Select, Purify, and Exchange: A Multisource Unsupervised Domain Adaptation Method for Building Extraction

Shuang Wang, *Senior Member, IEEE*, Qi Zang, *Student Member, IEEE*, Dong Zhao, *Student Member, IEEE*, Chaowei Fang, Dou Quan, *Member, IEEE*, Yutong Wan, Yanhe Guo, *Member, IEEE*, and Licheng Jiao, *Fellow, IEEE*

***Abstract*— Accurately extracting buildings from aerial images has essential research significance for timely understanding human intervention on the land. The distribution discrepancies between diversified unlabeled remote sensing images (changes in imaging sensor, location, and environment) and labeled historical images significantly degrade the generalization performance of deep learning algorithms. Unsupervised domain adaptation (UDA) algorithms have recently been proposed to eliminate the distribution discrepancies without re-annotating training data for new domains. Nevertheless, due to the limited information provided by a single-source domain, single-source UDA (SSUDA) is not an optimal choice when multitemporal and multiregion remote sensing images are available. We propose a multisource UDA (MSUDA) framework SPENet for building extraction, aiming at selecting, purifying, and exchanging information from multisource domains to better adapt the model to the target domain. Specifically, the framework effectively utilizes richer knowledge by extracting target-relevant information from multiple-source domains, purifying target domain information with low-level features of buildings, and exchanging target domain information in an interactive learning manner. Extensive experiments and ablation studies constructed on 12 city datasets prove the effectiveness of our method against existing state-of-the-art methods, e.g., our method achieves 59.1% intersection over union (IoU) on Austin and Kitsap ⟶ Potsdam, which surpasses the target domain supervised method by 2.2%. The code is available at https://github.com/QZangXDU/SPENet.**

***Index Terms*— Building extraction, deep learning, information constraint, multisource, unsupervised domain adaptation (UDA).**

I. Introduction

THE building extraction task aims to identify the building region in images, which generally assigns a class label for each pixel. In recent decades, the automatic extraction of buildings from aerial images has been a hot research topic. It is crucial in many applications, such as urban planning [37], natural resource protection [33], and land resource monitoring [59]. With the continuous development of remote sensing technology, many airborne and spaceborne images are available for learning building extraction models. Thus, data-dependent deep learning is gradually being applied to tackle this task, showing more powerful representation capabilities than traditional methods based on artificially designed features.

The success of deep learning depends heavily on having vast amounts of labeled training data for optimization. Meanwhile, the test data usually have a similar distribution with the training data, which ensures the deep learning methods have good generalization performances on test data. However, due to variations of imaging mechanisms [optical and synthetic aperture radar (SAR)], imaging sensors (spectrum and resolution), imaging environments (illumination and climate), and imaging locations (city and urban), remote sensing images show significant discrepancies between the labeled training data and test data. In addition, the artificial buildings of different regions exhibit significant cultural variation in style and structural features. These differences significantly harm the generalization capacity of deep learning methods. Thus, adapting models learned on labeled training data to test data without extra data annotation is one of the main challenges for applying deep learning to building extraction.

Unsupervised domain adaptation (UDA) is an effective method for solving the problem of distribution differences (also known as domain shift) between a labeled source domain and an unlabeled target domain, making the model trained on the source domain adapt well to the target domain. For the semantic segmentation of natural images, many UDA methods have been proposed to reduce the shift between two domains, via aligning the distribution at three levels, i.e., input level [31], [35], [51], feature level [8], [9], [48], and output level [40], [46], [61], or using pseudo-labels of the target domain generated by models pretrained on the source domain to retrain models, i.e., self-training [22], [60], [64]. Inspired by these works, in remote sensing semantic segmentation, some works [17], [44], [53] have gradually been proposed to narrow

Manuscript received 6 April 2022; revised 6 March 2023; accepted 20 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62271377, Grant 62201407, and Grant 61977052; in part by the Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-06 and Grant 2022ZDLGY01-12; in part by the National Key Research and Development Program of China under Grant 2021ZD0110404; in part by the Key Scientific Technological Innovation Research Project by Ministry of Education; in part by the China Postdoctoral Science Foundation under Grant 2022M722496; and in part by the Xidian University—SenseTime Joint Laboratory of Smart Health Care. *(Corresponding author: Shuang Wang.)*

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi 710071, China (e-mail: shwang@mail.xidian.edu.cn; qzang@stu.xidian.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3291876.

Digital Object Identifier 10.1109/TNNLS.2023.3291876

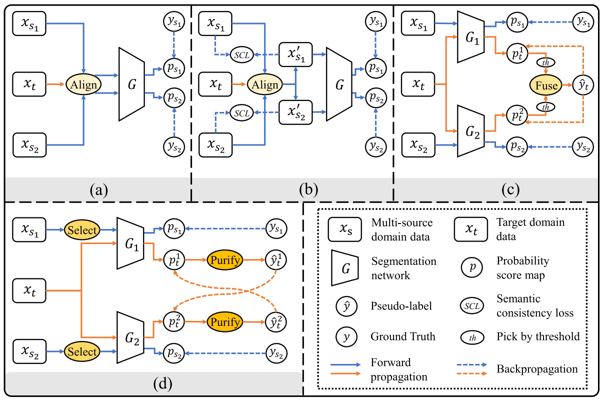


Fig. 1. Schematic of the framework comparison between the proposed method and the existing MSUDA methods. (a) ColorMapGAN [45]. (b) MADAN [63]. (c) Multisource domain adaptation with collaborative learning (MSDACL) [11]. (d) Ours (SPENet).

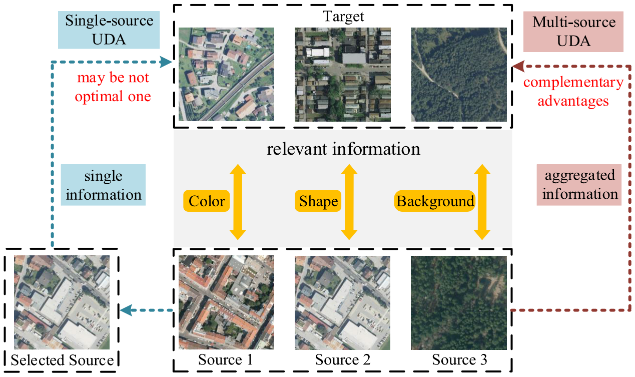


Fig. 2. Comparison of adaptation using single-source domains versus multiple-source domains. A randomly selected source domain may not be suitable due to the large disparity with the target and limited information, whereas multisource domains contain rich target-relevant information that is beneficial to adaptation.

such domain shifts by aligning at different levels. Nevertheless, these works merely focus on exploring source data from a single domain. The information provided by single-source domain is limited, so that aligning distributions is hard to achieve, especially for remote sensing scenes with larger shifts, such as those caused by sensor or cultural changes.

Fortunately, multitemporal and multiregional remote sensing images are available, providing much richer labeled source data. Training with multiple-source domains can further encourage the model to learn essential information, enhancing its adaptability to the target domain. A straightforward approach is to use multiple-source datasets by combining them as a unified source domain. In that case, the model ignores the complementary information contained in different domains, so that it is restricted from learning stronger domain adaptability. Hence, the current single-source UDA (SSUDA) building extraction framework is no longer suitable for the multisource scene.

Effective exploration of complementary information is the key to multisource UDA (MSUDA). Several MSUDA methods have recently been proposed for semantic segmentation of natural and remote sensing images. They can be classified into two streams: 1) *domain alignment*—align distributions of multiple-source and target domains at a certain level and 2) *self-training*—use pseudo-labels of the target domain from different networks to retrain each network. The existing alignment-based methods [45], [63] align multiple domains at the input level by translating the image style, as shown in Fig. 1(a) and (b). However, such alignment is not sufficient to narrow shifts, because building scenes have variations in shape and background except for style, and easily causes distortion of complex remote sensing data. Besides, they ignore the mining of target-specific knowledge (e.g., pseudo-label). Thus, this article follows the self-training-based method. He et al. [11] fuse high-quality pseudo-labels picked by the threshold to retrain networks, weakening the interference of label noise caused by shifts to training, as shown in Fig. 1(c). However, the predictions of unadapted networks are so unreliable that false high-quality ones are picked. Furthermore, the model is forced to learn all data from multisource domains, introducing some information irrelevant to the target domain, with negative effects on the network's learning efficiency. Therefore, it is essential to adaptively select out information relevant to the target domain from multisource domains.

To address these problems, this article innovatively proposes an MSUDA framework called SPENet of building detection for the first time, **s**electing, **p**urifying, and **e**xchanging information from multisource domains to better adapt to the target. Our observation shows the case of target-relevant information in building extraction. As shown in Fig. 2, the color and shape of the building in the target are similar to the first and second source domains, and the background is similar to the third source domain. This information provides positive influence for adaptation. In contrast, the redundant (source-specific) information from the task-irrelevant factors may bring negative effects for adaptation and produce uninformative gradients. Moreover, as the number of source domains increases, more source-specific information may be introduced, making it challenging to extract target-relevant information from different source domains. A multisource domain selection module with domain recognition capabilities is designed to guide the network to focus on more relevant regions in the multisource domains. Due to the diversity of multisource domains, the models trained on target-relevant information of each source domain have specific advantages for the target domain, making their predictions on the target data complementary. An interactive learning strategy is proposed to aggregate this complementary information. During interactive learning, to purify noisy predictions of the target domain, low-level building structural features are exploited, and an uncertainty estimation module is designed to provide explicit information constraints, generating reliable supervision signals for networks. The main contributions can be summarized as follows.

1) A novel MSUDA framework SPENet is developed for building extraction. The framework improves the efficient utilization of rich information by selecting, purifying, and exchanging information from multisource domains, boosting the model's adaptation performance to the target domain.
2) In this framework, a multisource domain selection module is designed to provide positive support for adaptation by select the target-relevant information from multisource domains. An uncertainty estimation module is designed to purify pseudo-labels of the target domain by exploiting building structural features.
3) Extensive experiments are conducted on multiple public datasets, proving that the proposed framework significantly outperforms the existing state-of-the-art methods. Our framework is more robust to the target domain and improves the generalization ability of the network.

II. Related Work

A. Building Extraction Task

In the early years, many methods separate buildings from the background based on low-level features, such as geometry and texture [1], [5], [7], [15], [20], [43]. Among these methods, Huertas and Nevatia [15] assume that the building's shape is a regular rectangle "L" or "T" and employ geometric constraints to extract buildings from aerial images. Lorette et al. [27] propose a chain-based Gaussian model to analyze texture information and incorporate the obtained texture parameter into the Markov segmentation model. Sirmacek and Unsalan [43] use color features to extract shadow information, which is subsequently used to determine building locations. Li et al. [21] leverage the relationship between buildings and their cast shadows to implement the extraction of buildings. These methods promote the development of building extraction technology. However, the features extracted manually in traditional methods are less discriminative and cannot perform well in complex scenes.

In recent years, deep learning has gradually been applied in many methods [2], [3], [14], [16], [24], [26], [36], [38], [54], [58]. Li et al. [24] use adversarial learning for building extraction from remote sensing images. They adopt an autoencoder network as a discriminator to stably learn the high-order structural features of buildings. Hui et al. [16] apply the separable convolution module Xception to U-Net to independently consider remote sensing images' spectral and spatial correlation. Yuan [58] integrates multiple activation layers and uses the signed distance function of building contour as the final output. Ye et al. [57] introduce reweighted attention to adaptively integrate shallow features into deep features, which enhances the usage of shallow features in convolutional neural networks. Although various improvements promote the performance of building extraction, the trained network could not be directly generalized to unlabeled images with domain differences from the training data. Our method aims to address the problem of how the model extracts buildings from the unlabeled image when there is a domain shift between unlabeled images and labeled images.

B. Single-Source UDA

The SSUDA aims to narrow the domain gap between the labeled source dataset and the unlabeled target dataset. Some methods directly align the two domains' distributions in the original input space, based on style transfer techniques [10], [23], [31], [35], [51], [55]. These studies reduce the image style discrepancy between different domains. On the other hand, Hoffman et al. [12] propose to align the global distribution of the source and target domains at the feature space by adversarial training, but the improvement is relatively limited. Inspired by this idea, many works try to improve the feature alignment method in semantic segmentation. One improvement direction is to alleviate the problem of difficult alignment in high-dimensional space. In [13] and [41], they propose to align the two domains in the transformed low-dimensional space. Then, Tsai et al. [46], [47] directly treat the output of the classifier as the transformed space, expecting to align the class layout of the two domains at the output space. Another improvement direction is to align the feature distribution in a fine-grained fashion. Chen et al. [4] and Du et al. [6] turn the feature distribution into a class distribution and design multiple class discriminators to align the feature distributions of each class separately. Wang et al. [49] turn it into an instance distribution and align the distribution by matching the distribution statistics of the features. The abovementioned works focus on inter-domain alignment, and some methods also use the self-training strategy for intra-domain adaptation, which uses pseudo-labels of the target domain obtained from network inference as training supervision [22], [25], [60], [64].

Recently, some methods have been proposed for semantic segmentation in the remote sensing. Yan et al. [53] first apply the UDA method for remote sensing semantic segmentation to align distributions of source and target domains at the output level. However, since the specific spatial layout structure in natural images does not exist in remote sensing images, it is hard to find a reliable optimization goal for aligning distributions at the output level. Thus, Iqbal and Ali [17] propose to learn the weak label of target data (whether there are buildings in the image) for building extraction and align distributions at the feature level and the output level. Weak labels provide additional constraint for improving the representation capacity of the network, but their requirement for manual annotation still increases the cost. Tang et al. [44] align distributions at the input level and output level, but the style translate for input distribution alignment easily causes distortion for remote sensing data captured in diverse sensors and complex scenes. Besides, the information from one domain exploited by these methods is limited. The multitemporal and multiregion characteristics of remote sensing make data in multiple domains available, providing richer information. Thus, we use data from multiple-source domains to improve the adaptability of the network to the target domain.

C. Multisource UDA

Involving multiple-source domain datasets can expand the information repository for fostering the segmentation model in the target domain. Recently, some MSUDA methods for image classification have been proposed [34], [52], [62]. The UDA algorithms designed for image classification perform poorly on the semantic segmentation task, because the latter requires dense pixel-level prediction, which increases the difficulty of adaptation. Zhao et al. [63] propose a multisource adversarial domain aggregation network (MADAN) for semantic segmentation. The process first translates each source domain style to the target and then aggregates them. The final aggregated domain and the target are aligned in the feature space. Tasar et al. [45] design a color mapping generative adversarial networks (ColorMapGANs), which makes each source domain and target domain have similar spectral distribution. He et al. [11] align the distribution in the input space through style transfer and force multiple-source models to produce consistent predictions on the target domain.

The above methods align the styles of multiple domains for adaptation, which is not enough for building extraction, because building scenes have variations in shape and background except for style differences. Also, these methods force the model to learn from all source domains data, creating a risk that irrelevant information will negatively influence the model's learning. Our method devises a information selection module to explore target-relevant information from each source domain. Furthermore, inspired by traditional unsupervised methods for extracting buildings using low-level features, we introduce low-level features into deep neural networks to focus on building structures in different domains.

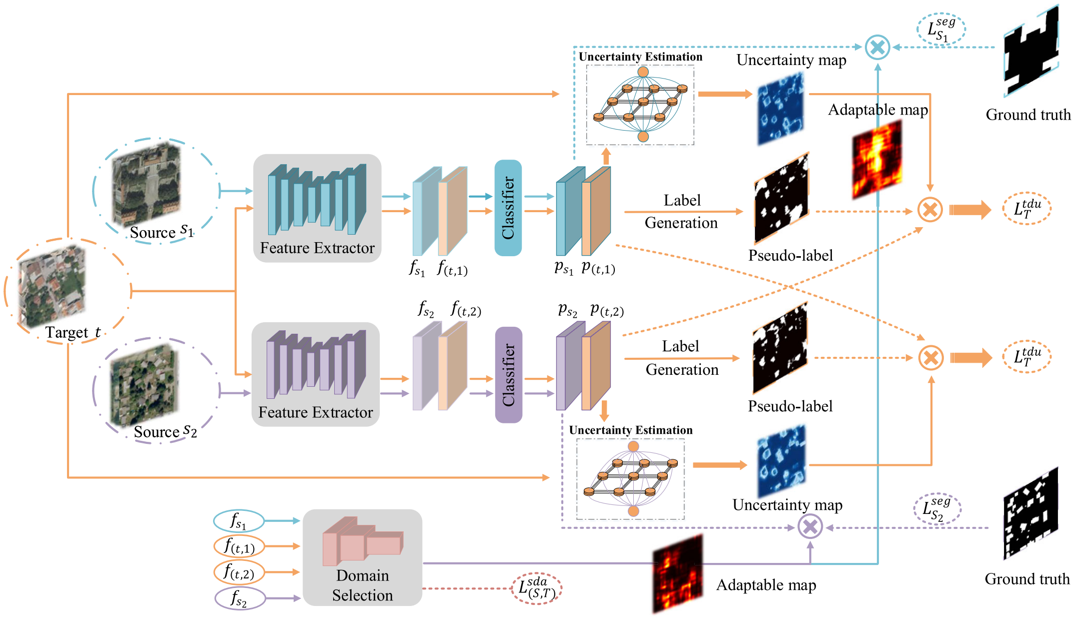


Fig. 3. An overview of the proposed **SPENet**. First, the encoded features of the source and target images are fed into the domain selection network. Then, the domain selection network generates adaptable maps for source images, and these adaptable maps are weighted on the losses $\mathcal{L}^{\text{seg}}_{S_1}$, $\mathcal{L}^{\text{seg}}_{S_2}$. Second, the pseudo-labels generated for the target images by segmentation networks trained on multisource domains are exchanged in an interactive learning way. Third, the uncertainty estimation module further purifies the pseudo-labels by generating uncertainty maps for them, and these maps are used to reweigh the loss $\mathcal{L}^{\text{tdu}}_T$. Besides, the subscripts of f and p indicate the feature maps and predicted probability scores produced by different segmentation networks for source or target images, respectively. The solid lines in different colors indicate the forward inference processes on different input images, and the dotted lines indicate the calculation flow of different losses. During testing, we integrate the results of all networks $y = (\hat{y}_i + \hat{y}_j)/2$ as the prediction result.

III. Proposed Method

In this section, we introduce the technical details of the proposed MSUDA framework for building extraction. The formula setting is as follows.

Preliminaries: We assume the training data consist of I source domains $\mathbb{D}_{s_i}(i \in [1, I])$ and a target domain $\mathbb{D}_t$. The ith source domain contains samples $\{x_{(s_i,j)}\}_{j=1}^{N_{s_i}}$ and their corresponding pixel-level labels $\{y_{(s_i,j)}\}_{j=1}^{N_{s_i}}$. The target domain contains samples $\{x_{(t,k)}\}_{k=1}^{N_t}$. N_{s_i} and N_t are the number of samples in the ith source domain and the target domain, respectively. The categories in each source domain are the same as those in the target domain. Given these training data, we aim to adapt I segmentation networks $\{\mathcal{G}_i\}_{i=1}^{I}$ trained on each source domain to the target domain. In general, $\mathcal{G}$ consists of a feature extractor $\mathcal{F}$ and a classifier $\mathcal{C}$. $\mathcal{F}$ maps the sampled image $x \in \mathbb{D}_{s_i} \cup \mathbb{D}_t$ to the feature space and obtains a feature map $f = \mathcal{F}(x) \in \mathbb{R}^{h\times w\times c}$, where h, w, and c stand for the height, width, and channel numbers of f. Then, $\mathcal{C}$ classifies each point in f to obtain a probability score map $p = \mathcal{C}(\mathcal{F}(x)) \in \mathbb{R}^{H\times W}$, where $H \times W$ stands for the size of the input image x.

A. Overview of Framework

To effectively explore multiple-source domains, we propose a novel framework for building extraction based on **s**electing, **p**urifying, and **e**xchanging information from different source domains, namely, **SPENet**. As shown in Fig. 3, for each source domain, SPENet builds a segmentation network with the same structure but different parameters to extract diverse domain information. Three components corresponding to the above three goals are introduced as follows.

1) *Source Domain Information Selection:* We first design a domain selection network, which is optimized to identify domain identifiers of input features. For each source image, an adaptable map is then generated to evaluate its relevance against the target domain, ensuring each segmentation network focuses on target-relevant information (the technical details will be introduced in Section III-B).
2) *Information Exchange Among Domains:* An interactive strategy is devised to aggregate complementary information of different domains, through making segmentation models teach each other on unlabeled target domain images (the technical details will be introduced in Section III-C).
3) *Purification of Target Domain Information:* We design an uncertainty estimation method to purify the target domain information, which is used for relieving the influence of uncertain predictions for each network (the technical details will be introduced in Section III-D).

B. Source Domain Information Selection

In general, for the network of each source domain, information extraction is achieved by forcing the distribution of the output class probability and the corresponding label to be close, i.e., optimizing the pixel-level cross-entropy loss $\mathcal{L}^{\text{seg}}_S$

$$\mathcal{L}^{\text{seg}}_S = -\frac{1}{|\mathbb{D}_s|}\sum_{i=1}^{I}\sum_{j=1}^{N_{s_i}}[y_{(s_i,j)}\log(p_{(s_i,j,i)}) + (1 - y_{(s_i,j)}) \times \log(1 - p_{(s_i,j,i)})]$$

where

$$p_{(s_i,j,i)} = \mathcal{C}_i(\mathcal{F}_i(x_{(s_i,j)})). \quad (1)$$

$p_{(s_i,j,i)}$ is the prediction generated by the ith network trained on the source domain s_i for the jth image in s_i. Cross entropy is a well-known measure of information. The principle of minimizing cross entropy is the inference of an unknown distribution p under the guidance of a prior distribution y [42]. Also, by this means, all source domain information is regarded equally. However, the information in each source domain may be only partially relevant to the target domain (e.g., color, shape, and background), which is conducive to adaptation. Thus, we focus on extracting target-relevant information in multisource domains. A source domain selection network $\mathcal{D}$ with domain recognition capabilities is designed, generating an attention map that has larger values on target-relevant regions for each source image. To achieve this, $\mathcal{D}$ is forced to perform a pixel-level domain recognition task, in which the features in the input feature map f are identified as the source or target.

Specifically, the domain selection network $\mathcal{D}$ is composed of four convolutional blocks with sliding step of 2 and a sigmoid layer. Given the training samples $\mathbb{D}_s$ of multisource domains and the training samples $\mathbb{D}_t$ of the target domain, $\mathcal{D}$ classifies the domain to which the feature maps of these samples belong. Formally, the domain selection network $\mathcal{D}$ is trained with the source domain adaptable loss $\mathcal{L}^{\text{sda}}_{(S,T)}$

$$\mathcal{L}^{\text{sda}}_{(S,T)} = -\frac{1}{|\mathbb{D}_t|}\sum_{i=1}^{I}\sum_{k=1}^{N_t}\log(\mathcal{D}(\mathcal{F}_i(x_{(t,k)}))) - \frac{1}{|\mathbb{D}_s|}\sum_{i=1}^{I}\sum_{j=1}^{N_{s_i}}\log(1-\mathcal{D}(\mathcal{F}_i(x_{(s_i,j)}))). \quad (2)$$

Under this supervision, $\mathcal{D}$ has the ability to evaluate the probability score of the domain identifier corresponding to the input feature. Then, for source domain features that are predicted to be the target domain with high probability scores, they are more relevant to the target domain and should be stressed with higher importance during the adaptation process.

Based on this, the probability score maps output by $\mathcal{D}$ is further used as adaptable maps $A_{s_i} = \{a_{(s_i,j)}\}_{j=1}^{N_{s_i}} \in \mathbb{R}^{H\times W}$ for each image $x_{(s_i,j)}$ of multisource domains

$$a_{(s_i,j)} = \beta + \mathcal{D}(\mathcal{F}_i(x_{(s_i,j)})) \quad (3)$$

where β is a hyperparameter to balance the network's learning of target-irrelevant information and maintain a neutral focus on them. This map indicates the relevance between each pixel in images and the target domain. When the probability value in the map is larger, it indicates that the pixel at that position is more relevant to the target domain. In this way, target-relevant information is effectively selected from multisource domains by the network $\mathcal{D}$.

Then, we combine the generated adaptable maps A_{s_i} to reweight the loss function $\mathcal{L}^{\text{seg}}_S$. The modified $\mathcal{L}^{\text{seg}}_S$ is used to supervise labeled samples from multisource domains, which is defined as follows:

$$\mathcal{L}^{\text{seg}}_S = -\frac{1}{|\mathbb{D}_s|}\sum_{i=1}^{I}\sum_{j=1}^{N_{s_i}} a_{(s_i,j)}[y_{(s_i,j)}\log(p_{(s_i,j,i)}) + (1-y_{(s_i,j)}) \times \log(1-p_{(s_i,j,i)})]. \quad (4)$$

In this way, the regions whose representations are more adaptable are weighted with larger attention values, thus allowing the segmentation networks $\mathcal{G}_i$ to focus on those more relevant regions during learning. Meanwhile, target-irrelevant information is adaptively filtered out to prevent the networks from being interfered by redundant information.

C. Information Exchange Among Domains

Different source domains have diverse building appearances and backgrounds, so the target-relevant information selected from multisource domains may be complementary. Therefore, we propose an interactive learning strategy to take advantage of the complementary target-relevant information from different domains. This strategy enables the segmentation network $\mathcal{G}_i$ trained on a certain source domain to teach other networks. Since different networks learn the segmentation capabilities of different target-relevant regions from multiple-source domains, the proposed interaction method can fully aggregate the specific advantages of different networks and improve the adaptability of each network to the target domain.

Specifically, the target domain information (i.e., target domain pseudo-label) learned from the segmentation network $\mathcal{G}_i$ is exchanged to supervise other networks during training. Given an image $x_{(t,k)}$ in the target domain $\mathbb{D}_t$, it is fed into all segmentation networks $\{\mathcal{G}_i\}_{i=1}^{I}$, and the corresponding pseudo-labels $\{\hat{y}_{(t,k,i)}\}_{i=1}^{I} \in \mathbb{R}^{H\times W}$ are calculated with (5)

$$\hat{y}^{(h,w)}_{(t,k,i)} = \begin{cases} 1, & \text{if } p_{(t,k,i)} \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

where

$$p_{(t,k,i)} = \mathcal{C}_i(\mathcal{F}_i(x_{(t,k)})). \quad (5)$$

$p_{(t,k,i)}$ is the prediction generated by the ith network trained on arbitrary source domain s_i for the kth target image. Then, in the interactive learning process, the pseudo-labels generated by other networks $\{\mathcal{G}_i\}_{i=1,i\neq l}^{I}$ for the target image $x_{(t,k)}$ supervise the network $\mathcal{G}_l$'s prediction of it. For the lth network $\mathcal{G}_l$, the objective function $\mathcal{L}^{\text{td}}_T$ is as follows:

$$\mathcal{L}^{\text{td}}_T = -\frac{1}{|\mathbb{D}_t|}\sum_{i=1,i\neq l}^{I}\sum_{k=1}^{N_t}[\hat{y}_{(t,k,i)}\log(p_{(t,k,l)}) + (1-\hat{y}_{(t,k,i)}) \times \log(1-p_{(t,k,l)})]$$

where

$$p_{(t,k,l)} = \mathcal{C}_l(\mathcal{F}_l(x_{(t,k)})). \quad (6)$$

As the process of interactive learning continuously advances, the gap between each source domain and the target domain is gradually bridged. Furthermore, multiple segmentation networks $\{\mathcal{G}_i\}_{i=1}^{I}$ are also encouraged to maintain consistent predictions for the same target images, thus cooperating with multiple networks to boost the network performance.

D. Purification of Target Domain Information

Due to domain differences, pseudo-labels generated by the network of each source domain are impure, which seriously affects the performance of interactive learning. A simple approach to purification is to pick high-quality pseudo-labels by setting a confidence threshold. However, the output probability of the unadapted network is unreliable, and the threshold is hard to set. In this article, we utilize low-level structural features of the building to design a novel uncertainty estimation method to purify the pseudo-labels of the target domain.

Buildings have low-level structural characteristics, such as small pixel variances in building blocks and large pixel variances between buildings and backgrounds. Thus, some traditional unsupervised building extraction works [15], [21] adopt pixel intensity and spatial location of the building as structural features for describing the pixelwise visual features. Such strong structural constraints and accurate building edge information implied in low-level structural features can provide reliable cues for purification.

Inspired by these works, we introduce low-level structural features of the building into deep neural networks in the purification process and utilize their consistency to estimate the accuracy of the network prediction. Specifically, given a target domain image $x_{(t,k)}$, we first construct an affinity matrix $W_k \in \mathbb{R}^{HW \times HW}$ to indicate the affinity between pixels in the image. For two pixels (x_n, y_n) and (x_m, y_m) in $x_{(t,k)}$, the affinity between their structural features is defined as follows:

$$W_k^{nm} = \exp\Bigg(-\frac{(x_n - x_m)^2 + (y_n - y_m)^2}{2\sigma_s^2} - \frac{\|I(x_n, y_n) - I(x_m, y_m)\|^2}{2\sigma_i^2}\Bigg) \quad (7)$$

where σ_s and σ_i represent the standard deviation of the spatial domain and the intensity domain, respectively. $I(x_n, y_n)$ and $I(x_m, y_m)$ denote the intensity of pixels n and m, respectively. Based on the affinity matrix W_k, an uncertainty matrix $M_{(t,k,i)} \in \mathbb{R}^{HW \times 2}$ is generated by modeling the W_k and the probability score map $p_{(t,k,i)}$ with (8), which estimates unreliable regions of the pseudo-label $\hat{y}_{(t,k,i)}$

$$\begin{aligned} P_{(t,k,i)} &= [p_{(t,k,i)}, 1 - p_{(t,k,i)}] \\ M_{(t,k,i)} &= (W_k P_{(t,k,i)}) * P_{(t,k,i)} \end{aligned} \quad (8)$$

where $*$ represents elementwise multiplication and $P_{(t,k,i)}$ is a constructed matrix and is resized to $HW \times 2$ for calculating the uncertainty matrix. Specifically, the element $M_{(t,k,i)}^{(h,w)}$ with the coordinate (h, w) in the matrix $M_{(t,k,i)}$ is obtained by dot multiplying the elements in row h in W_k with the elements in column w in $P_{(t,k,i)}$ and then adding them, which can be represented as follows:

$$\begin{aligned} M_{(t,k,i)}^{(h,w)} &= \big[\big(W_k^{h1} P_{1w} + W_k^{h2} P_{2w} + \cdots + W_k^{h(HW)} P_{(HW)w}\big) \\ &\quad \times P_{hw}\big] / W_k^{h1} + W_k^{h2} + \cdots + W_k^{h(HW)} \\ &= \big(W_k^{h1} P_{1w} P_{hw} + W_k^{h2} P_{2w} P_{hw} + \cdots + W_k^{h(HW)} \\ &\quad \times P_{(HW)w} P_{hw}\big) / W_k^{h1} + W_k^{h2} + \cdots + W_k^{h(HW)} \end{aligned} \quad (9)$$

where W_k^{h1} is the element at coordinate $(h, 1)$ in affinity matrix W_k and P_{1w} is the element at coordinate $(1, w)$ in constructed matrix $P_{(t,k,i)}$. As shown in (9), for a pixel in the target image $x_{(t,k)}$, the affinity matrix W_k imposes a positive constraint on whether the network $\mathcal{G}_i$ classifies the pixel and other pixels into the same class. That is, if $\mathcal{G}_i$ correctly classifies both pixels, then the constrained value will be larger. Otherwise, the value becomes smaller. We sum the elements of the matrix $\{M_{(t,k,i)}\}_{k=1}^{N_t}$ by rows and reshape them into $H \times W$ as the uncertain maps $U_{(t,k)} = \{u_{(t,k,i)}\}_{(i=1,i\neq l)}^{I}$. Based on the maps $U_{(t,k)}$, the $\mathcal{L}_T^{\text{td}}$ is modified to generate the following objective function $\mathcal{L}_T^{\text{tdu}}$:

$$\begin{aligned} \mathcal{L}_T^{\text{tdu}} = -\frac{1}{|\mathbb{D}_t|} \sum_{i=1, i\neq l}^{I} \sum_{k=1}^{N_t} u_{(t,k,i)} [\hat{y}_{(t,k,i)} \log(p_{(t,k,l)}) \\ + (1 - \hat{y}_{(t,k,i)}) \log(1 - p_{(t,k,l)})] \end{aligned} \quad (10)$$

where $u_{(t,k,i)}$ is the generated uncertainty map for the kth target image. Optimizing the $\mathcal{L}_T^{\text{tdu}}$ suppresses the negative effects of uncertain pseudo-labels. Therefore, during the information exchange, the target domain information can be positively constrained to generate reliable supervision signals for other segmentation networks.

E. Optimization Objective

For the proposed SPENet, the final optimization process is composed of two parts, including the simultaneous optimization of domain selection network $\mathcal{D}$ and semantic segmentation network $\mathcal{G}_i$. Specifically, the loss $\mathcal{L}_{(S,T)}^{\text{sda}}$ is used for optimizing the domain selection network and learns the ability to identify the domain identifier of features. The loss $\mathcal{L}_S^{\text{seg}}$ on the labeled source domains and the loss $\mathcal{L}_T^{\text{tdu}}$ on the unlabeled target domain are used for optimizing the segmentation network. The overall objective function for optimization can be expressed as follows:

$$\arg\min_{\theta_{\mathcal{D}}} \lambda^{\text{sda}} \mathcal{L}_{(S,T)}^{\text{sda}} + \arg\min_{\theta_{\mathcal{G}_i}} \left(\mathcal{L}_S^{\text{seg}} + \mathcal{L}_T^{\text{tdu}}\right) \quad (11)$$

where $\{\theta_{\mathcal{G}_i}\}_{i=1}^{I}$ is the parameters of corresponding segmentation networks $\{\mathcal{G}_i\}_{i=1}^{I}$, θ_d is the parameters of the network $\mathcal{D}$, and λ^{sda} is the trade-off coefficient of $\mathcal{L}_{(S,T)}^{\text{sda}}$ to avoid the interference caused by its poor recognition ability (especially early training) on adaptation.

In Algorithm 1, we illustrate the detailed training procedure of **SPENet**. After the training is terminated, a set $\mathcal{G}$ of multiple optimal segmentation networks and an optimal selection network will be obtained. During inference, the outputs of $\mathcal{G}$ are averaged as the final prediction for the target domain.

IV. Experiments

In this section, we conduct extensive experiments on five datasets to evaluate the performance of the proposed method. We first describe the datasets and the experimental setup. Then, we carry out comparisons with the existing state-of-the-art methods and elaborate ablation studies to validate the effectiveness of each component.

A. Datasets and Experimental Setup

1) Datasets: Considering the differences in scene, sensor, and resolution, we use 12 cities from five datasets to validate the proposed method, i.e., Massachusetts dataset [30], Village Finder dataset [32], Potsdam dataset [18], institut national de recherche en infomatique et automatique (INRIA) dataset [29], and Gaofen-2 (GF-2) dataset.

a) Massachusetts buildings dataset: This dataset consists of 151 aerial images in the Boston area, the size of each image is 1500 × 1500, and the resolution is 1.0 m. There are differences in sensors, terrain, and resolution between this dataset and the other four datasets. The Massachusetts dataset is a single-source high-resolution dataset covering only built-up areas in a single geographical location. We use 137 images as the training set, four images as the validation set, and 17 images as the testing set. These datasets are available freely at the website.[1]

b) Village finder dataset: The village finder dataset is collected over areas covering more than 100 km^2, containing nucleated villages in 15 countries and spread over four continents. This dataset has 60 satellite images of size 2400 × 2400 captured by different sensors from Google Earth; 25 images are provided for training, ten images for validation,

[1] [Online]. Available: https://www.cs.toronto.edu/~vmnih/data/

TABLE I

Details of Publicly Available and Our Own Datasets Before Preprocessing (m/p: Number of Square Meters per Pixel and Image Size: the Original Size Provided by the Dataset)

Dataset	Sub-Dataset	Total Images	Training Images	Validation Images	Testing Images	Image Size	Resolution (m/p)
Massachusetts [30]	–	151	137	4	10	1500 × 1500	1.0
Village Finder [32]	–	60	25	10	25	2400 × 2400	0.54
ISPRS-Potsdam [18]	–	38	16	2	18	6000 × 6000	0.05
INRIA [29]	Austin	36	19	2	15	5000 × 5000	0.3
	Kitsap	36	19	2	15	5000 × 5000	0.3
	Tyrol	36	19	2	15	5000 × 5000	0.3
	Chicago	36	19	2	15	5000 × 5000	0.3
	Vienna	36	19	2	15	5000 × 5000	0.3
GF-2	Beijing	64	35	5	24	2500 × 2500	0.81
	Hangzhou	64	35	5	24	2500 × 2500	0.81
	Wuhan	64	35	5	24	2500 × 2500	0.81
	Chongqing	64	35	5	24	2500 × 2500	0.81

Algorithm 1 Proposed **SPENet**

```
Input:
    Multi-source domain data $\mathbb{D}_{s_i} = \{x_{(s_i,j)}\}_{j=1}^{N_{s_i}}$ $(i \in [1, I])$
    and labels $Y_i = \{y_{(s_i,j)}\}_{j=1}^{N_{s_i}}$, target domain data
    $\mathbb{D}_t = \{x_{(t,k)}\}_{k=1}^{N_t}$, total epoch: $E$.
Output:
    Optimized multiple segmentation networks and a source
    domain selection network: $\mathcal{G} = \{\mathcal{G}_i\}_{i=1}^{I}$, $\mathcal{D}$
    $\mathcal{G}_i = (\mathcal{F}_i, \mathcal{C}_i)$.
Firstly, train multiple basic segmentation networks $\mathcal{G}$ with $\mathbb{D}_s$;
for $e = 1$ to $E$ do
    Extract feature maps $f_{(s_i,j)}$, $f_{(t,k)}$ using the feature extractor $\mathcal{F}$;
    Get the prediction maps $p_{(s_i,j,i)}$, $p_{(t,k,i)}$ from the networks $\mathcal{G}$;
    Source domain selection: generate the adaptable map $\{A_{s_i}\}_{i=1}^{I}$ by the selection network $\mathcal{D}$;
    Calculate the source domain adaptable loss $\mathcal{L}_{(S,T)}^{sda}$ based on $f_{(s_i,j)}$, $f_{(t,k)}$ with Eq. (2);
    Exchange among domains: utilize the $p_{(t,k,i)}$ to obtain pseudo-labels $\hat{y}_t$ of the target domain data with Eq. (5);
    Calculate the segmentation loss $\mathcal{L}_S^{seg}$ based on $p_{(s_i,j,i)}$, $\{A_{s_i}\}_{i=1}^{I}$, $y_{s_i}$ with Eq. (4);
    Target domain purifying: generate the uncertain map $U_{(t,i)}$ by modeling with Eq. (8) - (9);
    Calculate the target domain uncertainty loss $\mathcal{L}_T^{tdu}$ based on $p_{(t,k,i)}$, $U_{(t,i)}$, $\hat{y}_t$ with Eq. (10);
    $\mathcal{L}$ back propagation, update parameters.
end for
Return Trained model $\mathcal{G}$
```

and 25 for testing. These are very high-resolution images with a resolution of 0.54 m. These datasets are available freely at the website.[2]

c) ISPRS 2-D semantic labeling Potsdam: The International Society for Photogrammetry and Remote Sensing (ISPRS) dataset is provided by the ISPRS 2-D Semantic Segmentation Competition and has two sub-datasets. We only use the Potsdam dataset in the ISPRS, containing 38 aerial images collected in the Potsdam area. Each image has a size of 6000 × 6000 and the highest resolution of 5.0 cm; 16 images are used for training, two images for validation, and 18 images for testing. This dataset can be downloaded from the ISPRS official website.[3]

d) INRIA aerial image labeling dataset: The INRIA dataset is a very high-resolution dataset and has ten sub-datasets covering the United States and Australia. Each sub-dataset has 36 aerial images with 0.3-m resolution, and the original image size is 5000 × 5000. Five sub-datasets in the INRIA are used for our experiments, including high-density (Chicago and Vienna) and low-density (Austin, Kitsap, and Tyrol) city residential areas. All images are obtained via the same sensor but with regional and built-up structure differences. For each sub-dataset, 19 images are used as the training set, two images as the validation set, and 15 images as the testing set. These datasets are available freely at the website.[4]

e) GF-2 built-up regions dataset: The GF-2 dataset is a high-resolution dataset covering the four cities of Beijing, Hangzhou, Wuhan, and Chongqing in China. This dataset is provided by Aerospace Hongtu Information Technology Company Ltd. and collected by the GF-2 satellite, which is different from the other four datasets. GF-2 satellite is the first civil optical remote sensing satellite independently developed by China with a spatial resolution better than 1 m. It is equipped with two high-resolution 1.0-m panchromatic and 4.0-m multispectral cameras. Due to changes in regions and sensors, this dataset is fundamentally different from other public datasets. The varying building structure size and appearance make it very challenging. Each city consists of 64 satellite images with 0.81-m resolution, and the size of each image is 2500 × 2500. We use 35 images as the training set, five images as the validation set, and 24 images as the testing set.

To verify the stability and generalization of our method, training and testing are conducted on the data of the above 12 cities. We randomly select two from 12 cities as the data of multisource domains (i.e., Austin and Kitsap, Tyrol and Chicago, and Kitsap and Vienna), and the other ten are used as the target domain. The images of each dataset are further cropped to 512 × 512 patches by a sliding window as the

[2] [Online]. Available: http://cvlab.lums.edu.pk/villagefinder/

[3] [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/semantic-labeling.html/

[4] [Online]. Available: https://project.inria.fr/aerialimagelabeling/

standard image size for this work. Besides, there are different domain gaps between the 12 cities. It can be found that Austin and Chicago have dense buildings, narrow roads, and sparse vegetation. Kitsap has abundant vegetation, and buildings are scattered. A small proportion of the images in the Tyrol are densely distributed buildings, while most of them are green plains. Vienna and Massachusetts have many wide roads and buildings. Potsdam has many roads and viaducts but fewer buildings. The buildings in the Village Finder dataset are mostly nucleated villages with diverse distributions. Four cities of the GF-2 dataset have very irregular construction patterns, with buildings next to rivers, forests, roads, and so on, and their structures vary. The standard evaluation metric intersection over union (IoU) is applied as the evaluation metric.

2) Network and Experimental Setup:

a) Segmentation network: U-Net [39] is a famous fully convolutional network in remote sensing, which has been verified to be effective in building detection tasks [17], [19], [50]. It forms a symmetrical U-shape using skip connections, combining the shallow and deep features of the network to aggregate features of different scales. Benefitting from this, we adopt U-Net as the primary segmentation network for implementing UDA in building extraction. For a fair comparison, we also reproduce other current UDA methods based on the U-Net network and perform sufficient parameter tuning.

b) Domain selection network: The domain selection network is designed based on a fully convolutional structure similar to [46]. The network consists of five convolutional layers with the kernel size of 4×4, where the number of channels is 64, 128, 256, 512, and 1. The stride of the first four layers is set to 2, and the stride of the last layer is set to 1. Each of the first four layers is followed by a LeakyReLU layer [28] parameterized by 0.2. Neither pooling layer nor batch normalization is used.

c) Implementation details: The proposed algorithm is implemented in PyTorch 1.8 on an NVIDIA Tesla P100 GPU with 16 GB of memory. The batch size is set to 4. Stochastic gradient descent (SGD) optimizer is used; the weight decay rate is set to 1×10^{-8}, and momentum is set to 0.9. The initial learning rate is set to 5×10^{-3} and is decayed following a polynomial learning rate scheduling with a power of 0.9 during training. The hyperparameters $\lambda_{(S,T)}^{\text{sda}}$ and β are set to 0.1 and 0.5, respectively.

B. Comparison With the Existing Methods

In this section, our method SPENet is compared with the existing SSUDA methods [17], [48], [53], [56] and MSUDA methods [11], [45], [63]. All SSUDA methods in Table II are reproduced under two settings. The first setting is to train on two source domains, respectively. The second setting is to mix two source domains as one domain for training. In Tables III and IV, we use the best setting (i.e., mix two source domains) in Table II to show the performance of SSUDA methods. Fig. 4 demonstrates the visual comparison results. We analyze and discuss all the above experimental results below.

1) Austin and Kitsap → Ten Targets: In the single source of Table II, the performance of different source domains adapting to the target domain is different in SSUDA methods. For example, when Austin is used as the source domain, the average IoU of SSUDA methods on ten target domains is 41.6%, 40.5%, 43.0%, and 41.1%, respectively. When Kitsap is used as the source domain, SSUDA methods only achieve the average IoU of 34.8%, 33.4%, 33.7%, and 33.8%, respectively. This shows that the domain gap between the ten target domains and Kitsap is larger than that between them and Austin. In the source-combination setting, the average performance of the four SSUDA methods [i.e., triplet adversarial domain adaptation (TriADA), weakly supervised domain adaptation (WSDA), Advent, and Fourier domain adaptation (FDA)] is slightly better than training with a single-source domain (i.e., only 1.8%, 1.8%, 1.7%, and 1.1%), which illustrates that the mixed domain provides richer information. However, the information cannot be fully mined in such a mixed manner.

In the multisource setting, MSUDA methods, ColorMapGAN [45] and MADAN [63], adapt to the target domain by changing the style of the data. Their performance in the building extraction task is not satisfactory, only achieving the average IoU of 40.3% and 39.6%. This occurs because the input-level style transfer easily leads to image distortion on remote sensing images from different sensors, and lacks adaptation to diverse backgrounds. The recently proposed MSDACL [11] is a multisource domain UDA method based on collaborative learning in computer vision community, which shows better stability than other methods. Compared with MSDACL, our method SPENet further boosts the performance. We think the reason is that we select target-relevant information from multiple-source domains, instead of learning all the information indiscriminately, and utilize low-level structural features of buildings to reduce the interference of label noise. As discussed in Section IV-A1, the distribution of buildings in Village Finder and GF-2 datasets is irregular (nucleated or variable), so the performance gains of our method on these two datasets are limited compared with other datasets. In summary, SPENet shows the best performance on each target domain, and its average performance outperforms the three MSUDA methods by 13.4%, 14.1%, and 4.7%, respectively. In particular, when using Potsdam as the target domain, our method even surpasses the supervised method by 2.2%. As shown in Fig. 4, our method can better recognize buildings of different sizes and diversified backgrounds and is more robust than other methods. The above analyses and results illustrate the effectiveness of our method SPENet.

2) Tyrol and Chicago → Ten Targets: We also conduct experiments by using Tyrol and Chicago as source domains. The results are shown in Table III. The four SSUDA methods trained on the mixed domain bring performance improvements, and their average performance on the ten target domains surpasses the mixed source-only model by 5.4%, 5.2%, 6.6%, and 3.4%. In the three MSUDA methods, ColorMapGAN and MADAN cannot be effectively used for the building extraction task. Their performance improvement is limited, and the average IoU is slightly lower than the four SSUDA methods. MSDACL shows better adaptation performance than ColorMapGAN and MADAN. Compared with these two methods, the performance of MSDACL is improved by 6.9% and 5.8%, respectively. In contrast, our method SPENet outperforms the existing SSUDA and MSUDA methods and achieves 48.3% average IoU. Compared with the three MSUDA methods, the average performance on the ten target domains is improved by 13.0%, 11.9%, and 6.1%, respectively. Even though the domain gap among the target domain Kitsap and source domains (Tyrol and Chicago) is large, SPENet can maintain better building extraction performance. This verifies our method's effectiveness in selecting target-relevant information from source domains and fully aggregating them.

3) Kitsap and Vienna → Ten Targets: Table IV reports the performance of each method using Kitsap and Vienna as

TABLE II

COMPARISON RESULTS OF ADAPTING FROM AUSTIN AND KITSAP TO TYROL, CHICAGO, VIENNA, MASSACHUSETTS, POTSDAM, VILLAGE FINDER, BEIJING, HANGZHOU, WUHAN, AND CHONGQING, RESPECTIVELY. AUSTIN, KITSAP, TYROL, CHICAGO, VIENNA, MASSACHUSETTS, POTSDAM, VILLAGE FINDER, BEIJING, HANGZHOU, WUHAN, AND CHONGQING ARE ABBREVIATED AS "au," "ki," "ty," "ch," "vi," "ma," "po," "vf," "bj," "hz," "wh," AND "cq," RESPECTIVELY. † DENOTES THAT WE REMOVE THE WEAK SUPERVISION OF IMAGE-LEVEL LABELS IN WSDA [17] FOR A FAIR COMPARISON

Austin and Kitsap ⟶ Ten Targets (IoU)												
Forms	Methods	Source Domain ⟶ Target Domain										
		$au \to ty$	$au \to ch$	$au \to vi$	$au \to ma$	$au \to po$	$au \to vf$	$au \to bj$	$au \to hz$	$au \to wh$	$au \to cq$	Average
Single-Source	Source Only	29.0	41.2	48.3	34.1	49.8	27.9	26.1	27.6	25.9	30.0	34.0
	TriADA [53]	57.1	52.1	56.1	43.6	53.4	29.7	28.9	32.5	30.1	32.6	41.6
	†WSDA [17]	56.3	49.2	55.2	42.6	50.3	29.3	28.0	32.8	29.3	31.5	40.5
	Advent [48]	56.7	52.8	57.2	54.8	55.6	31.8	27.2	30.9	29.8	33.4	43.0
	FDA [56]	56.1	51.3	55.8	43.5	51.7	30.0	28.7	31.6	29.6	32.9	41.1
		$ki \to ty$	$ki \to ch$	$ki \to vi$	$ki \to ma$	$ki \to po$	$ki \to vf$	$ki \to bj$	$ki \to hz$	$ki \to wh$	$ki \to cq$	Average
Single-Source	Source Only	13.0	33.6	46.7	29.8	36.2	18.2	16.7	24.4	23.4	21.4	26.3
	TriADA [53]	47.6	44.6	55.9	40.1	43.1	22.6	17.1	28.0	27.6	21.7	34.8
	†WSDA [17]	48.1	41.6	54.7	37.9	41.7	17.3	18.6	29.3	22.4	22.2	33.4
	Advent [48]	48.5	43.4	56.7	41.6	44.2	21.2	18.1	28.4	27.2	26.0	33.7
	FDA [56]	48.2	42.1	55.1	38.8	42.2	18.5	21.0	26.6	24.5	20.5	33.8
		$au, ki \to ty$	$au, ki \to ch$	$au, ki \to vi$	$au, ki \to ma$	$au, ki \to po$	$au, ki \to vf$	$au, ki \to bj$	$au, ki \to hz$	$au, ki \to wh$	$au, ki \to cq$	Average
Source-Combination	Mixed Source Only	37.1	46.5	48.6	34.5	40.1	19.1	20.9	28.8	23.8	24.1	32.4
	TriADA [53]	58.7	51.2	58.0	44.1	51.8	34.8	30.1	33.7	34.5	37.2	43.4
	†WSDA [17]	57.5	50.0	56.7	43.9	50.9	33.4	31.5	33.0	30.6	35.1	42.3
	Advent [48]	59.7	53.7	59.8	50.9	53.7	34.0	33.1	34.8	31.3	35.6	44.7
	FDA [56]	58.1	50.9	56.5	44.1	51.2	33.5	30.2	31.8	30.9	34.4	42.2
Multi-Source	Source Only	36.7	34.4	49.4	36.6	37.7	18.1	19.8	27.9	24.0	23.4	30.8
	ColorMapGAN [45]	60.6	50.1	52.0	35.6	30.1	35.7	31.7	35.8	33.4	37.9	40.3
	MADAN [63]	62.4	45.1	38.9	40.1	37.2	34.1	31.4	36.3	32.7	37.3	39.6
	MSDACL [11]	60.7	54.1	63.9	45.7	55.3	44.1	39.3	41.2	43.6	42.1	49.0
	SPENet (ours)	**65.8**	**58.1**	**68.4**	**49.1**	**59.1**	**49.9**	**45.8**	**43.7**	**48.9**	**48.2**	**53.7**
Target-Only	Supervised	70.1	68.4	75.3	65.6	56.9	69.7	60.1	54.6	57.8	63.7	64.2

TABLE III

COMPARISON RESULTS OF ADAPTING FROM TYROL AND CHICAGO TO AUSTIN, KITSAP, VIENNA, MASSACHUSETTS, POTSDAM, VILLAGE FINDER, BEIJING, HANGZHOU, WUHAN, AND CHONGQING, RESPECTIVELY. SYMBOL ABBREVIATIONS ARE CONSISTENT WITH THOSE IN TABLE II. † DENOTES THAT WE REMOVE THE WEAK SUPERVISION OF IMAGE-LEVEL LABELS IN WSDA [17] FOR A FAIR COMPARISON

Tyrol and Chicago ⟶ Ten Targets (IoU)												
Forms	Methods	Source Domain ⟶ Target Domain										
		$ty, ch \to au$	$ty, ch \to ki$	$ty, ch \to vi$	$ty, ch \to ma$	$ty, ch \to po$	$ty, ch \to vf$	$ty, ch \to bj$	$ty, ch \to hz$	$ty, ch \to wh$	$ty, ch \to cq$	Average
Source-Combination	Mixed Source Only	46.9	29.7	52.4	38.0	36.1	24.9	25.6	22.7	23.0	28.4	32.8
	TriADA [53]	50.7	37.1	63.9	43.2	44.9	26.5	28.4	27.2	27.8	32.3	38.2
	†WSDA [17]	49.9	35.6	62.8	42.5	42.8	26.6	28.5	27.9	28.8	34.2	38.0
	Advent [48]	48.7	38.7	60.8	44.1	49.9	27.6	30.0	28.2	29.9	35.8	39.4
	FDA [56]	48.5	35.8	62.3	43.6	42.5	25.7	25.8	23.9	24.0	30.3	36.2
Multi-Source	Source Only	35.2	26.7	54.5	36.8	25.7	22.1	25.0	19.8	22.3	27.9	29.6
	ColorMapGAN [45]	50.1	41.5	58.2	37.3	32.4	26.6	27.6	24.0	24.8	30.7	35.3
	MADAN [63]	49.8	43.6	52.9	46.1	35.5	27.2	28.3	24.1	24.9	31.4	36.4
	MSDACL [11]	52.2	39.1	64.8	45.6	51.6	36.5	34.3	30.0	30.9	37.3	42.2
	SPENet (ours)	**56.1**	**45.3**	**72.0**	**49.8**	**60.0**	**40.2**	**41.6**	**36.6**	**38.1**	**43.4**	**48.3**
Target-Only	Supervised	71.8	58.8	75.3	65.6	56.9	69.7	60.1	54.6	57.8	63.7	63.4

TABLE IV

COMPARISON RESULTS OF ADAPTING FROM KITSAP AND VIENNA TO AUSTIN, TYROL, CHICAGO, MASSACHUSETTS, POTSDAM, VILLAGE FINDER, BEIJING, HANGZHOU, WUHAN, AND CHONGQING, RESPECTIVELY. SYMBOL ABBREVIATIONS ARE CONSISTENT WITH THOSE IN TABLE II. † DENOTES THAT WE REMOVE THE WEAK SUPERVISION OF IMAGE-LEVEL LABELS IN WSDA [17] FOR A FAIR COMPARISON

Kitsap and Vienna ⟶ Ten Targets (IoU)												
Forms	Methods	Source Domain ⟶ Target Domain										
		$ki, vi \to au$	$ki, vi \to ty$	$ki, vi \to ch$	$ki, vi \to ma$	$ki, vi \to po$	$ki, vi \to vf$	$ki, vi \to bj$	$ki, vi \to hz$	$ki, vi \to wh$	$ki, vi \to cq$	Average
Source-Combination	Mixed Source Only	44.2	16.1	45.6	38.7	43.3	20.1	23.9	26.1	24.3	30.9	31.3
	TriADA [53]	50.2	57.9	49.7	45.1	51.8	24.0	27.5	29.9	28.9	34.9	40.0
	†WSDA [17]	51.0	55.2	48.5	44.2	50.9	24.2	26.8	30.2	27.8	33.4	39.2
	Advent [48]	47.5	57.9	51.0	45.7	53.7	25.1	29.2	31.7	29.3	36.2	40.7
	FDA [56]	49.8	56.1	49.3	44.7	50.4	22.6	26.7	28.9	26.6	33.3	38.8
Multi-Source	Source Only	42.8	19.1	41.3	37.1	42.6	19.6	22.1	24.0	23.9	28.8	30.1
	ColorMapGAN [45]	47.0	66.2	52.0	48.3	43.9	23.7	27.8	31.6	28.6	34.5	40.4
	MADAN [63]	46.2	61.4	51.9	43.4	46.8	26.7	30.5	33.1	30.2	36.1	40.6
	MSDACL [11]	48.7	65.6	52.1	48.8	47.8	30.4	34.1	36.4	36.0	40.4	44.0
	SPENet (ours)	**51.1**	**66.3**	**56.2**	**49.7**	**57.8**	**34.2**	**39.2**	**41.4**	**40.7**	**45.4**	**48.2**
Target-Only	Supervised	71.8	70.1	68.4	65.6	56.9	69.7	60.1	54.6	57.8	63.7	63.9

source domains. From Table IV, we can see that the proposed method SPENet achieves the best performance among all the comparison methods. Regardless the differences between the target domain and each source domain, the performance of our method is stable and maintains the best average IoU. These results further demonstrate the effectiveness of our method SPENet on the building extraction task.

In addition, from the results in Table II, the performance of training on the mixed domain may be poorer than that on a single-source domain. For example, when source domains

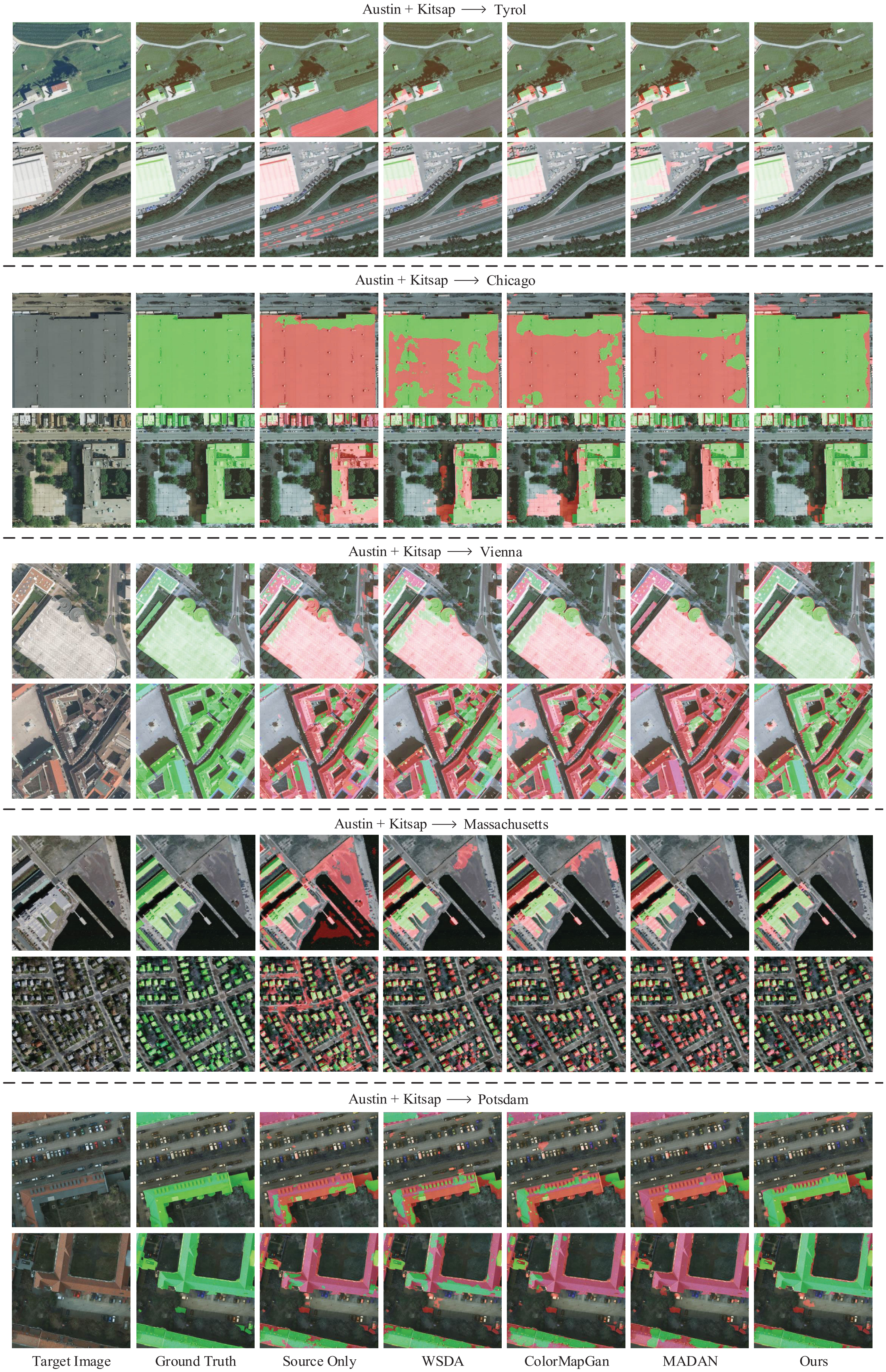


Fig. 4. Qualitative comparison between our method and source only, WSDA [17], ColorMapGAN [45], and MADAN [63] when source domains Austin and Kitsap are adapted to five target domains Tyrol, Chicago, Vienna, Massachusetts, and Potsdam, respectively.

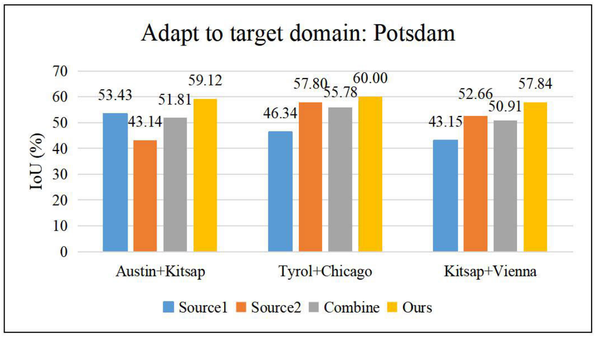


Fig. 5. Compare the results of our method and the SSUDA method TriADA in source1, source2, and the mixed source domain.

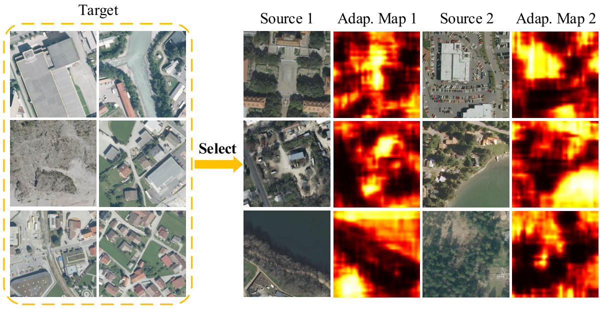


Fig. 6. Visualization of information relevant to the target domain (Tyrol) selected from source 1 (Austin) and source 2 (Kitsap). The "Adap.Map" is the abbreviation of adaptable map. The brighter region in Adap.Map indicates the higher relevance with the target domain.

Austin and Kitsap are mixed, the performance of the SSUDA method TriADA on the target domain Potsdam is reduced. In order to facilitate observation, we make a histogram, as shown in Fig. 5, in which the three settings of two source domains are used to adapt to the target domain Potsdam, respectively. When one of the source domains has poor adaption to the target domain, the performance in the three settings is reduced by 1.62%, 2.02%, and 1.75% after mixing source domains. This indicates that introducing much information irrelevant to the target domain negatively affects domain adaptation. In contrast, our method focuses on target-relevant information in multisource domains and achieves the best performance, which outperforms the three performances trained on a single-source domain by 5.69%, 2.20%, and 5.18%, respectively. This further supports the conclusions drawn in Tables II–IV.

C. Ablation Studies

In this section, we conduct adequate experiments with several settings of multisource domain and target domain to analyze the effectiveness of each component in the proposed method.

1) Performance Impact of Each Module: Given that our method consists of three modules, we first report the contribution of each module to the performance of our method, as shown in Table V. As the results shown, training on the combination of source domains (Austin and Kitsap) achieves limited performance on three target domains (Tyrol, Chicago, and Potsdam), i.e., 37.12%, 46.51%, and 40.10%, respectively. Replacing the combination of source domains with the exchange of information among domains can improve performance to some extent, but the performance is unstable. Adding the purification of target domain information further boosts the performance on three target domains, which is improved by 14.46%, 12.14%, and 22.68% on Tyrol, Chicago, and Potsdam, respectively. Adding the selection of source domain information also improves the adaptation performance on three target domains by 15.15%, 13.70%, and 26.62%, respectively. This result proves that the multisource domain selection module brings more improvements than purification of target domain information, especially when source domains are very relevant to the target domain. Furthermore, the full version of our method SPENet achieves the best performance on target domains, i.e., 65.83%, 58.14%, and 59.12%, respectively. Thus, we can conclude that the three modules are complementary, and each module contributes to these improvements.

TABLE V
ABLATION STUDIES FOR EACH COMPONENT OF OUR PROPOSED METHOD

Austin and Kitsap → Three Targets (IoU)						
Source Comb.	Exchange among Dom.	Source Select.	Target Purifi.	Tyrol	Chicago	Potsdam
✓				37.12	46.51	40.10
	✓			45.52	42.01	31.03
	✓	✓		60.67	55.71	57.65
	✓		✓	59.98	54.15	53.71
	✓	✓	✓	65.83	58.14	59.12

TABLE VI
PERFORMANCE VERIFICATION OF UP BRANCH, DOWN BRANCH, AND ENSEMBLE MODEL FOR THE PROPOSED METHOD

Austin and Kitsap → Four Targets (IoU)				
Model	Tyrol	Chicago	Vienna	Massachusetts
A-Network	62.09	57.14	67.55	48.27
K-Network	61.60	56.31	66.89	47.66
Ensemble	65.83	58.14	68.42	49.13

2) Analysis of Each Network in the Model: Since our method consists of multiple segmentation networks, we then report the performance of each network and the final ensemble network on four target domains (i.e., Tyrol, Chicago, Vienna, and Massachusetts). We define the networks trained on the source domain Austin and Kitsap as A-Network and K-Network, respectively. From Table VI, we can observe that A-Network and K-Network achieve similar performance. The IoU of the two networks on four target domains differs by less than 1.00%. This proves that multiple networks achieve consistent predictions for the same target domain data. In addition, the ensemble network further improves performance on four target domains by 3.74%, 1.00%, 0.87%, and 0.86%, respectively. Therefore, we report the performance of the ensemble model in all experiments. Besides, one of multiple networks can be randomly selected as the test network to save time and memory for inference.

3) Visualization of the Proposed Modules Performance: To verify the effectiveness of the information selection and purification modules, we visualize the output of the two modules in Austin + Kitsap → Tyrol task, respectively, as shown in Figs. 6 and 7. In the Adap.Map of Fig. 6, these maps shows the relevance between the images of source domains (Austin and Kitsap) and the target domain (Tyrol). It can be seen that the target-relevant information in Austin and Kitsap is effectively selected, e.g., rivers, trees, and buildings. In the Uncer.Map of Fig. 7, these maps show the estimation of the uncertainty

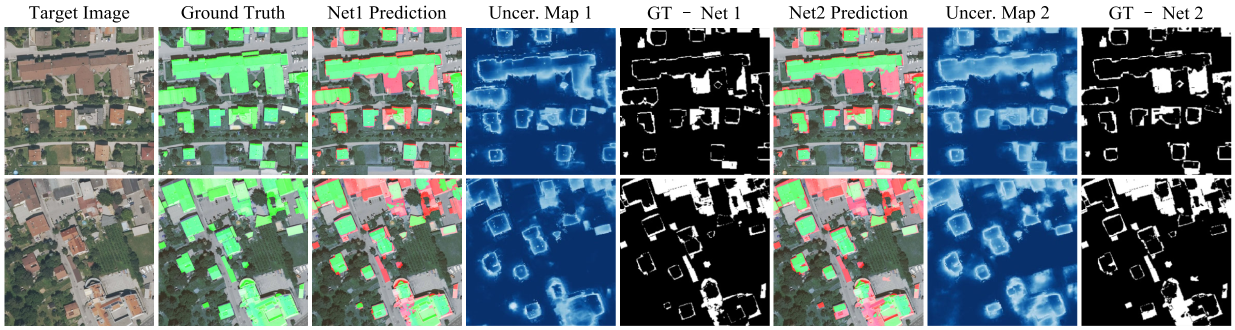


Fig. 7. Visualization of information with low prediction reliability for the target domain (Tyrol). The “Net1” and “Net2” are the networks trained on source domains Austin and Kitsap, respectively. The “Uncer.Map” is the abbreviation of uncertain map. The brighter region in the Uncer.Map indicates the less reliable prediction.

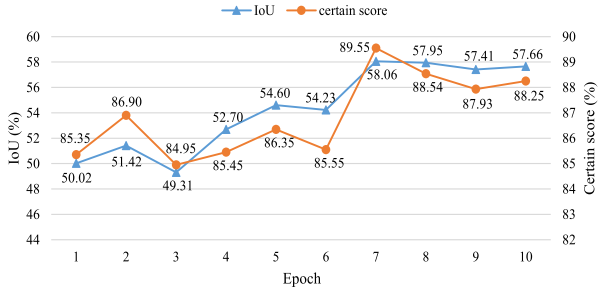


Fig. 8. Trend of IoU with the uncertainty estimation module’s estimation for the prediction results of the target domain.

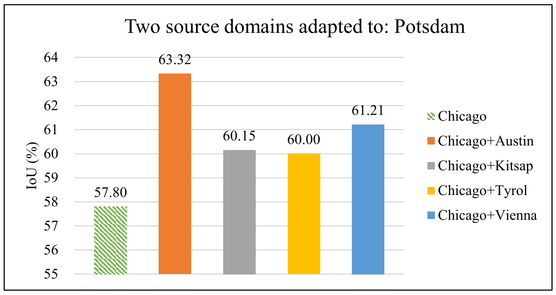


Fig. 9. Performance of our method when source domain Chicago is introduced into other four source domains (Austin, Kitsap, Tyrol, and Vienna), respectively.

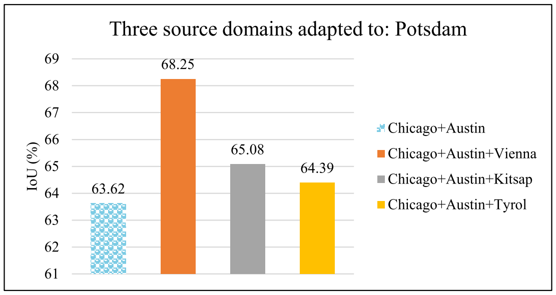


Fig. 10. Performance of our method when source domains Chicago and Austin are introduced into other three source domains (Kitsap, Tyrol, and Vienna), respectively.

of the target domain information (i.e., pseudo-label) by the information purification module. By observing ground truth and our prediction, we can find that the region where the prediction is wrong corresponds to the brighter region in Uncer.Map. This further illustrates that the information purification module accurately estimates the unreliable regions of pseudo-labels.

4) Quantitative Uncertainty Estimation of Prediction: In order to further illustrate the performance of the information purification module, we quantify this module’s estimation of pseudo-labels in the target domain. We define the percentage of pixels estimated to be above 0.5 in the pseudo-label as a quantitative metric, i.e., the certainty score. As shown in Fig. 8, we draw the line chart of IoU changes with certainty score in Austin + Kitsap → Chicago task. These two curves have similar changing trends. The performance of our method is better when the certainty score is higher. This shows that the information purification module can effectively estimate unreliable regions of pseudo-labels in the target domain to generate reliable supervision for other networks.

5) Verification of Model Generalization: To validate the generalization of our method when introducing a new source domain, we do the following experiments. From Fig. 5, we can observe that the source domain Chicago is most conducive to adapting the network to the target domain Potsdam. Therefore, based on the source domain Chicago, we introduce the other four source domains (Austin, Kitsap, Tyrol, and Vienna), respectively. As shown in Fig. 9, the performance of the four two-source settings on Potsdam further surpasses the best single-source domain trained on Chicago by 5.52%, 2.35%, 2.20%, and 3.41%, respectively. Our method can further improve performance when introducing other source domains. This shows that our method effectively explores more target-relevant information from the two source domains and fully aggregates them to better adapt to the target domain.

Similarly, we conduct experiments of introducing a new source domain into two source domains. As shown in Fig. 10, based on the best two-source domain settings (i.e., Chicago and Austin), we introduce the other three source domains (i.e., Vienna, Kitsap, and Tyrol), respectively. Our method further boosts the performance on the target domain Potsdam, which is improved by 4.63%, 1.46%, and 0.77%. The above sufficient experiments illustrate that our method effectively utilizes target-relevant information in multisource domains and has strong generalization.

6) Sensitivity Analysis of Hyperparameter $\lambda^{\text{sda}}_{(S,T)}$: Because the source domain adaptable loss is adjusted by the hyperparameter $\lambda^{\text{sda}}_{(S,T)}$, we further analyze the sensitivity of $\lambda^{\text{sda}}_{(S,T)}$ to the performance of our method. In this experiment, we use Austin and Kitsap as source domains and randomly select three datasets (i.e., Massachusetts, Potsdam, and Chicago) as target domains, respectively. As can be seen from Fig. 11, our method is not sensitive to the parameter $\lambda^{\text{sda}}_{(S,T)}$ in the range of $\{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. Moreover, the proposed method achieves the best performance when the parameter $\lambda^{\text{sda}}_{(S,T)}$ is 0.1. Therefore, in all other experiments of this article, the value of hyperparameter $\lambda^{\text{sda}}_{(S,T)}$ is set to 0.1.

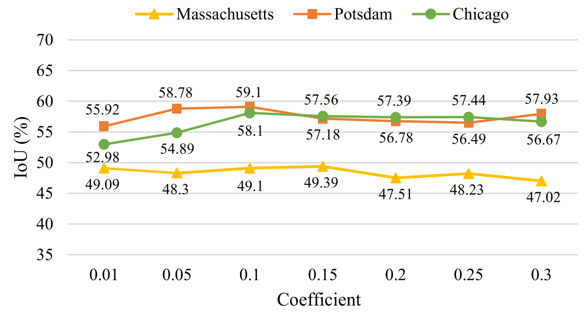


Fig. 11. Sensitivity analysis for coefficient $\lambda^{\text{sda}}_{(S,T)}$ of the source domain adaptable loss on three target domains, respectively.

TABLE VII

SENSITIVITY ANALYSIS FOR HYPERPARAMETER β ON THREE TARGET DOMAINS, RESPECTIVELY

Target Domain	β=0.1	β=0.3	β=0.5	β=0.7	β=0.9	β=1.1
Massachusetts	48.2	48.9	49.1	46.0	45.4	43.1
Potsdam	57.5	58.6	59.1	55.2	53.0	52.3
Chicago	56.7	57.8	58.1	55.6	54.2	53.7

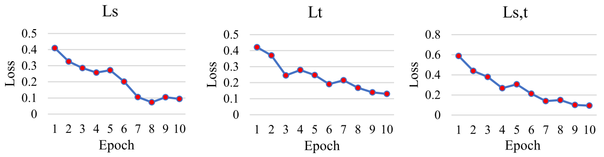


Fig. 12. Trend of the three losses, respectively, decreasing as training progresses.

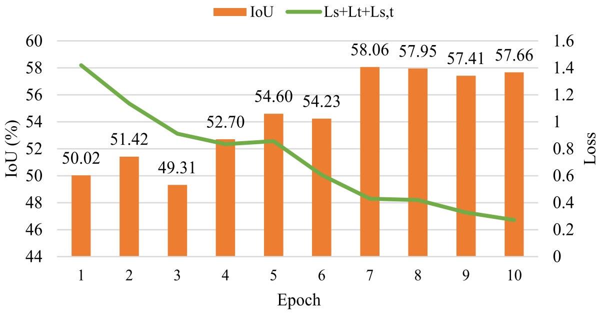


Fig. 13. Trend of the proposed method's IoU and total loss as training progresses.

7) Sensitivity Analysis of Hyperparameter β: The hyperparameter β is introduced to enlarge the score of the adaptable map for the source image. Thus, we construct experiments to verify the sensitivity of β on three tasks, i.e., Austin + Kitsap → Massachusetts/Potsdam/Chicago. The results are shown in Table VII. Our method is not sensitive to β, while β lies between 0.3 and 0.5. When β is 0.1, the performance of the model drops slightly. The performance drops clearly, while β lies between 0.7 and 1.1. This demonstrates that learning too much or barely learning from information that is extremely irrelevant to the target domain is not conducive to model optimization.

8) Stability of Training Process: We discuss the convergence of losses in our method during training and the effectiveness of joint training with the proposed three losses to adapt to the target domain. As shown in Figs. 12 and 13, we draw the loss curves with epochs and the IoU column corresponding to each epoch in Austin + Kitsap → Potsdam task. The $\mathcal{L}_S^{\text{seg}}$, $\mathcal{L}_T^{\text{tdu}}$, and $\mathcal{L}_{(S,T)}^{\text{sda}}$ curves drop rapidly in the early training and gradually tend to stabilize after the eighth epoch. From the curve and IoU column in Fig. 13, we can see that the total loss can converge stably when three losses are jointly trained, and the IoU gradually increases as the total loss decreases. This shows that the three losses proposed in our method have strong adaptability and effectiveness to the target domain.

V. CONCLUSION

In this article, we present an effective MSUDA framework for building extraction. The framework utilizes multitemporal and multiregional remote sensing images to enrich information and boost the performance of building extraction across diverse aerial imagery datasets. During the training process, multiple segmentation networks focus on target-relevant information in multisource domains to influence adaptation positively. Target-relevant information is complementary due to the diversity of multiple-source domains. To aggregate complementary information, pseudo-labels of target images inferred by one segmentation network supervise the learning of other networks in an interactive learning manner. The segmentation capabilities of different networks are fully utilized. In addition, to overcome the label noise generated by the difference between the target domain and the multisource domains, the low-level features of the building are introduced as a priori in the process of purifying pseudo-labels, thus estimating the unreliable regions and boosting the performance of the interactive learning process. Sufficient experiments are constructed on 12 city datasets with different resolutions to evaluate the performance of our framework. The proposed framework evidently outperforms the existing state-of-the-art methods and even approaches the supervised method in some settings.

ACKNOWLEDGMENT

The Potsdam dataset is provided by the German Society for Photogrammetry and Remote Sensing and is published by the U.S. Geological Survey. Furthermore, the authors would like to greatly appreciate Aerospace Hongtu Information Technology Company Ltd., Beijing, China, for providing Gaofen-2 (GF-2) data assistance.

REFERENCES

[1] S. Ahmadi, M. J. V. Zoej, H. Ebadi, H. A. Moghaddam, and A. Mohammadzadeh, "Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 12, no. 3, pp. 150–157, Jun. 2010.

[2] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.

[3] K. Bittner, S. Cui, and P. Reinartz, "Building extraction from remote sensing data using fully convolutional networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XLII-1/W1, pp. 481–486, May 2017.

[4] Y. Chen, W. Chen, Y. Chen, B. Tsai, Y. F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2011–2020.

[5] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.

[6] L. Du et al., "SSF-DAN: Separated semantic feature based domain adaptation network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 982–991.

[7] A. Fischer et al., "Extracting buildings from aerial images using hierarchical aggregation in 2D and 3D," *Comput. Vis. Image Understand.*, vol. 72, no. 2, pp. 185–203, Nov. 1998.

[8] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[9] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Jan. 2016.

[10] R. Gong, W. Li, Y. Chen, and L. Van Gool, "DLOW: Domain flow for adaptation and generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2472–2481.

[11] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," 2021, *arXiv:2103.04717*.

[12] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNs in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv:1612.02649*.

[13] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1335–1344.

[14] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 96–115, Feb. 2022.

[15] A. Huertas and R. Nevatia, "Detecting buildings in aerial images," *Comput. Vis., Graph., Image Process.*, vol. 41, no. 2, pp. 131–152, Feb. 1988.

[16] J. Hui, M. Du, X. Ye, Q. Qin, and J. Sui, "Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 786–790, May 2019.

[17] J. Iqbal and M. Ali, "Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 263–275, Sep. 2020.

[18] ISPRS. (2018). *ISPRS Benchmark Datasets: PotsDAM*. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-potsdam

[19] 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. 2019.

[20] K. Karantzalos and N. Paragios, "Recognition-driven two-dimensional competing priors toward automatic and accurate building detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 133–144, Jan. 2009.

[21] E. Li, J. Femiani, S. Xu, X. Zhang, and P. Wonka, "Robust rooftop extraction from visible band images using higher order CRF," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4483–4495, Aug. 2015.

[22] G. Li, G. Kang, W. Liu, Y. Wei, and Y. Yang, "Content-consistent matching for domain adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 440–456.

[23] P. Li, X. Liang, D. Jia, and E. P. Xing, "Semantic-aware grad-GAN for virtual-to-real urban scene adaption," 2018, *arXiv:1801.01726*.

[24] X. Li, X. Yao, and Y. Fang, "Building-A-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3680–3687, Oct. 2018.

[25] Q. Lian, L. Duan, F. Lv, and B. Gong, "Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6757–6766.

[26] Y. Liu et al., "Multilevel building detection framework in remote sensing images based on convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3688–3700, Oct. 2018.

[27] A. Lorette, X. Descombes, and J. Zerubia, "Texture analysis through a Markovian modelling and fuzzy classification: Application to urban area extraction from satellite images," *Int. J. Comput. Vis.*, vol. 36, no. 3, pp. 221–236, 2000.

[28] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1. Princeton, NJ, USA: Citeseer, 2013, p. 3.

[29] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.

[30] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.

[31] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4500–4509.

[32] K. Murtaza, S. Khan, and N. Rajpoot, "VillageFinder: Segmentation of nucleated villages in satellite imagery," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.

[33] J.-F. Pekel, A. Cottam, N. Gorelick, and A. S. Belward, "High-resolution mapping of global surface water and its long-term changes," *Nature*, vol. 540, no. 7633, pp. 418–422, Dec. 2016.

[34] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1406–1415.

[35] C. Qin, L. Wang, Y. Zhang, and Y. Fu, "Generatively inferential co-training for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1055–1064.

[36] D. Quan et al., "Element-wise feature relation learning network for cross-spectral image patch matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3372–3386, Aug. 2022.

[37] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the Internet of Things using big data analytics," *Comput. Netw.*, vol. 101, pp. 63–80, Jun. 2016.

[38] C. Ren, P. Ge, P. Yang, and S. Yan, "Learning target-domain-specific classifier for partial domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 1989–2001, May 2021.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[40] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Adversarial dropout regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2018, pp. 1–15.

[41] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Learning from synthetic data: Addressing domain shift for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3752–3761.

[42] J. Shore and R. Johnson, "Properties of cross-entropy minimization," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 4, pp. 472–482, Jul. 1981.

[43] 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.

[44] Z. Tang, B. Pan, E. Liu, X. Xu, T. Shi, and Z. Shi, "Srda-net: Super-resolution domain adaptation networks for semantic segmentation," 2020, *arXiv:2005.06382*.

[45] 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.

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

[47] Y. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1456–1465.

[48] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2512–2521.

[49] Z. Wang et al., "Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12632–12641.

[50] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[51] Z. Wu, X. Wang, J. Gonzalez, T. Goldstein, and L. Davis, "ACE: Adapting to changing environments for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2121–2130.

[52] R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin, "Deep cocktail network: Multi-source unsupervised domain adaptation with category shift," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3964–3973.

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

[54] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.

[55] Y. Yang, D. Lao, G. Sundaramoorthi, and S. Soatto, "Phase consistent ecological domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9008–9017.

[56] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4084–4094.

[57] Z. Ye, Y. Fu, M. Gan, J. Deng, A. Comber, and K. Wang, "Building extraction from very high resolution aerial imagery using joint attention deep neural network," *Remote Sens.*, vol. 11, no. 24, p. 2970, Dec. 2019.

[58] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[59] Q. Zhang and K. C. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2320–2329, Sep. 2011.

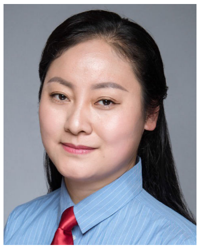
[60] Y. Zhang, P. David, H. Foroosh, and B. Gong, "A curriculum domain adaptation approach to the semantic segmentation of urban scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1823–1841, Aug. 2020.

[61] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2039–2049.

[62] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 8559–8570.

[63] S. Zhao et al., "Multi-source domain adaptation for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7287–7300.

[64] Y. Zou, Z. Yu, X. Liu, B. V. K. V. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5981–5990.

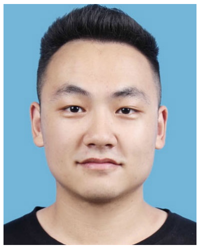


Shuang Wang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi'an, China, in 2000, 2003, and 2007, respectively, all in circuits and systems.

She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University. Her research interests include sparse representation, image processing, synthetic aperture radar (SAR) automatic target recognition, remote sensing image captioning, and polarimetric SAR data analysis and interpretation.

Qi Zang (Student Member, IEEE) received the B.S. degree in electronic information engineering from Northwest Normal University, Lanzhou, China, in 2019. She is currently pursuing the Ph.D. degree in computer science and technology with the School of Artificial Intelligence, Xidian University, Xi'an, China.

Her research interests include semantic segmentation, unsupervised domain adaptation, and domain generalization.

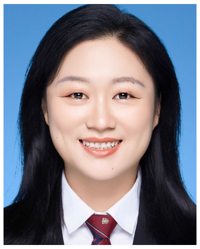


Dong Zhao (Student Member, IEEE) received the B.S. degree in electronic information engineering from the University of Science and Technology Liaoning, Anshan, China, in 2018. He is currently pursuing the Ph.D. degree in control science and engineering with the School of Artificial Intelligence, Xidian University, Xi'an, China.

His research interests include semantic segmentation and domain adaptation.

Chaowei Fang received the bachelor's degree from Xi'an Jiaotong University, Xi'an, China, in 2013, and the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2019.

He is a Lecturer affiliated with the School of Artificial Intelligence, Xidian University, Xi'an. He has authored or coauthored several publications featured in prestigious journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON MEDICAL IMAGING (TMI), and *Medical Image Analysis* (MIA), and renowned conferences, such as IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), Association for the Advancement of Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), and ACM International Conference on Multimedia (ACM MM). Moreover, he took on the role of organizing the third International Workshop on Human-Centric Multimedia Analysis, held at ACM MM 2022. His research interests include various domains, including low-level image processing, medical image analysis, and machine learning.

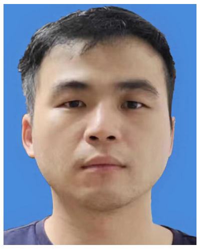


Dou Quan (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2015 and 2021, respectively.

From 2019 to 2020, she was a Joint Ph.D. along with Prof. Jocelyn Chanussot at the Research Center of Inria Grenoble-Rhone-Alpes, Montbonnot-Saint-Martin, France. She is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University. Her research interests include machine learning, deep learning, and metric learning; image matching; image registration; and image classification.

Yutong Wan received the B.S. degree in Internet of Things engineering from Xidian University, Xi'an, China, in 2020, where she is currently pursuing the M.S. degree in computer science and technology with the School of Artificial Intelligence.

Her research interests include semantic segmentation and domain adaptation.



Yanhe Guo (Member, IEEE) received the B.S. degree in intelligent science and technology and the M.S. degree in electronic engineering from Xidian University, Xi'an, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the School of Artificial Intelligence.

His current research interests include machine learning and polarimetric symmetric radar data classification.

Licheng Jiao (Fellow, IEEE) was born in Shaanxi, China, in October 1959. He received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

From 1984 to 1986, he was an Assistant Professor with the Civil Aviation Institute of China, Tianjin, China. In 1990 and 1991, he was a Post-Doctoral Fellow with the Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education. His research interests include signal and image processing, nonlinear circuits and systems theory, learning theory and algorithms, optimization problems, wavelet theory, and machine learning.