

Spatial Reconstruction and Joint Training in Transformer Network for Cross-Domain Remote Sensing Images Semantic Segmentation

Junying Zeng¹, Member, IEEE, Senyao Deng¹, Yikui Zhai², Senior Member, IEEE, Xudong Jia¹, Chuanbo Qin¹, Pasquale Coscia³, Angelo Genovese³, Senior Member, IEEE, Vincenzo Piuri³, Fellow, IEEE, Fabio Scotti³, Senior Member, IEEE

Abstract—Recently, Unsupervised Domain Adaptation (UDA) methods have attracted considerable attention in Remote Sensing Images (RSI) semantic segmentation. However, cross-domain RSI exhibit diverse scales, imbalanced distributions within domains, and significant inter-domain variations. In response to these challenges, we combine Spatial reconstruction and Joint training with the Transformer Network (SJT-Net). This framework introduces a spatial reconstruction method to address the issue of inconsistent ground sampling distances in cross domain RSI, which is rarely considered in existing approaches. Transferring domain knowledge at a similar spatial scale improves the spatial representation ability of UDA models. Unlike traditional adversarial training using ResNet for feature extraction, the SJT-Net employs Segformer, which enhances the model’s ability to capture in-class features across domains and improves global dependency modeling. Transmitting these refined features to the discriminator allows for more precise feature-level domain alignment. To enhance feature decoding, an interactive global-local decoder is constructed to efficiently capture both global relationships and local details of landform objects. Our framework leverages adversarial training to generate highly confident model weights and pseudo-labels for self-training in the target domain. Through iterative updates, the model’s generalization capability is gradually improved, eventually achieving optimal segmentation performance. Experimental results demonstrate that SJT-Net outperforms current UDA approaches and accomplishes state-of-the-art (SOTA) segmentation accuracy. The repository can be accessed at <https://github.com/AnsonD0820/SJT-Net>.

Index Terms—unsupervised domain adaptation, remote sensing images, semantic segmentation, joint training, Transformer.

I. INTRODUCTION

OWING to the swift promotion of remote sensing technology, the remote sensing big data era has arrived with its increasing availability and utility [1]. High-definition Remote Sensing Images (RSI) are critical in various disciplines, including landform mapping [2, 3], city layout and construction [4, 5], environmental monitoring [6, 7] and crop assessment [8, 9], etc. Recently, thanks to rapid advancements in computer vision technology, semantic segmentation algorithms have become essential tools for interpreting remote sensing images [10–12]. However, densely annotated training samples are a major determinant of current RSI semantic segmentation techniques. They not only consume considerable manpower and material resources but also necessitate the use of specialized software for annotating high-resolution RSI. How to use existing annotated remote sensing data for training and apply trained models to newly acquired data from diverse time periods and geographical regions, has emerged as a pivotal focus in global remote sensing research. Transferring domain knowledge can effectively address this challenge. However, domain shifts frequently arise due to variations in feature distribution between domains. Performing domain alignment is a common optimization technique in UDA. It involves the use of a specific optimization paradigm, and the adoption of certain training methodologies, like adversarial training [13–16] or alternatively self-training [17, 18].

The basic idea of adversarial training involves the simultaneous training of a feature extractor alongside a discriminator. The former endeavors to confound the latter by generating samples that are hard to distinguish into a specific domain. When the discriminator struggles to ascertain domain origins of these samples, it indicates substantial similarity across both domains. Many successful approaches in UDA adversarial training have been put forth and applied to cross-domain semantic segmentation tasks of natural images. For instance, Tsai et al. [13] put forward an adversarial network with multiple levels and Wang et al. [14] suggested a fine-grained adversarial methodology for learning. They both showed good segmentation results on cross domain street scenes. But remote sensing scenes have distinct attributes than natural scenes.

Manuscript received 21 February 2025; revised 13 May 2025 and 3 July 2025; accepted 11 August 2025. This paper was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011576, in part by Guangdong, Hong Kong, Macao and the Greater Bay Area International Science and Technology Innovation Cooperation Project under Grant 2021A050530080, in part by Key Research Projects for Universities of Guangdong Provincial Education Department under Grant 2020ZDZX3031 and in part by Special Project in key Areas of Artificial Intelligence in Guangdong Universities under Grant 2019KZDZX1017. (Junying Zeng, Senyao Deng, Yikui Zhai and Xudong Jia contributed equally to this work.) (Corresponding author: Yikui Zhai; Xudong Jia.)

Junying Zeng, Senyao Deng, Yikui Zhai, Chuanbo Qin are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, Guangdong 529020, China (e-mail: zengjunying@126.com; AnsonD0820@163.com; yikuizhai@163.com; tenround@163.com)

Xudong Jia is with the School of Engineering, Science, Technology, Central Connecticut State University, New Britain, CT 06050, USA. Also, as a visiting scholar, he is with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, Guangdong, 529020, China (e-mail: xjia@ccsu.edu)

Pasquale Coscia, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti are with the Department of Computer Science and the Dipartimento di Informatica, Università Degli Studi di Milano, 20133 Milano, Italy (e-mail: pasquale.coscia@unimi.it; angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it)

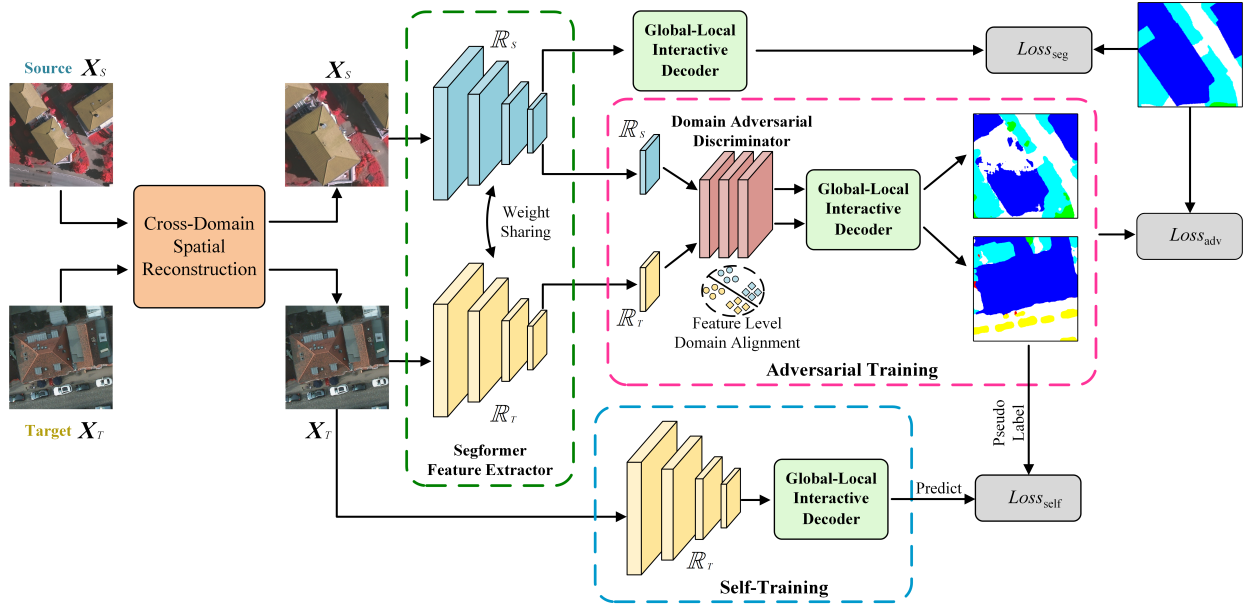


Fig. 1. SJT-Net framework. The optimizer of SJT-Net comprises three losses: segmentation loss in the source domain (L_{seg}), domain adversarial loss (L_{adv}) and self-training segmentation loss in the target domain (L_{self}). Throughout the training process, only annotated labels from the source domain guide the training, adhering to the principle that supervised information from the target domain is unavailable in UDA experiments. X_S and X_T represent the images of the source and target domains, respectively, while $\mathbb{R}_S$ and $\mathbb{R}_T$ represent their corresponding features.

These images exhibit significant variability in feature distribution, which can be attributed to differences in region, season, sensor type, and the altitude at which they were captured. Meanwhile, they encompass complex terrain structures and intricate spatial relationships. Moreover, existing adversarial UDA methods have a shortcoming that cannot be ignored, i.e., the limited receptive field in the segmentation network. The majority of adversarial frameworks predominantly employ the network architecture from the DeepLab series [19–21], using the ResNet101 [22] for feature extraction. However, its convolutional kernel receptive field is relatively limited to acquire comprehensive global contextual information, which makes it incapable of processing complex RSI. In contrast, the Transformer architecture [23–27], leveraging multi-head self-attention, proves its proficiency in capturing detailed information across diverse scales. Moreover, it adeptly comprehends the spatial and semantic relationships across diverse locations within RSI, facilitated by positional encoding.

Self-training, as a non-adversarial UDA method, has a huge difference in its training mechanism when compared with adversarial training. Using annotated source domain samples, the self-training strategy first trains an initial model to predict target domain’s unlabeled samples. To further enhance the target domain’s segmentation accuracy, these high-assurance predictions are used as pseudo-supervisions to iteratively update the model weights. DAFormer [17] integrates the Transformer with a self-training framework to perform cross-domain street scenes semantic segmentation. HRDA [18] enhances DAFormer by a multi-resolution self-training approach, and a method for integrating cropping and splicing data enhancement method for pseudo-labels to enhance segmentation effectiveness. Nevertheless, self-training exists an inherent shortcoming: segmentation performance is heavily contingent upon the generalizability of its original

model and the precision of pseudo-labels. Subpar initial model performance or notable disparities in sample distributions can result in low-quality pseudo-labels. In subsequent iterative training processes, erroneous pseudo-supervisory information accumulates continuously, ultimately resulting in segmentation outcomes that do not achieve the necessary accuracy in practical applications.

To address the limitations of two UDA training strategies, the key contributions are as follows.

- 1) This research suggests a framework (SJT-Net) that integrates adversarial training and self-training. It utilizes the Transformer to capture global context and long-range dependencies and improves cross-domain RSI segmentation accuracy by employing adversarial training to generate reliable pseudo-labels for self-training.
- 2) A cross-domain spatial reconstruction method is proposed to alleviate RSI resolution gaps and align spatial scales, thereby facilitating cross-domain feature alignment.
- 3) To address small-target mis-segmentation and scale variations, a global-local interactive decoder is proposed which expands the decoding-stage receptive field and enables synergistic integration of local and global features. Comparative experimental results show that SJT-Net achieves SOTA segmentation performance.

II. RELATED WORK

A. Semantic Segmentation with Convolutional Neural Network (CNN) for Remote Sensing Images

Semantic segmentation makes predictions at the pixel level, which is crucial for applications involving RSI interpretation. Since the Fully Convolutional Network (FCN) [28] was introduced, end-to-end image semantic segmentation has been employed extensively. Many well-known convolutional neural

networks have been extensively utilized as benchmarks for semantic segmentation tasks. Compared to natural images, objects in RSI contain complex textures, irregular shapes, and diverse scales. Researchers overcome these challenges by incorporating attention modules. Zheng et al. [29] proposed a multiscale channel-wise cross attention network to facilitate the fusion of multi-channel features using an attention mechanism. By employing an attention perception module with synergistic integration and hierarchical structure, Li et al. [30] effectively mitigated visual attention bias. Zhang et al. [31] used truncated local regions to learn from related-category information and improved the recognition of global contextual information through an attention cross-fusion module.

B. Semantic Segmentation with Transformer for Remote Sensing Images

Following the significant success of Vision Transformer (ViT) [23] in image classification tasks, SETR [24] introduced the first Transformer-based semantic segmentation framework by integrating ViT with an encoder-decoder architecture. Notably, the computational cost of self-attention in ViT increases significantly in processing large images. To address this issue, Xie et al. [27] introduced a more efficient hierarchical Transformer architecture, Segformer (MiT), specifically designed for image segmentation tasks. Building on this foundation, several studies have found that the Transformer architecture exhibits notable segmentation effects when processing high-resolution RSI. Zhang et al. [32] improved the ViT by effectively mining second-order information from RSI to establish global dependencies and guide its fusion with first-order information. Zeng et al. [33] employed cross-attention to enhance multi-scale feature connections in the encoder and developed a Transformer-based decoder that facilitates cross-window interactions. Zhang et al. [34] proposed the Long-Short-Range Transformer (LSRFormer), which can be incorporated after each CNN stage to enhance the model's ability to learn both local and global representations across different stages. Although Transformer-based models exhibit outstanding performance, they remain highly dependent on pixel-wise annotated remote sensing data, which is scarce. Therefore, effectively leveraging existing annotated data for transfer learning has emerged as a primary research focus.

C. Unsupervised Domain Adaptation (UDA) Semantic Segmentation Approach for Remote Sensing Images

Current UDA strives to minimize domain differences but frequently faces category distribution mismatches between target and source domains, commonly referred to as domain shift. Most UDA techniques for remote sensing scenes are built on adversarial training or self-training. Adversarial training approach attempts to conform feature distributions of source and target domains inside the Generative Adversarial Network (GAN) architecture. Based on the encoding-decoding framework, adversarial training approach can be divided into three UDA levels: input [35, 36], feature [37, 38], and output [39, 40]. Specifically, Ma et al. [38] proposed a decomposition-based generative adversarial network that decomposes encoder-generated features into high/low-frequency

components and aligns them within feature-level subspace. Similarly, self-training [41, 42], as a non-adversarial UDA approach, has garnered significant interest in semantic segmentation involving cross-domain RSI. For instance, Zhang et al. [42] proposed a self-training approach aimed at enhancing quality of pseudo-labels by leveraging a teacher model within a topological space. In addition, Liang et al. [43] combines the above two UDA approaches, further enhancing the models' generalization ability. However, most of these approaches rely on the DeepLab network, which is somewhat limited, when compared to the superior global modeling capabilities of Transformer. Furthermore, they often overlook the spatial scale discrepancies arising from differences in Ground Sample Distance (GSD) between the source and target domains.

III. METHODOLOGY

A. SJT-Net Framework

To lessen the detrimental effects of domain shift resulting from distribution imbalance across domains, this research proposes the SJT-Net. Fig. 1 depicts the overall SJT-Net framework. By aligning spatial category information with target domain, our framework performs cross-domain spatial reconstruction on source domain. To capture comprehensive contextual information during the adversarial training phase, the framework needs to derive characteristics from images in both source and target domains. Most existing UDA semantic segmentation methods rely on ResNet101 [22] for feature extraction. However, as a fully convolutional network, ResNet101 has a relatively smaller receptive field compared to Transformer-based models. The local convolutional sliding operation limits ResNet101's ability to capture long-range dependencies. Consequently, when applied to cross-domain semantic segmentation tasks, particularly those involving complex spatial relationships, this constraint further hinders the model's performance. This limits the development of cross domain RSI semantic segmentation. For overcoming the aforementioned limitation, our work develops a feature extractor built upon Segformer (MiT) [27] architecture for obtaining features from RSI in two domains, as depicted in Fig. 2. The Transformer model, initially introduced by Vaswani et al. [44], leverages the self-attention mechanism to capture the relationships between elements in an input sequence. Unlike traditional Convolutional Neural Networks (CNN), which rely on fixed local receptive fields, the Transformer calculates the relationships between each element and all others using self-attention. This enables the model to effectively capture long-range dependencies. Segformer (MiT) is an image segmentation model inspired by the Vision Transformer (ViT) [23], which has demonstrated impressive performance in image classification tasks. Building on this, Segformer improves the encoder by employing a lightweight encoding structure, thereby enhancing its performance in segmentation tasks. Compared to the Segformer encoder, the improved Segformer feature extractor in this research incorporates two sets of convolutional layers, normalization operations, and nonlinear activation functions, following the aggregation of the MLP (Multilayer Perceptron) layer. This design aims to

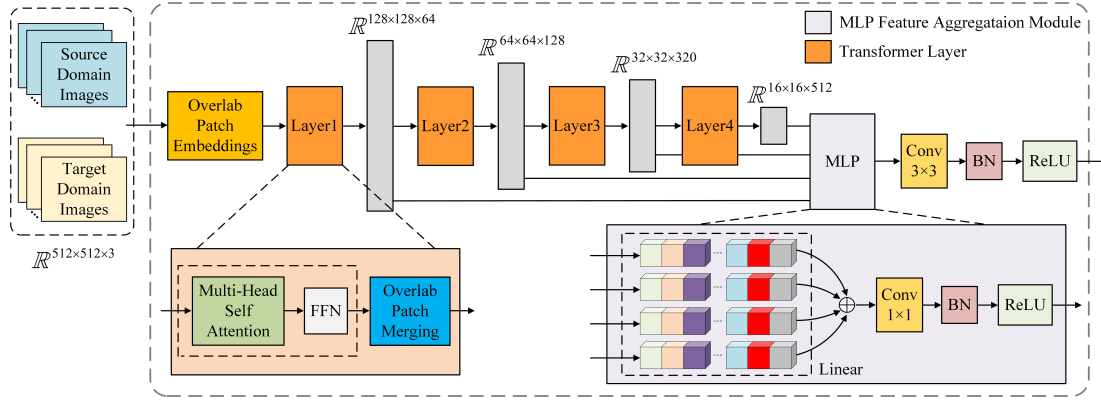


Fig. 2. Structure of Segformer Feature Extractor. The Segformer feature extractor comprises four Transformer layers. Each layer has a different number of multi-head self-attention heads: [1, 2, 5, 8]. The feed-forward network within a Transformer layer is denoted as FFN (Feed Forward Network). All linear flattening in this framework is achieved using fully connected layers.

match its output dimensions with those of the proposed global-local interactive decoder and domain discriminator. The final output feature dimensions of the improved Segformer Feature Extractor is $\mathbb{R}^{64 \times 64 \times 2048}$.

The Segformer Feature Extractor accepts input from the target and source domains with dimensions of $\mathbb{R}^{512 \times 512 \times 3}$. The Transformer architecture encounters challenges in processing large-size images directly. Therefore, a preprocessing step is required to partition the images into smaller patches before feeding them into the Transformer layer. Fig. 3 shows the detailed process of image patching and describes how self-attention operates with N heads. Where an input image is divided into smaller, overlapping patches, each patch is treated as an individual token. These patches are then flattened and linearly embedded into a high-dimensional feature space. The self-attention mechanism is subsequently applied across N heads, allowing the model to capture long-range dependencies and contextual relationships between patches. Each head in the multi-head self-attention mechanism independently attends to distinct input segments. The outputs are then concatenated and linearly transformed to generate the final representation, allowing the Segformer Feature Extractor to capture both local and global features. Each patch processed by a Transformer layer has the sizes $[128 \times 128, 64 \times 64, 32 \times 32, 16 \times 16]$ respectively. After each Transformer layer, the number of output channels is $[64, 128, 320, 512]$. To improve the representation capacity of semantic information, the Feed Forward Network (FFN) within each layer performs nonlinear transformations on the extracted features. FFN's mapping expression is as follows:

$$R_{out} = \text{MLP}(\text{GELU}(\text{Conv}_{3 \times 3}(\text{MLP}(R_{in})))) + R_{in} \quad (1)$$

where, R_{out} represents the features of the FFN output and R_{in} denotes the feature map extracted from the multi-head self-attention module. MLP consists of a series of fully connected layers that perform feature transformation on the input R_{in} . $\text{Conv}_{3 \times 3}$ refers to a convolution operation using a 3×3 kernel. As the multi-head self-attention is a global module, it is limited in its ability to capture local spatial information. Convolution operations, which captures local features, complement this limitation by conveying spatial structural information. GELU (Gaussian Error Linear Unit) is an activation function em-

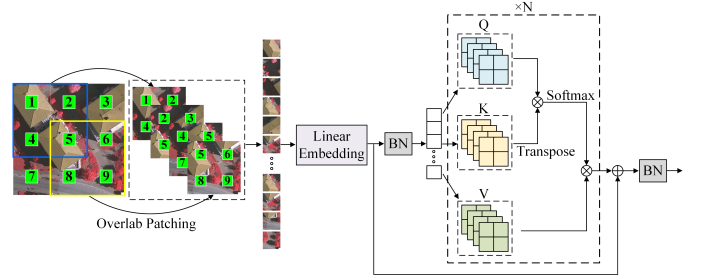


Fig. 3. Operation flow chart of overlap patching and multi-head self-attention mechanism.

ployed to apply nonlinear transformations to the convolutional outputs, thereby significantly enhancing the model's power in predictions. The second-layer MLP maps the nonlinear features to a lower-dimensional space, recovering more efficient feature representations from the high-dimensional space. Finally, the residual connection of R_{in} ensures the continuity of information flow, facilitating the optimization of the network. In the FFN, the dimensions of R_{out} and R_{in} are identical. Within each Transformer layer, the dimensions of R_{out} and R_{in} are $[\mathbb{R}^{16384 \times 64}, \mathbb{R}^{4096 \times 128}, \mathbb{R}^{1024 \times 320}, \mathbb{R}^{256 \times 512}]$.

In comparison to the ResNet101, the Segformer Feature Extractor in the proposed SJT-Net has a wider receptive field and is capable of extracting more refined features during the feature extraction on-stage. In the SegFormer Feature Extractor, the source and target domains share identical weights, facilitating effective transfer learning on the target domain and thereby minimizing the distributional gap between the two domains. The discriminator is then simultaneously treated and final layer feature maps from two domains are aligned. By engaging in adversarial training, the target domain's feature distribution probabilistically resembles that of the source one. To enhance the receptive field during decoding, a global-local interactive decoder is designed to upsample features after adversarial training, generating target domain segmentation masks. These masks serve as the target domain's pseudo-labels in self-training. Through iterative updates of model weights and pseudo-labels, the model achieves optimal segmentation performance. By combining adversarial training and self-training, we adopt a dual domain adaptation semantic segmen-

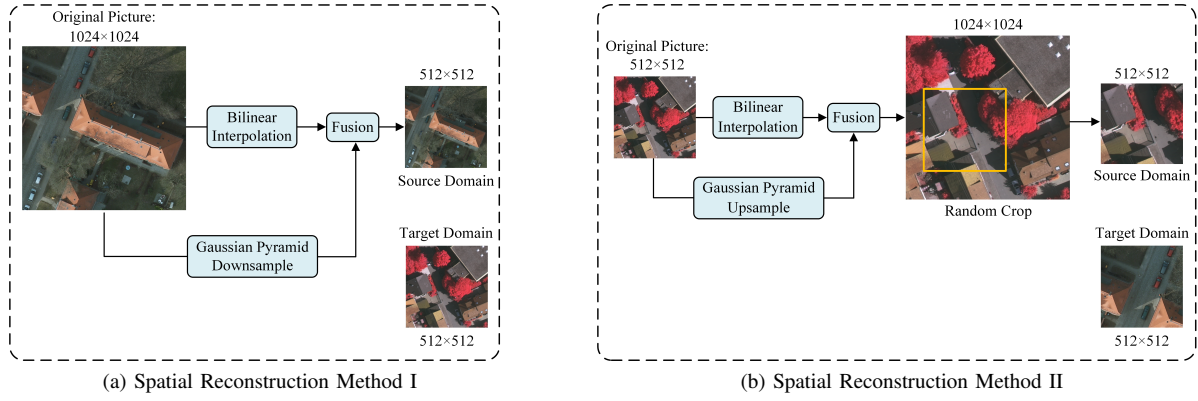


Fig. 4. (a) represents the spatial reconstruction when PotsdamRGB (or PotsdamIRRG) serves as the source domain and VaihingenIRRG serves as the target domain; (b) represents the spatial reconstruction when VaihingenIRRG serves as the source domain and PotsdamRGB (or FloodNetRGB) is used as the target domain. The reasons why they are set to 1024×1024 and 512×512 in Figure 2(a) and (b) are to follow formula (1), to consider the properties of Gaussian pyramid, and to satisfy the relationships of integer multiples during feature extraction downsampling. (Note: The datasets PotsdamRGB, PotsdamIRRG, VaihingenIRRG and FloodNetRGB are introduced in subsequent Section IV.)

tation strategy of “mainly adversarial training, supplemented by self-training”, which effectively reduces domain differences and improves the generalization performance of cross-domain segmentation in remote sensing scenes.

B. Cross-Domain Spatial Reconstruction

Previous works on UDA semantic segmentation technique for remote sensing scenario often overlooked the spatial resolution disparities between domains, which are usually measured in terms of Ground Sample Distance (GSD). Remote sensing data are commonly collected from various geographical regions and using diverse acquisition devices. Variations in shooting heights among acquisition devices result in distinct spatial distribution characteristics across datasets. Differences in GSD across datasets often result in an uneven distribution of categorical information between source and target domains, even when using cropped images of identical size (e.g., 512×512). Specifically, one domain may contain more semantic information than the other. The disparities in spatial distribution among the source and target domains might hinder feature extractor’s ability to adapt features from the former to the latter, consequently impacting the target domain’s segmentation effectiveness. To mitigate the adverse effects stemming from inconsistent spatial resolutions among cross-domain RSI, this research proposes a cross-domain spatial reconstruction method suitable for UDA RSI semantic segmentation. The proposed spatial reconstruction method employs the ratio of GSD between source and target domains as a reconstruction factor and integrates two image resampling techniques to perform spatial scale reconstruction. By combining bilinear interpolation [45, 46] and Gaussian pyramid sampling [47, 48], it effectively mitigates the degradation of spatial information during image reconstruction. Bilinear interpolation is a widely used technique in 2D image processing. It estimates the value of a target pixel by first performing linear interpolation along two horizontal segments to obtain intermediate values, followed by vertical interpolation of these values. Gaussian pyramid downsampling reduces image resolution through Gaussian blur at each level, creating a multi-scale hierarchical structure. Upsampling, an inverse process, reconstructs high-resolution images from low-resolution ones.

Let $\mathbf{X}_S \in \mathbb{R}^{H_S \times W_S \times C}$ denote the source domain images, $\mathbf{Y}_S \in \mathbb{R}^{H_S \times W_S \times K}$ denote the source domain labels and $\mathbf{X}_T \in \mathbb{R}^{H_T \times W_T \times C}$ denote the target domain images. Since supervised information is not employed by the target domain during the training period, $\mathbf{Y}_T$ is undefined. C indicates image’s channel quantity, while K indicates label’s category count. In this research, $C = 3$ and $K = 6$ (or $K = 5$). Let d_S represent the source domain’s GSD, and d_T represent the target one. The used cross-domain dataset just contains two types of GSD details. When $d_S = 5$ cm and $d_T = 9$ cm, the spatial reconstruction method shown in Fig. 4(a) is employed, and the reconstruction factor is 0.5. When $d_S = 9$ cm and $d_T = 5$ cm, as well as when $d_S = 9$ cm and $d_T = 3.75$ cm, the spatial reconstruction method shown in Fig. 4(b) is employed, and the reconstruction factor is 2. The image size and GSD generally adhere to the following:

$$\frac{d_S}{d_T} \approx \frac{H_S}{H_T} = \frac{W_S}{W_T} \quad (2)$$

The ultimate goal of the UDA approach is to boost the target domain’s segmentation accuracy. Therefore, spatial reconstruction is applied exclusively to the source domain images within SJT-Net, with no preprocessing performed on the target domain. After the spatial reconstruction shown in Fig. 4(a) or (b), the source along with target domain images exhibit consistent sizes and closely matched spatial resolutions. This suggests that every reconstructed image in the source domain, with dimensions of 512 x 512 pixels, contains category information that closely matches with the target domain.

C. Global-Local Interactive Decoder

The objects in RSI typically exhibit complex structures, textures, and shapes. To tackle segmentation challenges such as confusions with small target categories, ambiguous semantic boundaries, and significant variations in object scales within remote sensing scenes, we present a global-local interactive decoder, as depicted in Fig. 5(a). Its goal is to decode the feature maps derived from adversarial training and self-training processes, thereby producing prediction maps with optimal segmentation performance. At the beginning of this decoder, the feature maps are partitioned into global and local

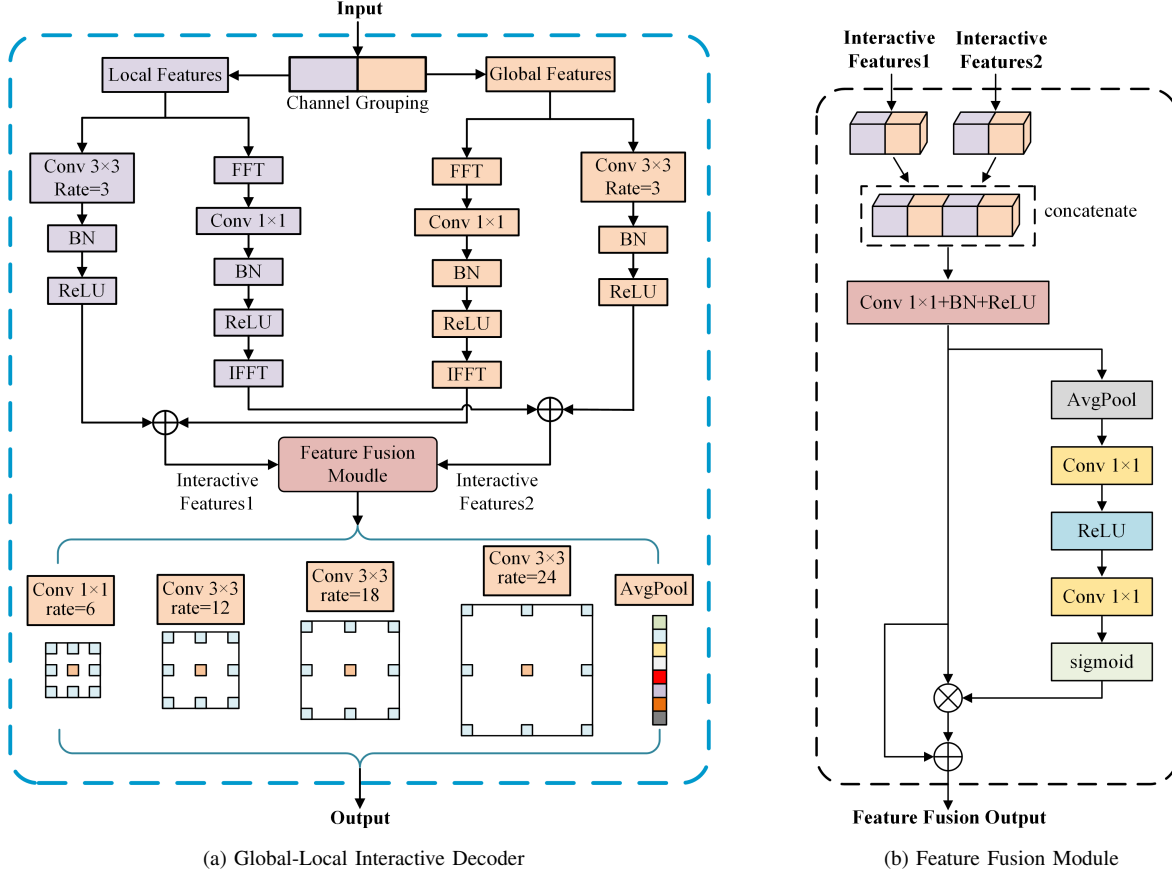


Fig. 5. (a) illustrates that the overall structure of the Global-Local Interactive Decoder, while (b) depicts the Feature Fusion Module within it. The FFT in the decoder signifies the Fast Fourier Transform, and the IFFT signifies the Inverse Fast Fourier Transform.

features. Initially, the input features $\mathbb{R}^{64 \times 64 \times 2048}$ are grouped along the channel dimension. Subsequently, global pooling operations and local feature processing are independently applied to each subset. Additionally, the global subset $\mathbf{X}_{\text{global}} \in \mathbb{R}^{64 \times 64 \times 1024}$ is merged to form a comprehensive feature map, followed by the same procedure for the local subset $\mathbf{X}_{\text{local}} \in \mathbb{R}^{64 \times 64 \times 1024}$, thereby yielding global and local feature maps $\mathbf{X} = \{\mathbf{X}_{\text{global}}, \mathbf{X}_{\text{local}}\}$ across varying scales. Chi et al. [49] demonstrated that the Fast Fourier Transform (FFT) can expand the receptive field of conventional convolution and facilitate cross-scale feature fusion across different branches. To acquire both global and local information, we design two convolutional paths, each with a distinct receptive field. One path employs a dilation rate of 3 for atrous convolution, while the other utilizes FFT convolution. In the FFT convolution path, the feature map is first transformed from the time domain to the frequency domain using the FFT, enabling the decoder to simultaneously process both high-frequency and low-frequency feature information at a global range. This process allows the convolution to overcome the constraints of the local receptive field, effectively integrating both global and local information through frequency domain processing. After the FFT, the real part feature R_{Re} and the imaginary part feature R_{Im} are obtained, each with dimensions of $\mathbb{R}^{64 \times 32 \times 1024}$. Subsequently, point-wise multiplication, normalization, and ReLU activation are applied to the real and imaginary components. Finally, the real and imaginary components are reintegrated

into time-domain features through the Inverse Fast Fourier Transform (IFFT), allowing this branch to effectively preserve global context information while simultaneously capturing local details.

Each path captures information from diverse receptive fields, enhancing the representation of a broader semantic relationship. By cross-aggregating features from both the atrous convolution path and the frequency domain transformation path, the decoder gains enhanced multi-scale interaction capabilities. Furthermore, we utilize a Feature Fusion Module [50], as depicted in Fig. 5(b), that effectively integrates feature maps from various branches, resulting in the fused features $\mathbb{R}^{64 \times 64 \times 2048}$. Introducing attention module facilitates the automatic learning of correlations among features across various scales, guiding the model's focus towards regions rich in semantic information. By thoroughly leveraging complementary information from multiple paths, this process generates more comprehensive and accurate segmentation masks in subsequent stages. Finally we introduce the Atrous Spatial Pyramid Pooling module (ASPP) [19–21] for feature decoding and enhances its dilation rates to [6, 12, 18, 24]. The purpose is to employ a wider spatial pyramid structure to enhance the preservation of semantic information and prevent the loss of multi-scale details during feature decoding, thereby enhancing the accuracy of semantic segmentation and ultimately yielding a precise segmentation mask $\mathbb{R}^{64 \times 64 \times K}$, where, the channel K represents the number of categories.

D. Loss Function

The SJT-Net is optimized via adversarial training, including the segmentation loss $L_{\text{seg}}(\mathbf{X}_S)$ of the source domain and the adversarial loss $L_{\text{adv}}(\mathbf{X}_S, \mathbf{X}_T)$. Additionally, self-training is employed, focusing on the segmentation loss $L_{\text{self}}(\mathbf{X}_T)$ of the target domain. Using the cross-entropy, the segmentation loss $L_{\text{seg}}(\mathbf{X}_S)$ of the source domain is as follows:

$$L_{\text{seg}}(\mathbf{X}_S) = - \sum_{n=1}^{N_S} \sum_{k=1}^K \mathbf{Y}_S^{(n,k)} \log \mathbf{F}(\mathbf{X}_S^{(n,k)}) \quad (3)$$

where, in the source domain, $\mathbf{X}_S^{(n,k)}$ signifies the k -th category of the n -th image, while $\mathbf{Y}_S^{(n,k)}$ signifies the k -th category of the n -th label. There exist a total of N_S images and K categories. $\mathbf{F}(\cdot)$ signifies the mapping function of the segmentation backbone architecture.

Based on minimax optimization strategy, the feature extractor's adversarial loss is as follows:

$$L_{\text{adv}_G} = \min_{\mathbf{G}} \max_{\mathbf{D}} \{-\mathbf{E}_{\mathbf{X}_T \sim \mathbf{P}(\mathbf{X}_T)} [\log \mathbf{D}(\mathbf{G}(\mathbf{X}_T))]\} \quad (4)$$

The discriminator's adversarial loss is as follows:

$$L_{\text{adv}_D} = \min_{\mathbf{G}} \max_{\mathbf{D}} \{\mathbf{E}_{\mathbf{X}_S \sim \mathbf{P}(\mathbf{X}_S)} [\log \mathbf{D}(\mathbf{X}_S)] + \mathbf{E}_{\mathbf{X}_T \sim \mathbf{P}(\mathbf{X}_T)} [\log (1 - \mathbf{D}(\mathbf{G}(\mathbf{X}_T)))]\} \quad (5)$$

The feature extractor and discriminator collectively make up the adversarial loss $L_{\text{adv}}(\mathbf{X}_S, \mathbf{X}_T)$, as follows:

$$L_{\text{adv}}(\mathbf{X}_S, \mathbf{X}_T) = \lambda_1 L_{\text{adv}_G} + \lambda_2 L_{\text{adv}_D} \quad (6)$$

where λ_1 and λ_2 are set to 0.5 and 0.001, correspondingly. $\mathbf{G}(\cdot)$ represents the generator, and $\mathbf{D}(\cdot)$ represents the discriminator. $\mathbf{X}_S$ signifies the source domain's images, while $\mathbf{X}_T$ refers to the target one.

The first stage's adversarial training L_{stage1} is as follows:

$$L_{\text{stage1}} = L_{\text{seg}}(\mathbf{X}_S) + L_{\text{adv}}(\mathbf{X}_S, \mathbf{X}_T) \quad (7)$$

The cross-entropy is used as the target domain's self-training loss $L_{\text{self}}(\mathbf{X}_T)$ for the second stage, as follows:

$$L_{\text{stage2}} = L_{\text{self}}(\mathbf{X}_T) = - \sum_{n=1}^{N_T} \sum_{k=1}^K \mathbf{Y}_T'^{(n,k)} \log \mathbf{F}(\mathbf{X}_T^{(n,k)}) \quad (8)$$

where, in the target domain, $\mathbf{X}_T^{(n,k)}$ signifies the k -th category of the n -th image, while $\mathbf{Y}_T'^{(n,k)}$ signifies the k -th category of the n -th pseudo-label. There exist a total of N_T images and K categories. $\mathbf{F}(\cdot)$ signifies the mapping function of the segmentation backbone architecture.

To outline the core optimization steps of the proposed SJT-Net model, we present a concise pseudocode that illustrates the entire training procedure, as shown in Algorithm 1.

IV. EXPERIMENTS

In this section our experimental details, encompassing the remote sensing dataset descriptions, model evaluation metrics, as well as training parameter configurations are described. Subsequently, by comparing with existing UDA semantic segmentation methods, we demonstrate the effectiveness of SJT-Net. Ultimately, we undertake ablation analysis to further explore the contribution of every component within SJT-Net to the improvement of model segmentation accuracy.

Algorithm 1 The training process of SJT-Net.

Inputs: Source domain: Images $\mathbf{X}_S \in \mathbb{R}^{H_S \times W_S \times 3}$, Labels $\mathbf{Y}_S \in \mathbb{R}^{H_S \times W_S \times K}$; Target domain: Images $\mathbf{X}_T \in \mathbb{R}^{H_T \times W_T \times 3}$ (unlabeled)

Outputs: Domain adaptation model and prediction results $\hat{\mathbf{Y}}_T \in \mathbb{R}^{H \times W \times K}$ for target domain

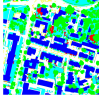
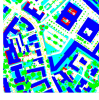
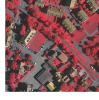
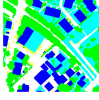
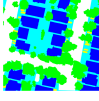
- 1: Cross-Domain Spatial Reconstruction: Resize $\mathbf{X}_S$ and $\mathbf{X}_T$ to a shared resolution (H, W) to align semantic content. $\mathbf{X}_S, \mathbf{X}_T \in \mathbb{R}^{H \times W \times 3}$
- 2: Source Domain Pretraining:
 - SegFormer Feature Extractor (**G**): $\mathbf{X}_S \rightarrow \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$
 - Global-Local Interactive Decoder as segmentation head (**H**): $\mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C} \rightarrow \mathbb{R}^{H \times W \times K_{cls}}$
 - Optimize **G**, **H** via cross-entropy loss: L_{seg}
- 3: Adversarial Training for Domain Alignment:
 - SegFormer Feature Extractor (**G**): $\mathbf{X}_S \rightarrow \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$, $\mathbf{X}_T \rightarrow \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$
 - Update Discriminator (**D**): L_{adv_D}
 - Update SegFormer Feature Extractor (**G**): L_{adv_G}
 - Joint Optimization: $L_{\text{adv}} = \lambda_1 L_{\text{adv}_G} + \lambda_2 L_{\text{adv}_D}$ (λ_1, λ_2 is weight)
- 4: Self-Training on Target Domain:
 - Generate pseudo-labels $\hat{\mathbf{Y}}_T$ for $\mathbf{X}_T$ using adversarial-trained model: $\mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C} \rightarrow \mathbb{R}^{H \times W \times K}$
 - Optimize **G**, **H** via cross-entropy loss: L_{self}
 - Repeat refine $\hat{\mathbf{Y}}_T$ iteratively

A. Dataset Details

To assess the validity of SJT-Net, this research initially conducted experiments using the RSI from the International Society for Photogrammetry and Remote Sensing (ISPRS) WG II/4: Potsdam 2D and Vaihingen 2D datasets [51]. Semantic labels are as follows: Background Clutter (Back. Cul.), Impervious Surface (Imp. Sur.), Car, Tree, Low Vegetation (Low. Veg.), and Building (Build.). As previously mentioned, RSI originate from various regions, and the variations in GSD during data collection across different research teams, which present challenges for cross-domain semantic segmentation. To assess the generalization capability of SJT-Net on RSI collected by other teams, we further conduct experiments on the FloodNet dataset [52]. The semantic categories in the FloodNet dataset are redefined to align with those of the ISPRS dataset, retaining the primary categories and classifying unrelated ones as background. The redefined categories are as follows: Background (Back.), Car, Tree, Grass, and Building (Build.).

ISPRS Potsdam 2D encompasses 38 orthoimages, each with a fixed 6000×6000 pixels and a spatial resolution of 5 cm. As depicted in Table I, the Potsdam dataset contains two imaging modality data, with spectral bands of three-channel IRRG and RGB. When Potsdam_IRRG (or Potsdam_RGB) serves as the source domain, it is cropped to 1024×1024 pixel size with a 512-pixel cropping stride, separated into about 2000 training images and 1000 test images. Conversely, when Potsdam_IRRG (or Potsdam_RGB) is utilized as the target

TABLE I
DETAILS OF THE CROSS-DOMAIN DATASET IN THE EXPERIMENT

Dataset	Size	Spatial Resolution	Band	Amount	Image	Label	Category
Potsdam	6000 × 6000	5 cm	IRRG	38			ISPRS  Back. Cul.  Imp. Sur.  Car  Tree  Low Veg.  Build.
			RGB	38			
Vaihingen	2000 × 2500	9 cm	IRRG	33			FloodNet  Back.  Car  Tree  Grass  Build.
FloodNet	4600 × 3000	3.75 cm	RGB	284			

domain, it is cropped to 512×512 pixel size with a 512-pixel cropping stride, separated into about 3000 training images and 1300 test images.

ISPRS Vaihingen 2D encompasses 33 orthoimages, each with an average resolution of about 2000×2500 pixels and a spatial resolution of 9 cm. As depicted in Table I, the Vaihingen dataset contains spectral bands of three-channel IRRG. Vaihingen_IRRG is cropped to a 512×512 pixel size with a 256-pixel cropping stride, whatever it functions as the source either target domain, separated into about 1200 training images and 500 test images.

FloodNet encompasses 284 images, each with an average resolution of about 4600 × 3000 pixels and a spatial resolution of 3.75 cm. As depicted in Table I, the FloodNet dataset contains spectral bands of three-channel RGB. FloodNet_RGB is cropped to a 512×512 pixel size with a 512-pixel cropping stride, separated into about 7700 training images and 1700 test images.

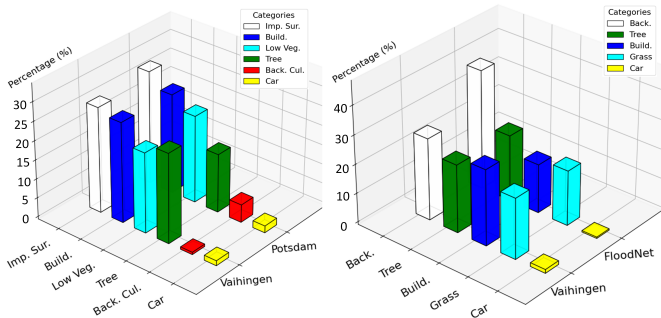


Fig. 6. Category distribution of cross-domain datasets.

Fig. 6 illustrates the pixel distribution across each category for different datasets. These datasets exhibit a common issue of class imbalance, with significantly lower proportions in the “Back. Clu.” and “Car” categories compared to others. The proportion of pixels for the “Back. Clu.” category in the Vaihingen dataset and the “Car” category in the FloodNet dataset is virtually insignificant. This imbalance stems from

factors like geographical location and terrain of aerial photography areas, posing some challenges in UDA segmentation task across remote sensing scenes.

B. Model Evaluation Metrics

This research evaluates the effectiveness of semantic segmentation for cross-domain remote sensing scenes by using three different metrics, namely mIoU, mF1, OA (Overall Accuracy). Similar evaluation metrics are applied during experimental assessments within studies [37], [38].

For each class k among the six different categories in the cross-domain dataset of this paper, the corresponding IoU and F1 scores are calculated as follows:

$$\text{IoU}_k = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (9)$$

$$\text{F1}_k = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (10)$$

mIoU involves computing the means of IoU for all category, mF1 involves computing the means of F1 for all category. The corresponding formulas are as follows:

$$\text{mIoU} = \frac{1}{K} \sum_{k=1}^K \text{IoU}_k \quad (11)$$

$$\text{mF1} = \frac{1}{K} \sum_{k=1}^K \text{F1}_k \quad (12)$$

Overall Accuracy (OA) measures the percentage of accurately classified pixels relative to the overall pixel count, as follows:

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (13)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative pixels, respectively.

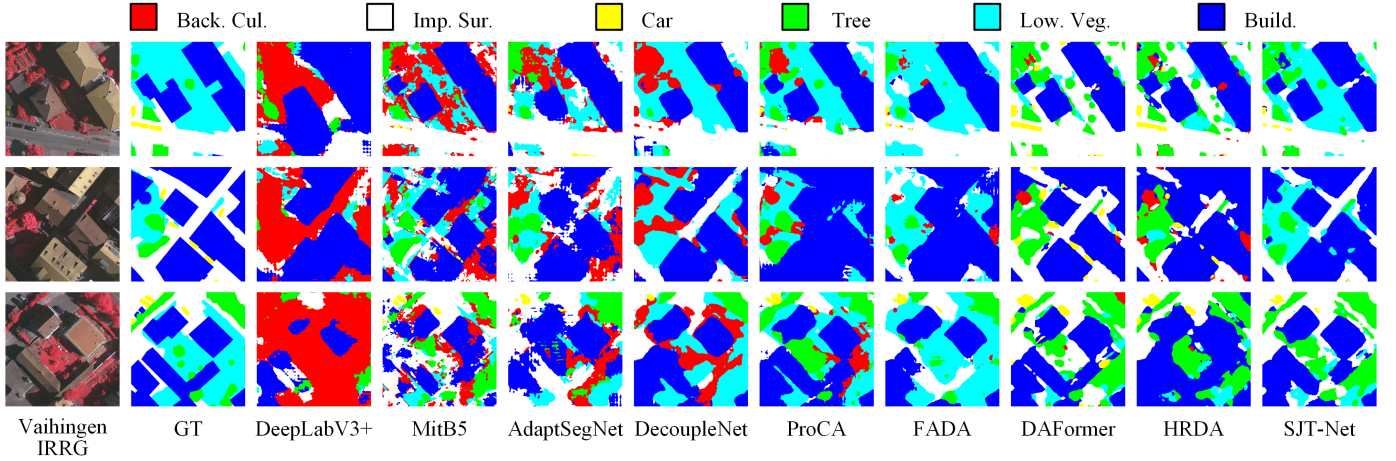


Fig. 7. Segmentation maps of varied models on the Pots_RGB → Vai_IRRG task

TABLE II
EVALUATION OUTCOMES OF VARIED APPROACHES ON THE POTS_RGB → VAI_IRRG TASK

Methods	Baseline		UDA Strategy		Class IoU %						mIoU %	mF1 %	OA %
	A	B	C	D	Back. Cul.	Imp. Sur.	Car	Tree	Low. Veg.	Build.			
DeepLabv3+	✓	×	×	×	0.62	33.49	29.93	37.56	1.99	54.73	26.39	34.22	52.59
Segformer	×	✓	×	×	1.15	45.08	41.16	55.11	17.46	61.19	36.86	46.84	57.02
AdaptSegNet	✓	×	✓	×	1.27	46.63	30.01	53.79	16.51	58.26	34.41	43.86	56.26
DecoupleNet	✓	×	✓	×	1.51	49.67	32.76	44.77	30.17	72.36	38.54	47.72	58.38
ProCA	✓	×	✓	×	10.44	47.39	47.45	60.12	21.25	71.98	43.11	52.99	60.77
FADA	✓	×	✓	×	11.08	49.36	46.46	51.19	30.22	75.14	43.91	54.06	61.73
DAFormer	×	✓	×	✓	16.37	57.30	41.75	69.00	15.93	84.81	47.53	60.14	63.62
HRDA	×	✓	×	✓	16.27	68.32	50.31	66.94	26.70	73.26	50.30	62.47	65.73
SJT-Net	×	✓	✓	✓	50.42	59.20	43.68	61.98	32.84	80.85	54.83	65.62	72.25

C. Training Parameter Configurations

All experiments in this research were conducted using PyTorch version 1.13.1. The experimental device is an NVIDIA GeForce RTX 8000 with 48GB of VRAM and an i9 12900K processor. With an initial learning rate of $5e-4$, momentum of 0.9, and weight decay of $1e-4$, our model employed the Stochastic Gradient Descent (SGD) optimizer for training the segmentation backbone architecture. The Adam optimizer was used to train the discriminator, setting an initial learning rate of $5e-4$, momentum parameters $\beta_1=0.9$ and $\beta_2=0.99$. The training batch size was 4, while the testing batch size was 1.

D. Comparative Experiments

For showing the excellence of the SJT-Net, we conducted four cross-domain semantic segmentation experiments, as follows:

- ① Source Domain: Potsdam_RGB, Target Domain: Vaihingen_IRRG, denoted as: Pots_RGB → Vai_IRRG
- ② Source Domain: Potsdam_IRRG, Target Domain: Vaihingen_IRRG, denoted as: Pots_IRRG → Vai_IRRG
- ③ Source Domain: Vaihingen_IRRG, Target Domain: Potsdam_RGB, denoted as: Vai_IRRG → Pots_RGB
- ④ Source Domain: Vaihingen_IRRG, Target Domain: FloodNet_RGB, denoted as: Vai_IRRG → Flood_RGB

The cross-domain experiment excludes Vai_IRRG → Pots_IRRG. This is because experiments ① and ② demonstrate superior result using similar bands, consistent with findings by Ma et al. [38] as well as Li et al. [1].

In previous UDA research, the DeepLab has commonly been used as the baseline model, while with less usage of the Transformer. Thus, this research contrasts two segmentation networks: DeepLabv3+ [21] and Segformer [27]. In addition, this research compares the existing UDA methods, including AdaptSegNet [13], FADA [14], DecoupleNet [15], ProCA [16], DAFormer [17], and HRDA [18]. By evaluating the output metrics of various approaches in cross-domain tasks, the proposed SJT-Net is demonstrated to be superior.

1) *Cross-Domain Segmentation Comparative Analysis of Pots_RGB → Vai_IRRG*: Table II shows the comparative experimental evaluation outcomes of Pots_RGB → Vai_IRRG, while Fig. 7 exhibits the visual segmentation outcomes of it. In Table II, A signifies the baseline as DeepLabv3+, B signifies the baseline as Segformer, C signifies the domain strategy as adversarial training, and D signifies the domain strategy as self-training. Table II indicates that the proposed SJT-Net surpasses the suboptimal model in overall evaluation indicators compared to the current SOTA, specifically manifested as mean IoU (mIoU) exceeding 4.53%, mean F1 (mF1) exceed-

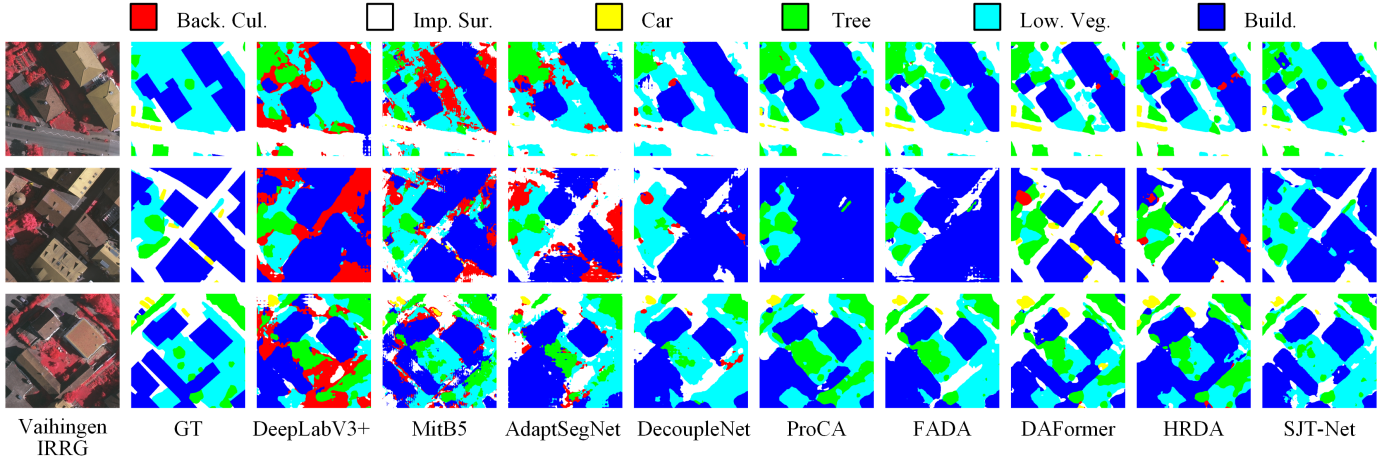


Fig. 8. Segmentation maps of varied models on the Pots_IRRG $\rightarrow$ Vai_IRRG task

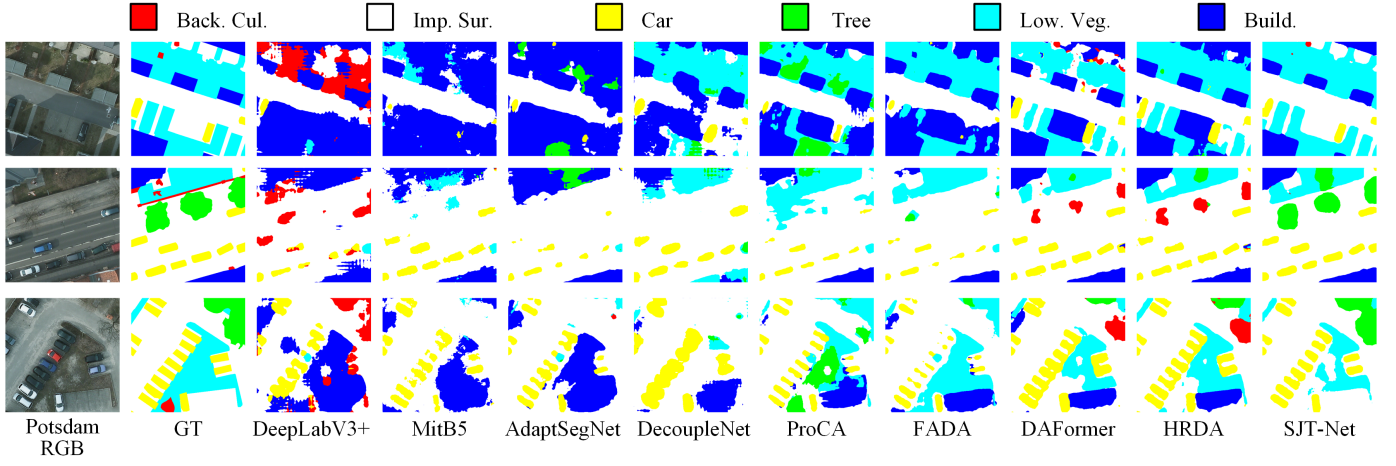
TABLE III
EVALUATION OUTCOMES OF VARIED APPROACHES ON THE POTS_IRRG $\rightarrow$ VAI_IRRG TASK

Methods	Baseline		UDA Strategy		Class IoU %						mIoU %	mF1 %	OA %
	A	B	C	D	Back. Cul.	Imp. Sur.	Car	Tree	Low. Veg.	Build.			
DeepLabv3+	✓	×	×	×	1.09	33.01	34.68	56.45	16.89	58.62	33.46	42.90	55.43
Segformer	×	✓	×	×	2.41	49.61	42.88	59.42	31.86	66.44	42.10	52.60	59.53
AdaptSegNet	✓	×	✓	×	1.47	47.47	28.49	51.07	27.81	58.34	35.78	45.64	56.70
DecoupleNet	✓	×	✓	×	7.15	48.40	43.37	41.06	28.97	73.80	40.46	50.39	59.06
ProCA	✓	×	✓	×	24.84	55.15	40.19	59.68	35.41	74.77	48.34	60.91	64.42
FADA	✓	×	✓	×	23.41	59.14	47.82	60.17	43.02	76.18	51.62	63.45	66.98
DAFormer	×	✓	×	✓	23.48	74.60	48.64	69.93	51.18	82.27	58.35	71.00	76.18
HRDA	×	✓	×	✓	11.87	73.49	51.05	69.21	50.93	83.19	56.62	68.94	72.98
SJT-Net	×	✓	✓	✓	54.66	69.13	52.03	63.16	51.85	80.63	61.91	74.16	78.81

ing 3.15%, and Overall Accuracy (OA) exceeding 5.52%, respectively. While some categories did not achieve the highest IoU, these categories still attained high scores. The SJT-Net model performed best in the imbalanced category “Back. Cul.” and the common category “Low. Veg.”. The SJT-Net model underperformed on the categories “Imp. Sur.”, “Car.”, “Tree”, and “Build.”, primarily due to pseudo-label generation within an adversarial training framework. The model lacks a mechanism to handle the boundaries between semantic categories, resulting in misclassifications, such as “Imp. Sur.”, “Car.” and “Build.” being mistaken for “Back. Cul.”, and “Tree” for “Low. Veg.” These categories share high visual similarity, making them difficult to distinguish without clear boundaries. As subsequent self-training iterations rely on these erroneous pseudo-labels, misclassification is compounded. Consequently, the SJT-Net model shows the improved performance in “Back. Cul.” and “Low. Veg.” relative to other models, but its performance in other categories remains suboptimal. Despite this, the SJT-Net model enhances recognition for imbalanced categories, achieving relatively balanced overall performance without severe errors in any category. The segmentation visualization depicted in Fig. 7 further illustrates that the SJT-Net model closely matches the ground truth (GT) labels. In the absence of UDA strategies, comparing the two baseline

models, Segformer outperforms DeepLabv3+ in terms of IoU in a single category, as well as overall mIoU, mF1 and OA. Overall, considering Table II and Fig. 7, the Transformer architecture’s strong segmentation performance is far superior to DeepLabv3+, and cross-domain segmentation experiments using domain adaptation strategies generally outperform those without domain adaptation strategies.

Although AdaptSegNet employs domain adaptation strategy, its effectiveness is hampered by its dependency on the antiquated fully convolutional network (DeepLabv3+), which compromises its advantages in domain alignment. Therefore, the segmentation outcomes of this method are marginally inferior to those of the experiment in which the baseline is Segformer without utilizing UDA. The SJT-Net model is built upon the Transformer structure, enabling the model to establish complex contextual dependencies on remote sensing categories during extraction of features. Additionally, our model employs adversarial domain adaptation strategies to produce high-caliber pseudo-labels, laying the foundation that supports further self-training, which outperforms methods that rely solely on self-training domain adaptation strategies. Finally, we iteratively update the pseudo labels and models using self-training to boost segmentation accuracy of the target domain. The SJT-Net model not only demonstrates

Fig. 9. Segmentation maps of varied models on the Vai_IRRG $\rightarrow$ Pots_RGB taskTABLE IV
EVALUATION OUTCOMES OF VARIED APPROACHES ON THE VAI_IRRG $\rightarrow$ POTS_RGB

Methods	Baseline		UDA Strategy		Class IoU %						mIoU %	mF1 %	OA %
	A	B	C	D	Back. Cul.	Imp. Sur.	Car	Tree	Low. Veg.	Build.			
DeepLabv3+	✓	×	×	×	4.07	34.29	41.20	13.82	5.84	29.45	21.45	26.67	48.34
Segformer	×	✓	×	×	15.44	36.90	61.48	22.81	18.98	33.72	31.56	38.32	53.19
AdaptSegNet	✓	×	✓	×	16.50	44.35	58.68	29.69	9.25	42.41	33.48	43.12	56.08
DecoupleNet	✓	×	✓	×	20.22	47.53	67.26	24.23	29.83	43.63	38.78	47.83	59.04
ProCA	✓	×	✓	×	25.50	44.19	62.56	21.59	36.47	60.40	41.79	51.36	59.17
FADA	✓	×	✓	×	26.68	48.91	55.08	26.22	38.60	57.81	42.22	52.66	60.05
DAFormer	×	✓	×	✓	20.47	65.68	64.76	51.91	58.34	76.65	56.30	68.12	72.92
HRDA	×	✓	×	✓	16.13	70.63	65.35	53.07	57.84	77.02	56.67	69.88	73.27
SJT-Net	×	✓	✓	✓	29.53	70.65	71.67	61.41	56.79	75.58	60.94	73.37	77.93

robustness in segmentation networks but also comprehensively integrates the benefits of both domain adaptation strategies, complementing each other and fully leveraging adversarial training and self-training advantages, thereby yielding superior cross-domain segmentation results.

2) *Cross-Domain Segmentation Comparative Analysis of Pots_IRRG $\rightarrow$ Vai_IRRG*: Table III shows the comparative experimental evaluation outcomes of Pots_IRRG $\rightarrow$ Vai_IRRG, while Fig. 8 exhibits the visual segmentation outcomes of it. In Table III, A signifies the baseline as DeepLabv3+, B signifies the baseline as Segformer, C signifies the domain strategy as adversarial training, and D signifies the domain strategy as self-training. Table III indicates that the SJT-Net model surpasses the suboptimal model in overall evaluation indicators compared to the current SOTA, specifically manifested as mean IoU (mIoU) exceeding 3.56%, mean F1 (mF1) exceeding 3.16%, and Overall Accuracy (OA) exceeding 5.03%, respectively. While some categories did not achieve the highest IoU, these categories still attained high scores, and the SJT-Net model achieved the best results in two imbalanced categories “Back. Cul.” and “Car” as well as the common category “Low. Veg.”. The segmentation visualization depicted in Fig. 8 also reveals that the SJT-Net model closely matches the GT labels. Similar conclusions regarding cross-domain seg-

mentation performance can be drawn by analyzing Table III and Fig. 8, consistent with findings from previous comparative experiments.

3) *Cross-Domain Segmentation Comparative Analysis of Vai_IRRG $\rightarrow$ Pots_RGB*: Table IV shows the comparative experimental evaluation outcomes of Vai_IRRG $\rightarrow$ Pots_RGB, while Fig. 9 exhibits the visual segmentation outcomes of it. In Table IV, A signifies the baseline as DeepLabv3+, B signifies the baseline as Segformer, C signifies the domain strategy as adversarial training, and D signifies the domain strategy as self-training. Table IV indicates that the SJT-Net model surpasses the suboptimal model in overall evaluation indicators compared to the current SOTA, specifically manifested as mean IoU (mIoU) exceeding 4.27%, mean F1 (mF1) exceeding 3.49%, and Overall Accuracy (OA) exceeding 4.66%, respectively. While some categories did not achieve the highest IoU, these categories still attained high scores, and the SJT-Net model achieved the best results in two imbalanced categories “Back. Cul.” and “Car” as well as two common categories “Imp. Sur.” and “Tree”. The segmentation visualization depicted in Fig. 9 also reveals that the SJT-Net model closely matches the ground truth labels. Similar conclusions regarding cross-domain segmentation performance can be drawn by analyzing Table IV and Fig. 9, consistent with findings from

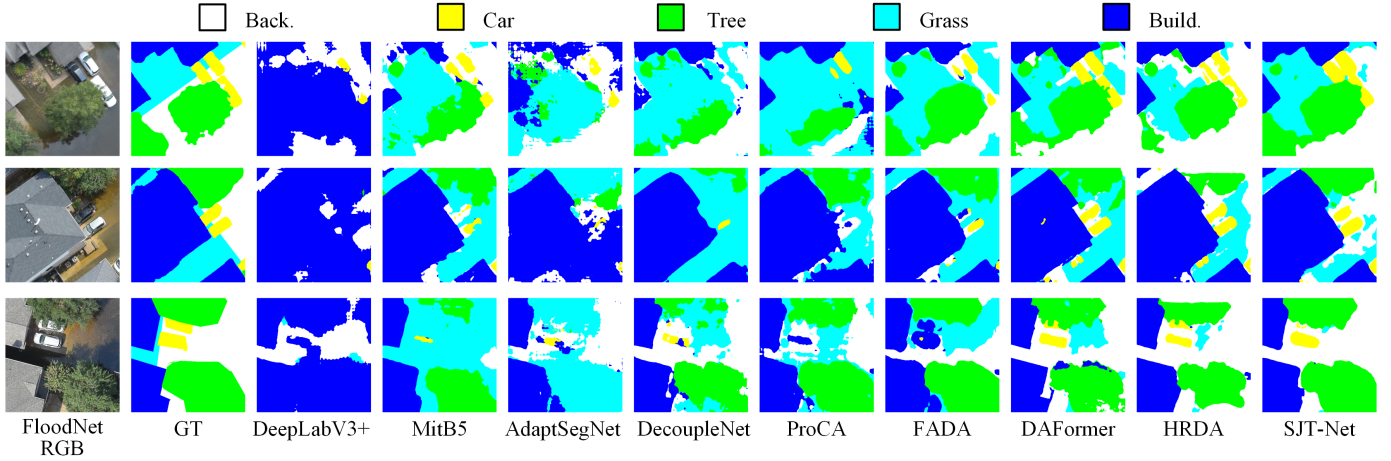


Fig. 10. Segmentation maps of varied models on the Vai_IRRG $\rightarrow$ Flood_RGB task

TABLE V
EVALUATION OUTCOMES OF VARIED APPROACHES ON THE VAI_IRRG $\rightarrow$ FLOOD_RGB

Methods	Baseline		UDA Strategy		Class IoU %					mIoU %	mF1 %	OA %
	A	B	C	D	Back.	Car	Tree	Grass	Build.			
DeepLabv3+	✓	×	×	×	18.11	37.68	23.56	18.82	45.60	28.75	31.68	36.65
Segformer	×	✓	×	×	20.42	42.15	35.02	26.19	57.40	36.24	42.20	52.97
AdaptSegNet	✓	×	✓	×	20.62	43.32	22.60	28.17	57.54	34.45	38.25	46.17
DecoupleNet	✓	×	✓	×	36.91	41.96	50.05	30.49	72.59	46.40	60.37	65.44
ProCA	✓	×	✓	×	34.94	41.73	62.62	31.11	69.65	48.01	62.73	68.04
FADA	✓	×	✓	×	35.40	42.11	64.38	32.95	70.82	49.13	64.57	69.76
DAFormer	×	✓	×	✓	38.53	44.88	60.80	35.51	73.56	50.66	66.07	71.00
HRDA	×	✓	×	✓	44.33	38.27	63.69	34.89	77.43	51.72	67.68	71.92
SJT-Net	×	✓	✓	✓	45.28	44.11	65.18	36.27	78.66	53.90	69.88	75.74

previous two comparative experiments.

4) *Cross-Domain Segmentation Comparative Analysis of Vai_IRRG $\rightarrow$ Flood_RGB*: Table IV shows the comparative experimental evaluation outcomes of Vai_IRRG $\rightarrow$ Flood_RGB, while Fig. 10 exhibits the visual segmentation outcomes of it. In Table V, A signifies the baseline as DeepLabv3+, B signifies the baseline as Segformer, C signifies the domain strategy as adversarial training, and D signifies the domain strategy as self-training. Table V indicates that the SJT-Net model surpasses the suboptimal model in overall evaluation indicators compared to the current SOTA, specifically manifested as mean IoU (mIoU) exceeding 2.18%, mean F1 (mF1) exceeding 2.20%, and Overall Accuracy (OA) exceeding 3.82%, respectively. The SJT-Net achieved the best results in “Back.”, “Tree”, “Grass” and “Build.”. While some categories did not achieve the highest IoU, these categories still attained high scores. The segmentation visualization depicted in Fig. 10 also reveals that the SJT-Net model closely matches the ground truth labels. Similar conclusions regarding cross-domain segmentation performance can be drawn by analyzing Table V and Fig. 10, consistent with findings from previous three comparative experiments.

5) *Visual Analysis of Confusion Matrices in Four Cross-Domain Segmentation Comparative Experiments*: This section

presents the visualization of confusion matrices for four sets of cross-domain RSI semantic segmentation comparative experiments, depicting both SJT-Net and suboptimal models, as illustrated in Fig. 11. Specifically, Figures (A1) and (A2) show the confusion matrices for SJT-Net and the suboptimal model, respectively, in the Pots_RGB $\rightarrow$ Vai_IRRG task. Similarly, (B1) and (B2) illustrate the confusion matrices for the same comparative models in the Pots_IRRG $\rightarrow$ Vai_IRRG task. Likewise, (C1) and (C2) show the matrices for the Vai_IRRG $\rightarrow$ Pots_RGB task, and (D1) and (D2) for the Vai_IRRG $\rightarrow$ Flood_RGB task. By comparing the confusion matrices between the SJT-Net model and the suboptimal model, we gain a more intuitive understanding of their performance disparities in diverse cross-domain semantic segmentation tasks. The confusion matrix analysis, depicted in Fig. 11, reveals that the SJT-Net model exhibits predominantly high values along its diagonal elements, contrasting with lower values in its off-diagonal counterparts. The confusion matrix of SJT-Net exhibits a higher concentration of its elements near the main diagonal. Conversely, the elements of suboptimal models display a more scattered distribution, a trend consistently observed across all four cross-domain segmentation experiments. This also demonstrates that the SJT-Net model possesses the most comprehensive segmentation performance

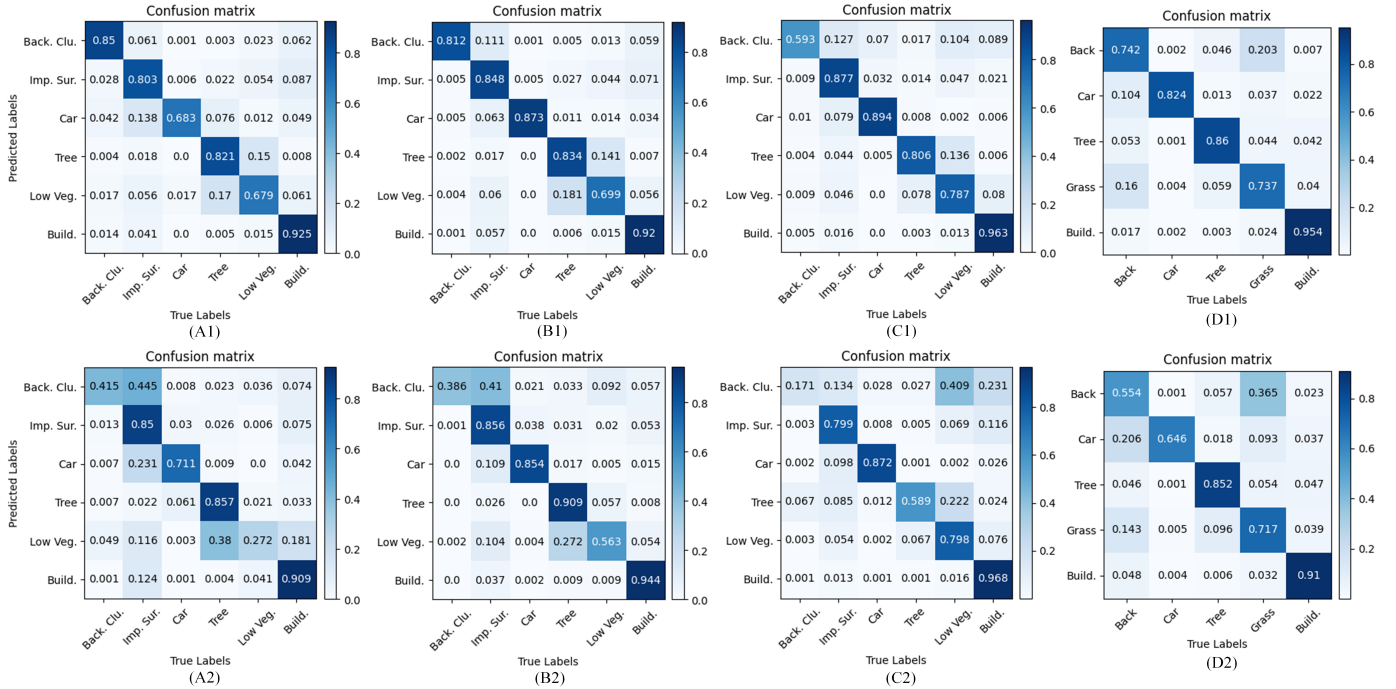


Fig. 11. The confusion matrices between SJT-Net and suboptimal models in four cross-domain segmentation comparative experiments.

and superior generalization performance when handling cross-domain semantic segmentation tasks.

E. Ablation Experiment

1) *Ablation Experiments of Different Improved Modules:* In demonstrating how each improved module in SJT-Net impacts model performance, various modules are applied in four cross-domain RSI semantic segmentation tasks for ablation analysis, as depicted in Table VI. CDSR denotes Cross-Domain Spatial Reconstruction, SFE denotes Segformer Feature Extractor, and GLID denotes Global-Local Interactive Decoder. In the Pots_RGB $\rightarrow$ Vai_IRRG task, compared with the SJT-Net model, when the CDSR is not used, the mIoU value drops by 1.41%. When the SFE is not used and ResNet101 is used as the feature extractor, the mIoU value drops by 8.26%. When the GLID is not used and the ASPP module is used to directly perform feature decoding, the mIoU value drops by 2.99%. Similarly, in the Pots_IRRG $\rightarrow$ Vai_IRRG task, the same ablation operations, when compared to the SJT-Net model, lead to a decrease in mIoU by 2.4%, 9.65%, and 4.96%, respectively. In the Vai_IRRG $\rightarrow$ Pots_RGB task, the same ablation operations lead to a decrease in mIoU by 2.53%, 10.26%, and 5.22%, respectively, compared to the SJT-Net model. In the Vai_IRRG $\rightarrow$ Flood_RGB task, the same ablation operations lead to a decrease in mIoU by 1.68%, 4.53%, and 3.05%, respectively, compared to the SJT-Net model. The evaluation of four cross-domain segmentation tasks shows that the improved modules enhance the SJT-Net's performance to varying degrees. Specifically, the SFE provides the most significant improvement, followed by the GLID, and then the CDSR.

2) *Ablation Experiments of Different Domain Adaptation Strategies:* This section presents ablation studies across four remote sensing tasks to assess the impact of various domain

adaptation strategies on the cross-domain segmentation performance of SJT-Net. The backbone of the SJT-Net model consists of an improved Segformer Feature Extractor and a global-local interactive decoder, with the different UDA strategies detailed in Table VII. The ablation analysis without domain adaptation involves training the segmentation backbone on source domain samples and directly applying the learned weights to the target domain. In the adversarial training ablation, the backbone is trained on the source domain, with the learned weights capturing the target domain's semantic characteristics. Features from both domains are aligned using a discriminator, and adversarial features are subsequently decoded into segmentation maps. In the self-training ablation, the backbone is trained on the source domain, and the learned weights are used to predict the target domain. These predictions are treated as pseudo-labels, which are iteratively refined to optimize segmentation performance on the target domain. In the Pots_RGB $\rightarrow$ Vai_IRRG task, when no domain adaptation strategy is applied, the mIoU value decreases by 14.04% compared to the SJT-Net model. With the use of adversarial training for feature-level domain alignment, the mIoU value decreases by 5.35%. When using self-training for feature-level domain alignment, the mIoU value decreases by 9.66%. Similarly, in the Pots_IRRG $\rightarrow$ Vai_IRRG task, the same ablation operations, lead to a decrease in mIoU by 14.76%, 4.05% and 7.66%, respectively. In the Vai_IRRG $\rightarrow$ Pots_RGB task, the same ablation operations lead to a decrease in mIoU by 19.68%, 6.46% and 10.78%, respectively. In the Vai_IRRG $\rightarrow$ Flood_RGB task, the same ablation operations lead to a decrease in mIoU by 15.66%, 5.74%, and 8.31%, respectively. Cross-domain RSI often exhibit significant class imbalance and scale variations. Evaluation across four cross-domain segmentation tasks reveals that direct segmentation, without domain adaptation, impairs the model's generalization ability,

TABLE VI
THE ABLATION OUTCOMES OF VARIED IMPROVED MODULES ON FOUR CROSS-DOMAIN SEGMENTATION TASKS

Tasks	Cross-Domain Spatial Reconstruction (CDSR)	Segformer Feature Extractor (SFE)	Global-Local Interactive Decoder (GLID)	mIoU %	mF1 %	OA %
Pots_RGB	×	✓	✓	53.42	64.46	68.61
→ Vai_IRRG	✓	×	✓	46.57	56.41	62.64
	✓	✓	×	51.74	63.72	67.72
	✓	✓	✓	54.83	65.62	72.25
Pots_IRRG	×	✓	✓	59.51	72.77	76.71
→ Vai_IRRG	✓	×	✓	52.26	63.83	68.09
	✓	✓	×	56.95	70.04	73.48
	✓	✓	✓	61.91	74.16	78.81
Vai_IRRG	×	✓	✓	58.41	71.42	76.42
→ Pots_IRRG	✓	×	✓	50.68	62.71	65.97
	✓	✓	×	55.72	66.23	72.56
	✓	✓	✓	60.94	73.37	77.93
Vai_IRRG	×	✓	✓	52.22	68.18	72.19
→ Flood_RGB	✓	×	✓	49.37	65.21	70.09
	✓	✓	×	50.85	66.32	71.33
	✓	✓	✓	53.90	69.88	75.74

TABLE VII
THE ABLATION OUTCOMES OF VARIED DOMAIN ADAPTATION STRATEGIES ON FOUR CROSS-DOMAIN SEGMENTATION TASKS

Tasks	Adversarial Training	Self- training	mIoU %	mF1 %	OA %
Pots_RGB	×	×	40.79	50.55	59.09
→ Vai_IRRG	✓	×	49.48	61.27	64.78
	×	✓	45.17	55.32	62.46
	✓	✓	54.83	65.62	72.25
Pots_IRRG	×	×	47.15	58.92	62.99
→ Vai_IRRG	✓	×	57.86	70.76	75.07
	×	✓	54.25	64.91	71.66
	✓	✓	61.91	74.16	78.81
Vai_IRRG	×	×	41.26	51.07	59.12
→ Pots_IRRG	✓	×	54.48	65.29	71.87
	×	✓	50.16	62.30	65.35
	✓	✓	60.94	73.37	77.93
Vai_IRRG	×	×	38.24	45.20	54.56
→ Flood_RGB	✓	×	48.16	63.05	68.23
	×	✓	45.59	59.68	64.88
	✓	✓	53.90	69.88	75.74

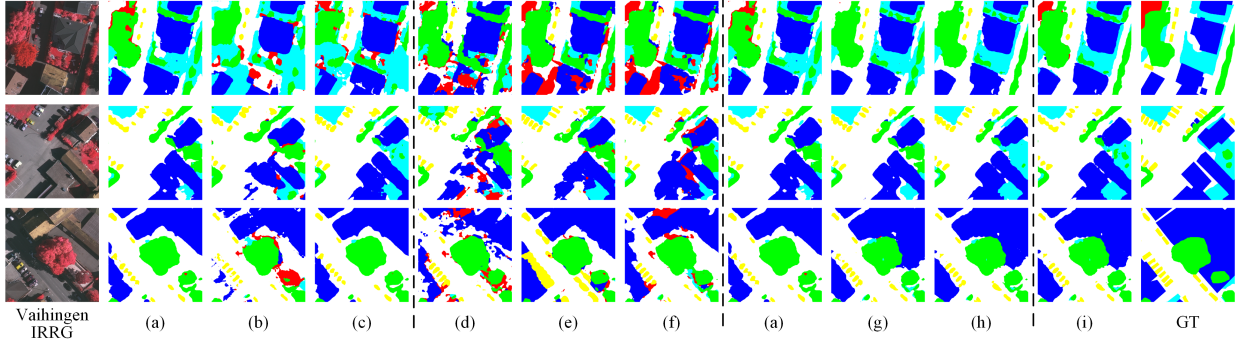
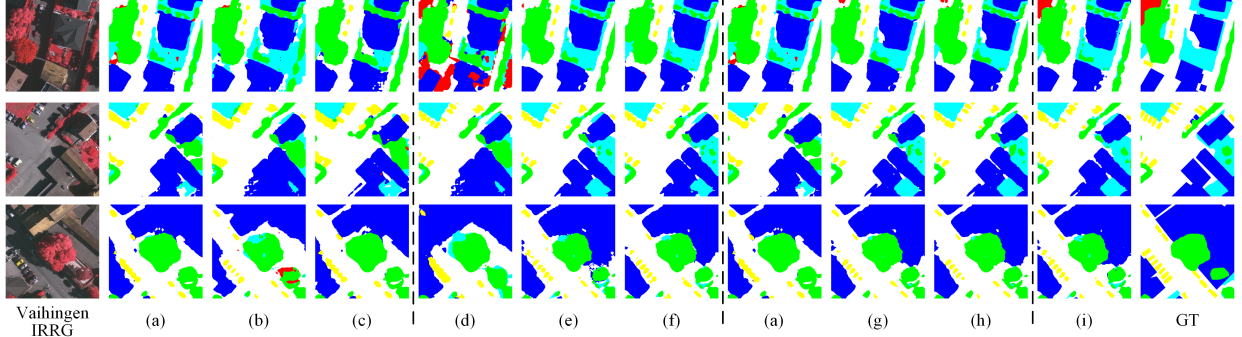
TABLE VIII
THE ABLATION OUTCOMES OF VARIED SPATIAL RECONSTRUCTION STRATEGIES ON FOUR CROSS-DOMAIN SEGMENTATION TASKS

Tasks	Gaussian Pyramid	Bilinear Interpolation	mIoU %	mF1 %	OA %
Pots_RGB	×	×	53.42	64.46	68.61
→ Vai_IRRG	✓	×	53.94	64.72	69.34
	×	✓	54.23	64.89	71.58
	✓	✓	54.83	65.62	72.25
Pots_IRRG	×	×	59.51	72.77	76.71
→ Vai_IRRG	✓	×	61.20	73.47	78.23
	×	✓	61.50	73.76	78.75
	✓	✓	61.91	74.16	78.81
Vai_IRRG	×	×	58.41	71.42	76.42
→ Pots_IRRG	✓	×	60.22	72.84	77.26
	×	✓	60.40	73.11	77.68
	✓	✓	60.94	73.37	77.93
Vai_IRRG	×	×	52.22	68.18	72.19
→ Flood_RGB	✓	×	53.14	69.28	72.37
	×	✓	53.55	69.57	73.53
	✓	✓	53.90	69.88	75.74

leading to suboptimal performance. Although both adversarial and self-training domain adaptation techniques can improve segmentation, their individual benefits are limited and fail to fully harness their potential. To overcome this limitation, we propose SJT-Net, a model that integrates adversarial and self-training strategies. Adversarial training generates precise pseudo-labels, which serve as reliable pseudo-supervision for the self-training process. This combined approach enhances

cross-domain segmentation performance by leveraging the strengths of both strategies.

3) *Ablation Experiments of Different Spatial Reconstruction Strategies:* This section demonstrates the effect of spatial reconstruction strategies in SJT-Net on cross-domain segmentation performance through ablation experiments conducted on four RSI tasks. Various image resampling techniques are employed to align the spatial resolution of source domain

Fig. 12. Segmentation maps of ablation experiment for Pots_RGB $\rightarrow$ Vai_IRRG taskFig. 13. Segmentation maps of ablation experiment for Pots_IRRG $\rightarrow$ Vai_IRRG task

images with that of the target domain, as detailed in Table VIII. In the absence of spatial reconstruction, both domain images are cropped to 512 $\times$ 512 pixels, resulting in resolution discrepancies between the domains, as illustrated in Table I. In the Pots_RGB $\rightarrow$ Vai_IRRG task, without spatial reconstruction, the mIoU decreases by 1.41%. Gaussian pyramid sampling reduces this decrease to 0.89%, while bilinear interpolation results in a 0.60% decrease. Similarly, in the Pots_IRRG $\rightarrow$ Vai_IRRG task, the same ablation operations, lead to a decrease in mIoU by 2.40%, 0.71% and 0.41%, respectively. In the Vai_IRRG $\rightarrow$ Pots_RGB task, the same ablation operations lead to a decrease in mIoU by 2.53%, 0.72% and 0.54%, respectively. In the Vai_IRRG $\rightarrow$ Flood_RGB task, the same ablation operations lead to a decrease in mIoU by 1.68%, 0.76%, and 0.35%, respectively. Analysis of evaluation results from four cross-domain segmentation tasks demonstrates that the SJT-Net model enhances performance through cross-domain spatial reconstruction. Bilinear interpolation outperforms Gaussian pyramid sampling slightly, and their combination yields smoother images while preserving spatial details, thereby enhancing image reconstruction quality and further improving overall segmentation performance.

4) *Visualization of Ablation Experiments*: This section presents visual segmentation maps from ablation experiments conducted across four cross-domain tasks. Specifically, Figs. 12–15 depict the following tasks: Pots_RGB $\rightarrow$ Vai_IRRG (Fig. 12), Pots_IRRG $\rightarrow$ Vai_IRRG (Fig. 13), Vai_IRRG $\rightarrow$ Pots_RGB (Fig. 14), and Vai_IRRG $\rightarrow$ Flood_RGB (Fig. 15). The ablation methods, represented by legends with matching subscripts in Figs. 12–15. Legends (a), (b), and (c) correspond to ablation experiments on improved modules (Section IV.E.1), where (a) indicates no Cross-Domain Spatial Reconstruction,

(b) indicates no Segformer Feature Extractor, and (c) indicates no Global-Local Interactive Decoder. Legends (d), (e), and (f) refer to domain adaptation strategies (Section IV.E.2), where (d) indicates no domain adaptation, (e) indicates adversarial training, and (f) denoting self-training. Legends (g) and (h) refer to spatial reconstruction strategies (Section IV.E.3), with (g) utilizing Gaussian pyramid sampling and (h) using bilinear interpolation. Legend (i) indicates the proposed SJT-Net, while "GT" indicates the ground truth labels. The visual results in Figs. 12–15 show that SJT-Net closely aligns with the GT across all tasks, showing superior segmentation accuracy.

V. CONCLUSION

To address domain distribution imbalance in UDA RSI, this research combines spatial reconstruction and joint training with the Transformer network. The proposed SJT-Net aligns spatial scales through spatial reconstruction, while synergistically unifying adversarial training and self-training, which overcomes adversarial training's limited segmentation modeling capacity and self-training's dependency to initial pseudo-label quality. Our global-local interaction decoder captures multi-scale features and complex spatial dependencies. Comprehensive experiments on four cross-domain benchmarks demonstrate SJT-Net's superiority over SOTA methods, with ablation studies validating the effectiveness of critical modules. In practical remote sensing interpretation, cross-domain segmentation from a single source to multiple targets and the utilization of multimodal information are both essential. However, these factors have not been addressed within our framework. Future work will focus on developing multi-target UDA frameworks and multimodal fusion strategies to enhance RSI interpretation utility.

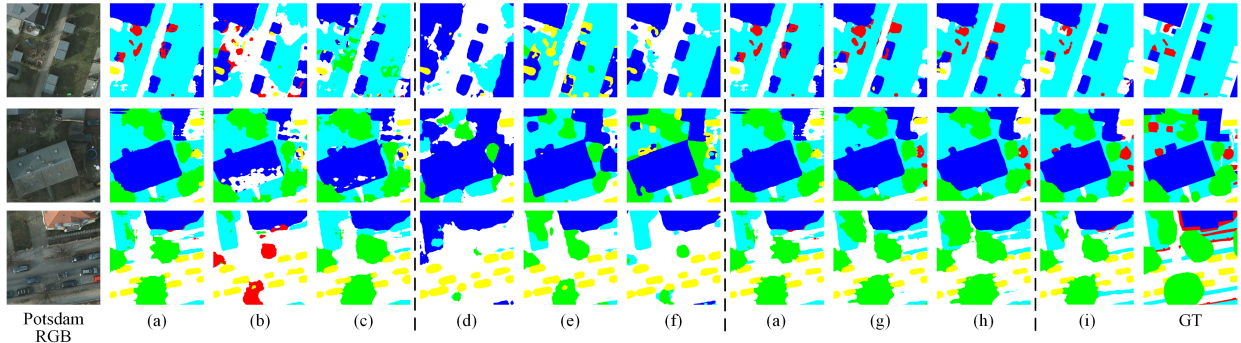


Fig. 14. Segmentation map of ablation experiment for Vai_IRRG $\rightarrow$ Pots_RGB task

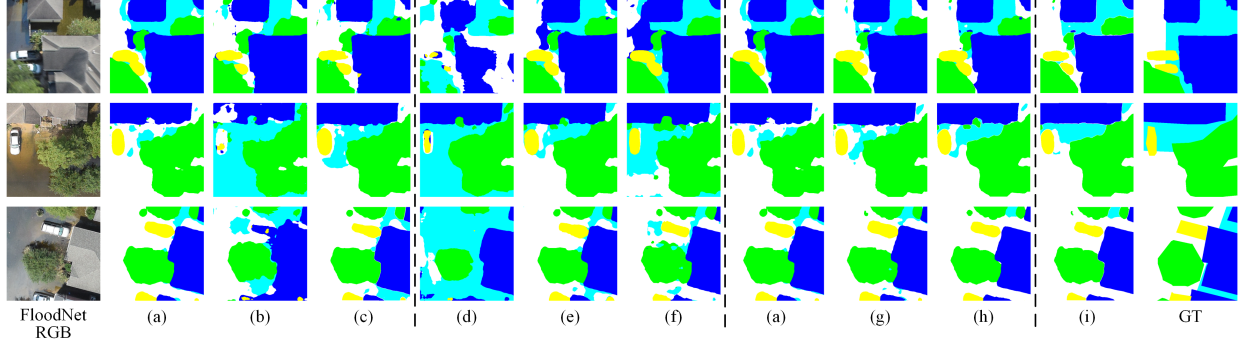


Fig. 15. Segmentation map of ablation experiment for Vai_IRRG $\rightarrow$ Flood_RGB task

REFERENCES

- [1] 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. 20-23, 2021.
- [2] R. Jalal, Z. Iqbal, M. Henry, G. Franceschini, M. S. Islam, M. Akhter, Z. T. Khan, M. A. Hadi, M. A. Hossain, M. G. Mahboob *et al.*, "Toward efficient land cover mapping: An overview of the national land representation system and land cover map 2015 of Bangladesh," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, pp. 3852-3861, 2019.
- [3] X. Y. Tong, G. S. Xia, Q. Lu, H. Shen, S. Li, S. You and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, p. 111322, 2020.
- [4] X. Huang and Y. Wang, "Investigating the effects of 3D urban morphology on the surface urban heat island effect in urban functional zones by using high-resolution remote sensing data: A case study of Wuhan, Central China," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 119-131, 2019.
- [5] L. Mu, L. Wang, Y. Wang, X. Chen and W. Han, "Urban land use and land cover change prediction via self-adaptive cellular based deep learning with multisourced data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 12, pp. 5233-5247, 2019.
- [6] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, p. 111716, 2020.
- [7] M. Zhao, F. Si, Y. Wang, H. Zhou, S. Wang, Y. Jiang and W. Liu, "First year on-orbit calibration of the Chinese environmental trace gas monitoring instrument onboard GaoFen-5," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8531-8540, 2020.
- [8] M. S. Rahman, L. Di, E. Yu, L. Lin and Z. Yu, "Remote sensing based rapid assessment of flood crop damage using novel disaster vegetation damage index (DVDI)," *Int. J. Disaster Risk Sci.*, vol. 12, pp. 90-110, 2021.
- [9] Y. Liu, Q. Yu, Q. Zhou, C. Wang, S. D. Bellingrath-Kimura and W. Wu, "Mapping the complex crop rotation systems in Southern China considering cropping intensity, crop diversity, and their seasonal dynamics," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 9584-9598, 2022.
- [10] L. Mou, Y. Hua and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comp. Vis. Pattern Recognit.*, 2019, pp. 12416-12415.
- [11] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-13, 2021.
- [12] 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, pp. 1-14, 2023.
- [13] Y. H. Tsai, W. C. Hung, S. Schuler, K. Sohn, M. H. Yang and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, 2018, pp. 7472-7481.
- [14] H. Wang, T. Shen, W. Zhang, L. Y. Duan and T. Mei, "Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 642-659.
- [15] X. Lai, Z. Tian, X. Xu, Y. Chen, S. Liu, H. Zhao, L. Wang and J. Jia, "Decouplet: Decoupled network for domain adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 369-387.
- [16] Z. Jiang, Y. Li, C. Yang, P. Gao, Y. Wang, Y. Tai and C. Wang, "Prototypical contrast adaptation for domain adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 36-54.
- [17] 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. Comp. Vis. Pattern Recognit.*, 2022, pp. 9924-9935.
- [18] L. Hoyer, D. Dai and L. Van Gool, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2022, pp. 372-391.
- [19] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 834-848, 2017.
- [20] L. C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
- [21] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801-818.

- [22] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, 2016, pp. 780-778.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [24] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comp. Vis. Pattern Recognit.*, 2021, pp. 6881-6890.
- [25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Conf. Comp. Vis. Pattern Recognit.*, 2021, pp. 10012-10022.
- [26] R. Strudel, R. Garcia, I. Laptev and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Conf. Comp. Vis. Pattern Recognit.*, 2021, pp. 7276-7272.
- [27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077-12090, 2021.
- [28] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, 2015, pp. 3431-3440.
- [29] J. Zheng, A. Shao, Y. Yan, J. Wu and M. Zhang, "Remote Sensing Semantic Segmentation via Boundary Supervision-Aided Multiscale Channelwise Cross Attention Network," in *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-14, 2023.
- [30] X. Li, F. Xu, F. Liu, X. Lyu, Y. Tong, Z. Xu and J. Zhou, "A synergistical attention model for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-16, 2023.
- [31] L. Zhang, Z. Tan, G. Zhang, W. Zhang and Z. Li, "Learn More and Learn Usefully: Truncation Compensation Network for Semantic Segmentation of High-Resolution Remote Sensing Images," in *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-14, 2024.
- [32] Y. Zhang, J. Cheng, Y. Su, C. Deng, Z. Xia and N. Tashi, "Global Adaptive Second-Order Transformer for Remote Sensing Image Semantic Segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-17, 2024.
- [33] Q. Zeng, J. Zhou, J. Tao, L. Chen, X. Niu and Y. Zhan, "Multiscale Global Context Network for Semantic Segmentation of High-Resolution Remote Sensing Images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-13, 2024.
- [34] R. Zhang, Q. Zhang and G. Zhang, "LSRFormer: Efficient Transformer Supply Convolutional Neural Networks With Global Information for Aerial Image Segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1-13, 2024.
- [35] D. Wittich and F. Rottensteiner, "Appearance based deep domain adaptation for the classification of aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 180, pp. 82-102, 2021.
- [36] 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, 2020.
- [37] L. Wu, M. Lu and L. Fang, "Deep covariance alignment for domain adaptive remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-11, 2022.
- [38] 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, pp. 1-18, 2024.
- [39] Z. Zheng, Y. Zhong, Y. Su and A. Ma, "Domain adaptation via a task-specific classifier framework for remote sensing cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-13, 2022.
- [40] 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. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 4109-4121, 2023.
- [41] Z. Zhang, K. Doi, A. Iwasaki and G. Xu, "Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 746-750, 2021.
- [42] F. Zhang, Y. Shi, Z. Xiong, W. Huang and X. X. Zhu, "Pseudo features guided self-training for domain adaptive semantic segmentation of satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1-14, 2023.
- [43] 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.
- [44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, 2017.
- [45] D. Han, "Comparison of commonly used image interpolation methods," in *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, Atlantis Press, 2013, pp. 1556-1559.
- [46] M. Patil, "Interpolation techniques in image resampling," in *Int. J. Eng. Technol.*, vol. 7, pp. 567-570, 2018.
- [47] J. A. Guerrero-Colón, L. Mancera and J. Portilla, "Image restoration using space-variant Gaussian scale mixtures in overcomplete pyramids," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 27-41, 2008.
- [48] Z. Li, H. Shu and C. Zheng, "Multi-scale single image dehazing using laplacian and gaussian pyramids," *IEEE Trans. Image Process.*, vol. 30, pp. 9270-9279, 2021.
- [49] L. Chi, B. Jiang and Y. Mu, "Fast fourier convolution," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 4479-4488, 2020.
- [50] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325-341.
- [51] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez and U. Breitkopf, "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 1, no. 1, pp. 293-298, 2012.
- [52] M. Rahmemonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari and R. R. Murphy, "FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding," *IEEE Access*, 2021.



Junying Zeng (Member, IEEE) received the M.S. degree in physical electronic from Yunnan University, Yunnan, China, in 2005, and the Ph.D. degree in physical electronics from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008. Since 2008, he has been working with the School of Electronics and Information Engineering, Wuyi University, Guangdong, China. He is a Professor with Wuyi University. His main research interests include deep learning, remote sensing image processing and medical image processing.



Senyao Deng received the B.S. degree from Guangdong Ocean University, Guangdong, China, in 2018, and the M.S. degree in Information and Communication Engineering from Wuyi University, Guangdong, China, in 2025. His main research focuses on deep learning, computer vision, remote sensing image semantic segmentation and unsupervised domain adaptation.



Yikui Zhai (Senior Member, IEEE) received the B.S. and M.S. degrees from Shantou University, Guangdong, China, in 2004 and 2007 respectively, and the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013. Since October 2007, he has been working with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, China, where he is a Professor. He has been a Visiting Scholar with the Department of Computer Science, University of Milan, Milan, Italy, since 2016. His research interests include image processing, deep learning, pattern recognition, object detection, Unmanned Aerial Vehicle (UAV) change detection, and self-supervised learning.



Xudong Jia received the B.S. and M.S. degrees from Beijing Jiaotong University, in 1983 and 1986 respectively, the second M.S. degree from the University of Toronto, Canada, in 1992, and the Ph.D. degree from the Georgia Institute of Technology, in 1996. He is a Professor and the Dean of the School of Engineering, Science, Technology, Central Connecticut State University, New Britain, USA. He is also a Visiting Scholar of Wuyi University, China. He is an associate editor of the IEEE Intelligent Transportation Systems Society and the Open Journal of IEEE Intelligent Transportation Systems. His research interests include intelligent transportation systems (ITS) standards, geographic information system (GIS) applications in transportation, traffic safety, transportation information systems, travel demand management, and air quality.



Chuanbo Qin received the B.S. and M.S. degrees from Wuyi University, Guangdong, China, in 2004 and 2008 respectively, and the Ph.D. degree from the Department of Automation Science and Engineering, South China University of Technology, in 2015. Since 2015, he has been working with the School of Electronics and Information Engineering, Wuyi University, Guangdong, China. He is an Associate Professor with Wuyi University. His main research interests are in deep learning, medical image segmentation and biometric recognition.



Pasquale Coscia received the Ph.D. degree in Industrial and Information Engineering from the Università degli Studi di Milano, Italy, in 2019. From 2019 to 2022, he was a post-doctoral researcher at the Università degli Studi di Padova. Dr. Coscia has received the 2021 Best Paper Award of the Elsevier's Image and Vision Computing Journal for his work on human motion prediction. Additionally, he is a member of the program committees for numerous IEEE international conferences and workshops. Since 2023, he has held the position of Co-chair for the Intelligent Measurement Systems Technical Committee (TC-22) within the IEEE Instrumentation and Measurement Society. His main research interests include the areas of applied aspects of computational intelligence for signal and image processing, computer vision, machine learning and probabilistic models applied to multiagents systems.



Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Università degli Studi di Milano, Milan, Italy, in 2014. He has been a Post-Doctoral Research Fellow in computer science with the Università degli Studi di Milano since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 articles in international journals, proceedings of international conferences, books, and book chapters, and patents. His current research interests include signal and image processing, 3-D reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems. Dr. Genovese is an Associate Editor of the Journal of Ambient Intelligence and Humanized Computing (Springer).



Vincenzo Piuri (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer engineering from Politecnico di Milano, Milan, Italy, in 1984 and 1988, respectively. He was the Department Chair with the University of Milan, Milan, from 2007 to 2012, where he has been a Full Professor since 2000. He was an Associate Professor with Politecnico di Milano from 1992 to 2000, a Visiting Professor with The University of Texas at Austin, Austin, TX, USA from 1996 to 1999, and a Visiting Researcher with George Mason University, Fairfax, VA, USA, from 2012 to 2016. He founded a startup company, Sensure srl, Bergamo, Italy, in the area of intelligent systems for industrial applications (leading it from 2007 to 2010) and was active in industrial research projects with several companies. His main research and industrial application interests are intelligent systems, computational intelligence, pattern analysis and recognition, machine learning, image processing, biometrics, intelligent measurement systems and cloud computing. Dr. Piuri is an ACM Fellow.



Fabio Scotti (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003. He has been a Professor with the Università degli Studi di Milano since 2020. His original results have been published in over 150 articles in international journals, proceedings of international conferences, books, book chapters, and patents. His research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, 3-D reconstruction, industrial applications, intelligent measurement systems, and high-level system design. Dr. Scotti is an Associate Editor with the IEEE Transactions on Human-Machine Systems and Soft Computing (Springer). He has been an Associate Editor with the IEEE Transactions on Information Forensics and Security, and a Guest Coeditor for the IEEE Transactions on Instrumentation and Measurement.