Unsupervised domain adaptation for VHR urban scene segmentation via prompted foundation model-based hybrid training joint-optimized network.

LYU, S., ZHAO, Q., SUN, Y., CHENG, G., HE, Y., WANG, G., REN, J. and SHI, Z.

2025

Unsupervised Domain Adaptation for VHR Urban Scene Segmentation via Prompted Foundation Model-Based Hybrid Training Joint-Optimized Network

Shuchang Lyu, *Member, IEEE*, Qi Zhao, *Member, IEEE*, Yaxuan Sun, Guangliang Cheng, Yiwei He, Guangbiao Wang, Jinchang Ren, *Senior Member, IEEE*, and Zhenwei Shi, *Senior Member, IEEE*

***Abstract*— Unsupervised domain adaptation for remote sensing semantic segmentation (UDA-RSSeg) is to adapt a model trained on the source-domain data to the target-domain samples, thereby minimizing the need for annotated data across diverse remote sensing (RS) scenes. In urban planning and monitoring, the task of UDA-RSSeg on very-high-resolution (VHR) images has garnered significant research interest. While recent deep learning techniques have demonstrated huge success in tackling the UDA-RSSeg task for VHR urban scenes, a persistent challenge in addressing the domain shift issue remains. Specifically, there are two primary problems: 1) severe inconsistencies in feature representation across diverse domains, characterized by notably differing data distributions; and 2) the domain gap problem due to the representation bias of the source-domain patterns when translating features to predictive logits. To solve these problems, we propose a prompted foundation model-based hybrid training joint-optimized network (PFM-JONet) for UDA-RSSeg on the VHR urban scene. Our approach integrates the notable "segment anything model" (SAM) as a prompted foundation model to leverage its robust generalized representation capabilities, thereby alleviating feature inconsistencies. Based on the feature extracted by the SAM-Encoder, we introduce a mapping decoder (MD) designed to convert SAM-Encoder features into predictive logits. Additionally, a prompted segmentor (PS) is employed to generate class-agnostic maps, which guide the MD's feature representations. To efficiently optimize the entire network in an end-to-end manner, we design a hybrid training scheme that integrates feature-level and logits-level adversarial training strategies alongside a self-training mechanism. This scheme enhances the model from diverse, compatible perspectives. To evaluate the performance of our proposed PFM-JONet, we conduct extensive experiments on urban scene benchmark datasets, including ISPRS (Potsdam/Vaihingen) and CITY-OSM (Paris/Chicago). On the ISPRS dataset, PFM-JONet surpasses previous SOTA methods by 1.60% in mean IoU value across four adaptation tasks. For CITY-OSM's adaptation task, it outperforms SOTA by 4.84% in the mean IoU value. These results demonstrate the effectiveness of our method. Furthermore, visualization and analysis reinforce the method's interpretability. The code of this article is available at https://github.com/CV-ShuchangLyu/PFM-JONet**

***Index Terms*— Hybrid training, prompted foundation model, semantic segmentation, unsupervised domain adaptation (UDA), urban scene, very-high-resolution (VHR) images.**

I. Introduction

REMOTE sensing semantic segmentation (RSSeg) has extensive application across a range of real-world scenarios, including land mapping [1], [2], [3], [4], urban planning and monitoring [5], [6], [7], [8], disaster evaluation [9], [10], [11], and various other applications. A variety of advanced methods [12], [13], [14], [15] have been developed to enhance its performance. However, despite these advances, the efficacy of RSSeg remains highly dependent on the similarity between the training (source) and testing (target) datasets. Significant discrepancies between these datasets can markedly degrade performance. To address this challenge and facilitate knowledge transfer across domains, the unsupervised domain adaptation for remote sensing semantic segmentation (UDA-RSSeg) task has emerged as a critical task.

In the UDA-RSSeg task on the urban scene, domain shift in very-high-resolution (VHR) images primarily arises from differences in ground sampling distance, variations in remote sensing (RS) sensors, and diverse geographical landscapes [16]. To address these challenges, several methods [17], [18], [19], [20] employ adversarial learning to align the source and target features. Additionally, other methods [21], [22], [23], [24] leverage self-training mechanisms to generate high-quality pseudo-labels for target annotations. Despite significant advancements, two critical problems shown in Fig. 1 persist: 1) the first problem is the severe inconsistencies in feature representation across diverse domains, characterized by notably differing data distributions. While adversarial learning

Received 23 December 2024; revised 13 April 2025; accepted 22 April 2025. Date of publication 24 April 2025; date of current version 6 May 2025. This work was supported by the National Natural Science Foundation of China under Grant 62072021. *(Corresponding authors: Qi Zhao; Guangbiao Wang.)*

Shuchang Lyu, Qi Zhao, Yaxuan Sun, and Guangbiao Wang are with the Department of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: lyushuchang@buaa.edu.cn; zhaoqi@buaa.edu.cn; sunyaxuan@buaa.edu.cn; wanggb@buaa.edu.cn).

Guangliang Cheng and Yiwei He are with the Department of Computer Science, University of Liverpool, L69 7ZX Liverpool, U.K. (e-mail: Guangliang.Cheng@liverpool.ac.uk; yiwei@liverpool.ac.uk).

Jinchang Ren is with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510640, China, and also with the National Subsea Centre, Robert Gordon University, AB10 7AQ Aberdeen, U.K. (e-mail: jinchang.ren@ieee.org).

Zhenwei Shi is with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3564216

aids in feature alignment, it does not fundamentally enhance generalized feature representation; and 2) the second problem lies in the bias in "feature-to-prediction" mapping. Even when the feature representations appear consistent, biases can still emerge during transferring this class-agnostic generalization knowledge to a specific class-aware downstream task. This highlights the need to maintain consistency not only in the feature space but also throughout the process of adapting these features to predictive logits.

To address these two issues, we propose PFM-JONet, a prompted foundation model-based hybrid training joint-optimized network built on the prompted foundation model. In this article, we adopt the segment anything model (SAM) [25] as the prompted foundation model. For the first problem, we integrate SAM into our architecture to leverage its generalized representation capabilities and mitigate feature inconsistency. Due to its extensive training on the large-scale dataset "SA-1B," SAM demonstrates exceptional proficiency in representing images from a wide range of domains. Since "SA-1B" primarily comprises natural images, we apply SAM to generate predictions for VHR urban scene images captured using different imaging modes, thereby showcasing its remarkable generalization capabilities for RS scenes. As shown in Fig. 2, it becomes intuitively evident that SAM has robust generalization representation capabilities and remarkable object localization abilities when applied to RS VHR images, even though it is trained on a large natural image corpus. Therefore, SAM is proved to be reliable and generalized to represent RS VHR images from different domains.

To solve the second problem, we design a hybrid training joint-optimization mechanism. First, we incorporate a mapping decoder (MD) to map the features extracted from the SAM encoder to predictions of a specific downstream task. However, directly fine-tuning the decoder using source-annotated samples will also encounter a significant domain shift issue. To mitigate this, we further utilize a logits-level adversarial discriminator to enhance the optimization of the MD. Second, we incorporate feature-level adversarial learning into the prompted segmentor (PS) to generate the prompted masks. This enables the SAM to automatically produce class-agnostic maps. With the guidance of this map, MD becomes more sensitive and adept at distinguishing between different categories in target images. Third, on predictions of MD and PS, we both design the self-training mechanism. This mechanism utilizes high-quality pseudo-labels of target-domain samples to guide the model, thereby mitigating the representation bias of source-trained networks. Finally, we freeze SAM and jointly optimize the architecture with hybrid training in an end-to-end manner.

Fig. 3 compares paradigms for the UDA-RSSeg task. Existing methods primarily use the "Segmentor Optimization" paradigm [Fig. 3(a)], where the encoder and the decoder of the segmentor are optimized end-to-end using various learning strategies (e.g., adversarial learning and self-training). Fig. 3(b) illustrates the "PFM-based Fine-tuning" paradigm, which utilizes SAM to generate more generalized features for target images. By optimizing the MD, these generalized features are mapped to predictive logits. In this article, we introduce the "PFM-based Joint Optimization" paradigm [Fig. 3(c)], where SAM provides generalized feature representation, and the PS offers class-agnostic prompted guidance. By jointly optimizing the PS and MD, our paradigm significantly enhances segmentation performance on target images. Compared to the "Segmentor Optimization" paradigm, our approach harnesses SAM's generalization capability to reduce feature inconsistency between source and target features. Compared to the "PFM-based Fine-tuning" paradigm, our method incorporates prompted guidance to address the domain gap between features and predictive logits.

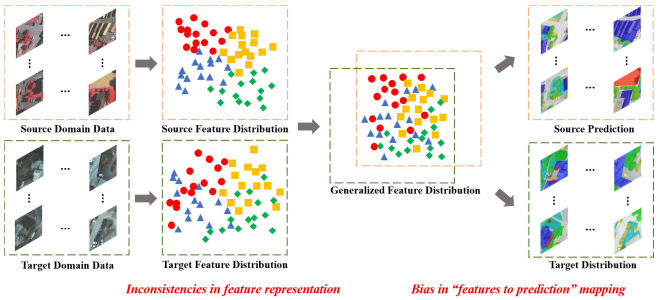

Fig. 1. Illustration of the two main problems addressed in our paper.

We conduct extensive experiments on prominent benchmark datasets, including ISPRS (Potsdam/Vaihingen) [26] and CITY-OSM (Paris/Chicago) [27]. Comparative analyses demonstrate that PFM-JONet outperforms previous state-of-the-art methods. Our visualization and analysis further highlight the interpretability of PFM-JONet. The key contributions of this work are summarized as follows.

1) We propose a PFM-JONet to tackle UDA-RSSeg on the urban scene. To the best of our knowledge, we are the first to introduce the "PFM-based Joint Optimization" paradigm in the realm of UDA-RSSeg.
2) Our proposed hybrid training scheme incorporates multi-level adversarial learning and self-training mechanisms, enhancing the model from diverse and compatible perspectives.
3) We propose a prompted guidance mechanism that integrates the optimization of a PS with an MD, which effectively bridges the gap between class-agnostic maps and class-aware predictions.
4) Our proposed PFM-JONet demonstrates superior performance compared to existing methods across multiple UDA-RSSeg tasks, utilizing prominent benchmark datasets for urban scenes.

The structure of this article is organized as follows. Section II presents a comprehensive review of recent relevant studies and their connections to our proposed method. In Section III, we elaborate on the technical details and implementation of our method. Section IV demonstrates the experimental results, including performance comparison, visualization outcomes analysis, and in-depth discussions. Finally, Section V concludes the article by summarizing our main contributions.

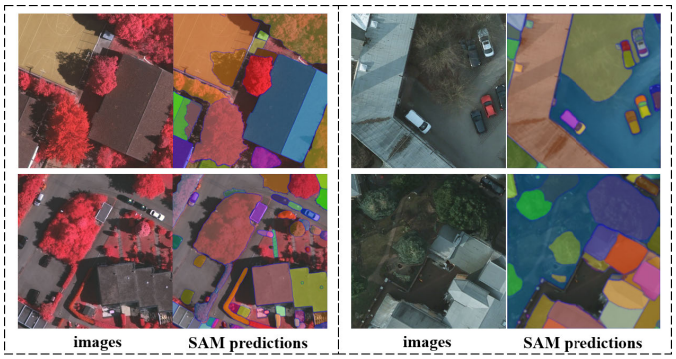

Fig. 2. Visualization of SAM's predictions on RS VHR images of the urban scene. The leftmost two cases are captured using the IR-R-G imaging mode, whereas the rightmost two cases utilize the R-G-B imaging mode. In each pair of images, the left and right images, respectively, represent the original image and segmentation predictions generated by SAM.

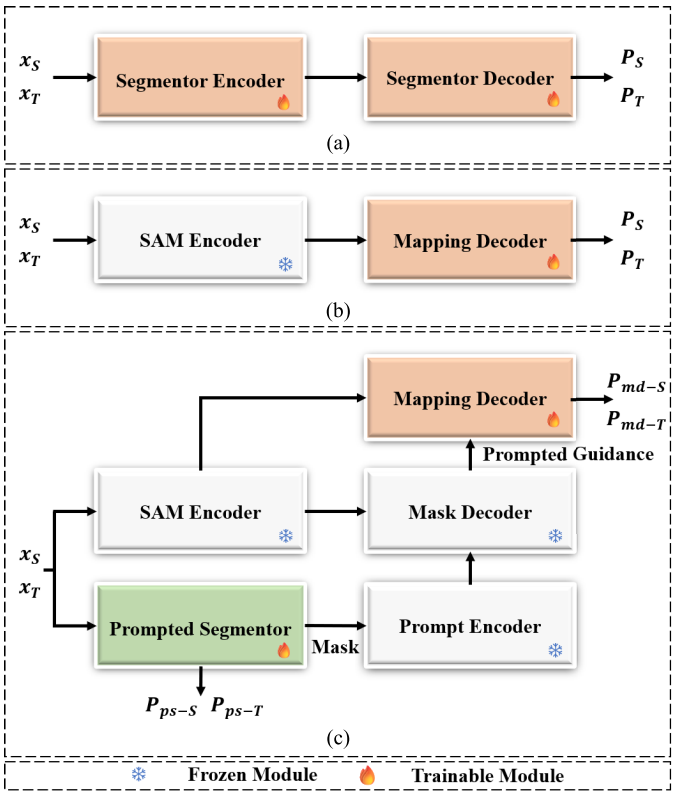

Fig. 3. Paradigm comparison on the UDA-RSSeg task. $\{x_S, x_T\}$ and $\{P_S, P_T\}$, respectively, denote the source/target images and predictions. $\{P_{md-S}, P_{md-T}\}$ and $\{P_{ps-S}, P_{ps-T}\}$, respectively, denote the source/target predictions of MD and PS. (a) "Segmentor Optimization" paradigm for UDA-RSSeg. (b) "PFM-based Fine-tuning" paradigm for UDA-RSSeg. (c) "PFM-based Joint Optimization" paradigm for UDA-RSSeg.

II. Related Work

A. Remote Sensing Semantic Segmentation

The semantic segmentation task aims to categorize target objects at the pixel level. Fully convolutional networks (FCNs) [28] is a pioneer deep learning-based method on this task. Following FCNs, many notable methods [29], [30], [31], [32] are proposed, which significantly promote the development of this task.

On RSSeg is widely applied to geographical element analysis, urban/rural planning, disaster assessment, and so on. Semantic segmentation on the RS scene mainly faces the challenge of complex landscapes on large geographical regions and large intraclass variance by different grounding sampling distances. To address these issues, many notable methods are proposed. Some methods [33], [34], [35], [36] utilize the abundant information from multiple hierarchy features to achieve strong segmentation performance. DPFANet [37] and BSNet [12] integrate the adaptive feature fusion network and the edge optimization block to enhance the representation ability from local to global features. With the development of Transformers, many methods [38], [39], [40], [41], [42] design effective Transformer-based networks to exploit self-attention information for the RSSeg task.

B. UDA Semantic Segmentation for RS

Unsupervised domain adaptation (UDA) aims to adapt the knowledge of source-trained models to target samples. In the natural scene UDA semantic segmentation task, some methods have made huge progress. Adversarial learning is frequently adopted in many excellent methods. Some methods utilize image generalization techniques to align image appearance between source and target images. Chen et al. [43], Yang and Soatto [44], and Guo et al. [45] employed image-level adaption in the first step and then trained the segmentation networks with cross-domain synthetic data. The authors [46], [47], [48], [49] insert discriminators into networks for consistency alignment on intermediate feature maps or output entropy maps. As another typical nonadversarial UDA paradigm, self-training has attracted much attention in cross-domain semantic segmentation tasks. Pan et al. [50], Zou et al. [51], and Hoyer et al. [52] promoted the adaption ability by generating reliable, consistent, and class-balanced pseudo-labels. Domain-generalized semantic segmentation presents a more challenging task compared to conventional UDA semantic segmentation, as it emphasizes the model's ability to generalize across multiple unseen target domains. Bi et al. [53], Yi et al. [54], and Ding et al. [55] developed robust models that can adapt to diverse and potentially heterogeneous data distributions without prior knowledge of the target domain. It is noteworthy that FADA [56] pushes the boundaries by integrating vision foundation models (VFMs) through a fine-tuning mechanism, significantly advancing the field of domain-generalized semantic segmentation.

For UDA-RSSeg, many excellent works have been proposed in recent years. UDA-GAN [57] first introduces a GAN-based segmentation network to tackle the UDA-RSSeg task. Following this pioneering work, numerous methods have emerged, leveraging adversarial learning for domain alignment. Among these, some methods [58], [59], [60], [61] apply image-level adversarial learning between source and target images. Conversely, other methods [18], [19], [62], [63], [64] utilize

feature-level adversarial learning to improve the feature consistency. Although adversarial learning can mitigate domain shift issues, it often fails to prevent the tendency of source-trained models to favor source image characteristics. To address this limitation, the focus has shifted toward nonadversarial paradigms using self-training mechanisms. Self-training methods [16], [21], [22], [23] typically rely on the exponential moving average (EMA) technique to generate pseudo-labels for target images. By training on pseudo-labels, source-trained models can better adapt to target domains.

C. Prompted Foundation Models for RS

Large foundation models, including large language models (LLMs) [65], [66], [67] and vision–language models (VLMs) [68], [69], [70], have significantly transformed and advanced the field of artificial intelligence. The introduction of the SAM [25] has revolutionized the field of semantic segmentation. Leveraging prompted learning for guidance, SAM-based networks demonstrate remarkable generalization across diverse scenarios. In RS, several innovative methods have been developed using the SAM framework. MAF-SAM [71] harnesses SAM's generalized capabilities to effectively process multispectral images. SCD-SAM [72] utilizes SAM's robust generalization to enhance performance in semantic change detection tasks. Based on SAM, RingMo-SAM [15] enhances its functionality by designing the CDMDecoder and integrating a prompt encoder. This allows RingMo-SAM to achieve the ability to segment any object in both optical and SAR remote-sensing data. MeSAM [73] proposes an innovative fine-tuning model that is well-suited for adapting models to the requirements of semantic segmentation tasks involving RS images. The method presented in [74] leverages class-agnostic predictions that incorporate SAM-generated objects (SGOs) and SAM-generated boundaries (SGBs). By designing a boundary preservation loss and an object consistency loss, these methods improve the performance of semantic segmentation in RS scenes. On autonomous driving tasks in the natural scenario, SAM-EDA [75] integrates SAM into a UDA network for semantic segmentation under adverse weather conditions. However, the UDA-Seg task in the RS scene remains unexplored. This article introduces a new paradigm that leverages the generalized capabilities of SAM to ease the domain shift in UDA-RSSeg.

III. Proposed Method

As shown in Fig. 4, our proposed PFM-JONet primarily comprises three key modules: the PS, the prompted foundation model, and the MD. The source- and target-domain images are initially processed by the PS, which generates coarse predictions. These predictions are then utilized as prompt information and fed into the prompted encoder to produce prompted features. In the prompted foundation model (SAM), the source- and target-domain images pass through the SAM encoder to generate generalized feature maps. These feature maps, combined with the prompted features, are processed by the mask decoder to yield class-agnostic prompted maps. During the process of mapping generalized features to the class-aware downstream task, the generalized feature maps from both source- and target-domain images are fed into the "MD Neck" of the MD. The resulting feature maps are then guided by the class-agnostic prompted maps in the "MD Head," ultimately generating refined predictions.

A. Prompted Segmentor

As shown in Fig. 4, PS is designed to provide a prompted mask for SAM to generate the class-agnostic map. To generate the prompted mask, both source and target images $(\boldsymbol{x_S}, \boldsymbol{x_T})$ are processed sequentially through the "PS Encoder" (f_{ps-e}) and the "PS Decoder" (f_{ps-d}), which can be formulated in the following equations:

$$\boldsymbol{F_{ps-S}} = f_{ps-e}(\boldsymbol{x_S}), \quad \boldsymbol{F_{ps-T}} = f_{ps-e}(\boldsymbol{x_T}) \tag{1}$$

$$\boldsymbol{P_{ps-S}} = f_{ps-d}\left(\boldsymbol{F_{ps-S}}\right), \quad \boldsymbol{P_{ps-T}} = f_{ps-d}\left(\boldsymbol{F_{ps-T}}\right) \tag{2}$$

where $\boldsymbol{F_{ps-S}}$ and $\boldsymbol{F_{ps-T}}$ represent the extracted feature maps for source- and target-domain images, respectively. $\boldsymbol{P_{ps-S}}$ and $\boldsymbol{P_{ps-T}}$ correspond to the output logits for source and target predictions.

$\boldsymbol{P_{ps-S}}$ and $\boldsymbol{P_{ps-T}}$ undergo a channel-wise argmax operation, as detailed in (3), to generate the prompted masks $\boldsymbol{P_S}$ and $\boldsymbol{P_T}$. These source- and target-prompted masks are then used as inputs for the prompted encoder

$$\boldsymbol{P_S} = \arg\max_c\left(\boldsymbol{P_{ps-S}}\right), \quad \boldsymbol{P_T} = \arg\max_c\left(\boldsymbol{P_{ps-T}}\right). \tag{3}$$

B. Prompted Foundation Model

To consistently represent the source and target images, we leverage the generalization capability of SAM. As shown in Fig. 4, the "SAM Encoder" $(f_{\text{sam}-e})$ is used to extract features for source and target images

$$\boldsymbol{F_{s-S}} = f_{\text{sam}-e}(\boldsymbol{x_S}), \quad \boldsymbol{F_{s-T}} = f_{\text{sam}-e}(\boldsymbol{x_T}) \tag{4}$$

where $\boldsymbol{F_{s-S}}$ and $\boldsymbol{F_{s-T}}$ denote the output features from the "SAM Encoder." These features will serve as input for MD. Moreover, these features together with the features output from the "Prompt Encoder" (f_{sam-pe}) served as the input of the "Mask Decoder" (f_{sam-md}) to generate the class-agnostic maps, denoted as $\boldsymbol{M_S}$ and $\boldsymbol{M_T}$. This process can be formulated in the following equation:

$$\begin{cases} \boldsymbol{M_S} = f_{sam-md}(\boldsymbol{F_{s-S}}, f_{sam-pe}(\boldsymbol{P_S})) \\ \boldsymbol{M_T} = f_{sam-md}(\boldsymbol{F_{s-T}}, f_{sam-pe}(\boldsymbol{P_T})). \end{cases} \tag{5}$$

C. Mapping Decoder

As shown in Fig. 4, MD is designed to map the features of the "SAM Encoder" to predictive logits with the guidance of SAM's class-agnostic maps. MD contains "MD Neck" $(f_{\text{md}-n})$ and "MD Head" $(f_{\text{md}-h})$. As shown in (6), output features from the "SAM Encoder" [see (4)] first pass through "MD Neck" to map multiscale features into single-scale features

$$\boldsymbol{F_{md-S}} = f_{\text{md}-n}(\boldsymbol{F_{s-S}}), \quad \boldsymbol{F_{md-T}} = f_{\text{md}-n}(\boldsymbol{F_{s-T}}) \tag{6}$$

where $\{\boldsymbol{F_{md-S}}, \boldsymbol{F_{md-T}}\} \in \mathbb{R}^{C\times H\times W}$ denote the output features from the "MD Neck."

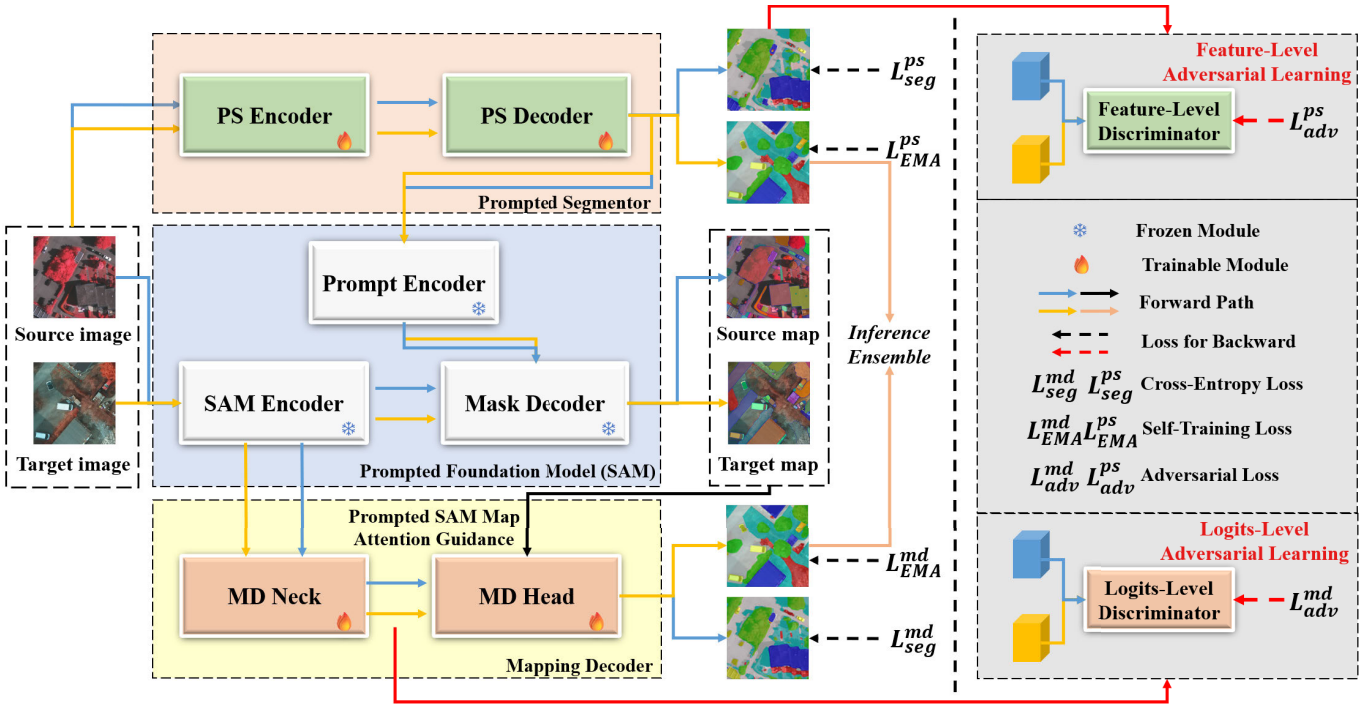

Fig. 4. The overview of PFM-JONet. The architecture contains three modules, which are PS, Prompted Foundation Model (SAM), and MD. Adversarial learning and self-training are applied for joint optimization.

Then, these features are guided by class-agnostic maps through the attention mechanism. As shown in Fig. 5, features are fed into a "conv block" and conducted matrix multiplication with $\{\boldsymbol{M_S}, \boldsymbol{M_T}\} \in \mathbb{R}^{H\times W}$ to generate channel-weighted vectors, denoted as $\{\boldsymbol{v_{md-S}}, \boldsymbol{v_{md-T}}\} \in \mathbb{R}^{C}$. This process is shown in the following equations:

$$v^{c}_{\mathrm{md}-S} = \frac{\exp\left(\sum_{h,w=1}^{H,W}(f_c(\boldsymbol{F_{md-S}}))^{c,h,w} \times \boldsymbol{M_S^{c,h,w}}\right)}{\sum_{c=1}^{C}\exp\left(\sum_{h,w=1}^{H,W}(f_c(\boldsymbol{F_{md-S}}))^{c,h,w} \times \boldsymbol{M_S^{c,h,w}}\right)} \tag{7}$$

$$v^{c}_{\mathrm{md}-T} = \frac{\exp\left(\sum_{h,w=1}^{H,W}(f_c(\boldsymbol{F_{md-T}}))^{c,h,w} \times \boldsymbol{M_T^{c,h,w}}\right)}{\sum_{c=1}^{C}\exp\left(\sum_{h,w=1}^{H,W}(f_c(\boldsymbol{F_{md-T}}))^{c,h,w} \times \boldsymbol{M_T^{c,h,w}}\right)} \tag{8}$$

where $\boldsymbol{v_{md-S}} = [v^{1}_{\mathrm{md}-S}, \ldots, v^{C}_{\mathrm{md}-S}]$ and $\boldsymbol{v_{md-T}} = [v^{1}_{\mathrm{md}-T}, \ldots, v^{C}_{\mathrm{md}-T}]$. These vectors set the ratio of features' channel, which can guide the features in a channel-selection manner

$$\boldsymbol{F'_{md-S}} = \boldsymbol{F_{md-S}} \times \boldsymbol{v_{md-S}}, \quad \boldsymbol{F'_{md-T}} = \boldsymbol{F_{md-T}} \times \boldsymbol{v_{md-T}}. \tag{9}$$

Finally, attention guided features, $\{\boldsymbol{F'_{md-S}}, \boldsymbol{F'_{md-T}}\} \in \mathbb{R}^{C\times H\times W}$ are mapped into predictive logits, denoted as $\{\boldsymbol{P_{md-S}}, \boldsymbol{P_{md-T}}\}$ by the "MD Head," formulated as follows:

$$\boldsymbol{P_{md-S}} = f_{\mathrm{md}-h}\left(\boldsymbol{F'_{md-S}}\right), \quad \boldsymbol{P_{md-T}} = f_{\mathrm{md}-h}\left(\boldsymbol{F'_{md-T}}\right). \tag{10}$$

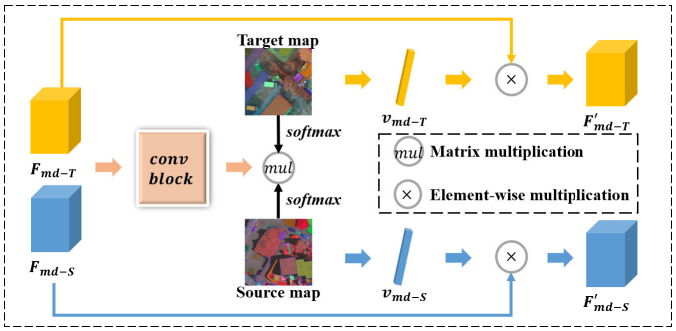

Fig. 5. Prompted SAM map attention guidance block. Here, the "conv block" indicates the [Conv-BN-ReLU] block.

D. Hybrid Training Joint-Optimized Mechanism

To jointly optimize PFM-JONet, we integrate adversarial learning and self-training strategies to build connections between generalized features and predictive logits.

1) Optimizing MD: To optimize MD, we involve a logits-level adversarial learning strategy to enhance the consistent feature representation of the "MD Neck" and "MD Head." Toward source and target predictive logits, we design a logits-level discriminator (D_l) for alignment. The adversarial loss ($\mathcal{L}^{\mathrm{md}}_{\mathrm{adv}}$) is calculated in (11). Here, θ_{md} and θ_{dl} denote the trainable parameters of the MD and the logits-level discriminator, respectively

$$\begin{aligned}\mathcal{L}^{\mathrm{md}}_{\mathrm{adv}}(\theta_{\mathrm{md}}, \theta_{dl}) = {} & \mathbb{E}_{x^s\sim X^s}\left[\log(D_l(\boldsymbol{P_{md-S}}))\right] \\ & + \mathbb{E}_{x^t\sim X^t}\left[\log(1 - D_l(\boldsymbol{P_{md-T}}))\right].\end{aligned} \tag{11}$$

In the MD, the focus shifts toward ensuring the consistency of feature mappings and prediction results as they are transferred to downstream tasks. Here, the emphasis is on refining the generalization of features and controlling any deviations

toward the source domain. By applying adversarial learning at the logits level, the decoder is encouraged to produce outputs that are more robust and less biased toward the source domain.

To further alleviate the representation tendency on source annotated samples, we embed a self-training mechanism into PFM-JONet to enhance the optimization effect on MD. The first step is to update "EMA-MD" ($f_{\text{ema-md}}$) by using the EMA technique. This process can be formulated in the following equation:

$$\theta^t_{\text{ema-md}} = \alpha\theta^{t-1}_{\text{ema-md}} + (1-\alpha)\theta^t_{\text{md}} \tag{12}$$

where α denotes the decay factors controlling the updating rate. $\theta^t_{\text{ema-md}}$ refers to trainable parameters of "EMA-MD" (including "EMA-MD Neck" and "EMA-MD Head") at the tth step.

The second step is to generate target pseudo-labels, denoted as $\hat{\boldsymbol{P}}_{\boldsymbol{md}}$, which is shown in the following equation:

$$\hat{\boldsymbol{P}}_{\boldsymbol{md}} = \arg\max_c f_{\text{ema-md}}(\boldsymbol{F}_{\boldsymbol{s-T}}, \boldsymbol{M}_{\boldsymbol{T}}). \tag{13}$$

The third step is to construct a self-training loss function to optimize MD, which is formulated in the following equation:

$$\mathcal{L}^{\text{md}}_{\text{EMA}}(\theta_{\text{md}}) = -\sum_{h,w=1}^{H,W}\sum_{c=1}^{C} \hat{\boldsymbol{P}}^{(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})}_{\boldsymbol{md}} \log\left(\boldsymbol{P}^{(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})}_{\boldsymbol{md-T}}\right). \tag{14}$$

2) Optimizing PS: A high-performance PS is crucial for providing a high-quality prompted mask and thus, enhancing its performance is also important. To achieve this, we implement feature-level adversarial learning to align the features of the "PS Encoder" shown in (15). Given that the "PS Encoder" lacks generalized feature representation capabilities, it is essential to improve the feature consistency between source and target images

$$\begin{aligned}\mathcal{L}^{\text{ps}}_{\text{adv}}(\theta_{ps-e}, \theta_{df}) = {} & \mathbb{E}_{x^s\sim X^s}\left[\log\left(D_f\left(\boldsymbol{F}_{\boldsymbol{ps-S}}\right)\right)\right] \\ & + \mathbb{E}_{x^t\sim X^t}\left[\log\left(1 - D_f\left(\boldsymbol{F}_{\boldsymbol{ps-T}}\right)\right)\right]\end{aligned} \tag{15}$$

where $\mathcal{L}^{\text{ps}}_{\text{adv}}$ denotes the adversarial loss. θ_{ps-e} and θ_{df}, respectively, denote the trainable parameters of the "PS Encoder" and the feature-level discriminator (D_f).

The PS aims to provide initial informative prompts to the prompted foundation model. However, its encoder may lack the ability to generalize across multiple domains in terms of feature representation. Therefore, aligning features between the source and target domains becomes crucial. Introducing a prediction-level adversarial loss in the PS is avoided due to potential uncontrollable optimization directions in adversarial training. These directions could negatively impact the positive effects of ensuring feature mapping consistency, leading to unstable training or degraded performance.

Similar to the optimization of MD [see (13) and (14)], we also apply self-training strategy on PS. The "EMA PS" ($f_{\text{ema-ps}}$) updating, pseudo-label generation, and loss function construction are, respectively, shown in the following equations:

$$\theta^t_{\text{ema-ps}} = \alpha\theta^{t-1}_{\text{ema-ps}} + (1-\alpha)\theta^t_{ps} \tag{16}$$

$$\hat{\boldsymbol{P}}_{\boldsymbol{ps}} = \arg\max_c f_{\text{ema-ps}}(\boldsymbol{x}_{\boldsymbol{T}}) \tag{17}$$

Algorithm 1 Joint-Optimized Paradigm on PFM-JONet

Input: Source and target images, $\boldsymbol{x}_{\boldsymbol{S}}$ and $\boldsymbol{x}_{\boldsymbol{T}}$. Source-label $\boldsymbol{y}_{\boldsymbol{S}}$. Trainable parameters, θ_{ps} and θ_{md}. Frozen parameters of PFM-JONet. The training iteration is set as K.

Output: Updated trainable parameters, θ'_{ps} and θ'_{md}.

for $k = 1, \ldots, K$ **do**

1st Stage Optimization on Prompted Segmentor:

Get feature maps, $\boldsymbol{F}_{\boldsymbol{ps-S}}$ and $\boldsymbol{F}_{\boldsymbol{ps-T}}$ using Eq. (1).

Get predictive logits, $\boldsymbol{P}_{\boldsymbol{ps-S}}$ and $\boldsymbol{P}_{\boldsymbol{ps-T}}$ using Eq. (2).

Get prompted masks, $\boldsymbol{P}_{\boldsymbol{S}}$ and $\boldsymbol{P}_{\boldsymbol{T}}$ using Eq. (3).

Calculate the segmentation cross-entropy loss, L^{ps}_{seg} with source-label.

Calculate the feature-level adversarial loss, L^{ps}_{adv} using Eq. (15). Use "$min-max$" criterion to optimize.

Calculate the self-training loss, L^{ps}_{EMA} using Eq. (16) $\sim$ Eq. (18).

Compute gradients by backward operation with combined loss, $L^{ps} = L^{ps}_{seg} + \gamma_1 L^{ps}_{adv} + \gamma_2 L^{ps}_{EMA}$.

Update trainable parameters, θ_{ps} by $step$.

Frozen prompted segmentor.

2nd Stage Optimization on Mapping Decoder:

Get generalized feature maps from SAM, $\boldsymbol{F}_{\boldsymbol{s-S}}$ and $\boldsymbol{F}_{\boldsymbol{s-T}}$ using Eq. (4).

Get class-agnostic maps, $\boldsymbol{M}_{\boldsymbol{S}}$ and $\boldsymbol{M}_{\boldsymbol{T}}$ using Eq. (5).

Get feature maps, $\boldsymbol{F}_{\boldsymbol{md-S}}$ and $\boldsymbol{F}_{\boldsymbol{md-T}}$ using Eq. (6).

Get attention guided feature maps, $\boldsymbol{F}'_{\boldsymbol{md-S}}$ and $\boldsymbol{F}'_{\boldsymbol{md-T}}$ using Eq. (7) $\sim$ Eq. (9).

Get predictions, $\boldsymbol{P}_{\boldsymbol{md-S}}$ and $\boldsymbol{P}_{\boldsymbol{md-T}}$ using Eq. (10).

Calculate the segmentation cross-entropy loss, L^{md}_{seg} with source-label.

Calculate the logits-level adversarial loss, L^{md}_{adv} using Eq. (11). Use "$min-max$" criterion to optimize.

Calculate the self-training loss, L^{md}_{EMA} using Eq. (12) $\sim$ Eq. (14).

Compute gradients by backward operation with combined loss, $L^{md} = L^{md}_{seg} + \gamma_3 L^{md}_{adv} + \gamma_4 L^{md}_{EMA}$.

Update trainable parameters, θ_{md} by $step$.

Frozen Mapping Decoder.

end for

$$\mathcal{L}^{\text{ps}}_{\text{EMA}}(\theta_{ps}) = -\sum_{h,w=1}^{H,W}\sum_{c=1}^{C} \hat{\boldsymbol{P}}^{(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})}_{\boldsymbol{ps}} \log\left(\boldsymbol{P}^{(\boldsymbol{h},\boldsymbol{w},\boldsymbol{c})}_{\boldsymbol{ps-T}}\right) \tag{18}$$

where $\theta_{\text{ema-ps}}$ and θ_{ps} denote the trainable parameters of "EMA-PS" and "PS," respectively. $\hat{\boldsymbol{P}}_{\boldsymbol{ps}}$ denotes the pseudo-label of target image.

3) Two-Stage Jointly Optimization With Combined Loss: To optimize PFM-JONet, we propose a two-stage jointly optimizing paradigm. In each iteration, the first stage is dedicated to optimizing the PS, where trainable parameters are updated through the backward propagation process. The second stage involves optimizing the MD. Notably, our proposed two-stage optimization framework operates seamlessly within an

end-to-end learning system. The whole optimization process is shown in Algorithm 1.

IV. Experiments and Analysis

A. Datasets and Metrics

1) Datasets: To evaluate PFM-JONet on the UDA-RSSeg task of the urban scene, we select two benchmark datasets containing cross-city VHR images: ISPRS [26] and CITY-OSM [27].

ISPRS [26] provides a rich collection of image samples for diverse tasks within the field of RS. Each image in this dataset is meticulously annotated at the pixel level across six categories: "Clutter," "Impervious Surfaces," "Car," "Tree," "Low Vegetation," and "Building." ISPRS contains two main subsets: Potsdam and Vaihingen. The Potsdam dataset consists of 38 VHR true orthophotos (VHR TOPs), each with dimensions of 6000 × 6000 pixels. This dataset includes images in three imaging modes: IR-R-G, R-G-B, and R-G-B-IR. The IR-R-G and R-G-B images are composed of three channels, whereas the R-G-B-IR images include four channels. For our experiments, we focus on the IR-R-G and R-G-B images. The Vaihingen dataset comprises 33 VHR TOPs, approximately sized at 2000 × 2000 pixels, and exclusively utilizes the IR-R-G imaging mode.

To ensure a fair comparison with previous methodologies [16], [17], [76], [77], we follow their data preprocessing technique, which involves cropping VHR images into smaller patches. These patches are standardized at a size of 512 × 512 pixels. For the cropping process, we apply strides of 512 for the Potsdam dataset and 256 for the Vaihingen dataset, resulting in totals of 4598 and 1696 patches, respectively. Additionally, we utilize the same strategy for the division of datasets, separating each into training and testing subsets. Consequently, the training subsets for Potsdam and Vaihingen include 2904 and 1296 images, respectively, while the testing subsets comprise 1694 and 440 images, respectively.

In this article, we design four UDA-RSSeg tasks, which are listed as follows.

1) Adapt Potsdam IR-R-G to Vaihingen IR-R-G (Potsdam IR-R-G → Vaihingen IR-R-G).
2) Adapt Vaihingen IR-R-G to Potsdam IR-R-G (Vaihingen IR-R-G → Potsdam IR-R-G).
3) Adapt Potsdam R-G-B to Vaihingen IR-R-G (Potsdam R-G-B → Vaihingen IR-R-G).
4) Adapt Vaihingen IR-R-G to Potsdam R-G-B (Vaihingen IR-R-G → Potsdam R-G-B).

CITY-OSM focuses exclusively on urban areas, capturing the intricate details of cities, including streets, buildings, parks, and other urban features. CITY-OSM [27] comprises several subsets, Berlin, Chicago, Zurich, Paris, and Tokyo. All images are annotated at the pixel level with three categories including "Background," "Road" and "Building." Following previous methods [16], [78], we select Paris and Chicago subsets to conduct experiments on the UDA-RSSeg task. Paris and Chicago datasets, respectively, have 725 and 457 images. Similar to the data-preprocessing approach on ISPRS, we also adopt cropping on CITY-OSM, where the stride and patch size are, respectively, set as 512 and 512 × 512. After cropping, Paris and Chicago datasets, respectively, have 22 500 and 13 710 images, where 70% are randomly selected as training images and the remaining 30% are selected as testing images.

TABLE I
Optimization Hyperparameters for Hybrid Training

Hyperparameters	Prompted Segmentor		Mapping Decoder	
	γ_1	γ_2	γ_3	γ_4
Values	0.005	0.75	0.001	0.75

In this article, we follow [16], [78] and conduct one UDA-RSSeg task, which is listed as follows.

1) Adapt Paris to Chicago (Paris → Chicago).

2) Metrics: To evaluate the model's performance, we adopt IoU/mIoU and $F1$-score/mF-score as metrics. Specifically, for a class i, IoU is formulated as $\text{IoU}_i = \text{tp}_i/(\text{tp}_i + \text{fp}_i + \text{fn}_i)$, where tp_i, fp_i, and fn_i denote true positive, false positive, and false negative, respectively. The mIoU is the mean value of all categories' IoU. Additionally, $F1$-score is defined as $F1\text{-score} = (2 \times \text{precision} \times \text{recall})/(\text{precision}+\text{recall})$. The mF-score is the mean value of all categories' $F1$-score.

B. Implementation Details

1) Architecture Details: PFM-JONet consists of three key modules, which are PS, prompted foundation model (SAM), and MD. For the PS, we select SegFormer-b5 [32]. The "PS Encoder" and "PS Decoder" indicate "mit-b5" and "ALL-MLP," respectively. For SAM, there are three main architecture types, "base," "large," and "huge." Each type indicates a specific ViT-based [79] "SAM Encoder." In this article, we select the "base" type for efficient training. For the MD, we select multilevel neck as "MD Neck" and select UperNet [80] as "MD Head." For the logits-level and feature-level discriminators, we select PatchGAN [81] to conduct adversarial learning. Specifically, two discriminators have four convolutional blocks with kernels of size 4 × 4. The stride settings are 2 for the first two blocks and 1 for the last two. The output channels for these blocks are set to 64, 128, 256, and 1, respectively. For the logits-level discriminator and the feature-level discriminator, the input channel is equal to the category's number and the output channel of the "PS Encoder," respectively.

2) Optimization Details: To implement the joint-optimized mechanism in PFM-JONet, we have developed separate multioptimizers for each critical component. For the "PS" and "MD," we employ AdamW [82] as the optimizer. The initial learning rate is set at 0.00006, with a weight decay of 0.01. For the two discriminators, we use Adam [83] as the optimizer, with an initial learning rate of 0.0001 and a weight decay of 0.01. As outlined in Algorithm 1, we use γ as optimization hyperparameters to adjust the intensity of different loss functions. As illustrated in Table I, for the first stage optimization on PS, γ_1 and γ_2 are set to 0.005 and 0.75, respectively, to achieve an optimal balance between the feature-level adversarial loss and the self-training loss. For the second-stage optimization on MD, γ_3 and γ_4 are,

TABLE II
UDA-RSSEG COMPARISON RESULTS (%) ON "POTSDAM IR-R-G → VAIHINGEN IR-R-G" TASK

Methods	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	$mIoU$	mF-score
AdaptSegNet [46]	4.60	8.76	54.39	70.39	6.40	11.99	52.65	68.96	28.98	44.91	63.14	77.40	35.02	47.05
ProDA [84]	3.99	8.21	62.51	76.85	39.20	56.52	56.26	72.09	34.49	51.65	71.61	82.95	44.68	58.05
DualGAN [17]	29.66	45.65	49.41	66.13	34.34	51.09	57.66	73.14	38.87	55.97	62.30	76.77	45.38	61.43
Zhang *et al.* [77]	20.71	31.34	67.74	80.13	44.90	61.94	55.03	71.90	47.02	64.16	76.75	86.65	52.03	66.02
Wang *et al.* [19]	21.85	35.87	**76.58**	**86.73**	35.44	52.33	55.22	71.15	49.97	66.64	82.74	90.56	53.63	67.21
DNT [58]	14.77	25.74	69.74	82.18	**53.88**	**70.03**	59.19	74.37	47.51	64.42	80.04	88.91	54.19	67.61
CIA-UDA [85]	27.80	43.51	63.28	77.51	52.91	69.21	64.11	78.13	48.03	64.90	75.13	85.80	55.21	69.84
JDAF [62]	38.65	55.75	68.76	81.49	42.76	59.90	58.38	73.72	47.39	64.30	77.19	87.13	55.52	70.38
CPCA [86]	-	60.18	-	84.92	-	68.10	-	79.75	-	61.99	-	91.76	60.75	74.45
DAFormer [52]	48.26	60.17	74.09	84.12	38.96	56.41	**70.88**	81.36	**57.53**	**71.48**	84.07	90.75	62.30	74.05
ST-DASegNet [16]	67.03	80.28	74.43	85.36	43.38	60.49	67.36	80.49	48.57	65.37	**85.23**	**92.03**	64.33	77.34
PFM-JONet (ours)	**68.88**	**81.57**	72.16	83.83	52.68	69.00	69.23	**81.82**	54.39	70.46	83.81	91.19	**66.86**	**79.65**

TABLE III
UDA-RSSEG COMPARISON RESULTS (%) ON "VAIHINGEN IR-R-G → POTSDAM IR-R-G" TASK

Methods	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	$mIoU$	mF-score
AdaptSegNet [46]	8.36	15.33	49.55	64.64	40.95	58.11	22.59	36.79	34.43	61.50	48.01	63.41	33.98	49.96
ProDA [84]	10.63	19.21	44.70	61.72	46.78	63.74	31.59	48.02	40.55	57.71	56.85	72.49	38.51	53.82
DualGAN [17]	11.48	20.56	51.01	67.53	48.49	65.31	34.98	51.82	36.50	53.48	53.37	69.59	39.30	54.71
DNT [58]	11.51	20.65	61.91	76.48	49.50	66.22	35.46	52.36	37.61	54.67	66.41	79.81	43.74	58.36
Zhang *et al.* [77]	12.31	**24.59**	64.39	78.59	59.35	75.08	37.55	54.60	47.17	63.27	66.44	79.84	47.87	62.66
Wang *et al.* [19]	11.65	19.47	73.43	84.55	63.86	77.85	32.68	47.36	47.69	63.45	76.32	87.43	50.94	63.31
JDAF [62]	**13.10**	23.17	67.70	80.74	63.22	77.47	36.21	53.17	51.19	67.72	76.36	86.59	51.30	64.81
CIA-UDA [85]	10.87	19.61	62.74	77.11	65.35	79.04	47.74	64.63	54.40	70.47	72.31	83.93	52.23	65.80
CPCA [86]	-	26.68	-	81.51	-	84.45	-	38.00	-	71.02	-	85.59	50.72	64.54
DAFormer [52]	2.56	5.02	68.42	79.07	65.20	79.31	**70.65**	**82.13**	56.39	72.48	78.94	87.64	57.03	67.61
ST-DASegNet [16]	0.18	0.35	**76.45**	**86.65**	73.54	84.76	62.89	77.22	**61.04**	**75.80**	83.81	**91.19**	**59.65**	69.33
PFM-JONet (ours)	5.72	8.44	75.97	85.78	**74.24**	**85.35**	68.62	80.91	58.35	75.61	**84.37**	90.18	**61.21**	**71.05**

respectively, set as 0.001 and 0.75 to balance the logits-level adversarial loss and the self-training loss. For further details on the optimization process, refer to our implementation available at https://github.com/CV-ShuchangLyu/PFM-JONet/

3) Experimental Environment: All experiments are conducted using the mmsegmentation framework [87], a leading open-source platform designed to advance research and development in semantic segmentation. Built on PyTorch, mmsegmentation offers a comprehensive suite of cutting-edge algorithms, pretrained models, and robust tools for researchers. For implementing adversarial learning and self-training within PFM-JONet, we also utilize resources from Mmagic [88]. All experiments are performed on two NVIDIA RTX 4090 GPUs, with a batch size of 3 per GPU. For more implementation details, refer to https://github.com/CV-ShuchangLyu/PFM-JONet/

C. Experimental Results

1) Comparison Experiments on the ISPRS Dataset: As shown in Tables II–V, we compare the PFM-JONet's performance with previous methods on the aforementioned four UDA-RSSeg task.

For the "Potsdam IR-R-G → Vaihingen IR-R-G" task, the model is trained using 2904 annotated training images from the Potsdam IR-R-G subset and 1296 unannotated training images from the Vaihingen IR-R-G subset. Evaluation is performed on 440 testing images from the Vaihingen IR-R-G subset. This task presents a significant domain gap due to the distinct geographical landscapes between the source (Potsdam) and target (Vaihingen) domains. The comparison results are shown in Table II. Compared to ST-DASegNet [16], PFM-JONet, respectively, achieves 2.53% and 2.31% improvement on mIoU and mF-score. Particularly in the "Clutter" category, PFM-JONet surpasses all previous methods. When distinguishing between the "Tree" and "Low vegetation" categories, PFM-JONet achieves second best, only outperformed by DAFormer.

For the "Vaihingen IR-R-G → Potsdam IR-R-G" task, the model is trained using 1296 annotated training images from the Vaihingen IR-R-G subset and 2904 unannotated training images from the Potsdam IR-R-G subset. Evaluation is conducted on 1694 testing images from the Potsdam IR-R-G subset. This task exhibits a domain-shift challenge similar to the first task. As shown in Table III, PFM-JONet surpasses previous SOTA method, ST-DASegNet [16] by 1.56% on the mIoU value and 1.72% on mF-score. In the "Car" category, PFM-JONet shows obvious superiority over previous methods. In the "Impervious surfaces" and "Building" categories, PFM-JOANet achieves performance close to that of the top-performing ST-DASegNet. For the "Tree" and "Low vegetation" categories, PFM-JOANet yields slightly lower results compared to DAFormer.

In the "Potsdam R-G-B → Vaihingen IR-R-G" task, the model is trained using 2904 annotated training images from the Potsdam R-G-B subset and 1296 unannotated training images from the Vaihingen IR-R-G subset. The evaluation is conducted on 440 testing images from the Vaihingen

TABLE IV
UDA-RSSEG COMPARISON RESULTS (%) ON "POTSDAM R-G-B → VAIHINGEN IR-R-G"

Methods	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	$mIoU$	mF-score
AdaptSegNet [46]	2.99	5.81	51.26	67.77	10.25	18.54	51.51	68.02	12.75	22.61	60.72	75.55	31.58	43.05
ProDA [84]	2.39	5.09	49.04	66.11	31.56	48.16	49.11	65.86	32.44	49.06	68.94	81.89	38.91	52.70
DualGAN [17]	3.94	13.88	49.16	61.33	40.31	57.88	55.82	70.66	27.85	42.17	65.44	83.00	39.93	54.82
Zhang *et al.* [77]	12.38	21.55	64.47	77.76	43.43	60.05	52.83	69.62	38.37	55.94	76.87	86.95	48.06	61.98
Wang *et al.* [19]	12.61	22.39	73.80	84.92	43.24	60.38	44.41	61.50	43.27	60.40	83.76	91.16	50.18	63.46
CIA-UDA [85]	13.50	23.78	62.63	77.02	52.28	68.66	63.43	77.62	33.31	49.97	79.71	88.71	50.81	64.29
JDAF [62]	32.71	49.30	64.33	78.29	45.87	62.90	51.99	68.41	42.16	59.31	75.53	86.06	52.10	67.38
DNT [58]	11.55	20.71	67.94	80.91	**52.64**	**68.97**	58.43	73.76	**43.63**	**61.05**	81.09	89.56	52.60	65.83
CPCA [86]	-	39.64	-	81.06	-	60.08	-	**81.50**	-	43.34	-	84.67	47.67	65.05
DAFormer [52]	22.57	33.72	67.44	79.65	45.60	60.13	66.27	80.41	40.49	54.93	81.34	90.07	53.95	66.49
ST-DASegNet [16]	**36.03**	**50.64**	68.36	**81.28**	43.15	60.28	64.65	78.31	34.69	47.08	**84.09**	**91.33**	55.16	68.15
PFM-JONet (ours)	32.56	47.83	**70.05**	81.16	49.61	65.88	**66.45**	79.50	39.76	57.65	81.01	89.44	**56.61**	**70.21**

TABLE V
UDA-RSSEG COMPARISON RESULTS (%) ON " VAIHINGEN IR-R-G → POTSDAM R-G-B"

Methods	Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	IoU	$F1$-score	$mIoU$	mF-score
AdaptSegNet [46]	6.11	11.50	37.66	59.55	42.31	55.95	30.71	45.51	15.10	25.81	54.25	70.31	31.02	44.75
ProDA [84]	11.13	20.51	44.77	62.03	41.21	59.27	30.56	46.91	35.84	52.75	46.37	63.06	34.98	50.76
DualGAN [17]	13.56	23.84	45.96	62.97	39.71	56.84	25.80	40.97	41.73	58.87	59.01	74.22	37.63	52.95
DNT [58]	8.43	15.55	56.41	72.13	46.78	63.74	36.56	53.55	30.59	46.85	69.95	82.32	41.45	55.69
Zhang *et al.* [77]	13.27	23.43	57.65	73.14	56.99	72.27	35.87	52.80	29.77	45.88	65.44	79.11	43.17	57.77
Wang *et al.* [19]	10.84	17.49	66.11	79.75	65.45	80.17	28.64	43.51	35.47	51.85	68.63	81.32	45.86	59.74
JDAF [62]	**18.09**	**30.93**	60.05	75.04	58.64	73.93	38.74	55.84	27.79	43.49	71.42	83.33	45.79	60.38
CIA-UDA [85]	9.20	16.86	53.39	69.61	63.36	77.57	44.90	61.97	43.96	61.07	70.48	82.68	47.55	61.63
DAFormer [52]	1.07	1.88	65.12	78.16	70.40	84.28	**61.25**	**76.59**	49.02	65.51	82.44	89.70	54.88	66.02
ST-DASegNet [16]	3.70	7.38	69.83	83.12	**75.99**	**87.89**	57.41	73.47	50.76	67.64	**83.46**	**90.67**	56.86	68.37
PFM-JONet (ours)	4.94	7.70	**73.31**	**84.60**	73.36	85.27	55.38	72.14	**59.88**	**74.09**	78.73	88.10	**57.60**	**68.65**

IR-R-G subset. This task is particularly challenging due to sensor variations and landscape discrepancies between the two datasets, making it more complex than the aforementioned tasks. Specifically, "Tree" is hard to distinguish, because the color of "Tree" is green in R-G-B images while red in IR-R-G images. In Table IV, we find that PFM-JONet achieves the best IoU value and the second best $F1$-score in this category. For overall performance, PFM-JONet also ranks top-1, which proves the generalized performance in more challenging situations.

In the "Vaihingen IR-R-G → Potsdam R-G-B" task, the model is trained using 1296 annotated training images from the Vaihingen IR-R-G subset and 2904 unannotated training images from Potsdam R-G-B subset. The evaluation is performed on 1694 testing images from the Potsdam R-G-B subset. This task encounters similar challenges as the third task, primarily due to sensor variations and landscape discrepancies between the datasets. In Table V, we find that PFM-JONet gains 0.74% and 0.28% on mIoU value and mF-score when compared to ST-DASegNet. Especially in the "Impervious surfaces" and "Low vegetation" categories, PFM-JONet surpasses previous methods by a large margin.

Overall, PFM-JONet demonstrates impressive performance across various UDA-RSSeg tasks in the ISPRS datasets. By leveraging generalized features extracted from SAM, the hybrid training mechanism of our proposed PFM-JONet is both logical and effective. When compared to various adversarial learning and self-training methods, PFM-JONet stands out as a superior option.

TABLE VI
UDA-RSSEG COMPARISON RESULTS (%) ON THE "PARIS → CHICAGO" TASK

Methods	Background (IoU)	Road (IoU)	Building (IoU)	$mIoU$
AdaptSegNet [46]	49.84	11.09	52.71	37.88
AdvEnt [89]	50.42	22.92	49.43	40.92
CLAN [90]	53.85	23.25	44.37	40.49
MaxSquare [91]	53.47	20.75	43.06	39.09
BDCA [92]	53.88	20.31	52.97	42.39
Wang *et al.* [93]	53.01	23.00	55.43	43.82
ColorMapGAN [94]	54.58	15.15	49.07	39.61
SAC [95]	55.69	20.86	55.09	43.88
DPL [96]	53.52	24.51	56.22	44.75
Chen *et al.* [78]	54.20	24.66	56.32	45.06
ST-DASegNet [16]	51.84	45.27	59.26	52.12
PFM-JONet (ours)	**59.54**	**46.80**	**64.56**	**56.96**

2) Comparison Experiments on the CITY-OSM Dataset: As shown in Table VI, we make a comparison between our proposed PFM-JONet with previous SOTA methods on the "Paris → Chicago" task. Obviously, PFM-JONet outperforms all previous methods. Specifically, PFM-JONet surpasses ST-DASegNet [16] by 4.84% in the mIoU value, thereby achieving new SOTA results. On three categories, PFM-JONet all show stronger performance. For the "Background" and "Building" categories, PFM-JONet significantly outperforms the second-best method by a considerable margin.

3) Ablation Study: To separately show the performance of each key component of PFM-JONet, we conduct ablation study on "Potsdam IR-R-G → Vaihingen IR-R-G" and "Potsdam R-G-B → Vaihingen IR-R-G" adaptation tasks. As shown in Table VII, we will analyze the following aspects.

TABLE VII

ABLATION STUDY WITH CATEGORY-LEVEL SEGMENTATION RESULTS ON "POTSDAM IR-R-G → VAIHINGEN IR-R-G" AND "VAIHINGEN IR-R-G → POTSDAM IR-R-G" (%). "PS" AND "ST," RESPECTIVELY, INDICATE PROMPTED SEGMENTOR AND SELF-TRAINING. "F-ADV" AND "L-ADV," RESPECTIVELY, INDICATE FEATURE-LEVEL AND LOGITS-LEVEL ADVERSARIAL LEARNING. $F1$ AND mF, RESPECTIVELY, DENOTE $F1$-SCORE AND mF-SCORE

Methods					Clutter		Impervious surfaces		Car		Tree		Low vegetation		Building		Overall	
Potsdam IR-R-G → Vaihingen IR-R-G																		
Baseline	PS	F-Adv	L-Adv	ST	IoU	$F1$	IoU	$F1$	IoU	$F1$	IoU	$F1$	IoU	$F1$	IoU	$F1$	$mIoU$	mF
✓	-	-	-	-	10.45	24.87	59.88	74.69	18.14	33.48	62.30	76.82	39.04	57.56	73.25	84.91	43.84	58.72
✓	✓	-	-	-	17.82	32.85	65.20	77.84	32.65	48.68	66.74	80.76	47.93	66.08	77.16	86.71	51.25	65.49
✓	✓	✓	-	-	30.73	42.11	66.83	77.93	35.28	49.74	65.85	78.43	51.59	67.90	79.37	87.04	54.94	67.19
✓	✓	-	✓	-	24.08	36.98	68.63	79.12	36.72	51.03	64.35	78.07	47.06	64.51	81.13	88.43	53.66	66.36
✓	✓	✓	✓	-	40.11	54.29	71.49	83.54	45.88	64.57	65.72	78.95	52.40	68.56	83.22	90.83	59.80	73.46
✓	✓	✓	✓	✓	68.88	81.57	72.16	83.83	52.68	69.00	69.23	81.82	54.39	70.46	83.81	91.19	**66.86**	**79.65**
Vaihingen IR-R-G → Potsdam IR-R-G																		
✓	-	-	-	-	0.42	0.68	60.81	75.26	50.74	67.83	26.65	48.19	42.69	60.18	68.06	82.74	41.56	55.81
✓	✓	-	-	-	6.35	15.38	66.89	80.15	53.78	71.42	38.09	55.36	50.40	67.97	72.40	83.29	47.99	62.26
✓	✓	✓	-	-	4.43	12.17	70.93	81.48	65.10	79.94	38.97	57.25	54.07	70.68	78.16	86.71	51.94	64.70
✓	✓	-	✓	-	2.70	7.91	69.45	81.34	67.44	82.58	36.97	54.87	53.04	70.24	73.29	85.59	50.48	63.76
✓	✓	✓	✓	-	5.12	8.69	74.25	84.01	72.99	83.27	49.37	65.62	55.18	71.85	82.80	89.63	56.62	67.18
✓	✓	✓	✓	✓	5.72	8.44	75.97	85.78	74.24	85.35	68.62	80.91	58.35	75.61	84.37	90.18	**61.21**	**71.05**

1) *Baseline Versus PFM-JONet:* In this article, the baseline model is designed as "PFM-based Fine-tuning" paradigm [Fig. 3(b)]. Compared to baseline models on two tasks, the PFM-JONet, respectively, improves by 23.02% and 19.65% on mIoU value and 20.93% and 15.24% on mF-score.
2) *Effectiveness of PS:* As shown in Table VII, the baseline model with a PS shows significant improvement, indicating that the MD benefits from the guidance of the prompted class-agnostic map.
3) *Effectiveness of Adversarial Learning:* In PFM-JONet, we integrate feature-level adversarial learning into the PS. Comparative results demonstrate that feature-level adversarial learning enhances the segmentor. Logits-level adversarial learning, embedded in the MD, also contributes to performance gains by reducing the discrepancy between features and predictive logits.
4) *Effectiveness of Self-Training:* Results in Table VII reveal that models employing self-training achieve further enhancements. This confirms that the self-training mechanism corrects the representation bias and stabilizes optimization during adversarial training, resulting in improved segmentation performance.

As shown in Table VII, we further provide the following specific analysis.

1) When integrating feature-level adversarial learning into the "Baseline + PS," the model exhibits significant improvement. This enhancement occurs because feature-level adversarial learning boosts the predictive capabilities of the PS. In turn, this enhanced prediction contributes to the refinement of the MD by supplying a high-quality prompted map.
2) From Table II, it is evident that feature-level and logits-level adversarial learning affect various categories differently. Feature-level adversarial learning is particularly beneficial for categories with fewer samples, enhancing their performance. Conversely, logit-level adversarial learning shows improved effectiveness in categories with a larger number of samples. Consequently, combining these two approaches results in a complementary and compatible two-scheme adversarial learning strategy that performs exceptionally well.
3) The self-training mechanism also demonstrates strong compatibility with adversarial learning, straightforwardly enhancing the model. Essentially, it helps correct representation bias and stabilizes the optimization process during adversarial training.
4) In the "Vaihingen IR-R-G → Potsdam IR-R-G" task, we observe that our proposed key components consistently struggle with the "Clutter" category. This issue stems from the scarcity of training samples featuring "Clutter." This scarcity exemplifies how the few-shot problem adversely affects the performance in the UDA-RSSeg task.

4) Sensitivity Analysis on Optimization Hyperparameters: As shown in Table I, we provide detailed information on the optimization hyperparameters (γ) used in hybrid training. Here, we conduct sensitivity analysis experiments to systematically evaluate the impact of different combinations of γ.

a) Rules of optimization hyperparameter configuration: 1) We follow the design principles of AdapSegNet [46] for the adversarial loss weighting factor (γ_1, γ_3), avoiding excessively large values to mitigate potential training instability. Specifically, γ_1 is set larger than γ_3. The reason is that the logits-level adversarial loss (for the MD) demonstrates greater training instability than its feature-level counterpart (for the PS), owing to its closer coupling with the segmentation objective; and 2) the weight factors for self-training (γ_2, γ_4) should not exceed those of the segmentation loss, as while target-domain pseudo-label supervision helps mitigate representation

TABLE VIII
SENSITIVITY ANALYSIS ON OPTIMIZATION HYPERPARAMETERS. WE CONDUCT UDA-RSSEG EXPERIMENTS ON THE "POTSDAM IR-R-G → VAIHINGEN IR-R-G" TASK

Methods				Optimization Hyperparameters				Overall	
F-Adv	L-Adv	PS-ST	MD-ST	γ_1	γ_2	γ_3	γ_4	$mIoU$	mF
Potsdam IR-R-G → Vaihingen IR-R-G									
-	-	-	-	0	0	0	0	51.25	65.49
✓	-	-	-	0.001	0	0	0	52.47	66.03
✓	-	-	-	0.01	0	0	0	53.05	66.37
✓	-	-	-	0.005	0	0	0	54.94	67.19
✓	-	✓	-	0.005	0.25	0	0	55.68	68.22
✓	-	✓	-	0.005	1.0	0	0	57.37	71.95
✓	-	✓	-	0.005	0.75	0	0	57.66	72.04
✓	✓	-	-	0.005	0	0.001	0	59.80	73.46
✓	✓	✓	-	0.005	0.75	0.0005	0	59.65	73.31
✓	✓	✓	-	0.005	0.75	0.005	0	60.92	74.23
✓	✓	✓	-	0.005	0.75	0.001	0	61.85	75.81
✓	✓	✓	✓	0.005	0.75	0.001	0.25	63.31	76.48
✓	✓	✓	✓	0.005	0.75	0.001	1.0	65.55	78.46
✓	✓	✓	✓	0.005	0.75	0.001	0.75	**66.86**	**79.65**

tendency toward source domain, it inevitably introduces some noisy signals.

b) Sensitivity analysis of optimization hyperparameters:

1) As shown in Algorithm 1, the PS and MD operate in different optimization loops. Notably, performance improvements in PS can enhance prompt information for MD. Accordingly, our experiments begin with a baseline implementation excluding both adversarial learning and self-training components and first evaluate the weight factors (γ_1, γ_3) of PS.
2) As shown in Table VIII, when γ_1 is lower, the adversarial effect becomes negligible, yielding results comparable to the nonadversarial baseline. Higher γ_1 will introduce training instability, degrading the segmentation accuracy of target domains. The self-training mechanism further enhances PS optimization when combined with feature-level adversarial learning. Performance scales with γ_2 (0.25 → 0.75 → 1.0 yields 55.68% → 57.66% → 57.37% in mIoU), indicating PS's robustness to pseudo-label noise. The slight drop at $\gamma_2 = 1.0$ suggests marginal overfitting.
3) Building upon the optimal performance of the PS, we conduct a comprehensive evaluation of the MD's optimization hyperparameters (γ_3, γ_4). As shown in Table VIII, the incorporation of both logits-level adversarial loss and self-training into the MD yields significant performance improvements. However, when these losses are jointly optimized with the segmentation loss, the model exhibits heightened sensitivity to variations in γ_3 and γ_4.
4) While the current implementation demonstrates promising results through coarse parameter tuning, comprehensive fine-grained optimization and deeper analysis of interloss tradeoffs warrant further systematic investigation.

5) Visualization and Analysis: To intuitively show the performance and interpretability of PFM-JONet, we conduct visualization experiments on the ISPRS and CITY-OSM datasets.

a) Qualitative visualization results on ISPRS: As shown in Fig. 6, we provide qualitative visualization results on all four tasks. The abundant visualization on segmentation prediction intuitively proves the effectiveness of PFM-JONet. With the guidance of a high-quality class-agnostic map, PFM-JONet shows a strong overall performance. In some specific categories, PFM-JONet can sometimes provide surprising results. Moreover, visualization results coincide with the segmentation results shown in Tables II–V.

b) Qualitative visualization results on CITY-OSM: To intuitively show the performance of PFM-JONet on the "Paris → Chicago" UDA-RSSeg task, we provide visualization results shown in Fig. 6. From the results, we analyze the following points: 1) compared to baseline models, PFM-JONet shows obvious overall superiority, especially in the "Road" category, which coincides with the results shown in Table V of the main manuscript; 2) it is clear that the prompted map shows clear boundary and accurate attention, which provides PFM-JONet with important guidance; and 3) from the results of the baseline model ("PFM-based Fine-tuning" paradigm), we find that even with generalized feature representation ability, it is still hard to fill the domain gap when mapping feature to predictive logits. In this article, we address this issue and propose PFM-JONet, which provides insights for UDA-RSSeg tasks in the urban scene.

c) Visualization analysis on prompted maps: As shown in Fig. 7, we provide detailed visualization analysis on prompted maps. Here, we will analyze the following aspects: 1) when guided by a mask, SAM can provide three prompted maps with different guidance. In Fig. 7, the images in "left-top," "right-top," and "left-bottom" are three prompted maps; 2) obviously, three prompted maps have different attention tendencies. Different prompted maps may focus on different categories or different instances. To maximally utilize the attention information of three prompted maps, we adopt channel-wise mean operation to generate an integrated prompted map for guidance. The image in "left-bottom" indicates the integrated prompted map, and 3) the segmentation results coincide with the attention tendency of prompted maps.

D. Discussion

1) Model Complexity Analysis: Table IX presents a comprehensive comparison of model complexity metrics between our method and two notable methods: ST-DASegNet [16] and AdapSegNet [46], utilizing SegFormer-b5 as the baseline segmentor. Our analysis reveals two key findings: 1) our proposed PFM-JONet demonstrates a characteristic performance-complexity tradeoff: while the incorporation of SAM leads to elevated parameter counts, computational costs (FLOPs), and reduced inference speed (FPS) compared to ST-DASegNet and AdapSegNet, it simultaneously achieves enhanced segmentation performance; and 2) our method achieves enhanced segmentation performance while maintaining training efficiency, as only approximately 60% of parameters require optimization during training due to SAM's frozen parameters. In contrast, ST-DASegNet, AdapSegNet,

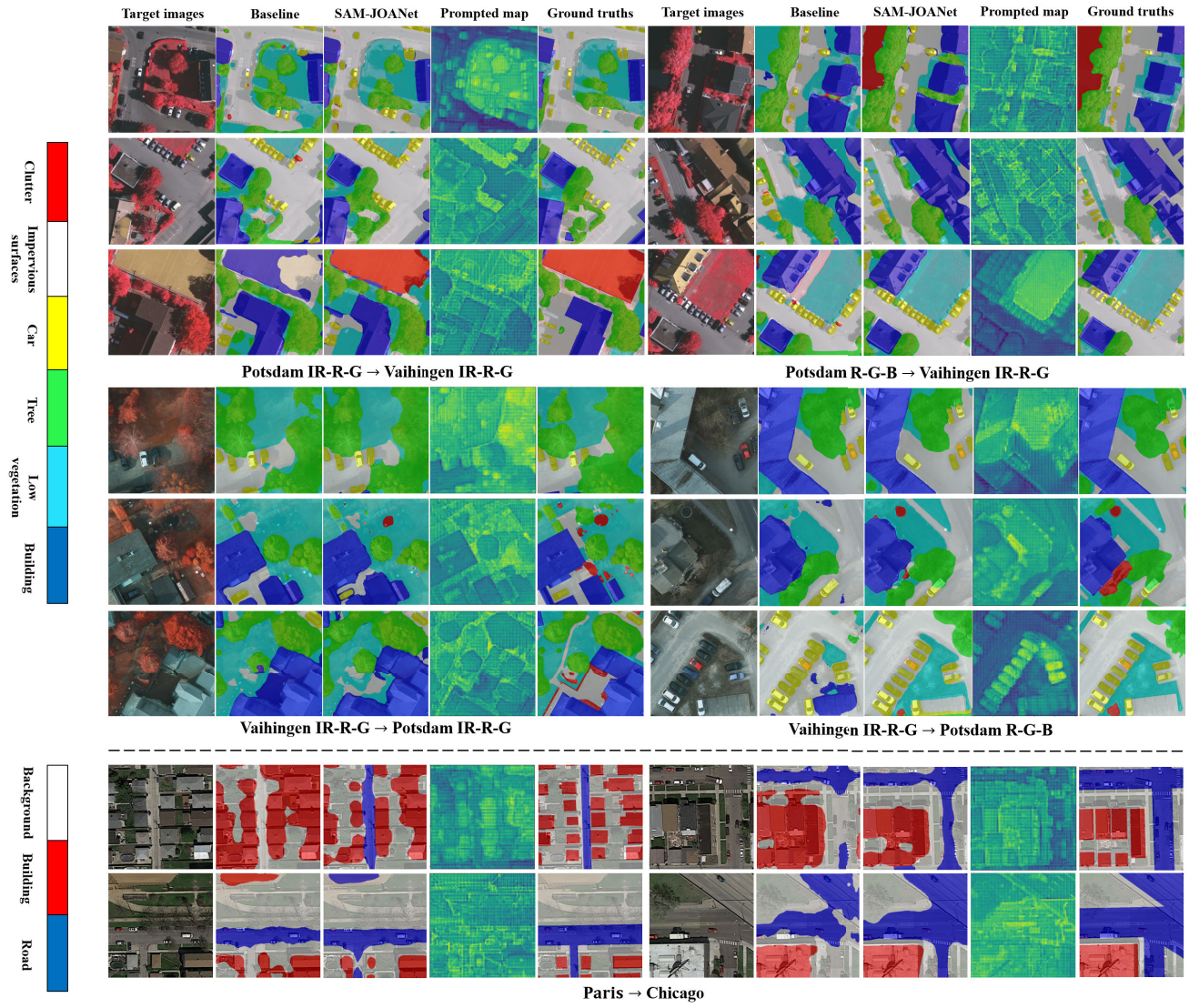

Fig. 6. Qualitative visualization results of the UDA-RSSeg task in the urban scene.

TABLE IX

MODEL COMPLEXITY ANALYSIS ON PFM-JONET. ALL MEASUREMENTS ARE CONDUCTED UNDER STANDARD CONDITIONS USING AN NVIDIA GEFORCE RTX 4090 GPU, WITH THE INPUT SIZE FIXED AT 512 × 512

Methods	Model complexity				
	Input shape	Total params (M)	Trainable params (M)	Inference FLOPs (G)	FPS
PFM-JONet	512×512	233.5	138.3	546.8	9.5
ST-DASegNet [16]		170.5	170.5	104.5	24.4
AdapSegNet [46]		82.0	82.0	39.7	49.7

and the majority of existing methods employ full-model optimization strategies to tackle the UDA-RSSeg task.

2) Analysis on Training Instability Issue: PFM-JONet is optimized with multiple optimization objectives. This design inevitably introduces training instability, which is a fundamental challenge inherent to UDA-RSSeg tasks. Indeed, among existing methods addressing domain shift, training instability remains a persistent issue. Here, we will provide an analysis of the following aspects.

1) Different optimization objectives operate through complementary mechanisms that collectively enhance model performance. The segmentation loss on source-domain annotations establishes fundamental representation learning for the entire framework. Adversarial learning operates at both the feature level and logit level to achieve cross-domain alignment. The self-training mechanism affects balancing the representation tendency between source- and target-domain images. Our ablation studies validate this synergistic interaction: domain-invariant features generated through adversarial learning enhance pseudo-label reliability, which subsequently improves target-domain representation through iterative self-training refinement. The integrated loss functions demonstrate strong compatibility with partially orthogonal relationships, collectively enhancing PFM-JONet's performance.
2) As shown in Algorithm 1, the PS and MD operate in different optimization loops. It means that the optimization

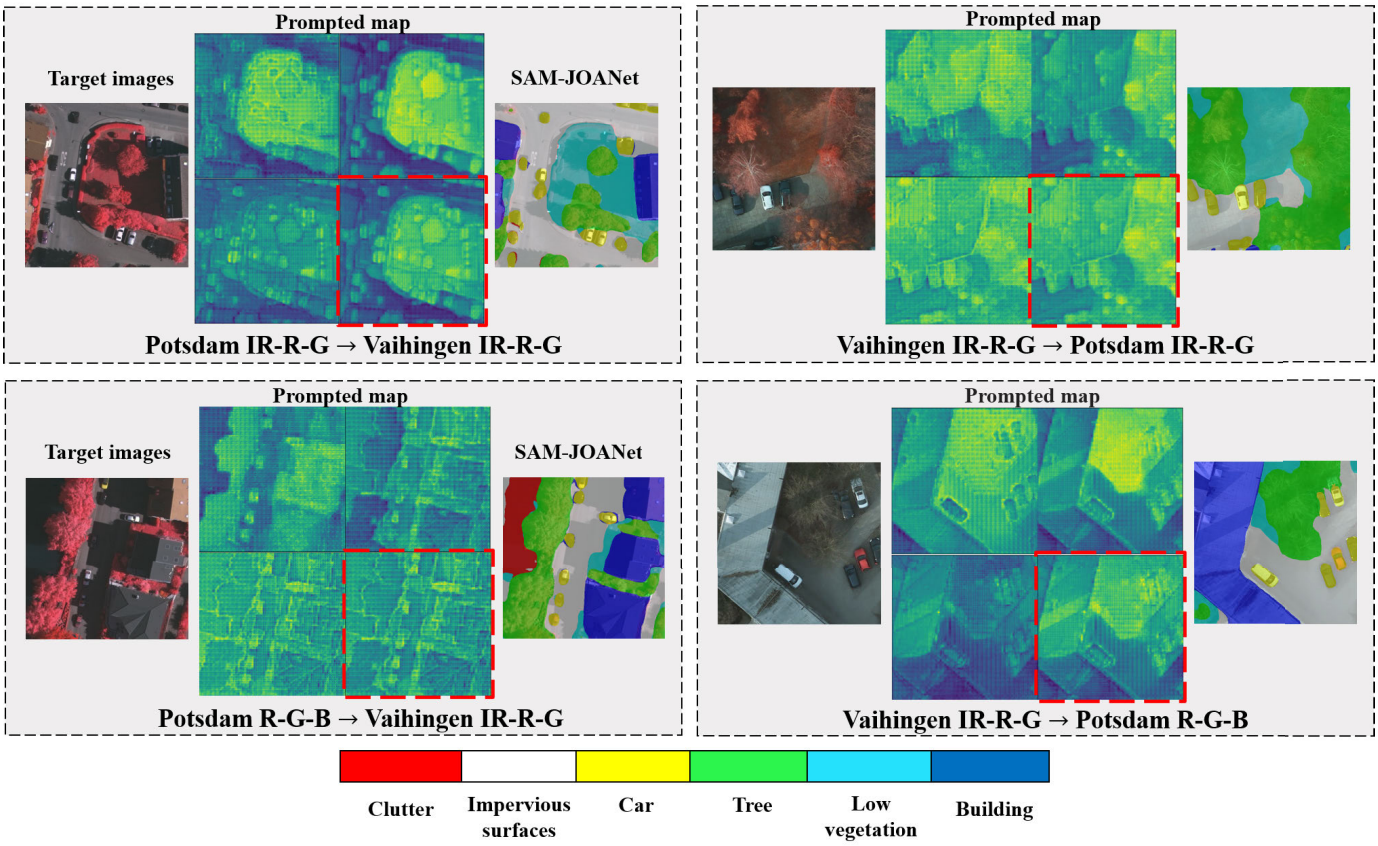

Fig. 7. Visualization analysis on prompted maps. The prompted maps without red doted boxes indicate the three prompted maps generated by SAM. The prompted maps with red dotted boxes indicate the integrated prompted map with channel-wise mean operation, which is used for guidance.

processes of these two key modules do not interfere with each other, which, to a certain extent, renders the PFM-JONet less prone to training instability than it might appear.

3) We introduce "pseudo_threshold" to filter out unreliable pseudo-labels. This design results in zero self-training loss during initial phases (first ∼2k/40k iterations) when the model cannot generate sufficiently confident pseudo-labels. This delayed activation ensures that self-training does not interfere with early-stage convergence. As training progresses, the gradual involvement of self-training effectively stabilizes the adaptation process.
4) In this task, empirical parameter-tuning skills also play a crucial role. In our released code, we provide detailed hyperparameters for each optimization objective.

3) Frozen SAM Versus LoRA Fine-Tuning: SAM is trained with natural corpus, and its feature extraction ability on RS images may still depend on certain fine-tuning strategies. To evaluate the fine-tuning performance, we insert low-rank adaptation (LoRA) [97] fine-tuning layer in PFM-JONet. In our implementations, LoRA is typically applied to query/key/value projections in self-attention and intermediate dense layers in the feed-forward network (FFN) blocks of the "Transformer Encoder" of the "SAM Encoder." The comparison of these two structures is shown in Fig. 8.

As shown in Table X, we conduct UDA-RSSeg experiments to evaluate the performance on "Potsdam IR-R-G → Vaihingen IR-R-G" and "Vaihingen IR-R-G → Potsdam IR-R-G" tasks. Overall, we find that these two models achieve comparable results, whereas the original model with SAM frozen shows a slight advantage. The analysis of the results can be presented as follows: 1) the results in Fig. 2 demonstrate that SAM's strong performance in natural image segmentation can be effectively extended to RS image segmentation tasks. This suggests that keeping the entire SAM model frozen may yield more stable and superior results; and 2) for downstream tasks with a small number of training samples in specialized datasets, the domain shifts disrupt the alignment between SAM modules, ultimately diminishing the benefits of RS features' informational advantages.

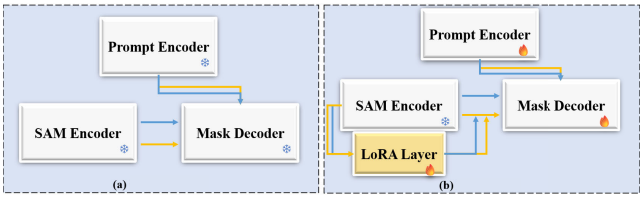

Fig. 8. Structure of (a) "original SAM" and (b) "SAM-LoRA" within PFM-JONet.

4) Limitations and Future Works: While the proposed method demonstrates promising results, it still has certain limitations that warrant further exploration. First, our proposed method mainly focuses on the UDA-RSSeg task configured as single-source and single-target data, which may restrict

TABLE X
"SAM VERSUS SAM-LORA." COMPARISON RESULTS (%) ON "POTSDAM IR-R-G → VAIHINGEN IR-R-G" AND "VAIHINGEN IR-R-G → POTSDAM IR-R-G" TASKS

Methods	$mIoU$	mF
Potsdam IR-R-G → Vaihingen IR-R-G		
PFM-JONet (original)	66.86	79.65
PFM-JONet (LoRA)	65.19	78.16
Vaihingen IR-R-G → Potsdam IR-R-G		
PFM-JONet (original)	61.21	71.05
PFM-JONet (LoRA)	60.74	70.72

its ability to fully leverage complementary information from diverse data domains. Second, the image representation utilized in this method may not fully capture the complexities of diverse scenarios, limiting its generalization capabilities across multiple target domains. Third, this article does not explicitly model complex relationships within the data. Fourth, the reliance on labeled data remains a challenge. Although our method effectively mitigates the issue of scarce target annotations, its performance remains dependent on the availability of a sufficient amount of labeled source-domain data.

In future work, we aim to expand our research in several promising directions. First, beyond the current "single-source and single-target" paradigm in UDA-RSSeg tasks, we will explore more complex scenarios, such as "multisource" settings. This will involve leveraging multimodality fusion techniques, as discussed in [98], to integrate additional modalities and further enhance model performance. Second, we will also focus on "multitarget" settings in UDA-RSSeg tasks. As highlighted in [99], the adoption of advanced representation methods could significantly improve the model's ability to generalize across diverse target-domain datasets. Third, we will exploit the use of latent features and graph neural networks, as proposed in [100], to further enhance feature consistency and representation capabilities. Fourth, we will we will conduct in-depth research on self-paced semi-supervised learning mechanisms, as proposed in [101]. This could mitigate this issue by effectively utilizing limited labeled data and improving segmentation accuracy.

V. CONCLUSION

In this article, we propose a PFM-JONet to tackle the UDA-RSSeg task in the urban scene. To essentially improve the feature consistency representation, we integrate SAM into our architecture to leverage its generalized representation capabilities. To bridge the domain gaps between features and predictive logits, we employ a feature-level adversarial PS for SAM to generate a prompted class-agnostic map as guidance. We also involve the logits-level adversarial learning and self-training mechanism for MD to maximally boost the optimization effectiveness. Compared to previous paradigms, we first propose a "PFM-based Joint Optimization" paradigm and provide insights into a hybrid training scheme on the UDA-RSSeg task. Extensive experiments on multiple urban scene VHR benchmark datasets and visualization analysis show that PFM-JONet outperforms existing methods in an interpretable manner.

The proposed method still has limitations: it focuses on single-source and single-target UDA-RSSeg tasks, limits multidomain and multitarget data utilization, and does not explicitly model complex data relationships. Additionally, it relies on labeled source-domain data despite addressing scarce target annotations. Future work will explore multisource and multitarget settings using multimodality fusion and advanced representation methods, investigate latent features and graph neural networks for feature consistency, and adopt self-paced semi-supervised learning to enhance segmentation accuracy with limited labeled data.

REFERENCES

[1] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[2] H. Alemohammad and K. Booth, "LandCoverNet: A global benchmark land cover classification training dataset," 2020, *arXiv:2012.03111*.

[3] J. Wang, A. Ma, Y. Zhong, Z. Zheng, and L. Zhang, "Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery," *Remote Sens. Environ.*, vol. 277, Aug. 2022, Art. no. 113058.

[4] T. Liu, Y. Liu, C. Zhang, L. Yuan, X. Sui, and Q. Chen, "Hyperspectral image super-resolution via dual-domain network based on hybrid convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5512518.

[5] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856.

[6] S. Hafner, Y. Ban, and A. Nascetti, "Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data," *Remote Sens. Environ.*, vol. 280, Oct. 2022, Art. no. 113192.

[7] Z. Mao, X. Huang, W. Niu, X. Wang, Z. Hou, and F. Zhang, "Improved instance segmentation for slender urban road facility extraction using oblique aerial images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 121, Jul. 2023, Art. no. 103362.

[8] W. Li, Q. Chen, G. Gu, and X. Sui, "Object matching of visible–infrared image based on attention mechanism and feature fusion," *Pattern Recognit.*, vol. 158, Feb. 2025, Art. no. 110972.

[9] Z. Wang et al., "Lightning-generated whistlers recognition for accurate disaster monitoring in China and its surrounding areas based on a homologous dual-feature information enhancement framework," *Remote Sens. Environ.*, vol. 304, Apr. 2024, Art. no. 114021.

[10] M. Li, Z. Jiang, L. Xiao, B. Su, S. Chen, and J. Zhu, "Innovative cross-hole grounded-wire source transient electromagnetic method for sensing geological disasters in urban underground spaces," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5931314.

[11] A. Sarkar, T. Chowdhury, R. R. Murphy, A. Gangopadhyay, and M. Rahnemoonfar, "SAM-VQA: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702716.

[12] J. Hou, Z. Guo, Y. Wu, W. Diao, and T. Xu, "BSNet: Dynamic hybrid gradient convolution based boundary-sensitive network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5604912.

[13] D. Wang, J. Zhang, B. Du, D. Tao, and L. Zhang, "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. Adv. Neural Inf. Process. Syst. (NeuralPS)*, Jan. 2023, pp. 8815–8827.

[14] J. Li, X. Wang, H. Zhao, S. Wang, and Y. Zhong, "Anomaly segmentation for high-resolution remote sensing images based on pixel descriptors," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, Jun. 2023, pp. 4426–4434.

[15] Z. Yan et al., "RingMo-SAM: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5625716.

[16] Q. Zhao, S. Lyu, H. Zhao, B. Liu, L. Chen, and G. Cheng, "Self-training guided disentangled adaptation for cross-domain remote sensing image semantic segmentation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 127, Mar. 2024, Art. no. 103646.

[17] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 20–33, May 2021.

[18] X. Ma, X. Zhang, Z. Wang, and M.-O. Pun, "Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5400515.

[19] L. Wang, P. Xiao, X. Zhang, and X. Chen, "A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4109–4121, 2023.

[20] J. Iqbal, A. Masood, W. Sultani, and M. Ali, "Leveraging topology for domain adaptive road segmentation in satellite and aerial imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 206, pp. 106–117, Dec. 2023.

[21] C. Liang, B. Cheng, B. Xiao, and Y. Dong, "Unsupervised domain adaptation for remote sensing image segmentation based on adversarial learning and self-training," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[22] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609413.

[23] X. Luo, W. Chen, Z. Liang, L. Yang, S. Wang, and C. Li, "Crots: Cross-domain teacher–student learning for source-free domain adaptive semantic segmentation," *Int. J. Comput. Vis.*, vol. 132, no. 1, pp. 20–39, Jan. 2024.

[24] Z. Fang, J. Ren, J. Zheng, R. Chen, and H. Zhao, "Dual teacher: Improving the reliability of pseudo labels for semi-supervised oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5602515.

[25] A. Kirillov et al., "Segment anything," in *Proc. Int. Conf. Comput. Vision (ICCV)*, 2023, pp. 3992–4003.

[26] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," ITC, Univ. Twente, Enschede, The Netherlands, Tech. Rep., 2015, doi: 10.13140/2.1.5015.9683.

[27] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler, "Learning aerial image segmentation from online maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6054–6068, Nov. 2017.

[28] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[31] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021.

[32] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[33] G. Li, L. Li, H. Zhu, X. Liu, and L. Jiao, "Adaptive multiscale deep fusion residual network for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8506–8521, Jun. 2019.

[34] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.

[35] Y. Liu, S. Shi, J. Wang, and Y. Zhong, "Seeing beyond the patch: Scale-adaptive semantic segmentation of high-resolution remote sensing imagery based on reinforcement learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16822–16832.

[36] H. Wei et al., "Spatio-temporal feature fusion and guide aggregation network for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5642216.

[37] J. Geng, S. Song, and W. Jiang, "Dual-path feature aware network for remote sensing image semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3674–3686, May 2024.

[38] D. Wang, Y. Chen, B. Naz, L. Sun, and B. Li, "Spatial-aware transformer (SAT): Enhancing global modeling in transformer segmentation for remote sensing images," *Remote Sens.*, vol. 15, no. 14, p. 3607, Jul. 2023.

[39] X. Zhou, L. Zhou, S. Gong, S. Zhong, W. Yan, and Y. Huang, "Swin transformer embedding dual-stream for semantic segmentation of remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 175–189, 2024.

[40] Z. Dong, G. Gao, T. Liu, Y. Gu, and X. Zhang, "Distilling segmenters from CNNs and transformers for remote sensing images' semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613814.

[41] Z. Qi, H. Chen, C. Liu, Z. Shi, and Z. Zou, "Implicit ray transformers for multiview remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4703115.

[42] Z. Wang, Z. Liao, B. Zhou, G. Yu, and W. Luo, "SwinURNet: Hybrid transformer-CNN architecture for real-time unstructured road segmentation," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–16, 2024.

[43] Y. Chen, W. Li, X. Chen, and L. Van Gool, "Learning semantic segmentation from synthetic data: A geometrically guided input–output adaptation approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1841–1850.

[44] 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.

[45] S. Guo et al., "Label-free regional consistency for image-to-image translation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[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] 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.

[48] H. Wang, T. Shen, W. Zhang, L. Duan, and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2020, pp. 642–659.

[49] G. Zeng et al., "Entropy guided unsupervised domain adaptation for cross-center hip cartilage segmentation from MRI," in *Proc. Med. Image Comput. Comput. Assist. Intervent (MICCAI)*, Jan. 2020, pp. 447–456.

[50] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3763–3772.

[51] 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.

[52] L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9914–9925.

[53] Q. Bi, S. You, and T. Gevers, "Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 38, Mar. 2024, pp. 801–809.

[54] J. Yi et al., "Learning spectral-decomposited tokens for domain generalized semantic segmentation," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8159–8168.

[55] J. Ding, N. Xue, G.-S. Xia, B. Schiele, and D. Dai, "HGFormer: Hierarchical grouping transformer for domain generalized semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15413–15423.

[56] Q. Bi et al., "Learning frequency-adapted vision foundation model for domain generalized semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NeuralPS)*, 2024, pp. 94047–94072.

[57] B. Benjdira, Y. Bazi, A. Koubaa, and K. Ouni, "Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images," *Remote Sens.*, vol. 11, no. 11, p. 1369, 2019.

[58] Z. Chen et al., "Joint alignment of the distribution in input and feature space for cross-domain aerial image semantic segmentation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, Dec. 2022, Art. no. 103107.

[59] Y. Zhao, P. Guo, Z. Sun, X. Chen, and H. Gao, "ResiDualGAN: Resize-residual DualGAN for cross-domain remote sensing images semantic segmentation," *Remote Sens.*, vol. 15, no. 5, p. 1428, 2023.

[60] Y. Cai et al., "BiFDANet: Unsupervised bidirectional domain adaptation for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 14, no. 1, p. 190, Jan. 2022.

[61] D. Biswas and J. Tešić, "Unsupervised domain adaptation with debiased contrastive learning and support-set guided pseudolabeling for remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 3197–3210, 2024.

[62] H. Huang, B. Li, Y. Zhang, T. Chen, and B. Wang, "Joint distribution adaptive-alignment for cross-domain segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5401214.

[63] 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, 2023, Art. no. 5603518.

[64] X. Ma, X. Zhang, X. Ding, M.-O. Pun, and S. Ma, "Decomposition-based unsupervised domain adaptation for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5645118.

[65] OpenAI et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[66] DeepSeek-AI et al., "DeepSeek LLM: Scaling open-source language models with longtermism," 2024, *arXiv:2401.02954*.

[67] DeepSeek-AI et al., "DeepSeek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," 2025, *arXiv:2501.12948*.

[68] Z. Wang, Z. Yan, S. Li, and J. Liu, "IndVisSGG: VLM-based scene graph generation for industrial spatial intelligence," *Adv. Eng. Informat.*, vol. 65, May 2025, Art. no. 103107.

[69] X. Song, D. Han, C. Chen, X. Shen, and H. Wu, "Vman: Visual-modified attention network for multimodal paradigms," *Vis. Comput.*, vol. 41, no. 4, pp. 2737–2754, Mar. 2025.

[70] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Nov. 2024, pp. 38–55.

[71] B. Song, H. Yang, Y. Wu, P. Zhang, B. Wang, and G. Han, "A multispectral remote sensing crop segmentation method based on segment anything model using multistage adaptation fine-tuning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4408818.

[72] L. Mei et al., "SCD-SAM: Adapting segment anything model for semantic change detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5626713.

[73] X. Zhou et al., "MeSAM: Multiscale enhanced segment anything model for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5623515.

[74] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5636916.

[75] Z. Wang et al., "Exploring semantic prompts in the segment anything model for domain adaptation," *Remote Sens.*, vol. 16, no. 5, p. 758, Feb. 2024.

[76] L. Bai, S. Du, X. Zhang, H. Wang, B. Liu, and S. Ouyang, "Domain adaptation for remote sensing image semantic segmentation: An integrated approach of contrastive learning and adversarial learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628313.

[77] B. Zhang, T. Chen, and B. Wang, "Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5611412.

[78] X. Chen, S. Pan, and Y. Chong, "Unsupervised domain adaptation for remote sensing image semantic segmentation using region and category adaptive domain discriminator," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4412913.

[79] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[80] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 432–448.

[81] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[82] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[83] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[84] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12414–12424.

[85] H. Ni, Q. Liu, H. Guan, H. Tang, and J. Chanussot, "Category-level assignment for cross-domain semantic segmentation in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5608416.

[86] J. Zhu, Y. Guo, G. Sun, L. Hong, and J. Chen, "Causal prototype-inspired contrast adaptation for unsupervised domain adaptive semantic segmentation of high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5640017.

[87] MMSegmentation Contributors. (2020). *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[88] MMagic Contributors. (2023). *MMagic: OpenMMLab Multimodal Advanced, Generative, and Intelligent Creation Toolbox*. [Online]. Available: https://github.com/open-mmlab/mmagic

[89] T.-H. 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.*, Jul. 2019, pp. 2517–2526.

[90] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2507–2516.

[91] M. Chen, H. Xue, and D. Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2090–2099.

[92] G. Yang, M. Ding, and Y. Zhang, "Bi-directional class-wise adversaries for unsupervised domain adaptation," *Appl. Intell.*, vol. 52, no. 4, pp. 3623–3639, 2022.

[93] 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.

[94] 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.

[95] N. Araslanov and S. Roth, "Self-supervised augmentation consistency for adapting semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15384–15394.

[96] Y. Cheng, F. Wei, J. Bao, D. Chen, F. Wen, and W. Zhang, "Dual path learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9062–9071.

[97] J. E. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[98] Y. Cai, X. Sui, G. Gu, and Q. Chen, "Multi-modal interaction with token division strategy for RGB-T tracking," *Pattern Recognit.*, vol. 155, Nov. 2024, Art. no. 110626.

[99] Y. Zhang, Z. Gao, X. Wang, and Q. Liu, "Image representations of numerical simulations for training neural networks," *Comput. Model. Eng. Sci.*, vol. 134, no. 2, pp. 821–833, 2023.

[100] X. Shi, Y. Zhang, A. Pujahari, and S. K. Mishra, "When latent features meet side information: A preference relation based graph neural network for collaborative filtering," *Expert Syst. Appl.*, vol. 260, Jan. 2025, Art. no. 125423.

[101] D. Guan, Y. Xing, J. Huang, A. Xiao, A. El Saddik, and S. Lu, "S2Match: Self-paced sampling for data-limited semi-supervised learning," *Pattern Recognit.*, vol. 159, Mar. 2025, Art. no. 111121.

Shuchang Lyu (Member, IEEE) received the B.S. degree from the Department of Communication and Information, Shanghai University, Shanghai, China, in 2016, and the M.E. and Ph.D. degrees from the Department of Electronic and Information Engineering, Beihang University, Beijing, China, in 2019 and 2024, respectively.

He is currently a Post-Doctoral Researcher with the Department of Electronic and Information Engineering, Beihang University. His current research interests include domain adaptation semantic segmentation, remote sensing image understanding, and high-efficiency network design.

Qi Zhao (Member, IEEE) received the Ph.D. degree in communication and information systems from Beihang University, Beijing, China, in 2002.

She was with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA, as a Visiting Scholar from 2014 to 2015. She is a Professor with Beihang University. Since 2016, she has been working on wearable device-based first-view image processing and deep learning-based image recognition. Her research interests include one-shot semantic segmentation, communication signal processing, and target tracking.

Yaxuan Sun is currently pursuing the B.S. degree with the Department of Electronic and Information Engineering, Beihang University, Beijing, China.

Her research interests include remote sensing image processing, communication signal processing, and multimodality image fusion.

Guangliang Cheng received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently a Reader (Associate Professor) at the Department of Computer Science, University of Liverpool, Liverpool, U.K. Before that, he was an Associate Research Director at SenseTime, Beijing, from 2019 to 2023. He was a Post-Doctoral Researcher at the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, from 2017 to 2019. His research interests include computer vision, autonomous driving, and remote sensing image processing.

Yiwei He received the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2019.

He is currently a Post-Doctoral Research Associate with the Department of Computer Science, University of Liverpool, Liverpool, U.K. During his Ph.D., he focused on transfer learning, contributing to methodologies that enhance the adaptability of machine learning models across domains. Before his current position, he was with Intel, Santa Clara, CA, USA, and Mobileye, Beijing, where he developed advanced systems for computer vision and autonomous driving. He has contributed to several projects and actively promoted open-source methodologies within the community. His research interests include computer vision, multimodal models, and machine learning.

Guangbiao Wang received the B.S. degree from the Department of Optical Information Science and Technology, Qingdao University, Qingdao, Shandong, China, in 2010, and the M.E. degree from the Department of Electronic and Information Engineering, Beihang University, Beijing, China, in 2013, where he is currently pursuing the Ph.D. degree.

His research interests include multimodal learning and RS image processing.

Jinchang Ren (Senior Member, IEEE) received the Ph.D. degree in electronic imaging and media communication from the University of Bradford, Bradford, U.K., in 2009.

He is a Professor with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K., and also a Visiting Professor with Guangdong Polytechnic Normal University, Guangzhou, China. He has published more than 350 articles. His research interests include computer vision and multimedia signal processing, especially in hyperspectral imaging, machine learning, and big data analytics.

Zhenwei Shi (Senior Member, IEEE) is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing, China. He has authored or co-authored more than 200 scientific articles in refereed journals and proceedings. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Prof. Shi serves as an Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition*, *ISPRS Journal of Photogrammetry and Remote Sensing*, and *Infrared Physics and Technology*. His personal website is http://levir.buaa.edu.cn/