Wuhan 430074, China (e-mail: liukejun@cug.edu.cn; wangchaofan@cug.edu.cn; zhanghongyan@whu.edu.cn; 13646346@qq.com).

Yuanyuan Liu is with the School of Computer Science, China University of Geosciences, Wuhan 430074, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: liuyy@cug.edu.cn).

Wujie Zhou is with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China (e-mail: wujiezhou@163.com).

Digital Object Identifier 10.1109/TGRS.2024.3376719

***Index Terms*— Cascaded instance extraction, multidomain feature approximation (MFA), multidomain style transfer (MST), multitarget domain adaptation building instance extraction (MD-BIE), remote sensing (RS).**

I. Introduction

THE continuous advancements in remote sensing (RS) and artificial intelligence technologies have led to the wide application of building instance extraction from high-resolution RS imagery [1], [2]. The study of building instance segmentation from RS images is important for human exploration and understanding of geography and ecology, especially in urban construction and planning [3], natural disaster and crisis management [4], smart cities [5], and other RS-related applications [6]. Previous methods on building instance extraction mainly focus on hand-crafted feature design and extraction, such as color [7], spectrum [8], edges [9], and texture [10]. However, these well-designed features are prone to bias and failure when dealing with various RS scenarios (i.e., target domains) because of potential semantic gaps in different RS data domains, including the gaps in illumination, conditions, sensor qualities, and scale variation. As a result, these artificial feature-based methods are only applicable to specific data sources [11], [12].

Recently, convolutional neural network (CNN) based-instance segmentation methods have achieved wide investigation on the building instance extraction task [13], [14], [15], [16], [17], [18]. For example, Fang et al. [16] used an attention-based feature pyramid subnetwork and a contour subnetwork to improve the performance of building instance extraction. Wen et al. [17] proposed an improved region CNN that can separate buildings from complex backgrounds and detect rotated bounding boxes. Tian et al. [18] employed an edge constraint-based multiscale U-Net method for building instance segmentation. More recently, with the rise of large vision models, such as segment anything model (SAM) [19], the advantages of large models can be significantly enhanced with an effective instance extraction method by training on a large amount of labeled data. Despite the progress, most of these methods assume that the training set and test set have the same data distribution, i.e., in the same data domain. However, in real-world RS applications, the training data distribution (source domain) usually differs from that of the test data (target domain) due to the potential gaps, called

domain shifts. In other words, we usually encounter a large amount of unlabeled target domain data and a small amount of labeled source domain data. Due to the domain shifts, it is difficult to directly apply these off-the-shelf methods to effectively generalize to unseen target domain data, leading to significant performance degradation.

To alleviate the domain shifts, researchers have developed domain adaptation building extraction methods that explore better generalization of features learned on the source domain to the unlabeled target domain [22], [23], [24], [25], [26]. For example, Peng et al. [24] proposed a novel full-level domain adaptation network (FDANet) for building extraction by combining image-, feature-, and output-level information, effectively. Tasar et al. [25], [26] employed generative adversarial network (GAN) [27] for cross-domain RS semantic segmentation via domain generation. Chen et al. [23] proposed a human memory mechanism-inspired domain adaptation network called MDANet for cross-domain building extraction. Despite the progress, these methods are primarily designed for achieving domain adaptation building extraction in a single target domain, and do not consider more realistic multiple target domain scenarios.

Compared with the single-target domain adaptation task, multitarget domain adaptation building instance extraction (MD-BIE) on RS imagery poses more challenges, making existing single-target domain adaptation methods less effective. Fig. 1 illustrates these challenges and intuitive results in the MD-BIE task. As shown in Fig. 1(a), MD-BIE includes more semantic gaps, such as building styles, densities, resolutions, locations, scales, and conditions, both between and within different domains. For example, *the interdomain semantic gaps* are reflected as: the source domain mainly collects images of Beijing with the dense building distribution, while the target domain 1 collects images of rural New Zealand with a sparse building distribution; meanwhile, the spatial resolution of the target domain 2 is lower compared to the source domain and target domain 1. Moreover, *the potential intradomain semantic gaps* can be also significant, such as the large-scale and illumination variation in different buildings. Current single-target domain adaptation approaches focus only on the gaps between a source and a single target domain (i.e., single interdomain shift), and cannot be sufficient to address inter- and intradomain shifts across the source domain and multiple target domains. It would be necessary to integrate multiple single-target domain adaptation models, respectively, trained on different domains to address multiple domain shifts. However, its main limitations are the excessive capacity of the model, the difficulty of convergence, and the difficulty of reuse when a new target domain is received [28]. Accordingly, these challenges further affect the building instance extraction performance on multiple RS target domains [see Fig. 1(b)].

To address these issues, we propose a novel, unified domain-common approximation learning (DAL) method for effectively modeling interdomain and intradomain adaptation, resulting in robust MD-BIE on RS imagery. More specifically, the method consists of three main modules: the multidomain style transfer (MST) module, the multidomain feature approximation module (MFA), and the multidomain cascaded instance

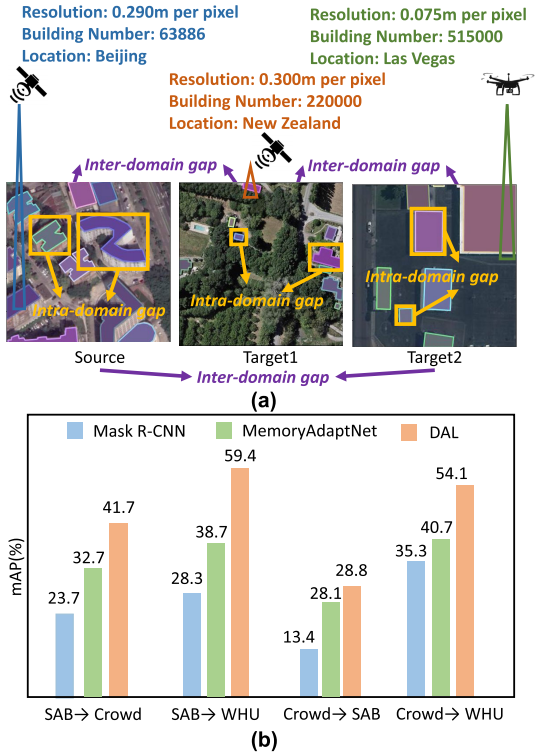

Fig. 1. Motivation of our method. (a) Challenges in the MD-BIE task, i.e., various semantic gaps between and within different domains, including building styles, densities, resolutions, locations, scales, and conditions. (b) Comparison of the results of our DAL, source only (Mask RCNN [20]), and single-target domain adaptation method (MemoryAdaptNet [21]). Compared with source only and single-target domain adaptation methods, our DAL achieved improvements of 18% and 9% on SAB → Crowd, as well as 31.1% and 20.7% on SAB → WHU, respectively.

extraction module (MCIE). For the interdomain adaptation, we first employ the MST to translate the features of source domain to multiple target-domain-like features that retain the styles of target domains and the content of the source domain, and then use the MFA to approximate the obtained features to a central domain-common space by introducing a novel central approximation loss. This process reduces the semantic gaps between multiple domains and facilitates knowledge transfer between multiple target domains, thus obtaining the domain-common semantic representations without interdomain shifts. Furthermore, we further devise the MCIE with hierarchical extraction losses for the intradomain adaptation to effectively extract refined building instance contours on each domain in a coarse-to-fine manner. Overall, by properly training these three components jointly, we can improve the performance of unsupervised building instance extraction in multiple target domains.

In summary, the article presents the following main contributions.

1) We propose an effective unsupervised MD-BIE method, called DAL. DAL is the first unified and effective MD-BIE method on RS imagery by modeling inter- and intradomain adaptation, and can be easily extended to more domains.
2) We employ an MST to generate target-domain-like features and use an MFA with its central approximation loss to learn the domain-common semantic representations

without the domain discrepancy. Both the two modules effectively suppress the semantic gaps between different RS domains for interdomain adaptation.

3) We devise an MCIE with hierarchical extraction losses for intradomain adaptation to further alleviate the semantic gaps within multiple domains, thus obtaining robust building instance extraction in each unlabeled RS target domain.
4) We conduct extensive experiments on three popular RS building datasets, including SAB, Crowd, WHU, and UBC datasets, and different popular MD-BIE tasks, including SAB → Crowd & WHU, Crowd → SAB & WHU, SAB → Crowd & SAB & WHU, and SAB → WHU. The results demonstrate that the proposed DAL method achieves improved performance over existing state-of-the-art methods in overcoming semantic gaps.

II. Related Work

A. Building Instance Extraction

The goal of building instance extraction is to automatically extract the geometric shapes and semantic information of individual building instances from images or point cloud data. Wei et al. [29] proposed a concentric loop CNN (CLP-CNN) method for building instance extraction from RS images. Zhu et al. [30] proposed an adaptive polygon generation algorithm (APGA), which directly parameterized the generated polygon output into a series of building vertices to represent each building instance. Xu et al. [31] proposed a two-stage instance segmentation network for building extraction named gated spatial memory and center of mass perception network (GSMC). The method consists of two modules: a gated spatial memory module (GSM) and a centroid-aware head (CH). He et al. [12] proposed a generative modeling framework by combining satellite images and spatial geometric features to improve the accuracy of building extraction. However, these current methods for building instance extraction focus on a specific data source, and are difficult to generalize to different data domains due to the fact that they do not take into account the semantic differences between different data domains.

B. Multitarget Domain Adaptation

Multitarget domain adaptation aims to transfer the knowledge learned by models in labeled source domains to multiple unlabeled target domains. Wu et al. [32] proposed a vector-decomposed disentanglement (VDD) method, which designs an extractor to separate domain-invariant representations from inputs. VDD promotes more domain-independent information in domain-invariant representations by enlarging the distance between domain-specific representations and domain-invariant representations, making it suitable for multitarget domain adaptation. Isobe et al. [33] proposed a collaborative learning framework for unsupervised multitarget domain adaptation. The framework first trains an expert model for each target domain data and further improves it by adding consistency regularization. Meanwhile, a student model is trained to mimic the output of the expert models, and their weights are regularized to be closer to each other. This method performs well on multiple target domains. Despite the progress, these method mainly address the domain adaptation in natural scenarios rather than the more challenging RS scenarios.

C. Domain Adaptation Building Extraction

Deep learning models have limited generalization ability when dealing with large-scale unlabeled RS data. In order to overcome this problem, recently, the cross-domain building extraction method can be employed. Tasar et al. [28] proposed a method called DAugNet for unsupervised multitarget domain adaptive building extraction. The approach comprises two main components: a classifier and a data augmenter. The data augmenter is capable of performing style transfer between multiple RS datasets in an unsupervised manner, and the classifier provides diverse data to make the model robust in data distribution across different domains. Cui et al. [34] proposed MDANet, which reduces the difference in image data distribution across different domains by projecting different images to the virtual center of the mixed domain. Zhu et al. [21] proposed MemoryAdaptNet for semantic segmentation of RS images. The method works by integrating invariant features obtained using adversarial learning in order to bridge the difference in domain distribution between the source and target domains. Lin et al. [35] proposed Duplex Alignment Networks with enhanced feature discrimination for building extraction, which removed interdomain differences through a pair of GANs and used a specialized classifier to improve the differentiation of extracted features from the target domain. Despite achieving the progress, most current methods for domain adaptation building extraction address the domain adaptation from a source domain to a target domain, overlooking the more realistic multiple RS target domains.

III. Proposed Approach

A. Overview: MD-BIE

The MD-BIE task is formulated as follows: there exist multiple domains, including a labeled source domain and multiple unlabeled target domains. The source domain contains rich labeled data, denoted as $\mathcal{S} = \{(x_i, y_i)|i = 1, \ldots, n^{\mathcal{S}}\}$, where x is a data sample, y is its corresponding mask label of buildings, i indexes the examples in $\mathcal{S}$, and $n^{\mathcal{S}}$ represents the size of $\mathcal{S}$. The multiple target domains, denoted as $T^m = \{(x_j^m)|j = 1, \ldots, n_m^T\}$, contain unlabeled data, where j indexes over the examples in the mth target domain T^m and n_m^T is the size of T^m. In the MD-BIE task, the $\mathcal{S}$ and T^m can differ in various aspects, such as the building style, data distribution, scale, building density, geographical location, and class labels. Therefore, the goal of MD-BIE is to obtain unsupervised building instance extraction in multiple target domains T^m, with great robustness and effectiveness.

For this purpose, we propose a novel DAL approach for the unsupervised MD-BIE on RS imagery. The DAL can effectively address various semantic gaps between and within the source domain and multiple target domains by introducing inter- and intradomain adaptation, leading to robust MD-BIE performance. The network structure of DAL for MD-BIE is

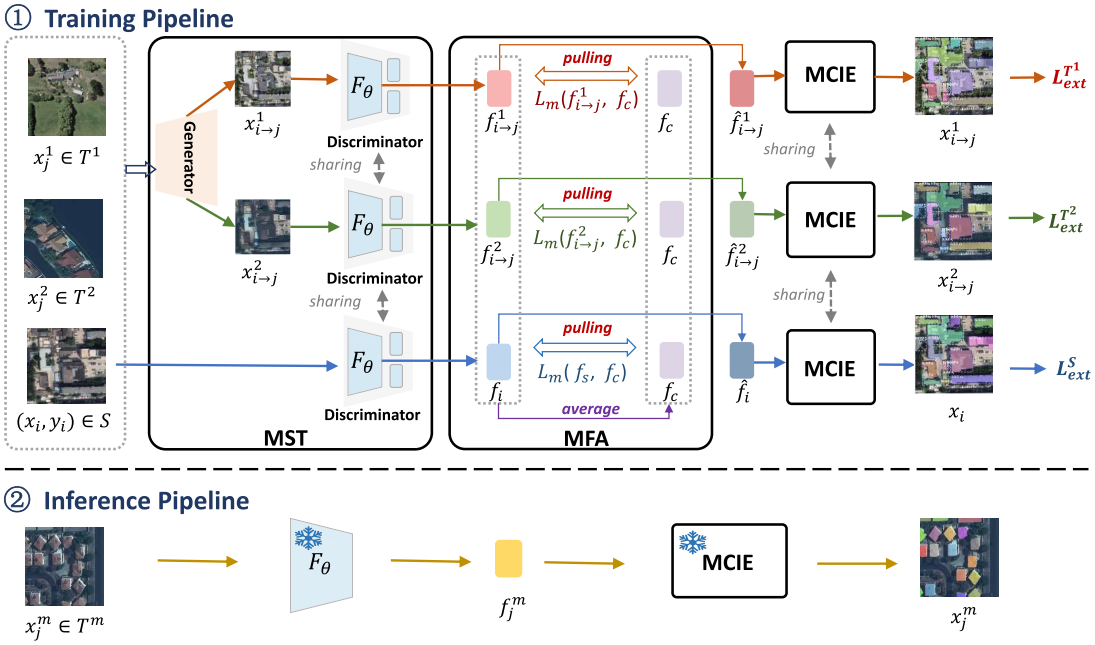

Fig. 2. Architecture of the DAL for MD-BIE. In the training pipeline, with the labeled source domain and unlabeled multiple target domains as inputs, we first use the MST module to generate multiple target-domain-like features, then introduce the MFA module to approximate domain-common semantic representations without domain differences, thus achieving the interdomain adaptation. Moreover, we further employ three parameter-shared MCIE for the intradomain adaptation, resulting in the accurate building instance extraction performance in a coarse-to-fine manner. It is worth noting that our DAL is easily extendable to a wider range of target domains. In the inference pipeline, given the target domain data as input, we first employ the trained discriminator subnetwork F_θ to extract the target domain data features and then use the MCIE module to extract the final building detection results.

illustrated in Fig. 2. Specifically, in the training pipeline, the DAL consists of three main modules: MST, MFA, and MCIE. In DAL, we first employ MST and MFA together for interdomain adaptation, reducing the semantic gaps between different domains to obtain domain-common semantic representations. Then, to alleviate the potential gaps within multiple domains, we further introduce the MCIE module with hierarchical detection losses for intradomain adaptation to extract accurate building contours from coarse to fine. By incorporating these three modules, the DAL approach enables robust inter- and intradomain adaptation in building instance extraction, effectively mitigating the challenges in MD-BIE. In the inference pipeline, taken the target domain data x_j^m as input, we first use the trained discriminator subnetwork F_θ in MST to get the target domain features f_j^m, then employ the trained MCIE module to extract and segment the building masks from the target domain features.

In Sections III-B–III-D, we will subsequently explain the proposed MST, MFA, and MCIE.

B. MST Module

Due to the lack of valid labels for effectively learning multiple target domain knowledge, we first employ the MST module to reduce the label dependency for transferring the source knowledge to unlabeled target domains, thus assisting in modeling interdomain adaptation. As shown in Fig. 3, taking the source image and multiple target images as inputs, the MST aims to translate the input images into target-domain-like features which preserves the source domain contents and the target domain styles, by introducing a typical generator and discriminator-based StarGAN framework [36].

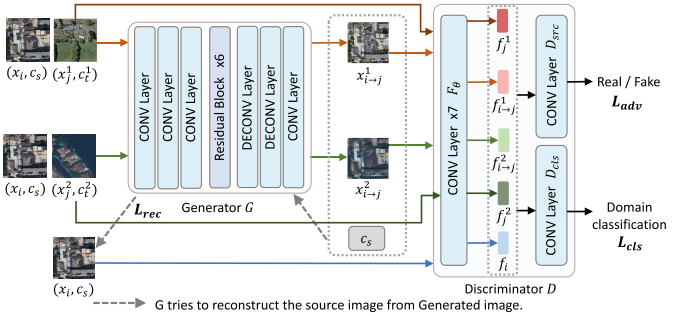

Fig. 3. Pipeline of the MST module. Taking the source image and multiple target images as inputs, the MST outputs the target-domain-like features which preserves the source domain contents and the target domain styles.

More specifically, following the StarGAN [36], given the image pairs $\{x_i, x_j^m\}$ as input, MST employs a typical GAN structure consisting of a generator G and discriminator D to generate the target-domain-like images $x_{i\to j}^m$ and the corresponding target-domain-like features $f_{i\to j}^m$. The generator G contains two convolutional layers for downsampling, six residual blocks, and two transposed convolutional layers for upsampling. Meanwhile, the discriminator D contains a feature extractor that consists of seven convolutional layers, denoted as F_θ, a real-fake image classifier containing one convolutional layer, denoted as D_{src}, and a domain classifier containing one convolutional layer, denoted as D_{cls}. Through the proper learning of MST, the $x_{i\to j}^m$ is consistent with the source domain image in content and the target domain images in style, and the target-domain-like features $f_{i\to j}^m$ can be given by $f_{i\to j}^m = F_\theta(x_{i\to j}^m)$. Similarly, we can also obtain the source domain features $f_i = F_\theta(x_i)$.

During training, an adversarial loss is first introduced into the D_{src} to make the generated target-domain-like features indistinguishable from real target domain features. Mathematically, the adversarial loss can be written as

$$L_{\text{adv}} = \frac{1}{n^{\mathcal{S}}}\sum_{i}\mathbb{E}_{f_i}\left[\log D_{\text{src}}(f_i)\right] + \sum_{m=1}^{M}\frac{1}{n^{\mathcal{S}}}\sum_{i}\mathbb{E}_{f^m_{i\to j}}\left[\log\left(1-D_{\text{src}}\left(f^m_{i\to j}\right)\right)\right] \quad (1)$$

where M represents the number of target domains and m indicates the target domain index. Then, to classify the generated features into the corresponding domain, we introduce a domain classification loss L_{cls} of real images (source domain images, target domain images) and the generated target-domain-like images. Formally, the L_{cls} is defined as

$$L_{\text{cls}} = \frac{1}{n^{\mathcal{S}}}\sum_{i}\mathbb{E}_{f_i,c_s}\left[-\log D_{\text{cls}}(c_s|f_i)\right] + \sum_{m=1}^{M}\frac{1}{n^{T}_{m}}\sum_{j}\mathbb{E}_{f^m_j,c^m_t}\left[-\log D_{\text{cls}}\left(c^m_t|f^m_j\right)\right] + \sum_{m=1}^{M}\frac{1}{n^{\mathcal{S}}}\sum_{i}\mathbb{E}_{f^m_{i\to j},c^m_t}\left[-\log D_{\text{cls}}\left(c^m_t|f^m_{i\to j}\right)\right] \quad (2)$$

where c_s and c^m_t are represented as the image features coming from the source domain and the mth target domain, respectively. This process is used to distinguish which domain an image feature comes from.

In addition, to guarantee the generated target-domain-like images preserve the content of its source images, we introduce a cycle consistency loss L_{rec} to the generator G to achieve this. Mathematically, the L_{rec} is defined as

$$L_{\text{rec}} = \sum_{m=1}^{M}\frac{1}{n^{\mathcal{S}}}\sum_{i}\mathbb{E}_{x_i,x^m_{i\to j},c_s}[\|x_i - G(x^m_{i\to j},c_s))\|_1]. \quad (3)$$

Finally, the total loss L_{MST} of the MST can be written as

$$L_{\text{MST}} = L_{\text{adv}} + L_{\text{cls}} + L_{\text{rec}}. \quad (4)$$

Overall, through the adversarial training in MST, the generator G learns to generate specific target-domain-like images, while the discriminator D tries to distinguish the generator images coming from which domain and learn the target-domain-like features. As a result, the generated target-domain-like features $f^m_{i\to j}$ have the consistent feature distributions of the target domains while remaining the contents and labels of the source domain.

C. MFA Module

With the target-domain-like features $f^m_{i\to j}$ and source features f_i, there still exist feature distribution differences in feature space, as shown in Fig. 4, which affects the performance of MD-BIE. To address this, the DAL further introduces the MFA module with the central approximation loss to approximate the $f^m_{i\to j}$, f_i toward a domain-common central feature space, by pulling the semantic distances between different domain features. This process aligns the features of

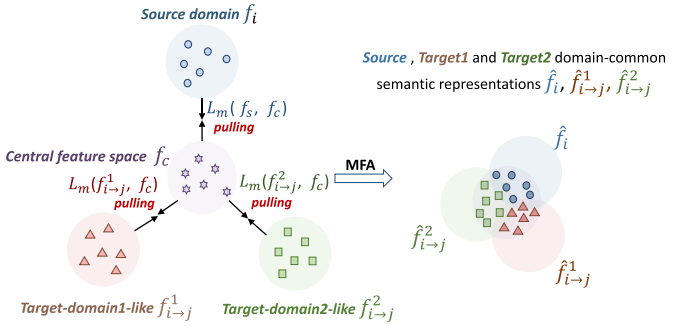

Fig. 4. Learning procedure of the MFA for the domain-common semantic representations. The MFA pulls the obtained features of different domains closer to the central feature space, thus further migrating the semantic gaps between the source and target-domain-like features.

different domains for modeling the interdomain adaptation and obtaining the domain-common semantic representations. The DAL procedure with the MAF is shown in Fig. 4, which consists of two steps, namely the central feature space averaging and the domain-common central feature space approximation.

First, the central feature space averaging is used to generate a unified central feature distribution, enabling that the multiple target-domain-like and source features have a consistent learning objective. This avoids the hard convergence issue caused by inconsistent domain objective directions for the multitarget domain adaptation task. Formally, given the multiple target-domain-like features $f^m_{i\to j}$ and the source domain feature f_i as input, we first average these three feature vectors to obtain the central feature distribution f_c

$$f_c = \frac{\left(f_i + \sum_{m=1}^{M} f^m_{i\to j}\right)}{M+1} \quad (5)$$

where M represents the number of target domains in the task and m indicates the target domain index.

Then, the MFA introduces $M+1$ central approximation losses to pull the source-domain and target-like-domain features toward the central feature distribution space f_c, thus minimizing the semantic feature distribution discrepancy between them. Mathematically, the central approximation loss L_{MFA} can be written as

$$L_{\text{MFA}} = L_m(f_i, f_c) + \sum_{m=1}^{M} L_m\left(f^m_{i\to j}, f_c\right) \quad (6)$$

where the $L_m()$ is ℓ_2-norm, namely the mean squared error loss [37], which is used to minimize the distribution discrepancy between the input features. The $L_m()$ is represented as

$$L_m(f_i, f_c) = \frac{1}{n^{\mathcal{S}}}\sum_{i}(f_i - f_c)^2$$
$$L_m\left(f^m_{i\to j}, f_c\right) = \frac{1}{n^{\mathcal{S}}}\sum_{i}\left(f^m_{i\to j} - f_c\right)^2. \quad (7)$$

Through MAF learning, we can obtain the source and target domain-common semantic representations as $\hat{f}_i$, $\hat{f}^m_{i\to j}$, respectively, by pushing these features from different domains as far as possible into the central feature space. This effectively addresses the feature semantic gaps between different domain

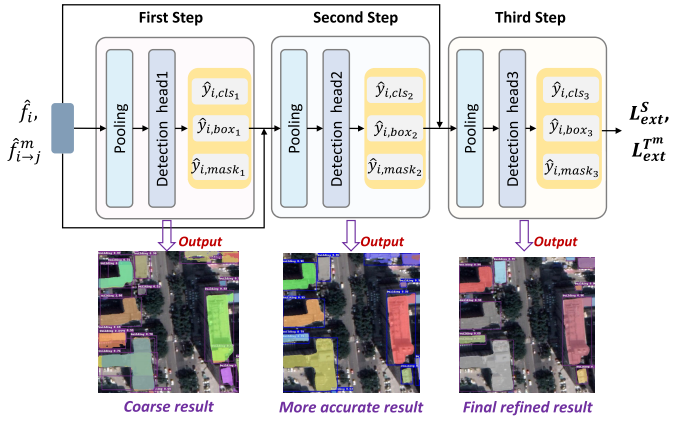

Fig. 5. Pipeline of the MCIE module from coarse to fine. Using the domain-common semantic representations $\hat{f}_i$, $\hat{f}^m_{i\to j}$, respectively, as input, we apply three cascaded detectors to extract precise building contours in a coarse-to-fine manner. The extraction results for each detector are shown at the bottom of the figure. Obviously, through such intradomain adaptation, we can obtain the fine, scale-invariant contours of the building instances.

features and keeps the optimization direction consistent across the network.

D. MCIE Module

The domain-common semantic representations $\hat{f}_i$, $\hat{f}^m_{i\to j}$ address the semantic gaps between different domains, but it still suffers from potential semantic gaps within different domains, such as scale and lighting variations. To further address the issue of intradomain shifts, we introduce multiple parameter-shared MCIE for modeling intradomain adaptation, each of which consists of three cascaded detectors with a pooling layer and a detection head. Fig. 5 shows the hierarchical structure of the MCIE for coarse-to-fine building instance extraction. Through this coarse-to-fine learning, MCIE can suppress the potential intradomain gaps, so that extract fine building contours from the domain-common semantic representations.

Using each domain-common semantic representation, the MCIE involves three cascaded detectors for coarse-to-fine building instance extraction, as shown in Fig. 5. In particular, taking a source domain-common representation $\hat{f}_i$ as input, the module would perform the source domain adaptation building instance extraction. The cascaded extraction procedure is as follows. First, in the first detector, the input $\hat{f}_i$ is fed into the first-hierarchy pooling layer and detection head 1. This step produces coarse results, including building classification results as $\hat{y}_{i,\mathrm{cls}_1}$, bounding box results as $\hat{y}_{i,\mathrm{box}_1}$, and segmentation contours as $\hat{y}_{i,\mathrm{mask}_1}$. Then, taken both the $\hat{y}_{i,\mathrm{box}_1}$ and $\hat{f}_i$ as input, the second detector further employs the second-hierarchy pooling layer and detection head 2 to obtain more accurate classification results $\hat{y}_{i,\mathrm{cls}_2}$, bounding box results $\hat{y}_{i,\mathrm{box}_2}$, and segmentation contours $\hat{y}_{i,\mathrm{mask}_2}$. Finally, in the third detector, both the $\hat{y}_{i,\mathrm{box}_2}$ extracted by the second detector and $\hat{f}_i$ are input into the pooling layer and detection head 3. This step further refines the results, providing the final classification $\hat{y}_{i,\mathrm{cls}_3}$, fine bounding boxes $\hat{y}_{i,\mathrm{box}_3}$, and building contours $\hat{y}_{i,\mathrm{mask}_3}$. Through such learning from coarse to fine, the module gradually refines the building extraction results at each step to improve the contours. To make MCIE satisfy the above procedure, the hierarchical extraction objective L^S_{ext} of the source domain in MCIE is given by

$$L^S_{\mathrm{ext}} = \sum_{l=1}^{3}\left(L^l_{\mathrm{cls}}\left(\hat{y}_{i,\mathrm{cls}_l}, y_i\right) + L^l_{\mathrm{box}}\left(\hat{y}_{i,\mathrm{box}_l}, y_i\right) + L^l_{\mathrm{mask}}\left(\hat{y}_{i,\mathrm{mask}_l}, y_i\right)\right) \tag{8}$$

where l is the cascaded level index in MCIE. L^l_{cls} is classification loss at the lth detector, L^l_{box} is the regression loss at the lth detector for the bounding box regression, and L^l_{mask} is the mask loss at the lth detector for building contour extraction. Detailed description of each loss can be referred to the Mask RCNN [20].

Moreover, taking the target domain-common representations $\hat{f}^m_{i\to j}$ as input, the module also extracts precise results as $\hat{y}^m_{i\to j,\mathrm{cls}_3}$, $\hat{y}^m_{i\to j,\mathrm{box}_3}$, and $\hat{y}^m_{i\to j,\mathrm{mask}_3}$ from the target domain-common representations, by introducing the hierarchical extraction objective $L^{T^m}_{\mathrm{est}}$ of target domains. Mathematically, the $L^{T^m}_{\mathrm{est}}$ of target domains can be written as

$$L^{T^m}_{\mathrm{est}} = \sum_{m=1}^{M}\sum_{l=1}^{3} L^l_{\mathrm{cls}}\left(\hat{y}^m_{i\to j,\mathrm{cls}_l}, y_i\right) + L^l_{\mathrm{box}}\left(\hat{y}^m_{i\to j,\mathrm{box}_l}, y_i\right) + L^l_{\mathrm{mask}}\left(\hat{y}^m_{i\to j,\mathrm{mask}_l}, y_i\right) \tag{9}$$

where M represents the number of multiple target domains, j represents the target domain samples, and m denotes the index of target domains. It is worth noting that the module can effectively leverage the source domain labels y_i for optimization due to target-domain-like feature translation with MST.

In general, the MCIE uses the concatenation of three-level detectors for intradomain adaptation, to further alleviate the potential gaps within multiple domains. In MCIE, the subsequent detector can perform more accurate detection based on the preliminary results obtained by the previous detector, resulting in effectively suppressing the intradomain semantic gaps. This coarse-to-fine strategy effectively adapts to building gaps within different domains, thus improving the quality of building instance extraction.

E. Overall Objective Function

Overall, our proposed model with these three modules is trained in turn in an end-to-end manner. The total objective function consists of four parts, namely, the MST loss, the MFA loss, source-domain hierarchical extraction loss, and target-domain hierarchical extraction loss. Formally, the overall optimization objective function in the training phase is defined as

$$L = \sum_i L^S_{\mathrm{ext}} + \sum_i L^{T^m}_{\mathrm{ext}} + \lambda L_{\mathrm{MFA}} + L_{\mathrm{MST}} \tag{10}$$

where λ represents the trade-off coefficient, which is used to balance the learning objectives in the multitask learning manner. i, j represent the sample indexes in source and target domains, respectively.

TABLE I
TYPICAL DOMAIN GAPS (NAMELY DIFFERENCES) BETWEEN DIFFERENT RS DATASETS

Dataset	Spatial resolution (m)	Image resolution (pixels)	Number of buildings (Each building)	Shooting Area	Filming Angle
SAB	0.290	500 × 500	63,886	Beijing, Shanghai, Shenzhen, Wuhan	Orthographic, Non-orthographic
Crowd	0.300	300 × 300	515,000	Las Vegas, Paris, Shanghai, Khartum	Orthographic
WHU	0.075	512 × 512	220,000	Christchurch, New Zealand	Orthographic
UBC	0.5-0.8	600 × 600	41,586	Beijing, Munich	Orthographic

TABLE II
KEY TRAINING PARAMETERS INVOLVED IN THIS WORK

Parameters	Settings
Optimizer	Stochastic gradient descent (SGD)
Initial learning rate	0.001
learning rate decay rate	0.1
Batch Size	1
Training times (Epoch)	10

IV. RESULTS AND DISCUSSION

A. Datasets

To evaluate our approach, four RS building instance datasets were used: Chinese Typical Urban Building Instance Dataset (SAB) [13], CrowdAI Mapping Challenge dataset (Crowd) [14], WHU Aerial Image Dataset (WHU) [15], and UBC Satellite Image Dataset (UBC) [38]. The more detailed information about the four datasets is shown in Table I.

1) Chinese Typical Urban Building Instance Dataset [13]: The Chinese Typical Urban Building Instance Dataset (referred to as SAB) selected four Chinese cities: Beijing, Shanghai, Shenzhen, and Wuhan, using Google Earth. The dataset covers an area of approximately 1.2 million km^2. The dataset includes orthonormal imagery and nonorthonormal RS imagery, capturing sparsely and densely distributed areas of buildings. Each RS image in SAB has dimensions of 500 × 500 pixels and a ground resolution of 0.29 m. There are a total of 7260 images and 63 886 building instances in the dataset. Among them, 5985 images are used for training, while the remaining 1275 images are used for prediction. Fig. 6 shows some examples from the SAB dataset.

2) CrowdAI Mapping Challenge Dataset [14]: The CrowdAI Mapping Challenge Dataset (hereinafter referred to as Crowd) is a simplified version of the SpaceNet dataset. It comprises original multiband satellite images with a resolution of 0.3 m from various cities, including Las Vegas, Paris, and Shanghai. Crowd specifically focuses on the RGB channels, making it suitable for the segmentation tasks. The dataset consists of a training set containing 280 741 images and a test set containing 60 697 images. Fig. 6 provides some examples from the Crowd dataset.

3) WHU Aerial Image Dataset [15]: The WHU Aerial Image Dataset (hereinafter referred to as WHU) encompasses an area of 450 km^2 in Christchurch, New Zealand. This area covers various types of buildings such as rural areas, residential areas, cultural areas, and industrial areas. The spatial resolution is 0.075 m. Each image has dimensions of 512 × 512 pixels. The WHU dataset is split into a training set consisting of 4736 images and a test set containing 2416 images. Fig. 6 displays some examples from the WHU dataset.

4) UBC Satellite Image Dataset [38]: The UBC Satellite Image Dataset (hereinafter referred to as UBC) consists of two main cities: Beijing in China and Munich in Germany. The spatial resolution is 0.5–0.8 m. The size of each image is 600 × 600 pixels. The WHU dataset is divided into a training set containing 560 images and a test set containing 160 images. Fig. 6 shows some examples from the UBC dataset.

In addition, Table I lists some differences of three above-mentioned datasets. These differences, including city, style, spatial resolution, image resolution, number of buildings, shooting areas, and shooting angles, contribute to the semantic gap between different RS datasets. The presence of these gaps poses significant challenges for cross-domain building instance extraction, as models trained on one dataset may struggle to generalize well to others due to these differences.

5) MD-BIE Experiment Settings: Following [28], [39], we performed experiments under two popular MD-BIE settings, i.e., the definitions of the source domain and the target domains, as below in the form of source → target. The experiment settings of the MD-BIE includes several cross-domain scenes, such as city-to-city, sensor-to-sensor, as well as dataset-to-dataset.

1) *SAB → Crowd & WHU:* We selected the SAM as the source domain and selected Crowd and WHU as two target domains for evaluation.
2) *Crowd → SAB & WHU:* We selected the Crowd as the source domain and the SAB and WHU as two target domains to evaluate our method.
3) *SAB → Crowd & WHU & UBC:* We selected the Crowd as the source domain and the Crowd, WHU, and UBC as three target domains to evaluate our method.
4) *SAB → WHU:* We selected the SAB as the source domain and the WHU as one target domain to evaluate our method.

The experimental environment for this study was a Windows operating system with 16 GB of RAM, an AMD Ryzen 5 3600 CPU, and a single NVIDIA GeForce RTX 2070 SUPER graphics card. All experiments were implemented in the Paddle deep learning framework.[1] The key parameters for network training are shown in Table II.

We conducted extensive comparative experiments between our method and several state-of-the-art building instance extraction methods proposed in recent years. The compared methods include Source only methods and domain adaptation methods on building instance extraction. The Source only methods indicate that the model is trained only on the source domain data and tested directly on the target domain data, such as with Mask R-CNN [20], PBE [40], RSPrompter [41], Mask2Former [42], SOLOv2 [43], and

[1] https://github.com/PaddlePaddle/Paddle

Fig. 6. Examples and their annotations in (a) SAB dataset, (b) Crowd dataset, (c) WHU dataset, and (d) UBC dataset. SAB consists of orthonormal imagery and nonorthonormal RS imagery from four Chinese cities, namely Beijing, Shanghai, Shenzhen, and Wuhan. Crowd contains multiband original satellite images from three cities, namely Las Vegas, Paris, and Shanghai. WHU covers an area of 450 km^2 in Christchurch, New Zealand. UBC contains satellite images from two cities, namely Beijing and Munich.

Cascade Mask RCNN [44]. Mask R-CNN [44] is the baseline method in the study, which is widely used for building contour extraction by simultaneously predicting object bounding boxes and masks. PBE [40] adds frame field output to a neural network to improve the quality of building extraction. RSPrompter [41] is a SAM-based method that combines semantic class information to better extract building instances. Mask2Former [42] extracts local features with constraints and predicts cross-attention within the masked region for building instance extraction. SOLOv2 [43] further improves building extraction by dynamically learning the mask head to obtain accurate positions. Cascade Mask RCNN [44] proposes a multistage object detection architecture to extract high-quality features of buildings. In addition, we also compared with the state-of-the-art domain adaptation methods, including MemoryAdaptNet [21], SWDA [45], and VDD [32]. MemoryAdaptNet [21] uses the adversarial learning-derived invariant features to bridge the domain distribution gap between the source and target domains, thus improving the accuracy of building extraction. VDD [32] is a domain adaptation method that introduces VDD to extract domain-invariant object representations, thereby enabling better extraction of building instances. SWDA [45] proposed a detector adaptation method that utilizes strong local alignment and weak global alignment to improve the quality of building extraction.

B. Evaluation Protocol

To evaluate the performance of the proposed method, we utilized four standard metrics commonly employed for building instance segmentation evaluation: AP@0.5, AP@0.75, mean average precision (mAP) [46], and mean average recall (mAR) [47]. These metrics enable the quantification of model accuracy and recall in instance segmentation tasks. Higher values for each metric indicate improved accuracy and segmentation performance. AP@0.5 and AP@0.75 measure the accuracy of the predicted segmentation results against the ground truth segmentation results at IoU thresholds of 0.5 and 0.75, respectively. IoU [48] is defined as the ratio of the intersection to the union between the predicted value and the ground truth value. It can be written as

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{11}$$

where TP represents the true positives, FP represents the false positives, TN represents the true negatives, and FN represents the false negatives. mAP [46] calculates the average precision across multiple IoU thresholds, offering a comprehensive evaluation of the model's performance at various IoU thresholds. The precision is defined as the ratio of TP to the sum of TP and FP, which is given by

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{12}$$

On the other hand, mAR [47] calculates the average recall across multiple IoU thresholds, where the Recall is calculated as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{13}$$

By utilizing these metrics, we can conduct a comprehensive evaluation of the proposed method's performance in the building instance segmentation task. The results obtained from these metrics will allow us to validate the effectiveness of our method and enable objective comparisons with other existing methods.

C. Analysis of Experimental Results

1) SAB → Crowd & WHU: We used the SAB dataset as the source domain, and Crowd and WHU datasets as multiple target domains. The performance results of the MD-BIE are shown in Table III. From the results, one can see that our proposed DAL achieved the best performance on the two target domains. For the Crowd dataset, the proposed method outperformed other methods on all evaluation indicators. RSPrompter utilized the SAM for building extraction, but did not take into account domain adaptation information and therefore exhibits suboptimal performance. Compared to the current optimal target-domain adaptation method, VDD, the proposed DAL also achieved a relative improvement of 6.4% on AP@0.5, 37.3% on AP@0.75, 5.8% on mAP, and 3.4% on mAR, respectively. For the WHU dataset, the proposed method achieved a significant lead on all evaluation indicators. Compared to the current optimal domain adaptation method, MemoryAdaptNet, the proposed model improved mAP and mAR by 20.7% and 5.9%, respectively. It also shows that our method is more suitable for extracting the precise contours of

TABLE III
DOMAIN ADAPTATION EVALUATION RESULTS OF DIFFERENT METHODS ON SAB → CROWD & WHU. THE BEST RESULTS ARE IN BOLD

	SAB → Crowd				
	Methods	AP@0.5	AP@0.75	mAP	mAR
Source only	Mask R-CNN(baseline)	49.5	19.9	23.7	28.7
	SOLOv2 [43]	52.0	22.5	25.3	42.9
	Cascade Mask RCNN [44]	55.5	28.9	29.6	47.0
	PBE [40]	58.0	21.2	26.4	41.6
	Mask2Former [42]	46.5	25.9	25.6	38.8
	RSPrompter [41]	54.0	21.2	25.9	40.5
	Our method	55.5	28.9	29.6	47.0
Damain adaptation	SWDA [45]	49.2	21.5	35.3	45.0
	VDD [32]	68.7	32.4	39.4	49.4
	MemoryAdaptNet [21]	61.2	30.1	32.7	49.4
	Our method	**73.1**	**44.5**	**41.7**	**51.1**
	SAB → WHU				
	Methods	AP@0.5	AP@0.75	mAP	mAR
Source only	Mask R-CNN(baseline)	53.8	27.5	28.3	42.1
	SOLOv2 [43]	52.4	19.8	24.1	35.5
	Cascade Mask RCNN [44]	67.9	46.3	42.2	**67.9**
	PBE [40]	52.5	18.6	23.5	42.3
	Mask2Former [42]	56.6	28.1	29.2	40.9
	RSPrompter [41]	54.6	27.1	28.7	48.0
	Our method	67.9	46.3	42.2	67.9
Damain adaptation	SWDA [45]	32.4	10.8	21.6	29.1
	VDD [32]	73.4	42.0	36.7	67.5
	MemoryAdaptNet [21]	72.28	48.4	38.7	59.9
	Our method	**84.7**	**68.5**	**59.4**	65.8

TABLE IV
DOMAIN ADAPTATION EVALUATION RESULTS OF DIFFERENT METHODS ON CROWD → SAB & WHU. THE BEST RESULTS ARE IN BOLD

	Crowd → SAB				
	Methods	AP@0.5	AP@0.75	mAP	mAR
Source only	Mask R-CNN(baseline)	29.7	11.1	13.4	28.0
	SOLOv2 [43]	33.9	11.6	14.9	32.6
	Cascade Mask RCNN [44]	37.9	17.4	19.0	40.5
	PBE [40]	47.3	19.4	21.7	38.1
	Mask2Former [42]	43.6	24.2	23.9	33.5
	RSPrompter [41]	53.3	23.3	26.1	40.8
	Our method	37.9	17.4	19.0	40.5
Damain adaptation	SWDA [45]	29.7	11.0	20.3	34.0
	VDD [32]	49.5	21.6	23.9	43.9
	MemoryAdaptNet [21]	51.2	23.9	28.1	45.7
	Our method	**54.0**	**29.0**	**28.8**	**47.8**
	Crowd → WHU				
	Methods	AP@0.5	AP@0.75	mAP	mAR
Source only	Mask R-CNN(baseline)	67.4	34.8	35.3	44.6
	SOLOv2 [43]	51.3	19.1	23.4	33.8
	Cascade Mask RCNN [44]	68.4	47.0	42.6	52.5
	PBE [40]	62.8	38.3	36.0	49.3
	Mask2Former [42]	62.8	42.1	38.6	45.6
	RSPrompter [41]	73.3	46.1	42.5	52.1
	Our method	68.4	47.0	42.6	52.5
Damain adaptation	SWDA [45]	27.6	8.90	18.2	26.3
	VDD [32]	69.3	53.2	42.7	57.9
	MemoryAdaptNet [21]	69.6	56.2	50.7	58.7
	Our method	**81.0**	**63.1**	**54.1**	**61.4**

building instances due to properly interdomain and intradomain adaptation. In addition, our method has a slight decrease in mAR compared with Cascade Mask RCNN (about 1.9%), the possible reason is imbalance in the training samples of two datasets.

2) Crowd → SAB & WHU: We evaluated DAL on different RS datasets using Crowd dataset as source domain and SAB and WHU datasets as target domains. The performance results of the MD-BIE are shown in Table IV. We can observed that our proposed DAL outperformed other compared methods and achieved the state-of-the-art performance on both the two target domains. For the SAB dataset, compared with VDD for extracting domain-invariant features, the proposed DAL also achieved a relative improvement of 9.1% on AP@0.5, 34.3% on AP@0.75, 20.5% on mAP, and 8.9% on mAR. This shows that the approach of pulling toward domain common features is more effective in confusing different domains and thus better in extracting building. For the WHU dataset, the proposed method outperformed other methods on all evaluation indicators. The proposed DAL achieves a relative improvement of 16.3%/12.3%/12.2%/6.3%/4.6% over the current optimal domain adaptation method MemoryAdaptNet at AP@0.5/AP@0.75/mAP/mAR, respectively. It is proved that DAL effectively solves the semantic gap problem between different RS domains and obtains accurate building instance segmentation performance on multiple target domains.

TABLE V
DOMAIN ADAPTATION EVALUATION RESULTS OF DIFFERENT METHODS ON SAB → CROWD & WHU & UBC. THE BEST RESULTS ARE IN BOLD

SAB → WHU				
Methods	AP@0.5	AP@0.75	mAP	mAR
SWDA [45]	36.2	13.2	16.4	24.8
VDD [32]	30.7	11.6	14.2	21.8
MemoryAdaptNet [21]	66.3	43.5	36.7	55.8
Our method	**82.5**	**65.3**	**55.8**	**61.6**
SAB → Crowd				
Methods	AP@0.5	AP@0.75	mAP	mAR
SWDA [45]	47.8	22.4	23.9	35.1
VDD [32]	48.4	21.8	24.1	35.4
MemoryAdaptNet [21]	52.4	24.5	32.3	45.2
Our method	**72.0**	**42.9**	**40.5**	**51.3**
SAB → UBC				
Methods	AP@0.5	AP@0.75	mAP	mAR
SWDA [45]	14.2	5.6	6.7	9.0
VDD [32]	13.2	5.9	6.7	8.9
MemoryAdaptNet [21]	20.6	10.3	13.5	19.4
Our method	**36.7**	**18.4**	**19.4**	**25.8**

3) SAB → CROWD & WHU & UBC: We employed the SAB dataset as the source domain, and Crowd, WHU, and UBC datasets as multiple target domains. The performance results of the MD-BIE are shown in Table V. From the results, one can see that our proposed DAL achieved the best performance on all three target domains. For the WHU dataset, the proposed method outperformed other methods on all evaluation indicators, obtaining the best AP@0.5 of 82.5%, the best AP@0.75 of 65.3%, the best mAP of 55.8%, and the best mAR of 61.6%. For the Crowd dataset, the proposed DAL achieved an improvement of 19.6%/18.4%/12.2%/8.2%/6.1% over the current optimal domain adaptation method MemoryAdaptNet at AP@0.5/AP@0.75/mAP/mAR, respectively. For the UBC dataset, the proposed method achieved significant increase on all evaluation indicators, obtaining the best AP@0.5 of 36.7%, the best AP@0.75 of 18.4%, the best mAP of 19.4%, and the best mAR of 25.8%. This indicates that our DAL can be better extended to more target domains.

4) SAB → Single-Target Domain WHU: We evaluated DAL on one source → one target domain task, i.e., using SAB dataset as the source domain and WHU dataset as the target domain. The performance results of the MD-BIE are shown in

TABLE VI

DOMAIN ADAPTATION EVALUATION RESULTS OF DIFFERENT METHODS ON THE SINGLE TARGET DOMAIN TASK, I.E., THE SAB → WHU. THE BEST RESULTS ARE IN BOLD

Methods	AP@0.5	AP@0.75	mAP	mAR
SWDA [45]	33.5	12.6	15.6	22.8
VDD [32]	33.0	13.6	15.8	23.1
MemoryAdaptNet [21]	67.8	41.2	35.5	51.0
Our method	**82.2**	**62.6**	**54.1**	**60.9**

TABLE VII

EFFECT OF DIFFERENT MODULES UNDER THE SAB → CROWD & WHU MULTIDOMAIN BIS-RS TASK. THE BEST RESULTS ARE IN BOLD

Baseline	MCIE	MST	MFA	Source: SAB → Target: Crowd		Source: SAB → Target: WHU	
				mAP	mAR	mAP	mAR
✓				23.7	28.7	28.3	42.1
✓	✓			30.7	48.2	40.1	60.8
✓	✓	✓		35.9	49.8	48.5	63.4
✓	✓	✓	✓	**41.7**	**51.1**	**59.4**	**65.8**

Table VI. We observed that our proposed DAL outperformed other compared method on the single target domain, obtaining the best AP@0.5 of 82.2%, the best AP@0.75 of 62.6%, the best mAP of 54.1%, and the best mAR of 60.9%. Compared to MemoryAdaptNet, the proposed DAL also achieved an improvement of 14.4% on AP@0.5, 21.4% on AP@0.75, 18.6% on mAP, and 9.9% on mAR. This shows that our DAL method can also be well applied for single-target domain.

D. Ablation Studies and Analysis

1) Effect of Different Modules: To better understand the role of each module in the proposed DAL, Table VII presents the ablation results of the gradual addition MCIE, MST, and MAF components to the baseline Mask R-CNN framework on the MD-BIE task. Compared to the baseline method, adding the MCIE module improved the results by 7.0% in mAP and 19.5% in mAR on the Crowd target domain, and by 1.8% and 18.7% on the WHU target domain, respectively. The further integration of the MST module improved the mAP and mAR to 35.8, 49.8, and 48.5, 63.4 on both two target domains, respectively, as the MST aids in extraction of target-domain-like features. Finally, thanks to learning the domain-common semantic representations for effectively modeling interdomain adaptation, the addition of the MFA module achieved further improvements of 5.8%, 1.3%, 10.9%, and 2.4%, respectively. Overall, the proposed model's effectiveness and rationality were fully demonstrated.

2) Effect of the Key Parameters: To evaluate the effect of key parameters on the DAL, we adjusted different values for the loss weighting coefficient λ [see (10)] on the model performance and provided the results in Fig. 7. The results on both two target domains exhibit the same trend, i.e., as the coefficient values increased, the corresponding mAP and mAR values first increased and then decreased. From the figure, when the coefficient value is set to 0.001, the model achieved the best performance on both evaluation metrics of mAP and mAR. This is similar to the results in [49], showing that setting a relatively small loss weight (e.g., 0.001) to define the loss weight of the target domain loss could achieve improved performance.

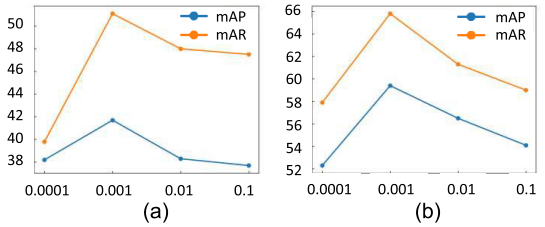

Fig. 7. Effect of different λ on model performance. The orange line indicates the mAP results for different λ values, and the blue line indicates the mAR results. Obviously, the model achieves the best performance when $\lambda = 0.001$. (a) SAB → Crow. (b) SAB → WHU.

TABLE VIII

EFFECT OF DIFFERENT FEATURE APPROXIMATION LOSSES ON MODEL PERFORMANCE UNDER THE SAB → CROWD & WHU. THE BEST RESULTS ARE IN BOLD

Feature approximation Losses	Source: SAB → Target: Crowd		Source: SAB → Target: WHU	
	mAP	mAR	mAP	mAR
w/o approximation loss	35.9	49.8	48.5	63.4
ℓ_1	39.7	49.0	56.1	61.1
Smooth ℓ_1	40.4	49.2	53.2	58.0
ℓ_2	**41.7**	**51.1**	**59.4**	**65.8**

3) Effect of Different Loss Functions: To evaluate the effect of different loss functions used in the MFA module for multidomain central approximation learning, Table VIII shows the impact of using different feature approximation losses in L_{MFA} [see (7)] on the model performance. We conducted experiments using three commonly used loss functions for feature approximation: ℓ_1 loss [50], Smooth ℓ_1 loss [51], and ℓ_2 loss [37], on the SAB → Crowd & WHU multidomain task, respectively. Compared to the results without using any approximation loss, our model with any of the feature approximation loss functions improved the results, demonstrating that the MFA loss can alleviate the semantic gap between different domains. Among the three different feature approximation losses, the model achieved the best results using the ℓ_2 loss. The results show that the ℓ_2 loss effectively mitigates data distribution differences between different RS datasets (domains), thus performing well in each target domain.

4) Experimental Complexity Analysis: Table IX shows the comparative results of the model complexity experiments on the SAB → Crowd & WHU domain adaptation task. We used two widely used indicators for model complexity evaluation, including the inference speed (fps) and model size (MB). The inference speed of the proposed DAL is much faster (more than 4.5 fps) than the current domain adaptation methods, such as MemoryAdaptNet and VDD, showing that our method obtains a great balance of accuracy and efficiency. Moreover, the DAL has the same size of 274.1 MB as the models of source-only methods such as Mask R-CNN and cascade Mask R-CNN. This indicates that our model significantly improves accuracy without introducing additional parameters for training. Although the inference speed is slightly slower than the Source only methods such as Mask R-CNN, there

TABLE IX
RESULTS OF THE COMPLEXITY EXPERIMENTS OF DIFFERENT METHODS ON THE SAB → CROWD & WHU DATASET. THE BEST RESULTS ARE IN BOLD

Methods	Inference speed (fps)	Model size (MB)
Mask R-CNN(baseline)	7.3	274.1
Cascade Mask RCNN [44]	7.0	274.1
VDD [32]	5.6	405.2
MemoryAdaptNet [21]	2.4	666.1
Our method	6.9	274.1

TABLE X
EFFECT OF DIFFERENT CENTRAL FEATURE SELECTION METHODS ON MODEL PERFORMANCE UNDER THE SAB → CROWD & WHU. THE BEST RESULTS ARE IN BOLD

Methods	Source: SAB → Target: Crowd		Source: SAB → Target: WHU	
	mAP	mAR	mAP	mAR
Sum	38.7	48.3	51.1	56.9
To Source	38.9	48.8	55.4	61.9
Ours	**41.7**	**51.1**	**59.4**	**65.8**

is a significant improvement in accuracy without increasing the training resources, further proving the effectiveness and efficiency of our method.

5) Effect of Different Central Feature Selection Methods: In order to evaluate the effect of different central feature selection methods used in the MFA module, Table X shows the results of three different methods for central feature selection on the SAB → Crowd & WHU multidomain task. In the table, "Sum" represents that the central feature distribution f_c is obtained by summing the three features; "To Source" represents that the central feature distribution f_c is derived from the source domain features; "Ours" represents the average of the three features as the central feature distribution f_c. Among these three methods, the average strategy obtains the best performance, showing that the average central feature distribution effectively alleviates the inconsistent convergence in the multiobjective domain adaptive learning.

6) Effect of Different Layers and Inputs to the Next Detector in MCIE: On the SAB → Crowd & WHU task, we compared the experiments with different inputs to the next detector in MCIE, i.e., masks & boxes and only bounding boxes. As illustrated in Table XI, the results indicate that using bounding boxes as input yields superior performance compared to using mask results. The reason is that the mask results from the previous detector may contain additional noise, adversely affecting the performance of the next detector. Conversely, utilizing bounding box results enables the detection model to progressively enhance the building extraction result, with a focus on refining the building outline. Moreover, for a more comprehensive understanding of MCIE, we conducted an ablation study on different layers within MCIE on the SAB → Crowd & WHU task, as presented in Table XII. The results further demonstrate that building extraction performance continues to improve as the layers increase, underscoring the effectiveness of MCIE for our cross-domain task.

TABLE XI
EVALUATION RESULTS ON DIFFERENT INPUTS TO THE NEXT DETECTOR IN THE MCIE MODULE UNDER THE SAB → CROWD & WHU. THE BEST RESULTS ARE IN BOLD

SAB → WHU				
Input	AP@0.5	AP@0.75	mAP	mAR
Mask & box	82.3	60.2	52.5	60.4
Box (Ours)	**84.7**	**68.5**	**59.4**	**65.8**
SAB → Crowd				
Input	AP@0.5	AP@0.75	mAP	mAR
Mask & box	71.0	41.2	39.3	48.8
Box (Ours)	**73.1**	**44.5**	**41.7**	**51.1**

TABLE XII
EVALUATION RESULTS ON DIFFERENT LAYERS IN THE MCIE MODULE UNDER THE SAB → CROWD & WHU. THE BEST RESULTS ARE IN BOLD

SAB → WHU				
Numbers	AP@0.5	AP@0.75	mAP	mAR
Layers=1	82.1	61.2	51.5	60.2
Layers=2	82.5	61.7	52.9	58.6
Layers=3 (Ours)	**84.7**	**68.5**	**59.4**	**65.8**
SAB → Crowd				
Input	AP@0.5	AP@0.75	mAP	mAR
Layers=1	65.8	37.0	36.1	50.0
Layers=2	69.2	39.3	37.9	48.4
Layers=3 (Ours)	**73.1**	**44.5**	**41.7**	**51.1**

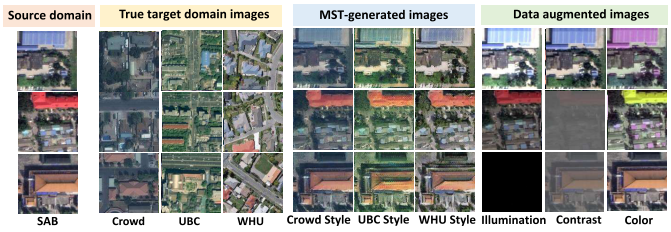

Fig. 8. Comparison of MST-generated images and data augmentation images.

7) Comparison of MST and Data Augmentation: To clarify that our MST does not simply get data augmentation, we added the visualization and quantitation experiments with using our MST and data augmentation. Fig. 8 shows the comparison of style transferring by using MST and the current data augmentation methods. From the figure, the generated images through data augmentation (i.e., illumination, contrast and color) exhibit a high degree of randomization and are not specifically styled for the target domain. The images converted by MST are more similar to the corresponding target domains. In addition, Table XIII lists the building extraction results of data augmentation and our MST on the SAB → CROWD & WHU & UBC task. Obviously, our DAL with MST outperforms data augmentation in extracting buildings across all three datasets.

E. Visual Experimental Analysis

1) Feature Distribution Visualization: To analysis the use of different methods for feature alignment, Fig. 9 shows the

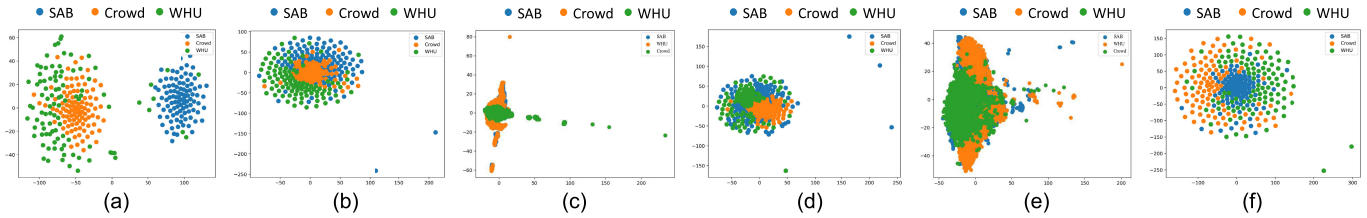

Fig. 9. Comparison of the feature space distributions using different methods on the SAB → Crowd & WHU task. From the results, the more confusing the learned features of different domains are, i.e., the better the domain adaptation is obtained. Obviously, our DAL with MAF and MCIE modules obtain more effective feature alignment by modeling inter- and intradomain adaptation. (a) Source only. (b) VDD. (c) MemoryAdaptNet. (d) DAL with MCIE. (e) DAL with MFA. (f) DAL with both MFA and MCIE.

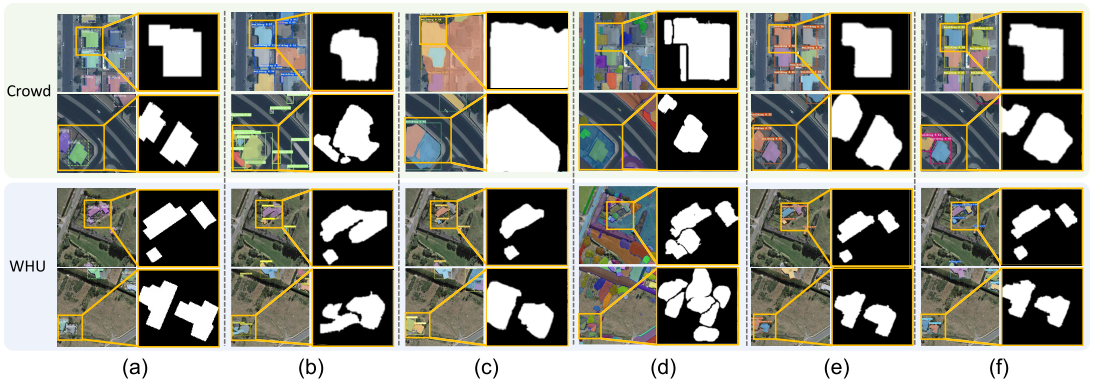

Fig. 10. Visualization of different methods on the SAB → Crowd & WHU MD-BIE task. For the extracted results by each method, the left columns are the extracted results from the original images, while the binarized image on the right is the result of local zooming. (a) Ground truth. (b) Mask R-CNN. (c) Cascade Mask RCNN. (d) SAM. (e) MemoryAdaptNet. (f) Our DAL.

TABLE XIII

COMPARISON RESULTS OF OUR MST AND DATA AUGMENTATION ON DIFFERENT TARGET DOMAINS. THE BEST RESULTS ARE IN BOLD

Target domain	Methods	AP@0.5	AP@0.75	mAP	mAR
WHU	Data augmentation	59.3	25.1	28.9	41.3
	Our MST	**82.5**	**65.3**	**55.8**	**61.6**
Crowd	Data augmentation	56.6	26.1	28.4	43.2
	Our MST	**72.0**	**42.9**	**40.5**	**51.3**
UBC	Data augmentation	25.4	14.0	14.1	20.8
	Our MST	**36.7**	**18.4**	**19.4**	**25.8**

comparison of domain feature distributions using different methods on the SAB → Crowd & WHU task. Compared to Mask R-CNN [20], MemoryAdaptNet [21], and VDD [32], our method better aligned the feature distributions of the source domain and multiple target domains, obtaining the optimal domain adaptation. In addition, we further visualized the feature distributions with different modules in our method, namely DAL with the MFA and MCIE, respectively, in Fig. 9(d)–(f). We find that the feature distributions of both using MFA and MCIE are confounded together and more centralized than using one of two modules separately, suggesting that the proposed MFA and MCIE can help the model to obtain effective feature alignment by inter- and intradomain adaptation.

2) Visualization Results on SAB → Crowd & WHU: Fig. 10 shows the visualization results comparison between our method and other state-of-the-art methods on the SAB → Crowd & WHU MD-BIE task. The first two rows show the extraction results on the Crowd dataset, and the last two rows show the extraction results on the WHU dataset. It can be seen that the Crowd dataset mainly contains urban areas, while the WHU dataset contains more wild areas, with significant domain gaps between the two datasets. As shown in Fig. 10(b), the classical instance extraction method, Mask R-CNN [20], cannot adapt well to these gaps, with missed and false extraction results on both target domains. Fig. 10(c) shows the extraction results of Cascade Mask RCNN [44]. Compared with Mask R-CNN, this method reduces the missed results but performs poorly in dense building extraction, resulting in interconnections between building instances [see the first row of Fig. 10(c)]. Fig. 10(d) shows the extraction results of the latest large model SAM [19]. We observe that SAM is difficult to extract the complex and large building instance and tens to split the whole building into parts, such as the fourth row in Fig. 10(d). Fig. 10(e) shows the extraction results of MemoryAdaptNet [21], which obtains better extraction results than the previous methods through interdomain adaptation. Finally, as shown in Fig. 10(f), our DAL method extracted more accurate contours of building instances than the other methods on different target domains, by effectively eliminating the semantic gaps between and within multiple domains.

3) Visualization Results on Crowd → SAB & WHU: We compared the building instance extraction results of our model and other state-of-the-art methods on the Crowd → SAB & WHU task, as shown in Fig. 11. The first two rows show the extraction results on the SAB dataset, and the last two rows show the extraction results on the WHU dataset. Both Mask R-CNN [20] and Cascade Mask RCNN [44] perform poorly in both target domains, with many missed building instances in the SAB dataset and misclassifying trees as buildings in the WHU dataset [see Fig. 11(b) and (c)]. In addition, the

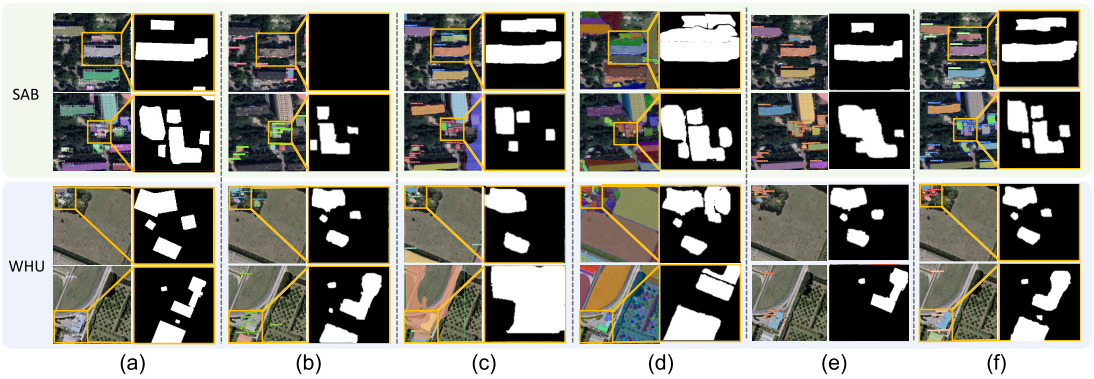

Fig. 11. Visualization of different methods in Crowd → SAB & WHU task. For the extracted results by each method, the left columns are the extracted results from the original images, while the binarized image on the right is the result of local zooming. (a) Ground truth. (b) Mask R-CNN. (c) Cascade Mask RCNN. (d) SAM. (e) MemoryAdaptNet. (f) Our DAL.

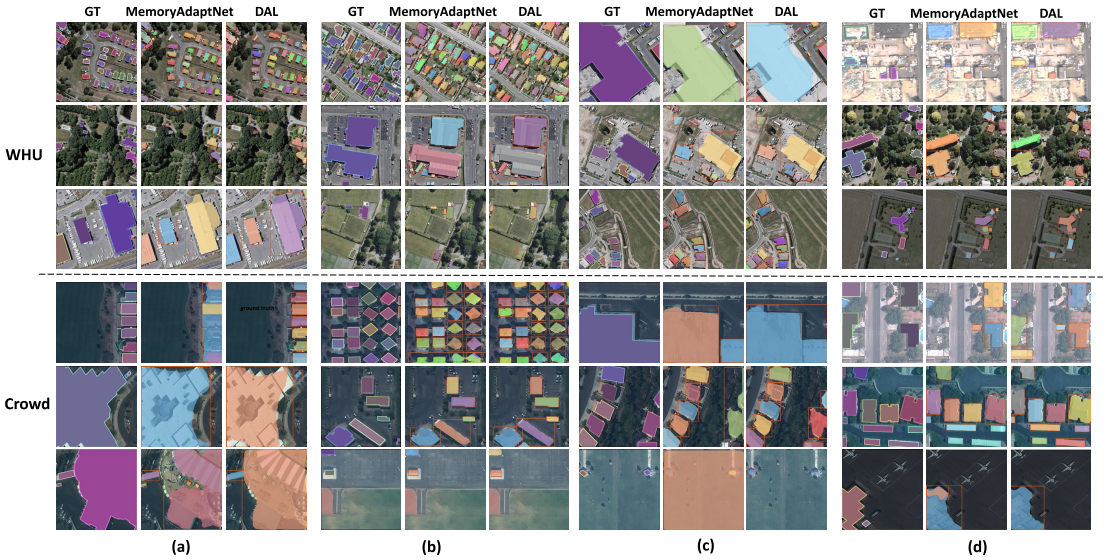

Fig. 12. Visualization results for MD-BIS under different complex scenarios with domain gaps on the SAB → Crowd & WHU. First column: ground truth. Second column: MemoryAdaptNet results. Third column: our DAL results. Compared to MemoryAdaptNet, which only performs interdomain adaptation, our approach achieves better performance, suggesting that both interdomain and intradomain adaptation are important for the MD-BIS task. (a) Different building styles. (b) Different building densities. (c) Different building scales. (d) Different illumination.

current large segmentation model, SAM [19], appears to split a building into two parts [see Fig. 11(d)]. MemoryAdaptNet [21] improved the detected results in different target domains but obtained the rougher building instance contours [see Fig. 11(e)]. Moreover, our proposed DAL extracted more accurate contours of building instances under different scales and conditions, as shown in Fig. 11(f). This indicates that the proposed DAL can effectively handle buildings with various sizes through both interdomain and intradomain adaptation.

4) Visualization Results Under Different Complex Scenarios With Domain Gaps: In order to provide a comprehensive evaluation of the proposed DAL method, we summarized four different challenging scenarios including different domain gaps, such as style, density, scale, and illumination, in Fig. 12. Fig. 12(a) shows the extraction results of different domain adaptation methods under significant style differences. Obviously, compared with MemoryAdaptNet, our methods can extracted the building contours more accurately. For different building densities, as shown in Fig. 12(b), our method extracted building instances well in terms of both building density and sparsity, while MemoryAdaptNet worked better only for building sparsity but missed detection in building density, e.g., the first and fourth rows. Furthermore, for building scale and illumination variation, Fig. 12(c) and (d) shows that our DAL method performed more effective building instance extraction for both large and small buildings as well as for different illumination conditions, compared to MemoryAdaptNet that only considers the interdomain adaptation. This further validates that our proposed DAL method can adapt well to inter- and intradomain gaps, resulting in better building instance extraction performance.

5) Visualization of MST-Generated Images and Features: In order to analyze the effectiveness of target domain-like images and features generated by the MST module, we conducted two experiments. Fig. 13 shows the target domain-like images generated by MST, where images from different source domains are effectively converted into the styles of multiple target domains by MST, facilitating the mitigation of interdomain differences. Moreover, we also compared the original feature distributions of three domains and MST-generated target

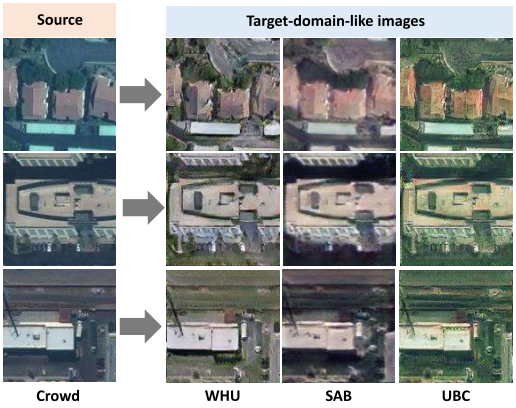

Fig. 13. Target-domain-like images generated by MST, i.e., from the source domain (Crowd) to multitarget domains (WHU, SAB, and UBC).

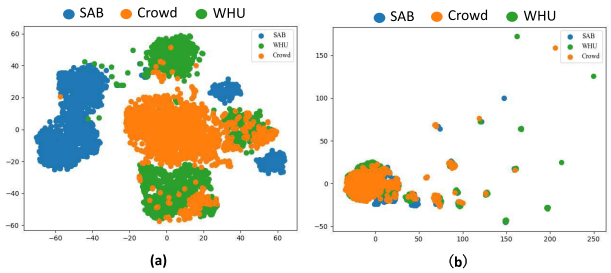

Fig. 14. Feature distributions of the three datasets with and without using MST. Obviously, our MST effectively aligns the feature distributions of source domain and two target domains.

domain-like feature distributions in the SAB → Crowd & WHU task. The results, as shown in Fig. 14, show that the MST-generated target domain-like feature distributions have better aggregation, which suggests that the proposed MST can help the model to obtain effective feature alignment through the interdomain adaptation.

V. Conclusion

In this article, we propose a novel DAL approach for MD-BIE, which effectively models interdomain and intradomain adaptations to suppress the semantic gaps between and within different RS domains. DAL can achieve robust unsupervised performance in multiple target domains by training a unified model. For interdomain adaptation, DAL first employ the MST to generate target-domain-like features in per target domain, then devise a novel MAF module to learn the domain-common semantic representations through pulling both source domain features and target-domain-like features toward the central domain-common feature space. This effectively reduces the feature distribution differences and semantic gaps between different domains. Moreover, for intradomain adaptation, an MCIE module is used to perform accurate instance segmentation of buildings with various scales on different domains in a coarse-to-fine manner.

Experiments and detailed analysis were conducted on three publicly available RS building datasets: SAB, Crowd, and WHU with two multiple cross-domain settings. On the SAB → Crowd & WHU and Crowd → SAB & WHU, the two multidomain BIS-RS tasks, the proposed method outperformed existing cross-domain instance extraction methods by a significant margin and achieved the state-of-the-art performance. Despite the effectiveness of our method, we found that there are still some rooms that our DAL could be improved. For example, our method still depends on the building labels on the source domain. In the future, we will further introduce contrastive learning to achieve unsupervised building instance segmentation without source domain guidance.

References

[1] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, "Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[2] J. Wang et al., "A knowledge-based method for road damage detection using high-resolution remote sensing image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 3564–3567.

[3] H. Liu, C. Lin, B. Gong, and D. Wu, "Extending the detection range for low-channel roadside LiDAR by static background construction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[4] S. Voigt, T. Kemper, T. Riedlinger, R. Kiefl, K. Scholte, and H. Mehl, "Satellite image analysis for disaster and crisis-management support," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1520–1528, Jun. 2007.

[5] Z. Chen et al., "Vehicle detection in high-resolution aerial images via sparse representation and superpixels," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 103–116, Jan. 2016.

[6] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[7] B. Sirmacek and C. Unsalan, "Building detection from aerial images using invariant color features and shadow information," in *Proc. 23rd Int. Symp. Comput. Inf. Sci.*, Oct. 2008, pp. 1–5.

[8] S.-H. Zhong, J.-J. Huang, and W.-X. Xie, "A new method of building detection from a single aerial photograph," in *Proc. 9th Int. Conf. Signal Process.*, Oct. 2008, pp. 1219–1222.

[9] G. Ferraioli, "Multichannel InSAR building edge detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1224–1231, Mar. 2010.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[11] C. Li, G. Cheng, G. Wang, P. Zhou, and J. Han, "Instance-aware distillation for efficient object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[12] L. He, J. Shan, and D. Aliaga, "Generative building feature estimation from satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.

[13] K. Wu et al., "A dataset of building instances of typical cities in China," *Chin. Sci. Data*, vol. 6, pp. 191–199, Mar. 2021.

[14] S. P. Mohanty et al., "Deep learning for understanding satellite imagery: An experimental survey," *Frontiers Artif. Intell.*, vol. 3, Nov. 2020, Art. no. 534696.

[15] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2018.

[16] F. Fang et al., "A coarse-to-fine contour optimization network for extracting building instances from high-resolution remote sensing imagery," *Remote Sens.*, vol. 13, no. 19, p. 3814, Sep. 2021.

[17] Q. Wen et al., "Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network," *Sensors*, vol. 19, no. 2, p. 333, Jan. 2019.

[18] Y. Tian et al., "Multi-scale U-net with edge guidance for multimodal retinal image deformable registration," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1360–1363.

[19] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[21] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.

[22] Z. K. L. Jingjing and M. Lichao, "A review of domain adaptation research," *Comput. Eng.*, vol. 47, no. 6, pp. 1–13, 2021.
[23] J. Chen et al., "Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
[24] D. Peng, H. Guan, Y. Zang, and L. Bruzzone, "Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
[25] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7178–7193, Oct. 2020.
[26] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "SEMI2I: Semantically consistent image-to-image translation for domain adaptation of remote sensing data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 1837–1840.
[27] I. Goodfellow, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[28] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "DAugNet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1067–1081, Feb. 2021.
[29] S. Wei, T. Zhang, and S. Ji, "A concentric loop convolutional neural network for manual delineation-level building boundary segmentation from remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
[30] Y. Zhu, B. Huang, J. Gao, E. Huang, and H. Chen, "Adaptive polygon generation algorithm for automatic building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[31] L. Xu, Y. Li, J. Xu, and L. Guo, "Gated spatial memory and centroid-aware network for building instance extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[32] A. Wu, R. Liu, Y. Han, L. Zhu, and Y. Yang, "Vector-decomposed disentanglement for domain-invariant object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9342–9351.
[33] T. Isobe et al., "Multi-target domain adaptation with collaborative consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8187–8196.
[34] H. Cui et al., "MDANet: Unsupervised, mixed-domain adaptation for semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[35] C. Lin, Y. Li, Y. Liu, X. Wang, and S. Geng, "Building damage assessment from post-hurricane imageries using unsupervised domain adaptation with enhanced feature discrimination," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
[36] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.
[37] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Dec. 2017.
[38] X. Huang et al., "Urban building classification (UBC)—A dataset for individual building detection and classification from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 2022, pp. 1412–1420.
[39] B. H. Ngo, Y. J. Chae, J. H. Park, J. H. Kim, and S. I. Cho, "Easy-to-hard structure for remote sensing scene classification in multitarget domain adaptation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
[40] N. Girard, D. Smirnov, J. Solomon, and Y. Tarabalka, "Polygonal building extraction by frame field learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5891–5900.
[41] K. Chen et al., "RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," 2023, *arXiv:2306.16269*.
[42] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17864–17875.
[43] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17721–17732.
[44] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[45] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6949–6958.
[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[47] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[48] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin Transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[49] H. Zhou, F. Jiang, and H. Lu, "SSDA-YOLO: Semi-supervised domain adaptive YOLO for cross-domain object detection," *Comput. Vis. Image Understand.*, vol. 229, Mar. 2023, Art. no. 103649.
[50] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

Fayong Zhang received the Ph.D. degree from China University of Geosciences, Wuhan, China, in 2009.

He is currently an Associate Professor with China University of Geosciences. He has authored or coauthored nearly 30 academic papers domestically and internationally. His research interests include spatio-temporal big data, urban underground pipelines, urban geology, and GIS and their applications.

Dr. Zhang is a member of the Smart City Working Committee of the Chinese Society for Geodesy Photogrammetry and Cartography.

Kejun Liu received the B.S. degree in network engineering from Henan University, Kaifeng, China, in 2022. She is currently pursuing the master's degree with China University of Geosciences, Wuhan, China.

Her research interests include remote sensing image segmentation.

Yuanyuan Liu received the Ph.D. degree from Central China Normal University, Wuhan, China, in 2015.

She was a Visiting Scholar with Nanyang Technological University, Singapore. She is currently an Associate Professor with China University of Geosciences, Wuhan. She has authored or coauthored various top conferences and journals, such as Conference on Computer Vision and Pattern Recognition (CVPR), ACM Multimedia (ACM MM), IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS (T-VCG), Pattern Recognition (PR), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), Conference on Information and Knowledge Management (CIKM), Information Sciences (INS), and Nature Communications (NC). Her research interests include computer vision and multimodal analysis.

Chaofan Wang received the M.S. degree in computer technology from China University of Geosciences, Wuhan, China, in 2023.

His research interests include object detection.

Wujie Zhou (Senior Member, IEEE) received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2012.

He was a Visiting Scholar funded by the China Scholarship Council at Nanyang Technological University, Singapore. Currently, he is an Associate Professor with Zhejiang University of Science and Technology. He has authored or coauthored over 70 academic papers as the first author in prestigious journals such as Association for the Advancement of Artificial Intelligence (AAAI), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE Transactions on Industrial Informatics (TII), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), and IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (JSTSP). His research mainly focuses on artificial intelligence and deep learning, machine vision and pattern recognition, and image processing.

Dr. Zhou is a Senior Member of the Communication Society, and a CCF Member and a member of the China Artificial Intelligence Society.

Hongyan Zhang (Senior Member, IEEE) received the Ph.D. degree from Wuhan University, Wuhan, China, in 2010.

Currently, he is the Dean of the School of Computer Science, China University of Geosciences, Wuhan, and the Director of the Key Laboratory of Intelligent Geoscience Information Processing, Hubei, China. He is a Visiting Professor with Ghent University, Ghent, Belgium, and a National-Level Young Talent. He has authored or coauthored over 110 articles in authoritative journals in the fields of information science, Earth science, remote sensing science, and agricultural science, including IEEE TRANSACTIONS ON IMAGE PROCESSING, *Earth System Science Data*, ISPRS *Journal of Photogrammetry and Remote Sensing*, and *Agricultural and Forest Meteorology*. His research mainly focuses on intelligent information processing and agricultural remote sensing monitoring.

Lizhe Wang (Fellow, IEEE) received the Ph.D. degree from Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2007.

Currently, he is a Professor with China University of Geosciences, Wuhan, China. His research interests include digital Earth theory, remote sensing information engineering, and geological information applications.

Dr. Wang is an Academician of the Academia Europaea and SPIE Fellow. He is a recipient of the National Distinguished Youth Science Fund. He serves as an Editorial Board Member for international journals such as *Scientific Data*, IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS (J-MASS), IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (J-STARS), and *International Journal of Design Engineering* (IJDE).