Domain-Incremental Learning for Remote Sensing Semantic Segmentation With Multifeature Constraints in Graph Space

Wubiao Huang, Mingtao Ding, *Member, IEEE*, and Fei Deng

Abstract—**The use of deep learning techniques for semantic segmentation in remote sensing has been increasingly prevalent. Effectively modeling remote contextual information and integrating high-level abstract features with low-level spatial features are critical challenges for semantic segmentation tasks. This article addresses these challenges by constructing a graph space reasoning (GSR) module and a dual-channel cross-attention upsampling (DCAU) module. Meanwhile, a new domain-incremental learning (DIL) framework is designed to alleviate catastrophic forgetting when the deep learning model is used in cross-domain. This framework makes a balance between retaining prior knowledge and acquiring new information through the use of frozen feature layers and multifeature joint loss optimization. Based on this, a new DIL of remote sensing semantic segmentation with multifeature constraints in graph space (GSMF-RS-DIL) framework is proposed. Extensive experiments, including ablation experiments on the ISPRS and LoveDA datasets, demonstrate that the proposed method achieves superior performance and optimal computational efficiency in both single-domain and cross-domain tasks. The code is publicly available at https://github.com/HuangWBill/GSMF-RS-DIL.**

Index Terms—**Cross attention, domain-incremental learning (DIL), graph space reasoning (GSR), remote sensing image, semantic segmentation.**

I. Introduction

SEMANTIC segmentation of remote sensing images is a key technology for ground object classification and a fundamental task in the field of remote sensing. With the rapid development of sensors such as optics, radar, and 3-D scanners, multimodal data such as images and point clouds have become the forefront of remote sensing, especially in target segmentation tasks [1], [2]. The performance bottleneck of single-modal semantic segmentation can be a breakthrough by integrating the advantages of various data sources, so as to obtain more diverse feature information. In recent years, solving remote sensing image problems by intelligent methods has become a hot research focus [3], [4]. Deep learning, as a data-driven technique, has been successfully applied in the fields of land-use change [5], [6], land cover mapping [7], [8], and landslide hazard identification [9], [10].

An inherent challenge in semantic segmentation tasks is that when individual pixels are considered in isolation, the pixels are difficult to classify due to the local image being fuzzy and noisy. Therefore, the model must be able to efficiently capture contextual information. Commonly used approaches are deep learning based on convolutional networks [11], [12]. However, a single convolutional layer can only capture local information due to the inherent limitation of its receptive field. The current approach that aggregates global context information by stacking multiple convolutional layers or using dilated convolutions [13] performs poorly on small objects. Some researchers have addressed this shortcoming by fusing multiscale features within the network [14], [15] or using transformer layers to model long-range dependencies [16], [17]. Recently, self-attention-based methods [18], [19] have been used to learn affinity maps for spatial locations and propagate information to neighboring spatial locations. However, pixels of the same object are not necessarily distributed in the same region, making it difficult to establish dependencies. To solve this issue, we project the feature representations into graph space, and we use the graph convolutional networks (GCNs) to effectively model contextual information for semantic segmentation through global relational inference.

Attention-based strategies have been widely used for deep learning feature fusion. Li et al. [20] efficiently fused spatial and spectral domains for hyperspectral image enhancement by constructing a cascaded attention module. Xia et al. [21] proposed a two-stage hierarchical cross-attention transformer module to fuse point-wise and voxel-wise features in a point cloud scene. To better leverage both deep and shallow features, we construct a dual-channel cross-attention upsampling (DCAU) module to replace the conventional direct upsampling

Received 6 July 2024; revised 4 September 2024 and 28 September 2024; accepted 14 October 2024. Date of publication 21 October 2024; date of current version 31 October 2024. This work was supported in part by the Jianbing Program of Zhejiang under Grant 2023C01040, in part by the Key Research and Development Projects in Hubei Province under Grant 2022BAA035, in part by the National Natural Science Foundation of China under Grant 42374027, in part by the National Key Research and Development Program of China under Grant 2021YFC3000400, in part by the Opening Fund of Key Laboratory of Smart Earth under Grant KF2023YB04-01, in part by Key Research and Development Program Projects in Zhejiang Province under Grant 2023C03177, and in part by the Fundamental Research Funds for the Central Universities under Grant 300102262203. *(Corresponding author: Fei Deng.)*

Wubiao Huang is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: huangwubiao@whu.edu.cn).

Mingtao Ding is with the College of Geological Engineering and Geomatics, Chang'an University, Xi'an 710054, China, also with the Key Laboratory of Loess, Xi'an 710054, China, and also with the Key Laboratory of Western China's Mineral Resource and Geological Engineering, Ministry of Education, Xi'an 710054, China (e-mail: mingtaoding@chd.edu.cn).

Fei Deng is with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China, and also with the Luojia Laboratory Hubei, Wuhan 430079, China (e-mail: fdeng@sgg.whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3481875

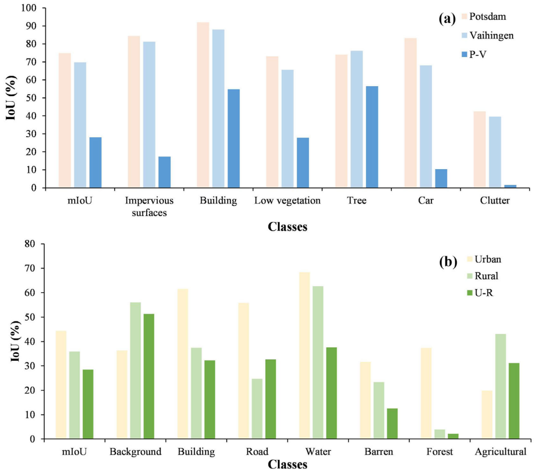

Fig. 1. Catastrophic forgetting in deep learning models. (a) ISPRS datasets and (b) LoveDA datasets. The light rectangle represents the *mIoU* values directly trained on a single domain and the *IoU* values of each class, while the dark rectangle represents the segmentation results of the second domain directly using the model trained on the first domain.

summation method, achieving richer and more representative multiscale feature representations.

In addition, most of the deep learning models are trained once using specific datasets and are typically domain-specific, performing poorly when applied to new domains. As shown in Fig. 1, the prediction results deteriorate significantly when a model trained on ISPRS Potsdam and LoveDA Urban scenarios is directly applied to new ISPRS Vaihingen and LoveDA Rural scenarios. Some researchers have proposed to use comparative learning to solve the problem of domain adaptation. Hong et al. [22] designed a high-resolution domain adaptive network that leverages adversarial learning to bridge the gap between different scenarios, facilitating effective knowledge transfer. Shen et al. [23] proposed an unsupervised domain adaptive method for the cross-domain semantic segmentation task of point clouds. However, this approach primarily focuses on the model's ability on a new task, often leading to a significant drop in performance on the old task, a phenomenon known as catastrophic forgetting [24]. This occurs because the traditional models assume that the data distribution is smooth, allowing the model to repeatedly see the same data across tasks. However, with the introduction of data from new domains, the data distribution becomes non-smooth. Continuous learning from this non-smooth distribution causes new knowledge to interfere with old knowledge, leading to a rapid decrease in model performance or even completely forgetting the previously learned knowledge [25], [26]. In the task of remote sensing semantic segmentation, different conditions, such as temporal changes, resolution variations, and weather conditions, can cause catastrophic forgetting. The simplest solution to catastrophic forgetting is to retrain the network parameters using all known data to adapt to distribution changes. Although retraining from scratch can completely solve catastrophic forgetting, it is highly inefficient. The emergence of large models and incremental learning has shown the potential in mitigating this challenge to some extent. Currently, several remote sensing large models, such as SpectralGPT [27] and Skysense [28], mainly focus on remote sensing image understanding. The research on downstream tasks is in the early stages because the training of large models is more demanding on equipment. Therefore, this article focuses on the incremental learning perspective. Incremental learning enables the model to learn features from the new domain without accessing the old domain data while preserving the performance on the old domain. It should be noted that this article focuses on the domain-incremental learning (DIL) for different domains with the same classes, which often occurs in the field of remote sensing.

To address the above problems, this article proposes a DIL framework of remote sensing semantic segmentation with multifeature constraints in graph space (GSMF-RS-DIL). The primary contributions of this article are as follows.

1) A novel DIL framework of remote sensing image semantic segmentation with multifeature constraints is designed. This framework addresses catastrophic forgetting in deep learning models without changing the model architecture by utilizing multifeature constraints loss.
2) A remote sensing image semantic segmentation model based on graph space transformation and attention upsampling is proposed. The graph space reasoning (GSR) module enlarges the receptive field, enhancing the modeling of global features. In addition, the DCAU module effectively leverages the disparity between high-level and low-level features to capture both global and local contextual information.
3) Extensive experiments are conducted on two representative datasets for both single-domain semantic segmentation and DIL to validate the performance of the proposed method.

The remainder of this article is organized as follows. Section II introduces the related work in graph-based semantic segmentation and DIL. Section III describes the details of the GSMF-RS-DIL framework. The experimental settings are presented in Section IV. Section V analyzes the experimental results and Section VI concludes this article.

II. Related Work

A. Graph-Based Semantic Segmentation

Graph-based methods have become very popular in recent years and demonstrate effectiveness in relational reasoning. Unlike traditional methods, graph-based methods need to construct a graph structure, followed by feature aggregation within the graph space. Currently, graph-based semantic segmentation networks are mainly divided into two types: one is the graph neural network combined with super-pixel segmentation, and the other is constructing a graph neural network in the feature space.

The former divides an image into many super-pixel blocks, where pixels within each block exhibit strong correlations. This method utilizes super-pixels instead of pixels as nodes, significantly reducing computational complexity. For instance, He et al. [29] integrated this approach with a multiscale DenseAtrousCNet, proposing a new two-stream deep neural network for remote sensing image semantic segmentation. Liu et al. [30] used convolutional neural network (CNN) and GCN branches to generate complementary spectral–spatial features at both pixel and super-pixel levels for hyperspectral image classification. Jian et al. [31] constructed spectral and spatial correlation graph structures based on super-pixel segmentation results and hyperspectral images and proposed the uncertainty-aware graph self-supervised learning method for unsupervised contrastive learning. Despite significantly reducing graph structure complexity, this approach often necessitates preliminary super-pixel segmentation and final results influenced by the segmentation accuracy.

The latter transforms features extracted by the convolutional network into graph space, performs graph convolution, and then recovers to the original space. For instance, Zhang et al. [32] achieved this transformation by modeling spatial relationships between pixels and interdependencies between channel dimensions. Chen et al. [33] introduced a global reasoning unit, which aggregates features globally over coordinate space and projects them into an interaction space for efficient relational reasoning, subsequently back-projecting relationship-aware features to the original coordinate space. Li et al. [34] proposed an improved Laplace formulation for graph inference. In addition, Lu et al. [35] constructed a graph model by using feature map pixels as vertices and defining region relationships as edges based on distances between features and self-attention mechanisms to establish associations and uncover object relationships. Jiang et al. [36] used K-nearest neighbors to construct the graph structure based on the assumption that two pixels may belong to the same category if they are similar in appearance and close in metric distance. Liu et al. [37] introduced the self-construct graph module, seamlessly integrating GNN and CNN by transforming the features into latent space through an encoding module and learning node similarities via a decoding module. This article adopts the idea of graph space transformation from the latter.

B. Domain-Incremental Learning

Incremental learning consists of three primary scenarios: task-incremental learning (TIL), DIL, and class-incremental learning (CIL) [38].TIL involves tasks with nonintersecting data labeling spaces, where task identifiers are provided during both training and testing. DIL entails tasks with the same data labeling spaces but differing input distributions. CIL tasks have disjointed data labeling spaces, and task identifiers are provided only during training. Currently, most of the research focuses on CIL tasks [39], [40], and fewer studies addressing DIL. This article primarily addresses the DIL problem.

The incremental learning methods are primarily divided into replay-based methods, regularization-based methods, and parameter isolation-based methods [25]. Replay-based methods are storing a portion of old data or training additional generators to generate pseudo-data for co-training with the new data [41], [42], which can generate storage pressure. Regularization-based methods utilize knowledge distillation or regularization terms in the loss function to balance the old and new tasks. This approach does not require previous data storage and extensive parameter introduction [43] but may result in suboptimal model convergence and poor performance. The LwF method proposed by Li and Hoiem [44] is one of the most representative methods and is currently the most commonly used. Parameter isolation-based methods usually freeze critical model parameters from old tasks and allow the model to introduce new parameters for subsequent tasks [45], [46]. This approach typically increases parameter count and computational effort.

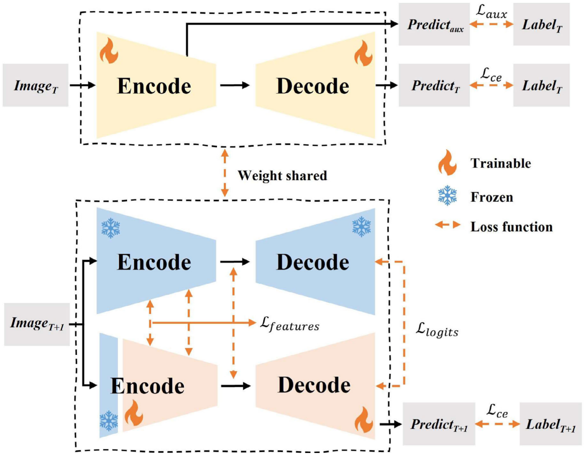

Fig. 2. Overall framework of GSMF-RS-DIL. The graph consists of two parts with shared weights. The gray boxes represent input and output data. The yellow trapezoid represents the single-domain training stage composed of trainable parameter networks, while the blue and orange represent the frozen and trainable parts of the DIL stage, respectively.

The GSMF-RS-DIL framework proposed in this article combines parameter isolation and regularization methods. This hybrid approach has shown promising results in recent studies. For example, Garg et al. [45] introduced a domain residual fitness block and distillation loss on top of lightweight efficient residual factorized ConvNet to solve the problem of DIL for semantic segmentation in computer vision. Rui et al. [26] applied this approach to the domain-incremental problem in remote sensing images. Michieli and Zanuttigh [47] incorporated distillation loss at different locations based on the DeepLabV2 network and explored the effect of different freezing parameters. In addition, Douillard et al. [43] and Kirkpatrick et al. [48] investigate how to improve the performance of DIL from the view of distillation loss function constraints. Beyond computer vision, this approach has also been applied in fire recognition [46], change detection [49], and medical segmentation [50].

III. Methodology

A. Overview

The overall flowchart of the GSMF-RS-DIL framework proposed in this article is shown in Fig. 2. Assuming that there are two different domains, T and $T+1$, the corresponding data samples for each domain are denoted as Image$_T$ and Image$_{T+1}$, and the corresponding labels are Label$_T$ and Label$_{T+1}$. The classes in Label$_T$ and Label$_{T+1}$ remain consistent.

First, the graph space semantic segmentation model is trained for domain T. To model performance, an auxiliary head loss function is added based on the original cross-entropy loss function [15]. Therefore, the prediction result of domain T includes the main output Predict$_T$ and the auxiliary output Predict$_{\text{aux}}$. Subsequently, the DIL model is trained on domain $T+1$ to obtain the output Predict$_{T+1}$. The incremental learning model consists of two branches: the first branch freezes the

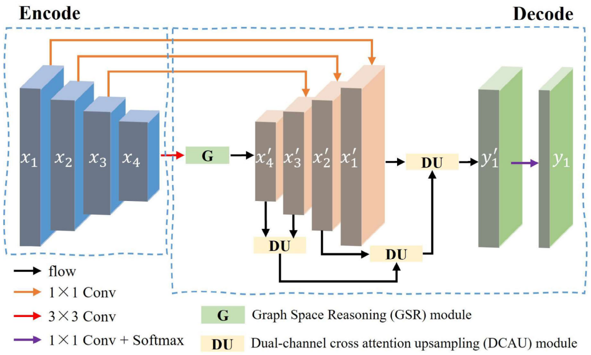

Fig. 3. Detailed architecture of the GS-AUFPN model.

model M trained on the first domain, while the second branch uses the weights of the model M as the pretraining weights and continues training. In addition, the encoder's first-stage feature extraction process is frozen in the second branch. During the entire DIL process, only the sample data and labels from domain $T+1$ are used for small batch training, and the feature layer loss function $\mathcal{L}_{\text{feature}}$, cross-entropy loss function $\mathcal{L}_{\text{ce}}$, and output layer distillation loss $\mathcal{L}_{\text{logits}}$ jointly constrain the training process.

The proposed graph space attention upsampling feature pyramid network (GS-AUFPN) adopts an encoder–decoder architecture, and the detailed network structure is shown in Fig. 3. The model uses the feature pyramid [51] as the main architecture, incorporating the classical ResNet-101 model as the encoder to obtain the outputs of four stages (x_1, x_2, x_3, x_4). A 3 × 3 convolution is applied to the deepest feature x_4 according to (1), followed by global feature modeling using the GSR module to obtain x'_4. Feature channel compression is performed using 1 × 1 convolution for each of the other three stages of features, resulting in (x'_1, x'_2, x'_3, x'_4) with the same number of feature channels. In addition, the original simple upsampling aggregation module is replaced with a DCAU module to synthesize the recovered image sizes of the features at each stage, yielding the feature map y'_1. The decoder output y_1 is obtained after 1 × 1 convolution and Softmax operation on the feature map y'_1 according to (2), which is the number of channels matching the number of classes

$$x'_i = \begin{cases} G(\text{Conv}(x_i, 3)), i = 4 \\ \text{Conv}(x_i, 1), i = 1, 2, 3 \end{cases} \tag{1}$$

$$y_1 = \text{Softmax}\big(\text{Conv}\big(y'_1, 1\big)\big). \tag{2}$$

B. GSR Module

To better model the global contextual relationships, this article transforms traditional feature graphs into a graph feature space and reasons about relationships on the graph. Inspired by Chen et al. [33], we propose the GSR module. As shown in Fig. 4, this module consists of three parts: projection to graph space, graph space convolution, and graph space back projection. To describe the process, we predefine the input deep features $\mathcal{F}$ with size (B, C, H, W), where B is the batch size, C is the number of channels, and H and W are the height and width, respectively. N is the number of graph nodes, and C' is the length of graph node features. The specific process is as follows.

1) Projection to Graph Space: The graph structure consists of nodes, edges, and node features, of which the most critical is the construction of graph nodes. For the input deep feature $\mathcal{F} \in \mathbb{R}^{B\times C\times H\times W}$, the feature F is projected to the graph space $\mathcal{S} \in \mathbb{R}^{B\times N\times C'}$ by constructing a projection function $f(\cdot)$ to obtain N nodes of length C. $f(\cdot)$ consists of a region aggregation function $\varphi(\cdot)$ and a dimension compression function $\theta(\cdot)$. To construct a learnable projection function, we model $\varphi(\cdot)$ and $\theta(\cdot)$ using 1 × 1 convolutional layers, respectively, computed as follows:

$$\varphi(\mathcal{F}) = \text{reshape}(\text{Conv}(\mathcal{F}, N, 1), (B, N, HW\ )) \tag{3}$$

$$\theta(\mathcal{F}) = \text{reshape}\big(\text{Conv}\big(\mathcal{F}, C', 1\big), \big(B, HW, C'\big)\big) \tag{4}$$

$$\mathcal{S} = f(\mathcal{F}) = \varphi(\mathcal{F}) \cdot \theta(\mathcal{F}). \tag{5}$$

2) Graph Space Convolution: To model the relationships between nodes and update the feature vectors, we apply the efficient graph convolution proposed by Kipf and Welling [52]. Specifically, the graph convolution process is defined as follows:

$$\mathcal{K} = Q\mathcal{S}W \tag{6}$$

where $\mathcal{K} \in \mathbb{R}^{B\times N\times C'}$ is the feature after graph convolution, Q is the node neighborhood matrix of size $N \times N$, and$W$ is the state update weight.

In the actual modeling process, the implementation is carried out by 1-D convolution in both the feature and node directions, as shown in the following equation:

$$\mathcal{K} = \text{Conv1D}\big((\text{Conv1D}(\mathcal{S}) + \mathcal{S})^T\big)^T. \tag{7}$$

3) Graph Space Back Projection: For $\mathcal{K} \in \mathbb{R}^{B\times C'\times N}$ after graph convolution, it needs to be backprojected to the original feature space to facilitate the subsequent decoding module. Similar to the "Projection to graph space," the feature $\mathcal{T}$ is backprojected to the original space $x'_4 \in \mathbb{R}^{B\times C\times H\times W}$ by constructing a projection function $g(\cdot)$. $g(\cdot)$ computes the dot product of $\mathcal{K}$ and $\varphi(\mathcal{F})$ and recovers the feature channel from C' to C by a 1 × 1 convolution, computed as follows:

$$x'_4 = g(\mathcal{K}) = \text{Conv}(\text{reshape}(\mathcal{K} \cdot \varphi(\mathcal{F})), C, 1) \tag{8}$$

where $\varphi(\mathcal{F})$ is the reuse of the region aggregation function in the first step, because it retains the positional coding information of the graph nodes and the original feature space, and $\varphi(\mathcal{F})$ reuse can reduce the number of parameters.

C. DCAU Module

To better exploit the intrinsic relationship between shallow and deep features, this article improves the feature fusion module for sampling features on the feature pyramid by introducing DCAU module.

As shown in Fig. 5, the module consists of two attention branches, which process the input low-level features ($\mathcal{X} \in \mathbb{R}^{B\times C\times W\times H}$) and high-level features ($\mathcal{Y} \in \mathbb{R}^{B\times C\times (W/2)\times (H/2)}$), respectively. First, the high-level features $\mathcal{Y}$ upsampling match the size of $\mathcal{X}$, resulting in $\mathcal{Y}'\in\mathbb{R}^{B\times C\times W\times H}$. Then, global average pooling (GAP) and two 1 × 1 convolution operations are applied to generate the

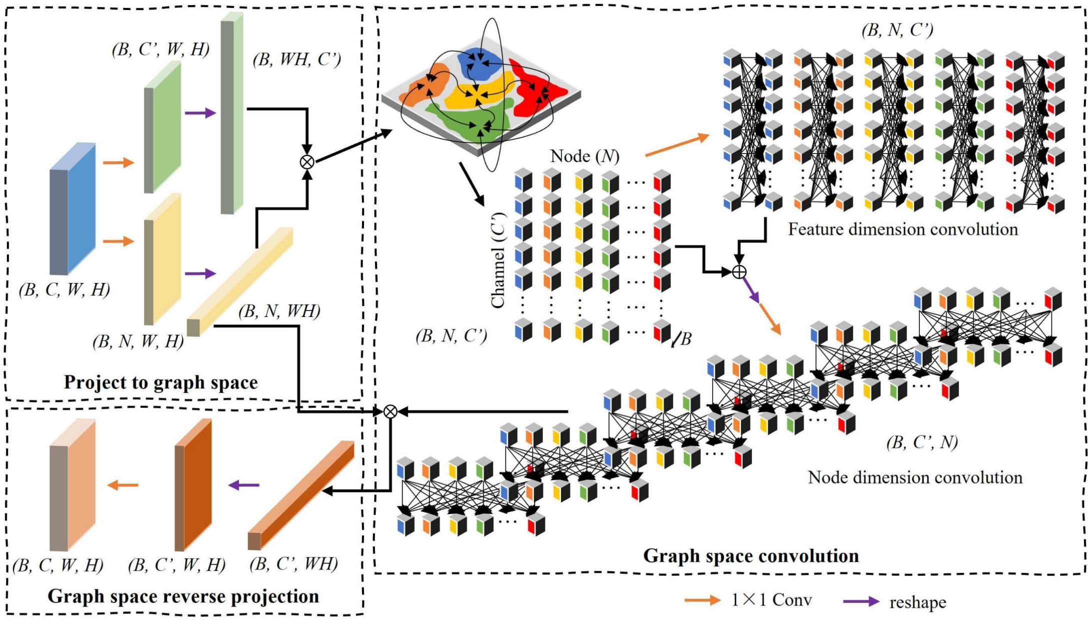

Fig. 4. GSR module architecture, including projection to graph space, graph space convolution, and graph space back projection. The different colors represent the features of different stages.

global context information of the deep features, resulting in $\mathcal{Y}^C \in \mathbb{R}^{B\times C\times 1\times 1}$. For $\mathcal{X}$, the average and maximum pooling are first performed, followed by a convolution operation to compress the channel dimensions, yielding the edge information spatial attention (SA) map $\mathcal{X}^S \in \mathbb{R}^{B\times 1\times W\times H}$. Subsequently, $\mathcal{Y}^C$ and $\mathcal{X}^S$ are multiplied with $\mathcal{X}$ and $\mathcal{Y}'$, respectively, to obtain $\mathcal{X}^C$ and $\mathcal{Y}^S$. Finally, $\mathcal{X}^C$, $\mathcal{Y}^S$, X, and $\mathcal{Y}'$ are summed to obtain the final fused feature map $\mathcal{Z}$. The specific formulas are as follows:

$$\mathcal{Y}^C = \text{Sigmoid}\left(\text{Conv}\left(\text{Conv}\left(\text{GAP}(\mathcal{Y}'), \frac{C}{r}, 1\right), C, 1\right)\right) \tag{9}$$

$$\begin{aligned}\mathcal{X}^S = {} & \text{Sigmoid}(\text{Conv}(\text{Concatenate} \\ & \times(\text{Maxpool}(\mathcal{X}), \text{Avepool}(\mathcal{X})), 1, j))\end{aligned} \tag{10}$$

$$\begin{aligned}\mathcal{Z} = {} & \mathcal{X} + \mathcal{Y}' + \mathcal{X}^C + \mathcal{Y}^S = \mathcal{X} + \mathcal{Y}' + \left(\mathcal{X} \times \mathcal{Y}^C\right) \\ & + \left(\mathcal{Y}' \times \mathcal{X}^S\right)\end{aligned} \tag{11}$$

where r is the channel attenuation coefficient, affecting the feature learning capability, and j is the convolution kernel size, determining spatial information aggregation ability.

D. Loss Functions

As shown in Fig. 2, the loss function used in the GSMF-RS-DIL framework is mainly divided into two parts, and the training and optimization strategies for each stage are independent of each other.

The first part is the semantic segmentation loss function $\mathcal{L}_{\text{SS}}$, used for optimizing the graph space semantic segmentation model over domain T. It consists of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{aux}}$

$$\mathcal{L}_{\text{SS}} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{aux}} \tag{12}$$

$$\mathcal{L}_{\text{ce}} = \mathcal{L}_{\text{aux}} = -\frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K} y_k^{(s)} \log \hat{y}_k^{(s)} \tag{13}$$

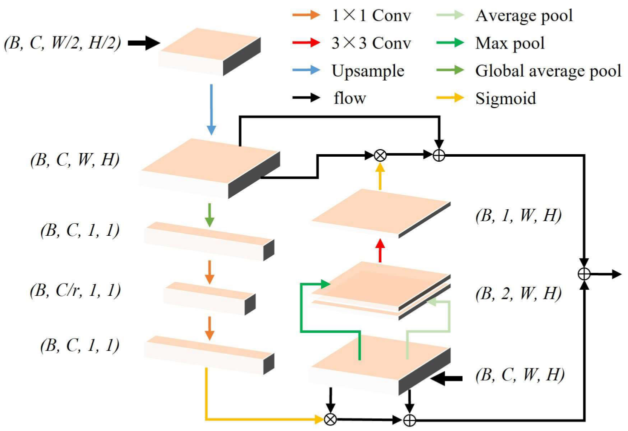

Fig. 5. Detailed architecture of the DCAU module.

where $s \in [1, 2, \ldots, S]$, S is the number of sampling points, and k is the number of object class. $\hat{y}_k^{(n)}$ is the one-hot value of the sample's prediction result, and $y_k^{(n)}$ is the true label value corresponding to this sample. λ is the weight parameter for balancing $\mathcal{L}_{\text{aux}}$ and is set to 0.4 in this article according to previous studies.

The second part is the DIL loss function $\mathcal{L}_{\text{DIL}}$ for model optimization over domain $T + 1$. It consists of $\mathcal{L}_{\text{ce}}$, $\mathcal{L}_{\text{logits}}$, and $\mathcal{L}_{\text{feature}}$

$$\mathcal{L}_{\text{DIL}} = \mathcal{L}_{\text{ce}} + \alpha \mathcal{L}_{\text{logits}} + \beta \mathcal{L}_{\text{feature}} \tag{14}$$

$$\mathcal{L}_{\text{logits}} = -\frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K} \frac{y_{1k}^{(s)}}{T_1} \log \frac{\hat{y}_{1k}^{(s)}}{T_1} \tag{15}$$

$$\mathcal{L}_{\text{feature}} = \frac{1}{I}\sum_{i=1}^{I} \frac{1}{S} T_2^2 \sum_{s=1}^{S}\sum_{k=1}^{K} KL\left(\frac{x_{ik}^{(s)}}{T_2}, \log \frac{\hat{x}_{ik}^{(s)}}{T_2}\right) \tag{16}$$

where $i \in I$, I is the encoder stage number used for computation, T_1 and T_2 are the temperature coefficients of $\mathcal{L}_{\text{logits}}$ and $\mathcal{L}_{\text{feature}}$, respectively, used to regulate the smoothing of the probability distribution. KL$(\cdot)$ denotes the computation of Kullback–Leibler divergence [53]. $\hat{y}_{1k}^{(s)}$ is the output logits of the trainable branch, and $y_{1k}^{(s)}$ is the output logits of the frozen branch. $\hat{x}_{ik}^{(s)}$ is the softmax value of the output feature map of the trainable branch encoder, and $x_{ik}^{(s)}$ is the softmax value of the output feature map of the frozen branch encoder. α and β are the balancing parameters for $\mathcal{L}_{\text{logits}}$ and $\mathcal{L}_{\text{feature}}$ weights.

IV. Experimental Settings

A. Datasets

We evaluate the performance of the proposed framework by experiments on two well-known open-source datasets: ISPRS datasets (http://www2.isprs.org/commissions/comm3/wg4/semantic-label-ing.html) and LoveDA datasets (https://github.com/Junjue-Wang/LoveDA). We conducted single-domain semantic segmentation experiments on ISPRS Potsdam scene and LoveDA Urban scene and then applied the trained models to the ISPRS Vaihingen scene and LoveDA Rural scene for cross-domain incremental learning experiments.

1) ISPRS Datasets: The datasets include the Vaihingen and Potsdam scenarios. The Potsdam scene is a typical historical city with large building clusters, narrow streets, and dense settlement structures. It consists of 38 remote sensing images with a spatial resolution of 0.05 m, each sized at 6000 × 6000 pixels. The Vaihingen scene is a small village with numerous individual buildings and small multistory buildings. It consists of 33 remote sensing images of different sizes, with a spatial resolution of 0.09 m. These scenarios have the same semantic labels: impervious surfaces, building, low vegetation, tree, car, and background. We used the near-infrared (NIR), red, and green channels and did not use other data. For the Potsdam scene, we utilized images with IDs: 2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11,4_12, 5_10, 5_11, 5_12, 6_10, 6_11, 6_12, 6_7,6_8, 6_9, 7_10, 7_11, 7_12, 7_7, 7_8, and 7_9 for training, and the remaining 14 images are used for testing. For the Vaihingen scene, we used images with IDs: 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, and 37 for training and the remaining 17 images for testing.

2) LoveDA Datasets: The LoveDA datasets [54] are challenging datasets with complex backgrounds. The dataset includes Google Earth images acquired in July 2016 from three cities: Nanjing, Changzhou, and Wuhan. Each image has a spatial resolution of 0.3 m and a size of 1024 × 1024 pixels. It includes the Urban and Rural scenarios and is classified into seven classes: buildings, road, water, barren, forest, agriculture, and background. The dataset provides three bands: red, green, and blue. The Urban scene training set comprises 1156 images, and the test set has 677 images. The Rural scene training set includes 1366 images, with 992 images in the test set.

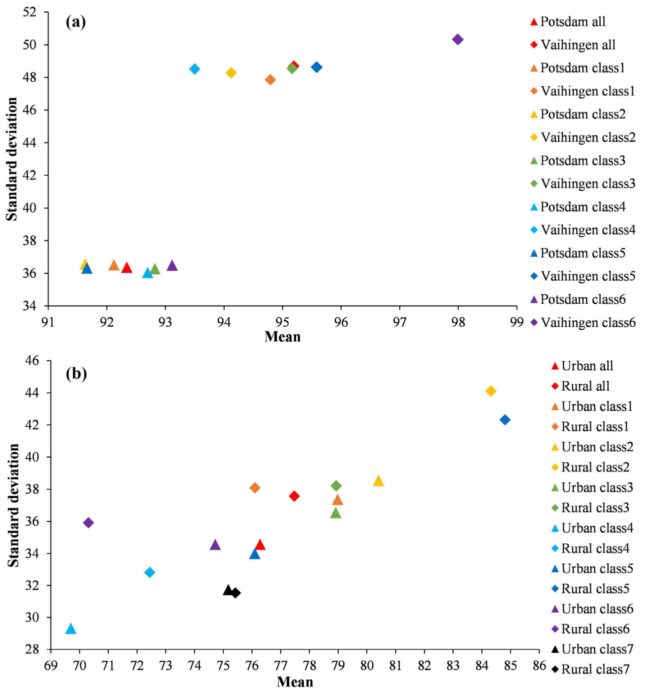

Fig. 6. Visualization results of the mean and standard deviation of all classes' bands in different scenes. (a) ISPRS datasets and (b) LoveDA datasets.

Fig. 6 shows the mean and variance of the bands for different classes of these two datasets. The same color scheme represents the same category, while the same shape indicates the same dataset. It can be seen that the distribution of different classes within the same scene is more concentrated in the ISPRS datasets, and the difference between different scenes is more obvious. Conversely, the distribution of different classes within the same scene and between different scenes is more dispersed in the LoveDA datasets.

B. Implementation Details

All experiments were conducted on a Linux PC with NVIDIA GeForce RTX 4090 24 G GPU installed. All code was implemented based on the PyTorch deep learning framework. The backbone network utilized the ResNet-101 model pretrained on the ImageNet dataset. The 7 × 7 convolution in the input layer was replaced by three 3 × 3 convolutions, and the last two downsampling operations were replaced by convolutional layers with dilation rates of 2 and 4. Due to hardware constraints, the batch size was set to 4.

For single-domain semantic segmentation model training, the maximum number of iterations was set to 80k for the ISPRS Potsdam scene dataset and 30k for LoveDA Urban scene dataset. The AdamW optimizer with weight decay was used, the initial learning rate was set to 0.0001, and the weight decay was set to 0.001. A "poly" learning rate strategy was used, with the formula $\text{lr} = \text{base\_lr} \times (1 - (\text{iteration}/\text{max\_iteration}))^{\text{power}}$, where base_lr denotes the initial learning rate, iteration denotes the current number of iterations, max_iteration denotes the total number of iterations, and power was set to 0.9. When training the incremental learning model cross domains, the maximum number of iterations was set to 10k. The initial learning rate of the optimizer was set to 0.00001.

Since the dataset images were too large, we cropped them to 512 × 512 pixels, with 128 overlapping pixels for the ISPRS dataset and no overlap for the LoveDA dataset.

TABLE I

EVALUATION METRICS FOR SINGLE-DOMAIN SEMANTIC SEGMENTATION ABLATION EXPERIMENTS ON THE ISPRS POTSDAM SCENE DATASET

	$F1/IoU$ (%)						OA	$mF1$	$mIoU$
	Impervious surfaces	Building	Low vegetation	Tree	Car	Clutter	(%)	(%)	(%)
Baseline	91.08/83.61	95.73/91.81	83.81/72.13	85.24/74.27	90.82/83.18	42.82/27.25	88.06	81.58	72.04
+ GSR	91.35/84.08	**95.83/92.00**	84.19/72.70	84.79/73.59	90.66/82.91	55.93/38.82	88.46	83.79	74.02
+ ASPP	91.31/84.01	95.68/91.72	84.27/72.82	84.95/73.84	90.81/83.17	55.03/37.96	88.44	83.68	73.92
+ PPM	91.32/84.02	**95.83/92.00**	84.22/72.74	85.23/74.26	90.78/83.11	54.06/37.05	88.51	83.57	73.86
+ CA	91.32/84.02	95.70/91.76	83.96/72.36	85.02/73.95	90.79/83.13	51.36/34.56	88.32	83.03	73.29
+ SA	91.27/83.95	95.77/91.88	83.84/72.17	85.21/74.23	90.77/83.10	48.65/32.15	88.24	82.59	72.91
+ DCAU	91.34/84.07	95.70/91.76	84.27/72.81	**85.34/74.43**	**90.93/83.37**	51.65/34.82	88.46	83.21	73.54
+ GSR + DCAU	**91.55/84.41**	95.80/91.94	**84.47/73.12**	85.10/74.07	90.82/83.19	**59.62/42.47**	**88.72**	**84.56**	**74.87**

TABLE II

EVALUATION METRICS FOR SINGLE-DOMAIN SEMANTIC SEGMENTATION ABLATION EXPERIMENTS ON THE LOVEDA URBAN SCENE DATASET

	$F1/IoU$ (%)							OA	$mF1$	$mIoU$
	Background	Building	Road	Water	Barren	Forest	Agricultural	(%)	(%)	(%)
Baseline	52.30/35.41	73.23/57.76	72.20/56.50	74.25/59.05	**51.82/34.97**	57.41/40.26	15.37/8.33	54.46	56.66	41.75
+ GSR	53.60/36.61	75.16/60.21	71.32/55.43	78.19/64.19	49.53/32.92	56.67/39.54	29.78/17.50	56.95	59.18	43.77
+ ASPP	**55.06/37.99**	75.10/60.13	**73.67/58.32**	80.74/67.70	43.58/27.86	51.03/34.26	28.21/16.42	56.75	58.20	43.24
+ PPM	53.22/36.26	75.79/61.02	71.91/56.14	80.99/68.05	43.81/28.05	56.28/39.16	27.75/16.11	56.75	58.54	43.54
+ CA	51.67/34.84	75.87/61.12	73.64/58.27	82.26/69.86	49.57/32.95	**59.28/42.13**	8.59/4.49	55.06	57.27	43.38
+ SA	52.33/35.44	76.06/61.37	72.43/56.78	**82.35/70.00**	48.15/31.70	58.78/41.63	12.06/6.42	55.37	57.45	43.33
+ DCAU	52.51/35.60	**76.52/61.96**	73.14/57.65	81.85/69.28	50.55/33.83	57.76/40.61	15.40/8.34	55.91	58.25	43.90
+ GSR + DCAU	53.27/36.31	76.19/61.53	71.62/55.79	81.21/68.36	48.03/31.60	54.42/37.38	**33.22/19.92**	**57.43**	**59.71**	**44.41**

During training, we used random scaling (with scales of [0.5,0.75,1.0,1.25,1.5,1.75,2.0]), random cropping, and random flipping for data enhancement.

C. Evaluation Metrics

1) Evaluation Metrics on Single-Domain Semantic Segmentation: To comprehensively evaluate the performance of our proposed model, overall accuracy (OA), mean $f1$-score (m$F1$), and mean intersection over union (mIoU) were used as evaluation metrics. Based on the accumulated confusion matrix, OA, m$F1$, and mIoU are computed as follows:

$$\mathrm{OA} = \frac{\sum_{k=1}^{N} \mathrm{TP}_k}{\sum_{k=1}^{N} \mathrm{TP}_k + \mathrm{FP}_k + \mathrm{TN}_k + \mathrm{FN}_k} \tag{17}$$

$$\mathrm{mF1} = \frac{1}{N}\sum_{k=1}^{N} \frac{2 \times \mathrm{Precision}_k \times \mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k} \tag{18}$$

$$\mathrm{Precision}_k = \frac{\mathrm{TP}_k}{\mathrm{FP}_k + \mathrm{TP}_k} \tag{19}$$

$$\mathrm{Recall}_k = \frac{\mathrm{TP}_k}{\mathrm{FN}_k + \mathrm{TP}_k} \tag{20}$$

$$\mathrm{mIoU} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{IoU}_k = \frac{1}{N}\sum_{k=1}^{N} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{21}$$

where $\mathrm{TP}_k, \mathrm{FP}_k, \mathrm{TN}_k$, and FN_k denote true positives, false positives, true negatives, and false negatives, respectively, for a particular object indexed as class k. N is the number of classes.

2) Evaluation Metrics on Cross-Domain Incremental Learning: Evaluation metrics for multidomain incremental learning tasks should consider performance both old and new domain. Therefore, we measure model performance by calculating the average decrease in the mIoU of the incremental learning model on each domain relative to the corresponding individual task baseline

$$\Delta_{\mathrm{mIoU}} = \frac{1}{B}\sum_{b=1}^{B} \Delta_{\mathrm{mIoU}}^{b} = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathrm{mIoU}_{d,b} - \mathrm{mIoU}_{a,b}}{\mathrm{mIoU}_{a,b}} \tag{22}$$

where B is the number of domains, $\Delta_{\mathrm{mIoU}}^{b}$ is the change of mIoU on the bth domain relative to the corresponding single task, $\mathrm{mIoU}_{a,b}$ is the mIoU value of the baseline model a on the bth domain, and $\mathrm{mIoU}_{d,b}$ is the mIoU value of the domain-incremental model d on the bth domain.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Ablation Study and Parameter Analysis of Single-Domain Semantic Segmentation

1) Ablation Study: In order to validate the performance of each proposed module, we used the most basic feature pyramid as a baseline and gradually added the GSR, channel attention module (CAM), and spatial attention module (SAM), conducting extensive ablation experiments. In addition, the GSR module is compared with the commonly used ASPP [15] and PPM [55] modules. The evaluation metrics of the ablation experiments on the ISPRS Potsdam scene dataset and LoveDA Urban scene dataset are given in Tables I and II,

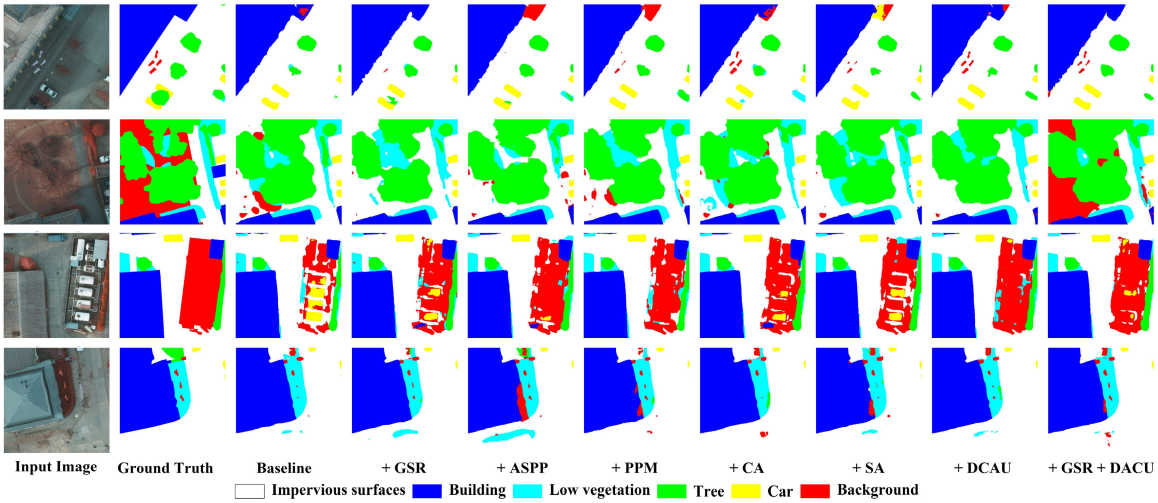

Fig. 7. Visualization results of some test samples for different modules on the ISPRS Potsdam scene dataset.

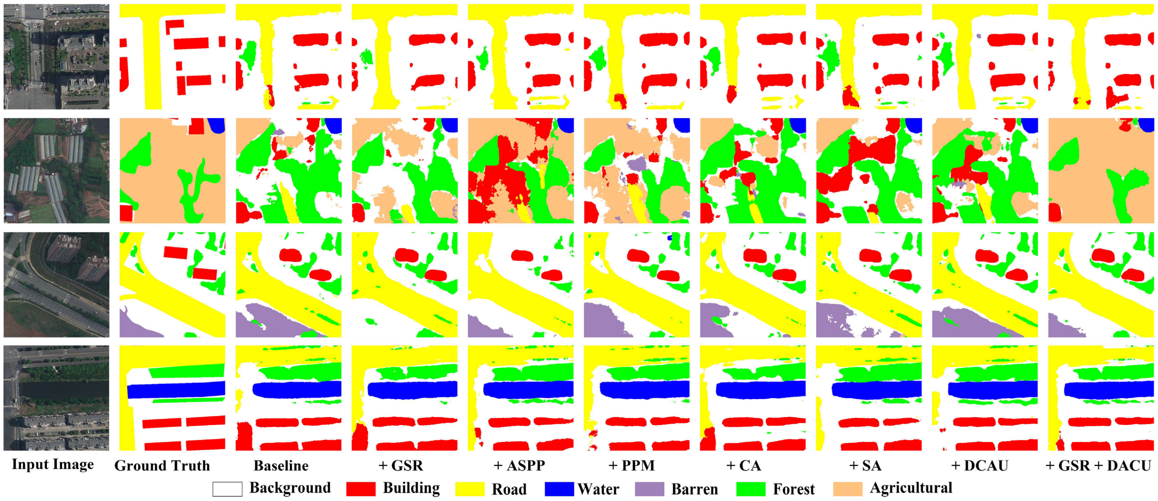

Fig. 8. Visualization results of some test samples for different modules on the LoveDA Urban scene dataset.

respectively. Figs. 7 and 8 illustrate the visualization results for some test samples on the ISPRS Potsdam scene dataset and LoveDA Urban scene dataset, respectively. Fig. 9 shows the floating-point operations (FLOPs) and parametric quantities (Params) for different modules. It can be seen that each proposed module improves the performance of the model to some extent. Specifically, our proposed model achieves an approximate 2.8% improvement in $mIoU$ with only 0.033 G additional FLOPs and 0.129 M additional Params compared to the baseline model.

1) *Ablation of GSR Module:*Compared with the baseline model, the introduction of the GSR module improves model performance in both datasets, enhancing the $mIoU$ metric by approximately 2% in each task. Notably, the GSR module demonstrates significantly lower FLOPs and Params than the ASPP and PPM modules while achieving slightly better performance. This indicates that the GSR module can effectively model the global contextual relationships by converting features into graph space, thereby reducing computational complexity without compromising accuracy.

2) *Ablation of DCAU Module:* Incorporating the DCAU module results in a model performance improvement of about 1.5%–2.0% over the baseline. The DCAU module consists of two branches: SA and channel attention (CA). Our analysis reveals that the inclusion of these branches positively impacts performance, with the CA branch providing a slightly greater improvement than the SA branch. Specifically, the SA branch enhances the segmentation of classes with distinct shape features, while the CA branch is effective in segmenting some complex features.

2) *Parameter Analysis:* In the GS-AUFPN framework, suitable parameter selection can significantly enhance the performance of the model. Therefore, we analyze the settings of these four parameters by the control variable method. The evaluation metrics for the proposed model on the ISPRS Potsdam scene dataset and LoveDA Urban scene dataset with different parameter values are presented in Tables III and IV, respectively. Our findings indicate that the model performance is optimal when N is set to 128, C' is set to 64, r is set to 16, and j is set to 3. When the data samples are relatively

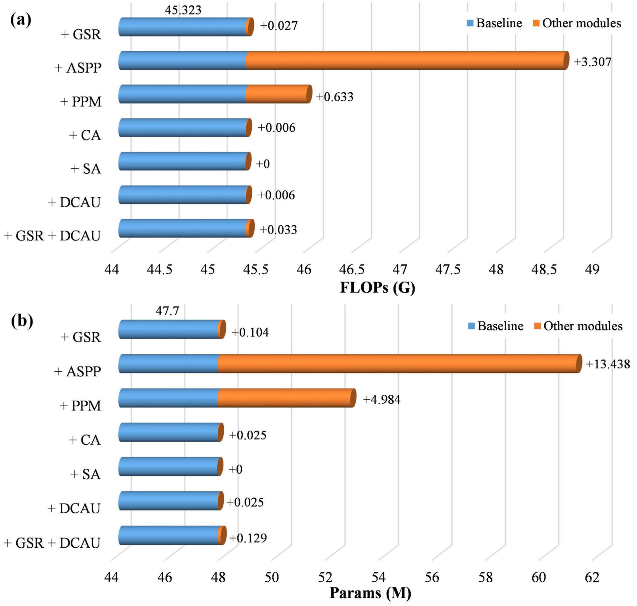

Fig. 9. (a) FLOPs and (b) Params for different modules on the ISPRS Potsdam scene dataset.

TABLE III

EVALUATION METRICS OF THE PROPOSED MODEL ON THE ISPRS POTSDAM SCENE DATASET WITH DIFFERENT PARAMETER VALUES

N	C'	r	j	OA (%)	$mF1$ (%)	$mIoU$ (%)
128	64	8	3	**88.49**	**84.05**	**74.26**
			5	88.49	83.91	74.13
			7	88.47	83.76	74.02
128	64	8	3	88.49	84.05	74.26
		16		**88.72**	**84.56**	**74.87**
		32		88.38	83.96	74.17
128	32	16	3	88.47	83.64	73.90
	64			**88.72**	**84.56**	**74.87**
	128			88.51	83.84	74.08
64	64	16	3	88.59	84.03	74.33
128				**88.72**	**84.56**	**74.87**
256				88.66	84.24	74.54

simple and the dataset is sufficiently large, variations in these parameters have a minimal impact on model performance. Among the parameters, C' has the most significant impact on performance, while j has the least. Conversely, when dealing with more complex data samples, adjustments in these parameters result in more pronounced changes in model performance.

B. Comparison With State-of-the-Art Methods of Single-Domain Semantic Segmentation

In this section, we compare the proposed GS-AUFPN approach with several state-of-the-art semantic segmentation models. These include Pointrend [56], which uses FPN as its framework, feature fusion networks such as PSANet [57], DANet [58], and CCNet [59], which are based on the attention mechanism, and DeepLabV3+ [15] and PSPNet [55], which incorporate the modeling of global contextual relationships.

TABLE IV

EVALUATION METRICS OF THE PROPOSED MODEL ON THE LOVEDA URBAN SCENE DATASET WITH DIFFERENT PARAMETER VALUES

N	C'	r	j	OA (%)	$mF1$ (%)	$mIoU$ (%)
128	64	8	3	**57.98**	**59.52**	**44.25**
			5	57.66	59.24	44.02
			7	56.40	58.34	43.38
128	64	8	3	**57.98**	59.52	44.25
		16		57.43	**59.71**	**44.41**
		32		57.21	59.46	44.07
128	32	16	3	56.56	58.21	43.29
	64			57.43	**59.71**	**44.41**
	128			**57.54**	59.53	44.23
64	64	16	3	56.80	58.44	43.21
128				**57.43**	**59.71**	**44.41**
256				57.37	59.09	43.76

The evaluation metrics on the ISPRS Potsdam scene dataset and LoveDA Urban scene dataset are presented in Tables V and VI, respectively. Figs. 10 and 11 provide the visualization results of some test samples on the ISPRS Potsdam scene dataset and LoveDA Urban scene dataset, respectively. The proposed method achieves best results on the ISPRS Potsdam scene dataset, obtaining OA values of 88.72%, mF1 values of 84.56%, and mIoU values of 74.87%. On the LoveDA Urban dataset, the proposed method attains the highest mF1 value (59.71%) and mIoU value (44.41%), while the OA value (57.43%) is slightly lower than that of the PSANet model (57.60%). In addition, our proposed model is in the upper-middle level for the classification of various objects, particularly achieving optimal performance on the classes with fewer samples and higher complexity.

Fig. 12 shows the scatter plots of the distribution of FLOPs and Params for different methods. It is evident that the GS-AUFPN and Pointrend models, which use FPN as the backbone, have significantly lower FLOPs and Params compared to other models. Although the Params of GS-AUFPN and Pointrend are nearly equal, the FLOPs of GS-AUFPN are fewer. Furthermore, considering the aforementioned evaluation metrics, the performance of the proposed model substantially surpasses that of the Pointrend model.

Although Tables V and VI demonstrate that our proposed method achieves the overall optimal performance in single-domain semantic segmentation, it does not consistently perform well across all classes. Consequently, when conducting specific segmentation tasks, it is necessary to consider the number of samples and the complexity of the features to select the model that performs best for the specific class.

C. Ablation Study and Parameter Analysis of Cross-Domain Incremental Learning

In this section, we discuss the influence of freezing different layers in the second branch of the GSMF-RS-DIL framework and analyze the effect of the parameters T_1, T_2, α, β, and I in $\mathcal{L}_{\text{DIL}}$. Specifically, we freeze only the first stage of the encoder in the second branch of the GSMF-RS-DIL framework. Fig. 13 presents six different incremental learning freezing frames: ① no freezing; ② freezing the first two stages of the encoder;

TABLE V

QUANTITATIVE COMPARISON RESULTS ON THE ISPRS POTSDAM SCENE DATASET WITH THE STATE-OF-THE-ART NETWORKS

Method	$F1/IoU$ (%)						OA (%)	$mF1$ (%)	$mIoU$ (%)
	Impervious surfaces	Building	Low vegetation	Tree	Car	Clutter			
DeepLabV3+	91.44/84.24	**95.95/92.22**	**84.55/73.24**	85.22/74.25	**91.29/83.98**	52.66/35.74	88.65	83.52	73.94
PSANet	91.49/84.32	95.82/91.98	84.27/72.82	85.20/74.21	91.10/83.66	55.63/38.53	88.62	83.92	74.25
PSPNet	**91.60/84.51**	95.88/92.08	84.49/73.15	85.32/74.40	91.05/83.56	54.28/37.25	88.69	83.77	74.16
Pointrend	91.17/83.77	95.83/91.99	83.92/72.29	85.22/74.25	90.69/82.97	45.89/29.78	88.21	82.12	72.51
DANet	91.37/84.12	95.73/91.81	84.26/72.80	85.29/74.36	91.05/83.57	50.36/33.66	88.44	83.01	73.38
CCNet	91.40/84.16	95.79/91.93	84.49/73.14	**85.37/74.48**	91.10/83.66	52.14/35.26	88.56	83.38	73.77
GS-AUFPN (ours)	91.55/84.41	95.80/91.94	84.47/73.12	85.10/74.07	90.82/83.19	**59.62/42.47**	**88.72**	**84.56**	**74.87**

TABLE VI

QUANTITATIVE COMPARISON RESULTS ON THE LOVEDA URBAN SCENE DATASET WITH THE STATE-OF-THE-ART NETWORKS

Method	$F1/IoU$ (%)							OA (%)	$mF1$ (%)	$mIoU$ (%)
	Background	Building	Road	Water	Barren	Forest	Agricultural			
DeepLabV3+	52.54/35.63	75.07/60.09	**74.05/58.79**	81.81/69.22	41.95/26.54	56.82/39.69	28.18/16.40	56.82	58.63	43.76
PSANet	**54.73/37.68**	75.41/60.53	73.57/58.19	81.60/68.92	44.44/28.57	**57.92/40.77**	26.42/15.22	**57.60**	59.16	44.27
PSPNet	54.29/37.26	74.88/59.85	73.50/58.10	80.48/67.34	42.22/26.76	56.38/39.26	26.28/15.12	56.91	58.29	43.38
Pointrend	51.49/34.67	74.87/59.84	73.40/57.97	81.85/69.28	48.56/32.07	57.89/40.74	6.36/3.29	54.43	56.35	42.55
DANet	53.13/36.17	75.32/60.42	72.77/57.19	79.88/66.50	**49.94/33.28**	57.58/40.43	20.40/11.36	56.21	58.43	43.62
CCNet	52.91/35.98	**76.75/62.28**	73.50/58.11	**82.36/70.00**	49.92/33.26	56.56/39.43	15.36/8.32	55.95	58.20	43.91
GS-AUFPN (ours)	53.27/36.31	76.19/61.53	71.62/55.79	81.21/68.36	48.03/31.60	54.42/37.38	**33.22/19.92**	57.43	**59.71**	**44.41**

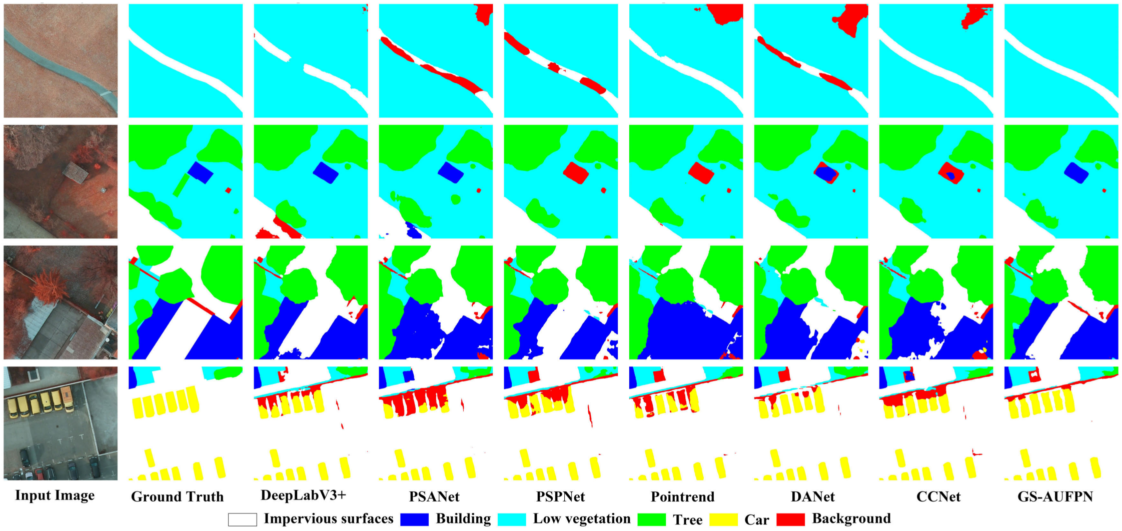

Fig. 10. Visualization results of some test samples on the ISPRS Potsdam scene dataset using various state-of-the-art networks.

③ freezing the first three stages of the encoder; ④ freezing the whole encoder; ⑤ freezing all the modules except the last convolutional layer; and ⑥ freezing the whole decoder. The corresponding parameter settings for $\mathrm{L}_{\mathrm{DIL}}$ are detailed in Table VII. For the analysis of parameter effects within the loss function, we selected frame ① as the base frame due to its no freezing, which allows for a more comprehensive assessment of each parameter setting.

We conducted extensive ablation experiments on the ISPRS and the LoveDA datasets, and the results are shown in Tables VII and VIII, respectively. In these experiments, the ISPRS Potsdam scene and LoveDA Urban scene are the old domains, while the ISPRS Vaihingen scene and LoveDA Rural scene are the new domains. Pretraining weights were adopted from the single-domain semantic segmentation results in Section V.

1) Parameter Effects in $\mathcal{L}_{DIL}$: We systematically investigated the influence of the parameters $T_1, T_2, \alpha, \beta,$ and I through ablation experiments using the control variable method. The experimental results indicate that T_1 and α

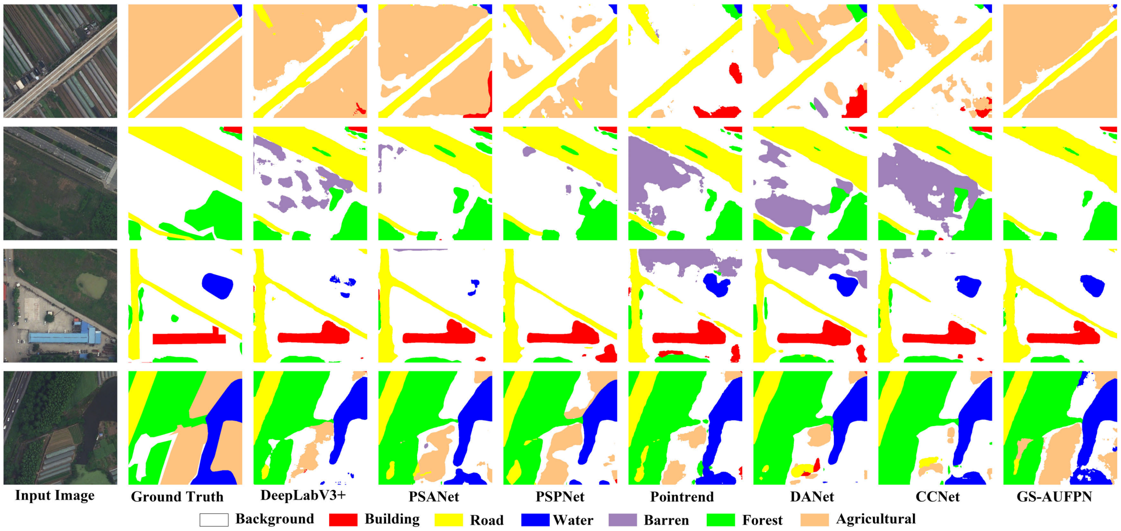

Fig. 11. Visualization results of some test samples on the LoveDA Urban scene dataset using various state-of-the-art networks.

TABLE VII

Evaluation Metrics of Ablation Experiments on DIL Semantic Segmentation With Different Freezing Frames and $\mathcal{L}_{\text{DIL}}$ Parameter Settings on the ISPRS Dataset

Frames	T_1	T_2	α	β	I	Potsdam		Vaihingen		Δ_{mIoU} (%)
						$mIoU$ (%)	Δ^b_{mIoU} (%)	$mIoU$ (%)	Δ^b_{mIoU} (%)	
Single task (A-A)	-	-	-	-	-	**74.87**	-	**69.77**	-	-
①	1.0	1.0	0.5	1.0	[0,1,2,3]	63.28	-15.48	62.61	-10.26	-12.87
①	2.0	1.0	0.5	1.0	[0,1,2,3]	**62.01**	**-17.17**	**66**	**-5.4**	**-11.29**
①	2.0	2.0	0.5	1.0	[0,1,2,3]	61.99	-17.2	65.99	-5.41	-11.31
①	2.0	1.0	0.5	0.5	[0,1,2,3]	61.99	-17.2	66.01	-5.39	-11.29
①	2.0	1.0	1.0	0.5	[0,1,2,3]	62.57	-16.43	63.8	-8.55	-12.49
①	2.0	1.0	1.0	1.0	[0,1,2,3]	62.6	-16.39	63.8	-8.55	-12.47
①	2.0	1.0	0.5	1.0	[1,2,3]	**61.98**	**-17.21**	**65.99**	**-5.41**	**-11.31**
①	2.0	1.0	0.5	1.0	[2,3]	61.96	-17.24	65.98	-5.43	-11.33
①	2.0	1.0	0.5	1.0	[3]	61.96	-17.24	65.97	-5.44	-11.34
GSMF-RS-DIL	2.0	1.0	0.5	1.0	[1,2,3]	**61.91**	**-17.31**	**65.81**	**-5.67**	**-11.49**
②	2.0	1.0	0.5	1.0	[2,3]	61.74	-17.53	65.19	-6.56	-12.05
③	2.0	1.0	0.5	1.0	[3]	62.47	-16.56	61.92	-11.25	-13.9
④	2.0	-	1.0	-	-	62.75	-16.18	57.56	-17.5	-16.84
⑤	2.0	-	1.0	-	-	61.74	-17.53	45.19	-29.55	-23.54
⑥	2.0	1.0	0.5	1.0	[0,1,2,3]	61.67	-17.63	65.45	-6.19	-11.91

significantly impact model performance, whereas T_2 and β have a smaller impact. The number of stages I used to calculate $\text{L}_{\text{feature}}$ has a minimal effect when the frame and other parameters are fixed. Optimal performance on both datasets was achieved with parameters $T_1 = 2.0$, $T_2 = 1.0$, $\alpha = 0.5$, $\beta = 1.0$, and $I = [0,1,2,3]$.

2) The Effect of Incremental Learning Freezing Frames: Among the freezing frames, frame ⑤ performed the worst, followed by frames ④ and ③. The difference between other frames was relatively minor, suggesting that feature extraction discrepancies between domains are mainly concentrated in the second, third, and fourth stages of the encoder, and the decoder also influences feature fusion.

On the ISPRS dataset, frame ① was the optimal, which is 0.2% higher than GSMF-RS-DIL in the Δ_{mIoU} metric. Conversely, on the LoveDA dataset, GSMF-RS-DIL is the optimal, exceeding frame ① by 0.58% in the Δ_{mIoU} metric. These findings indicate that freezing the first stage of the encoder is the most effective approach when considering both datasets.

D. Comparison With Other Cross-Domain Training Methods

In this section, we compare the proposed GSMF-RS-DIL framework with several cross-domain training methods, including single-task training, multitask training, fine-tuning,

TABLE VIII

Evaluation Metrics of Ablation Experiments on DIL Semantic Segmentation With Different Freezing Frames and $\mathcal{L}_{\mathrm{DIL}}$ Parameter Settings on the LoveDA Dataset

Frames	T_1	T_2	α	β	i	Urban		Rural		Δ_{mIoU} (%)
						$mIoU$ (%)	Δ^b_{mIoU} (%)	$mIoU$ (%)	Δ^b_{mIoU} (%)	
Single task (A-A)	-	-	-	-	-	**44.41**	-	**35.89**	-	-
①	1.0	1.0	0.5	1.0	[0,1,2,3]	48.98	+10.29	34.19	-4.74	+2.78
①	2.0	1.0	0.5	1.0	[0,1,2,3]	**51**	**+14.84**	**35.81**	**-0.22**	**+7.31**
①	2.0	2.0	0.5	1.0	[0,1,2,3]	50.87	+14.55	35.49	-1.11	+6.72
①	2.0	1.0	0.5	0.5	[0,1,2,3]	50.73	+14.23	35.6	-0.81	+6.71
①	2.0	1.0	1.0	0.5	[0,1,2,3]	49.71	+11.93	34.62	-3.54	+4.20
①	2.0	1.0	1.0	1.0	[0,1,2,3]	49.43	+11.3	34.53	-3.79	+3.76
①	2.0	1.0	0.5	1.0	[1,2,3]	50.73	+14.23	35.56	-0.92	+6.66
①	2.0	1.0	0.5	1.0	[2,3]	51.06	+14.97	35.58	-1.42	+6.78
①	2.0	1.0	0.5	1.0	[3]	**50.81**	**+14.41**	**35.78**	**-0.31**	**+7.05**
GSMF-RS-DIL	2.0	1.0	0.5	1.0	[1,2,3]	**51.02**	**+14.84**	**36.21**	**+0.89**	**+7.89**
②	2.0	1.0	0.5	1.0	[2,3]	50.55	+13.83	36.14	+0.7	+7.26
③	2.0	1.0	0.5	1.0	[3]	49.24	+10.88	34.54	-3.76	+3.56
④	2.0	-	1.0	-	-	47.16	+6.19	34.36	-4.26	+0.97
⑤	2.0	-	1.0	-	-	46.16	+3.94	33	-8.05	-2.06
⑥	2.0	1.0	0.5	1.0	[0,1,2,3]	50.55	+13.83	36.09	+0.56	+7.19

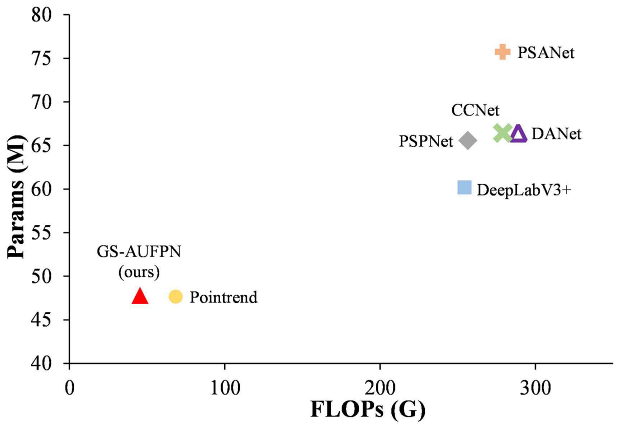

Fig. 12. Scatter plot of FLOPs and Params for different methods on the ISPRS Potsdam scene dataset.

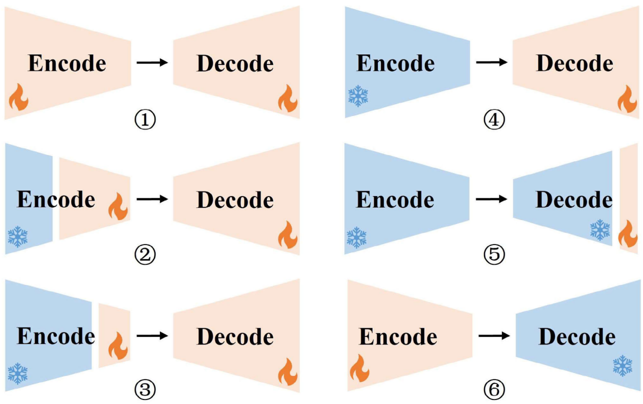

Fig. 13. Six different incremental learning freezing frames.

TABLE IX

Quantitative Comparison Results on the ISPRS Dataset With Other Cross Domain Training Methods

	Potsdam		Vaihingen		Δ_{mIoU} (%)
	$mIoU$ (%)	Δ^b_{mIoU} (%)	$mIoU$ (%)	Δ^b_{mIoU} (%)	
Single task (A-A)	74.87	-	69.77	-	-
Single task (A-B)	-	-	28.09	-59.74	-
Multi task	74.45	-00.56	66.84	-4.20	-2.38
Fine-tune	54.82	-26.78	**67.21**	**-3.67**	-15.22
LwF	**62.53**	**-16.48**	63.8	-8.55	-12.52
GSMF-RS-DIL (ours)	61.91	-17.31	65.81	-5.67	**-11.49**

and the incremental learning method LwF [44]. Multitask training is jointly training with data samples from both domains and represents the upper limit of accuracy for cross-domain training models. The fine-tuning model is retraining the old domain model using new domain samples at a smaller learning rate. Tables IX and X present the evaluation results on the ISPRS and LoveDA datasets, respectively. Figs. 14 and 15 show the visualization results for some samples on the ISPRS and LoveDA datasets, respectively.

The results indicate that directly applying the model trained on the old domain to the new domain yields very poor performance, with Δ^b_{mIoU} decreasing by 59.74% and 20.54% on the Vaihingen and Rural datasets, respectively. Several other cross-domain training methods significantly improve the model performance on the new domain. However, the fine-tuned method shows a 26.78% performance degradation on the old domain of the ISPRS dataset. Overall, our proposed GSMF-RS-DIL framework achieves the best results on both datasets, with Δ_{mIoU} decreasing by 11.49% on the ISPRS dataset and improving by 7.89% on the LoveDA dataset.

The performance of the DIL model is significantly impacted by the degree of distributional discreteness within the dataset.

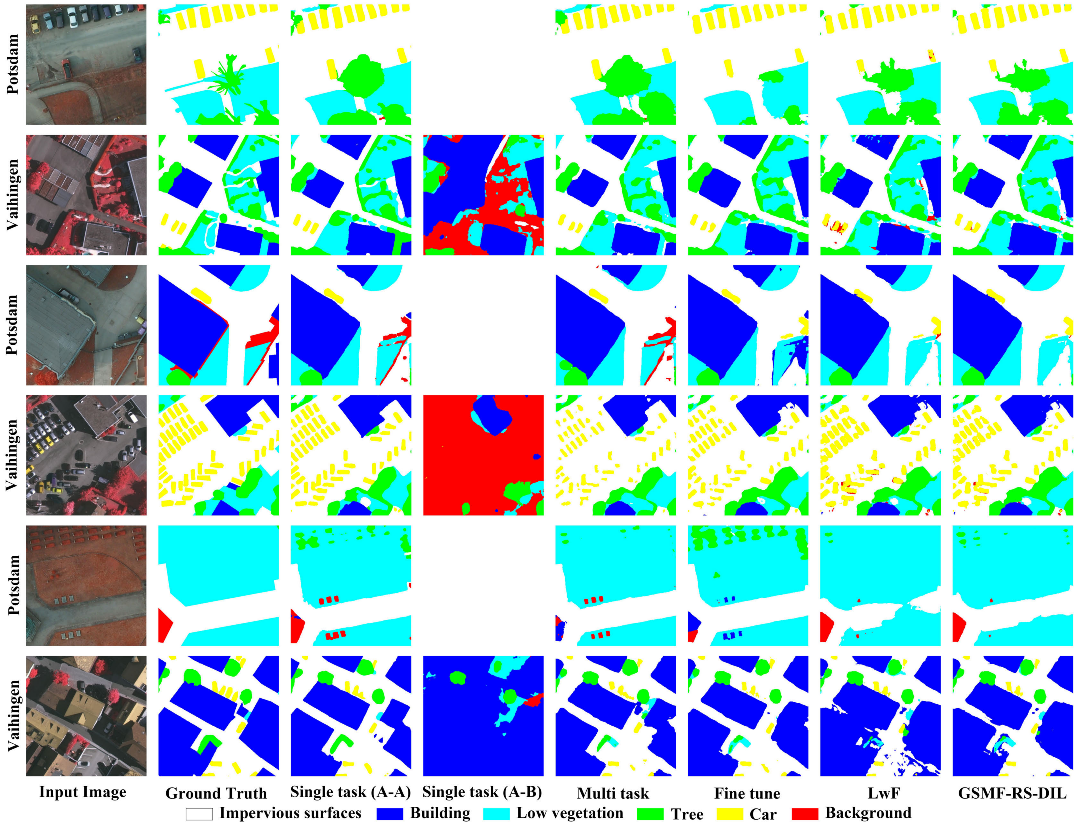

Fig. 14. Visualization results of some test samples on the ISPRS dataset for various cross-domain training methods.

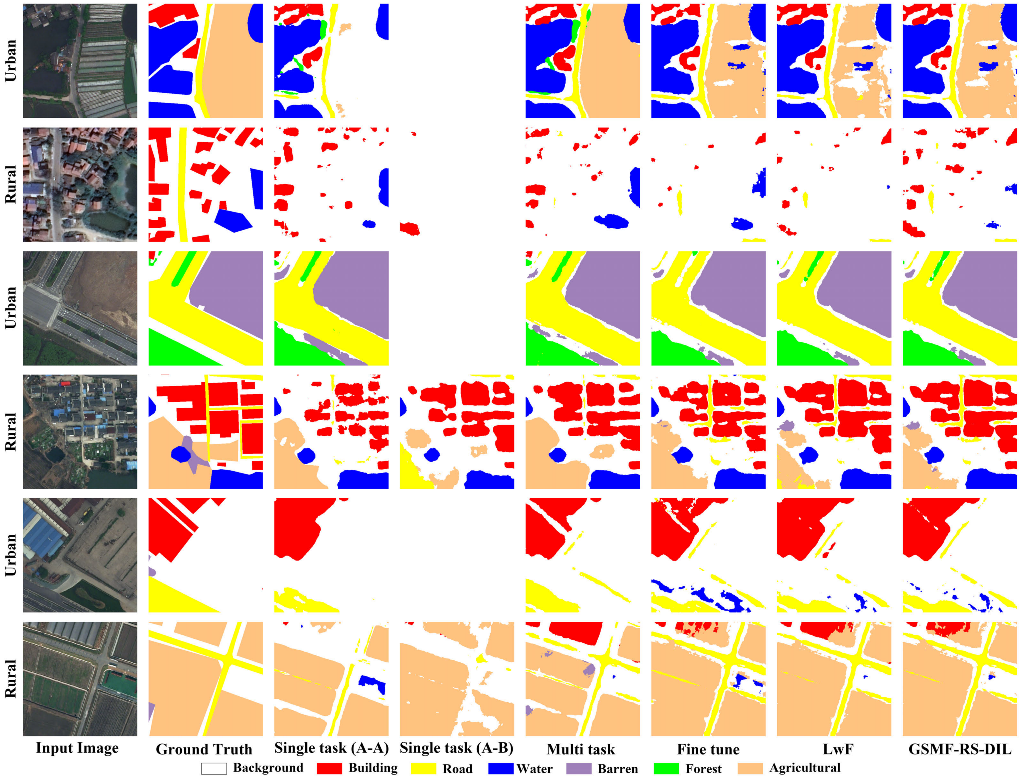

Fig. 15. Visualization results of some test samples on the LoveDA dataset for various cross-domain training methods.

Specifically, the addition of new domains generally enhances the accuracy of the old domains, particularly in the LoveDA dataset where the samples are more evenly distributed across both scenarios. However, in the ISPRS dataset, the sample

TABLE X

Quantitative Comparison Results on the LoveDA Dataset With Other Cross Domain Training Methods

	Urban		Rural		Δ_{mIoU} (%)
	$mIoU$ (%)	Δ^{b}_{mIoU} (%)	$mIoU$ (%)	Δ^{b}_{mIoU} (%)	
Single task (A-A)	44.41	-	35.89	-	-
Single task (A-B)	-	-	28.52	-20.54	-
Multi task	53.91	+21.39	35.71	-0.50	+10.45
Fine-tune	49.92	+12.41	35.26	-1.76	+5.33
LwF	49.92	+12.41	34.82	-2.98	+4.71
GSMF-RS-DIL (ours)	**51.02**	**+14.84**	**36.21**	**+0.89**	**+7.89**

distribution across the two scenarios is more discrete with minimal intersection between data distributions, leading to a significant decrease in the old domain accuracy when new domains are added. Future research could address this issue by exploring solutions from the perspective of large-scale models. Notably, in the cross-domain incremental learning task, our iteration number is set to 10k, which is significantly lower than the number of iterations for single-task direct training. Despite this, for a small sample dataset such as ISPRS Vaihingen scene dataset, our approach achieves performance that is not substantially different from multitask training.

VI. Conclusion

In this article, we propose a new framework for DIL of semantic segmentation in remote sensing images, namely the GSMF-RS-DIL. Within this framework, catastrophic forgetting due to domain shifts is addressed to a certain extent without using old domain data, through the means of frozen feature layers and a new multifeature knowledge distillation loss for co-constraints. The proposed GSR module employs graph convolution in graph space to expand the receptive field and extract contextual relationships within the image. The designed DCAU module effectively fuses high-level abstract features with low-level spatial features, enabling the network to autonomously learn valuable information from various feature types. Extensive experimental results on the ISPRS and LoveDA datasets demonstrate that the proposed method achieves the state-of-the-art performance in both single-domain and multidomain semantic segmentation tasks. However, the experiments in this article are limited to scenarios involving two domains. In the future, we will conduct experiments involving more domains and gradually advance incremental learning research across both domains and classes.

References

[1] W. Huang, P. Zou, Y. Xia, C. Wen, Y. Zang, and C. Wang, "OPOCA: One point one class annotation for LiDAR point cloud semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5701610.

[2] Y. Wang, Y. Wan, Y. Zhang, B. Zhang, and Z. Gao, "Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and LiDAR point clouds," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 385–404, Aug. 2023.

[3] L. Huang et al., "Deep learning-based semantic segmentation of remote sensing images: A survey," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 17, pp. 8370–8396, 2023.

[4] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[5] C. Chen, X. He, Z. Liu, W. Sun, H. Dong, and Y. Chu, "Analysis of regional economic development based on land use and land cover change information derived from Landsat imagery," *Sci. Rep.*, vol. 10, no. 1, p. 12721, Jul. 2020.

[6] H. He, J. Yan, D. Liang, Z. Sun, J. Li, and L. Wang, "Time-series land cover change detection using deep learning-based temporal semantic segmentation," *Remote Sens. Environ.*, vol. 305, May 2024, Art. no. 114101.

[7] H. Costa, P. Benevides, F. D. Moreira, D. Moraes, and M. Caetano, "Spatially stratified and multi-stage approach for national land cover mapping based on Sentinel-2 data and expert knowledge," *Remote Sens.*, vol. 14, no. 8, p. 1865, Apr. 2022.

[8] Q. Shi, D. He, Z. Liu, X. Liu, and J. Xue, "Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping," *J. Remote Sens.*, vol. 3, p. 0078, Jan. 2023.

[9] W. Shi, M. Zhang, H. Ke, X. Fang, Z. Zhan, and S. Chen, "Landslide recognition by deep convolutional neural network and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4654–4672, Jun. 2021.

[10] W. Huang et al., "Landslide susceptibility mapping and dynamic response along the sichuan-tibet transportation corridor using deep learning algorithms," *CATENA*, vol. 222, Mar. 2023, Art. no. 106866.

[11] 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.

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

[13] X. Cheng and H. Lei, "Semantic segmentation of remote sensing imagery based on multiscale deformable CNN and DenseCRF," *Remote Sens.*, vol. 15, no. 5, p. 1229, Feb. 2023.

[14] T. Li, Z. Cui, Y. Han, G. Li, M. Li, and D. Wei, "Enhanced multiscale networks for semantic segmentation," *Complex Intell. Syst.*, vol. 10, no. 2, pp. 2557–2568, Apr. 2024.

[15] 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. 833–851.

[16] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[17] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.

[18] Y. Xia et al., "SOE-net: A self-attention and orientation encoding network for point cloud based place recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11343–11352.

[19] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605116.

[20] C. Li et al., "CasFormer: Cascaded transformers for fusion-aware computational hyperspectral imaging," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102408.

[21] Y. Xia et al., "CASSPR: Cross attention single scan place recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8427–8438.

[22] 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.

[23] S. Shen, Y. Xia, A. Eich, Y. Xu, B. Yang, and U. Stilla, "SegTrans: Semantic segmentation with transfer learning for MLS point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[24] H. Liu, Y. Zhou, B. Liu, J. Zhao, R. Yao, and Z. Shao, "Incremental learning with neural networks for computer vision: A survey," *Artif. Intell. Rev.*, vol. 56, no. 5, pp. 4557–4589, May 2023.

[25] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, Aug. 2024.

[26] X. Rui, Z. Li, Y. Cao, Z. Li, and W. Song, "DILRS: Domain-incremental learning for semantic segmentation in multi-source remote sensing data," *Remote Sens.*, vol. 15, no. 10, p. 2541, 2023.

[27] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[28] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, and L. Ru, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 27672–27683.

[29] S. He et al., "RSI-Net: Two-stream deep neural network for remote sensing images-based semantic segmentation," *IEEE Access*, vol. 10, pp. 34858–34871, 2022.

[30] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021.

[31] P. Jian, Y. Ou, and K. Chen, "Uncertainty-aware graph self-supervised learning for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5509019.

[32] L. Zhang et al., "Dual graph convolutional network for semantic segmentation," 2019, *arXiv:1909.06121*.

[33] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, and Y. Kalantidis, "Graph-based global reasoning networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 433–442.

[34] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8947–8956.

[35] Y. Lu, Y. Chen, D. Zhao, B. Liu, Z. Lai, and J. Chen, "CNN-G: Convolutional neural network combined with graph for image segmentation with theoretical analysis," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 3, pp. 631–644, Sep. 2021.

[36] D. Jiang, H. Qu, J. Zhao, J. Zhao, and W. Liang, "Multi-level graph convolutional recurrent neural network for semantic image segmentation," *Telecommun. Syst.*, vol. 77, no. 3, pp. 563–576, 2021.

[37] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Self-constructing graph neural networks to model long-range pixel dependencies for semantic segmentation of remote sensing images," *Int. J. Remote Sens.*, vol. 42, no. 16, pp. 6184–6208, 2021.

[38] G. M. van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," *Nature Mach. Intell.*, vol. 4, no. 12, pp. 1185–1197, 2022.

[39] J. Li et al., "Class-incremental learning network for small objects enhancing of semantic segmentation in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5612920.

[40] X. Rong et al., "MiCro: Modeling cross-image semantic relationship dependencies for class-incremental semantic segmentation in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616218.

[41] M. Riemer et al., "Learning to learn without forgetting by maximizing transfer and minimizing interference," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–13.

[42] T. Kalb, M. Roschani, M. Ruf, and J. Beyerer, "Continual learning for class- and domain-incremental semantic segmentation," in *Proc. IEEE Intell. Vehicles Symp.*, Jul. 2021, pp. 1345–1351.

[43] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "PLOP: Learning without forgetting for continual semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4039–4049.

[44] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2017.

[45] P. Garg, R. Saluja, V. N. Balasubramanian, C. Arora, A. Subramanian, and C. V. Jawahar, "Multi-domain incremental learning for semantic segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 2080–2090.

[46] M. Wang, D. Yu, W. He, P. Yue, and Z. Liang, "Domain-incremental learning for fire detection in space-air-ground integrated observation network," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 118, Apr. 2023, Art. no. 103279.

[47] U. Michieli and P. Zanuttigh, "Knowledge distillation for incremental learning in semantic segmentation," *Comput. Vis. Image Understand.*, vol. 205, Apr. 2021, Art. no. 103167.

[48] K. James et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[49] L. Weng et al., "MDINet: Multidomain incremental network for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4402315.

[50] M. H. Vu, G. Norman, T. Nyholm, and T. Löfstedt, "A data-adaptive loss function for incomplete data and incremental learning in semantic image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1320–1330, Jun. 2022.

[51] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6392–6401.

[52] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–10.

[53] D. Kothandaraman, A. Nambiar, and A. Mittal, "Domain adaptive knowledge distillation for driving scene semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops*, Jan. 2021, pp. 134–143.

[54] J. Wang et al., "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. NeurIPS*, 2021, pp. 1–16.

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

[56] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9796–9805.

[57] H. S. Zhao et al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 270–286.

[58] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, and Z. Fang, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3141–3149.

[59] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

Wubiao Huang received the B.S. degree from Xiamen University of Technology, Xiamen, China, in 2020, and the M.Sc. degree in photogrammetry and remote sensing from Chang'an University, Xi'an, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include deep learning, landslide susceptibility mapping, semantic segmentation for remote sensing, and knowledge graph.

Mingtao Ding (Member, IEEE) received the B.S. and Ph.D. degrees in applied mathematics from Northwestern Polytechnical University, Xi'an, China, in 2007 and 2010, respectively.

From 2011 to 2012, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. In 2013, he joined the College of Geological Engineering and Geomatics, Chang'an University, Xi'an, where he is currently an Associate Professor of remote sensing science and technology. His research focuses on landslide detection, mapping, and monitoring with interferometric synthetic aperture radar and multispectral remote sensing. He also specializes in machine learning, artificial intelligence, and applied statistics.

Fei Deng received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, Hubei, China, in 1999, 2002, and 2006, respectively.

He is currently a Professor with Wuhan University. His research interests include 3-D reconstruction, parametric modeling, and machine learning for remote sensing images and point clouds.