Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

DLGCNet: Multimodal remote sensing semantic segmentation via dual diagonal low-rank adaptation and graph convolutional feature fusion

Junying Zeng [a,1], Jiahua Xu [b,1], Xudong Jia [c,1], Bin Deng [b,*], Yikui Zhai [b,*], Chuanbo Qin [b,*], Pasquale Coscia [d], Angelo Genovese [d], Xiaolin Tian [e]

[a] School of Emergency Technology and Management, Wuyi University, Jiangmen, 529020, Guangdong Province, China
[b] School of Electronics and Information Engineering, Wuyi University, Jiangmen, 529020, Guangdong Province, China
[c] School of Engineering, Science, and Technology, Central Connecticut State University, New Britain, 06050, CT, USA
[d] Department of Computer Science and the Dipartimento di Informatica, Universitá Degli Studi di Milano, Milano, 20133, Provincia di Milano, Italy
[e] Faculty of Innovation Engineering, Macau University of Science and Technology, 20133, Macao Special Administrative Region of China

ARTICLE INFO

Keywords:
Multimodal fusion
Remote sensing
Semantic segmentation
Parameter-efficient fine-tuning

ABSTRACT

In recent years, multimodal remote sensing images (MRSI) have demonstrated complementary cross-modal information owing to their heterogeneous characteristics, enabling more detailed and effective scene interpretation than single-modality data. To address challenges in existing multimodal fusion methods, this paper proposes DLGCNet, a multimodal segmentation network that jointly optimizes a dual diagonal low-rank adaptation (D^2LoRA) training framework for visual foundation models (VFMs) and a graph convolutional feature fusion (GCFF) module. To better adapt to the data distributions of MRSI, D^2LoRA introduces two trainable diagonal matrices that perform row-wise and column-wise transformations on the low-rank weight matrix, thereby improving the VFM's adaptability and feature extraction performance for MRSI. To overcome the limited cross-modal modeling capacity of convolutional neural network-based fusion and the excessive complexity of transformer-based fusion, GCFF dynamically adjusts the graph Laplacian according to modal information and establishes long-range cross-modal dependencies with lower computational complexity than transformer-based fusion methods. Experimental results demonstrate that, compared to current state-of-the-art multimodal data fusion methods, the proposed DLGCNet achieves optimal segmentation results on three datasets: Potsdam, Vaihingen, and WHU-OPT-SAR. The source code is accessible at https://github.com/2023xjh2023/DLGCNet.

1. Introduction

With the rapid advancement of remote sensing technology, the cost of acquiring remote sensing data has significantly decreased, while its data availability and application value have surged [1]. Remote sensing imagery plays a critical role across multiple disciplines, including geomorphological mapping [2–4], urban layout and infrastructure development [5–7], and environmental monitoring [8–10]. In recent years, driven by the rapid progress in computer vision technologies, deep learning-based semantic segmentation algorithms have become indispensable tools for interpreting remote sensing images [11]. Notably, multimodal remote sensing images (MRSI), leveraging the complementary advantages of heterogeneous data, provide richer features for complex scene parsing [12]. How to effectively leverage their data complementarity for more accurate land cover classification has become a key research direction in remote sensing image semantic segmentation. However, MRSI semantic segmentation faces significant challenges due to differences in feature distributions between modalities, leading to difficulties in feature alignment during cross-modal feature fusion and poor fusion results [13]. Therefore, establishing an efficient framework for cross-modal feature extraction and fusion has become a critical technology to break through the current bottleneck in MRSI semantic segmentation development [14].

The core of multimodal data fusion lies in extracting features from different modalities and using effective methods for fusion. In terms

* Corresponding authors.
E-mail addresses: zengjunying@126.com (J. Zeng), 15815762812@163.com (J. Xu), xjia@ccsu.edu (X. Jia), bindeng@wyu.edu.cn (B. Deng), yikuizhai@163.com (Y. Zhai), tenround@163.com (C. Qin), pasquale.coscia@unimi.it (P. Coscia), angelo.genovese@unimi.it (A. Genovese), xltian@must.edu.mo (X. Tian).

[1] Equal Contribution.

of feature extraction, existing networks often directly transfer pre-trained model encoders to enhance their feature extraction capabilities. Regarding fusion, current networks generally align and then fuse features using convolutional neural network (CNN)-based [15–19] or transformer-based [20–23] architectures. Numerous multimodal fusion methods have been developed for natural image segmentation. For example, Hazirbas et al. [17] proposed a dual-branch ResNet [24] for feature extraction combined with a convolution-based fusion network, and Chen et al. [17] introduced a dual-branch ResNet for feature extraction combined with a CNN-based attention fusion network, both of which demonstrated excellent performance in multimodal data segmentation for natural scenes. However, remote sensing images suffer from significant inter-modal distribution discrepancies caused by varying sensor parameters and imaging conditions, and intra-modal land cover features display pronounced multi-scale characteristics; these properties contrast markedly with natural-scene distributions [25], motivating the development of multimodal fusion methods tailored to the specific characteristics of MRSI.

In current remote sensing image semantic segmentation methods, visual foundation models (VFMs) such as ResNet [24] and ViT [26] are widely employed as encoders. However, owing to the poor adaptability of these natural-image pre-trained VFMs to remote sensing tasks, recent studies have trained VFMs on large-scale remote sensing datasets [27–30]. To effectively apply these visual models to downstream segmentation tasks, traditional approaches typically rely on full fine-tuning of all parameters [31], which demands substantial computational resources and dramatically increases training cost. Moreover, full fine-tuning often leads to the problem of catastrophic forgetting [32], where the network, during transfer learning, overwrites previously acquired knowledge while adapting to the new task. This may result in over-optimization toward the new objective, causing degraded performance on the original task and suboptimal results on the new one. Consequently, various parameter-efficient fine-tuning (PEFT) strategies have been proposed in recent years [31,33–36], which adapt the model to downstream tasks by freezing subsets of trainable parameters, thereby reducing computational overhead while retaining and leveraging the rich feature representations learned by the VFM. Existing PEFT studies for image segmentation (e.g., [31,37]) are mostly limited to the original LoRA formulation or to adapter-based approaches that require additional architectural modules; consequently, developing PEFT methods that yield stronger adaptability with fewer trainable parameters remains an important research direction. In addition, most current VFMs are pre-trained exclusively on red-green-blue (RGB) optical images [38,39], raising the question of how to effectively leverage such VFMs for MRSI. Given these limitations, it is necessary to employ more effective PEFT methods for training VFMs in MRSI semantic segmentation so as to exploit their powerful representational capacity for feature extraction.

On the other hand, VFMs with greater representational capacity are mostly transformer-based, which incurs high computational cost when used for feature extraction in MRSI semantic segmentation and thus motivates the design of more efficient multimodal fusion methods. Existing fusion approaches have notable shortcomings: CNN-based fusion struggles to capture long-range inter-modal dependencies due to a limited receptive field, while transformer-based fusion suffers from high computational complexity and large parameter counts that impede inference speed. Compared with CNNs, Graph Convolutional Networks (GCNs) effectively capture relationships between nodes and enable global contextual interaction via efficient information propagation [13,40]. Relative to transformers, GCNs offer lower parameter overhead and faster computation while still supporting rich feature interaction, making them an attractive choice for efficient fusion of MRSI information.

To further enhance the multimodal adaptability of VFMs and to address limitations such as the inability of CNN-based fusion methods to capture long-range inter-modal dependencies and the excessive computational complexity of transformer-based fusion, this paper proposes DLGCNet, a multimodal segmentation network incorporating a dual diagonal low-rank adaptation (D^2LoRA) framework and a graph convolutional feature fusion (GCFF) module. The contributions of this work are summarized as follows:

1. To further enhance the adaptability of VFMs, we propose D^2LoRA, which applies two additional trainable diagonal matrices to perform row- and column-wise transformations of the low-rank-adapted weight matrix, thereby strengthening the VFM's ability to accommodate new data distributions. We apply D^2LoRA during VFM training, and the resulting encoder demonstrates superior feature extraction capability.
2. To further improve multimodal fusion efficiency, we propose GCFF, which rapidly computes the graph Laplacian via inner products of channel-compressed modality features to effectively establish long-range inter-modal dependencies. Compared with transformer-based fusion, GCFF enables more efficient cross-modal integration and thereby improves segmentation performance.
3. We validate DLGCNet on three MRSI datasets—each comprising optical imagery paired with either Digital Surface Model (DSM) or Synthetic Aperture Radar (SAR) data. Experimental results demonstrate that DLGCNet significantly outperforms state-of-the-art (SOTA) MRSI segmentation methods.

The remainder of this paper is organized as follows. Section 2 reviews the related work on MRSI semantic segmentation and PEFT of foundation models. Section 3 describes the overall architecture of DLGCNet and presents flowcharts for D^2LoRA and GCFF. Section 4 reports extensive experiments and provides corresponding analyses. Finally, Section 5 concludes the paper and discusses directions for future work.

2. Related work

2.1. Semantic segmentation of remote sensing images

Semantic segmentation is crucial for the interpretation of remote sensing images, and many well-known CNNs have been widely adopted as benchmarks for segmentation tasks [41–43]. However, due to the limited receptive field of CNNs, and the unique characteristics of remote sensing images – such as diverse spatial relationships, complex textures, and highly variable shapes – segmentation methods developed for natural images cannot be directly transferred to the remote sensing field [44]. To address this, researchers have worked to improve semantic segmentation methods by tackling both the architectural limitations of CNNs and the distinctive features of remote sensing images. Li et al. [45] proposed a dual-attention CNN that efficiently enhances spatial and contextual feature interactions. Li et al. [46] also introduced a collaborative attention module that enables accurate spatial and channel information interaction, effectively mitigating visual attention bias. Although these efforts have alleviated the limited receptive field problem to some extent, they still fall short in fully addressing the fundamental weakness in global modeling capability. In contrast, ViT, leveraging the global modeling strengths of the self-attention mechanism [47,48], have been systematically introduced into the field of remote sensing image segmentation. Wang et al. [49] proposed transformer blocks with global-local attention mechanisms for the decoder, achieving efficient visual perception of remote sensing objects. Zhang et al. [50] employed multi-head attention mechanisms to jointly learn from pixel-level and coordinate information, thereby enhancing the robustness of extracted features. However, single-modality remote sensing image semantic segmentation methods are difficult to directly apply to MRSI, as they are unable to effectively leverage complementary information from multiple modalities for improved remote sensing image interpretation.

2.2. Multimodal semantic segmentation for remote sensing images

Most MRSI fusion approaches are based on CNN or transformer architectures, aligning features across modalities to fully exploit complementary information. Depending on the fusion stage, three main strategies are used: data-level, feature-level, and decision-level fusion. Specifically, Diakogiannis et al. [15] concatenated RGB optical and DSM images before network input, enabling direct segmentation via data-level fusion. Feature-level fusion methods typically employ separate encoders to extract features from each modality prior to integration; for example, Ma et al. [21] leveraged a transformer's cross-attention and channel-attention mechanisms for feature-level fusion. Ma et al. [23] proposed a multi-stage fusion scheme combining CNNs and ViTs with cross-attention for deep feature integration; Yan et al. [51] built a dual-branch Swin Transformer network and fused multimodal features via a depth-aware module, followed by an multi-layer perceptron (MLP) decoder to refine segmentation; Ni et al. [52] employed a Mamba-based feature extraction and fusion scheme to balance fusion speed and accuracy. Decision-level fusion uses multiple encoder–decoder streams to segment each modality independently before merging their outputs—for instance, Hosseinpour et al. [53] introduced a gated fusion module to effectively combine two modalities' features and achieved high-precision building extraction through multi-level feature fusion and a residual depthwise separable convolution (DSC) decoder. However, these fusion schemes predominantly rely on CNN or transformer architectures, the former being constrained by limited receptive fields that impede the capture of long-range inter-modal dependencies, while the latter suffering from excessive computational complexity that leads to prolonged fusion times. Therefore, a more efficient fusion strategy is required to establish long-range contextual relationships between modalities under reduced computational complexity.

2.3. Parameter-efficient fine-tuning

Current multimodal feature extraction methods seek to capture richer representations and therefore typically employ VFM encoders for feature extraction [54]. However, VFMs often exhibit poor adaptability and degraded performance when fully fine-tuned. Consequently, PEFT techniques have been adopted to mitigate this issue. PEFT techniques for VFMs fall into two main categories: LoRA and adapter-based methods. LoRA methods introduce trainable low-rank matrices into pre-trained VFMs to indirectly adjust model weights [34,35]; for example, Pu et al. [55] applied a low-rank approximation adaptation to update VFM weights and enhance feature extraction on remote sensing imagery, while Lu et al. [56] improved encoder adaptability by stacking multiple low-rank matrices within the VFM's linear layers. Adapter-based approaches insert lightweight networks into the VFM to enable rapid task-specific adaptation: Chen et al. [31] first integrated MLP-based adapters into VFMs for downstream tasks, and Zhou et al. [57] employed convolutional adapters in both attention and MLP modules of the VFM encoder to preserve high-frequency feature representations. Ma et al. [37] proposed MMLoRA and MMAdapter to efficiently fine-tune the SAM and, notably, were the first to leverage SAM's representational capacity for MRSI segmentation. However, most existing PEFT methods are derived from the original LoRA and Adapter paradigms [37,58–60] and do not leverage an even smaller parameter budget to further enhance VFM adaptability and enable more efficient training, thereby improving VFMs' multimodal feature extraction capability. Consequently, there is a need to develop more effective PEFT techniques to enhance VFMs' adaptability to MRSI.

3. Proposed method

3.1. Multimodal encoder

To further enhance VFMs' adaptability to MRSI and to address limitations such as VFMs' restricted modality adaptation, the inability of CNN-based fusion to capture long-range inter-modal dependencies, and the excessive computational complexity of transformer-based fusion, we propose DLGCNet, whose architecture is shown in Fig. 1. For MRSI fusion, we first perform feature extraction on each modality. Existing methods typically employ separate ResNet encoders for different modalities, but their inherently local receptive fields hinder the modeling of long-range dependencies; furthermore, VFMs pre-trained solely on RGB optical imagery fail to learn the complementary information present in other modalities, severely limiting performance in MRSI semantic segmentation. To overcome these limitations, we adopt the MTP encoder [30] – a transformer-based VFM specialized for remote sensing [61] – and optimize its training via D^{2}LoRA, thereby preserving pre-trained priors while enhancing adaptability to diverse datasets. Additionally, we employ DSC blocks to encode features from the DSM (SAR) modality, and then progressively fuse these heterogeneous features into the VFM encoder via the multimodal feature injectors (MFIs), ultimately generating cross-modal complementary representations. Specifically, the optical-modality image input to the VFM encoder is denoted as $\boldsymbol{X}_{\mathrm{VIS}} \in \mathbb{R}^{h\times w\times 3}$. First, $\boldsymbol{X}_{\mathrm{VIS}}$ is flattened via Patch Embedding to yield a feature map $\boldsymbol{X}_{\mathrm{V0}} \in \mathbb{R}^{h/16\times w/16\times 768}$, where h and w denote the image height and width, respectively. This is followed by four MTP blocks that extract features while keeping both spatial resolution and channel number constant. Each MTP block comprises three transformer blocks, each with 12 attention heads: the first two employ shifted window multi-head self-attention with a window size of 7, whereas the third uses global multi-head self-attention. During training of the VFM encoder, the proposed D^{2}LoRA is applied to further improve the encoder's adaptability to MRSI.

Given the substantial domain gap between DSM (or SAR) imagery and natural images, and to facilitate efficient extraction of DSM/SAR features, the DSM (or SAR) encoder does not employ a pre-trained backbone; instead, feature extraction is performed using DSC blocks [62]. For the DSM (SAR) modality, the input image is denoted as $\boldsymbol{X}_{\mathrm{DSM}} \in \mathbb{R}^{h\times w\times 1}$, which is processed by the DSC block shown in Fig. 2(a) to yield a feature map $\boldsymbol{X}_{\mathrm{D1}} \in \mathbb{R}^{h/4\times w/4\times 128}$. In each DSC block i ($i = 1,\ldots,4$), a DSC with kernel size 3 and channel expansion factor of 2 initially projects the features into a higher-dimensional space; following activation, a second DSC with kernel size 3 further enhances nonlinear transformation. Finally, batch normalization (BN) and ReLU activation functions are applied, and a residual connection [24] is employed to improve training stability. By stacking modules composed of DSCs, the network can rapidly and effectively capture the complex land-cover features present in DSM (SAR) imagery.

To address VFMs' limited multimodal adaptability, we implement the fusion structures from [16,23] as the MFI: a squeeze-and-excitation channel-attention module [63] that compresses and injects dual-modality features into the VFM encoder to enhance its MRSI feature representations. The architecture is shown in Fig. 2(b). In the MFI, input feature maps are $\boldsymbol{X}_{\mathrm{V}j} \in \mathbb{R}^{h/16\times w/16\times 768}$ and $\boldsymbol{X}_{\mathrm{D}j} \in \mathbb{R}^{h/2^{j+1}\times w/2^{j+1}\times C_j}$ for $j = 1,\ldots,4$. MFI first applies a Conv-BN-ReLU layer to match the spatial resolution and channel dimensionality of the DSM (SAR) feature map to those of $\boldsymbol{X}_{\mathrm{V}j}$, yielding $\boldsymbol{X}'_{\mathrm{D}j} \in \mathbb{R}^{h/16\times w/16\times 768}$. The two feature maps are then fed into separate SE attention modules to enhance cross-modal land cover feature representations and suppress channel-wise noise. The specific formulation is given by:

$$\mathrm{SE}(\boldsymbol{X}) = \sigma_{\mathrm{R}}(\mathrm{Conv}_{1\times 1}(\sigma_{\mathrm{R}}(\mathrm{GP}(\boldsymbol{X})))) \bullet \boldsymbol{X} \tag{1}$$

$$\boldsymbol{X}_{\mathrm{F}j} = \mathrm{SE}(\boldsymbol{X}_{\mathrm{V}j}) + \mathrm{SE}(\boldsymbol{X}'_{\mathrm{D}j}) \tag{2}$$

where GP denotes the global average pooling operator, σ_{R} represents the ReLU activation function, σ_{S} indicates the Sigmoid activation function, $\boldsymbol{X}_{\mathrm{F}j}$ denotes the fused feature, $\bullet$ signifies channel-wise multiplication. After injecting multimodal features via the MFI, fused features are fed into the next MTP block, enabling the VFM to incorporate additional modality information for contextual feature extraction in MRSI. Simultaneously, $\boldsymbol{X}_{\mathrm{F}j}$ is propagated via skip connections into the decoder composed of DSC blocks to preserve the multiscale spatial information of the MRSI.

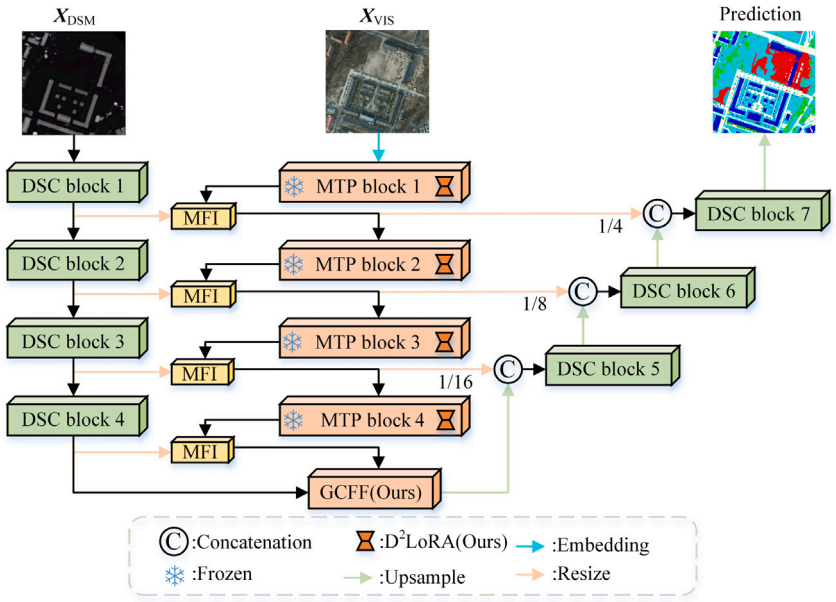

Fig. 1. Overall architecture of DLGCNet. The notations 1/4, 1/8, and 1/16 indicate the feature map resolutions relative to the input image size.

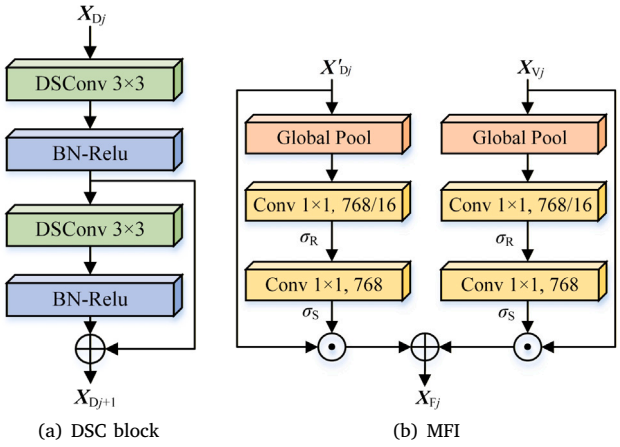

Fig. 2. (a) The DSC block in the multimodal encoder [62]. (b) The MFI [16].

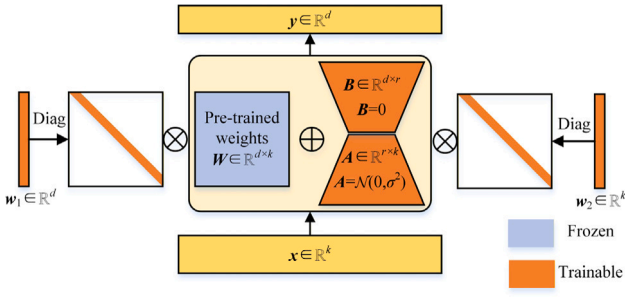

Fig. 3. Flowchart of the D²LoRA. The two diagonal matrices are initialized as identity matrices, ensuring that the VFM encoder's outputs in the first iteration are identical to those obtained without applying D²LoRA.

3.2. D²LoRA

Current MRSI semantic segmentation approaches commonly utilize VFMs as feature encoders and rely on full-parameter fine-tuning strategies for model optimization. However, this paradigm carries a significant risk of knowledge degradation. Specifically, when labeled data for downstream tasks is limited, full-parameter fine-tuning tends to make the model overly sensitive to training noise, thereby weakening the general feature representations acquired during VFM pre-training and reducing its ability to encode complex MRSI features. To further enhance the adaptability of VFMs to MRSI, we develop D²LoRA to perform PEFT of the VFM encoder, thereby improving its feature extraction capability for heterogeneous MRSI modalities. As illustrated in Fig. 3, D²LoRA restricts the optimization to a subspace defined by diagonal and low-rank matrices, enabling efficient adaptation to downstream tasks without compromising the VFM's inherent feature extraction capabilities. The proposed D²LoRA extends LoRA by introducing two additional trainable diagonal matrices $\boldsymbol{W}_1 \in \mathbb{R}^{d\times d}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{k\times k}$, which left- and right-multiply the original low-rank-adapted weight matrix $(\boldsymbol{W}+\alpha\boldsymbol{B}\boldsymbol{A})$ (with $\boldsymbol{W} \in \mathbb{R}^{d\times k}$), respectively. This modification enhances the network's adaptability to the complex and highly variable features present in MRSI. During backpropagation, the pre-trained weights are frozen and only the small set of parameters is updated, enabling efficient training with minimal parameter overhead. The D²LoRA computation is given as follows:

$$\boldsymbol{W}' = \boldsymbol{W}_1(\boldsymbol{W}+\alpha\boldsymbol{B}\boldsymbol{A})\boldsymbol{W}_2 \tag{3}$$

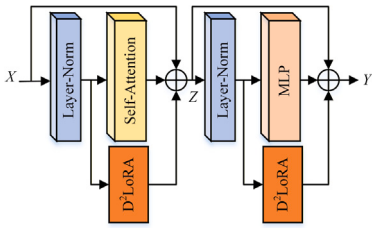

Fig. 4. Transformer architecture trained using the D²LoRA method.

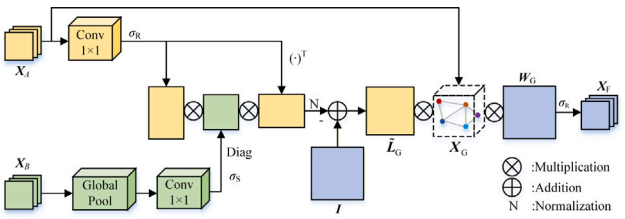

Fig. 5. Flowchart of the FGCN.

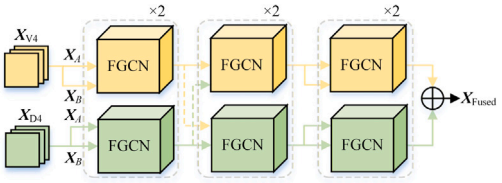

Fig. 6. Flowchart of the GCFF method.

where $\boldsymbol{B} \in \mathbb{R}^{d\times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r\times k}$ multiply to produce a rank r matrix. The scalar α is a stability parameter for training, typically set to 1. In the MTP encoder's MTP block used in this work, D²LoRA is applied to all linear layers, as illustrated in Fig. 4. The D²LoRA-augmented linear layers include the three projection layers (query, key, and value) in the multi-head attention module and the two linear layers in the feed-forward MLP. The single-head self-attention with D²LoRA in the MTP block is given by the following equations:

$$[\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}] = \mathrm{LN}(\boldsymbol{X})[\boldsymbol{W}'_Q, \boldsymbol{W}'_K, \boldsymbol{W}'_V] \tag{4}$$

$$\boldsymbol{Z} = \boldsymbol{X} + \mathrm{softmax}(\boldsymbol{Q}\boldsymbol{T}^\top/\sqrt{m})\boldsymbol{V} \tag{5}$$

where $\boldsymbol{W}'_Q \in \mathbb{R}^{768\times768}$, $\boldsymbol{W}'_K \in \mathbb{R}^{768\times768}$ and $\boldsymbol{W}'_V \in \mathbb{R}^{768\times768}$ are the weight matrices that project the input sequence into the query, key, and value matrices, respectively, and all three are trained using D²LoRA. LN(·) denotes layer normalization, $(\cdot)^\top$ indicates matrix transpose, and m is the normalization factor. Each MTP block also includes an MLP to enhance the model's nonlinear representational capacity; accordingly, the MLP within each block is likewise trained with D²LoRA. The output of the MTP block is then computed as follows:

$$\boldsymbol{Y} = \boldsymbol{Z} + \sigma_{\mathrm{G}}(\boldsymbol{Z}\boldsymbol{W}'_{\mathrm{FC1}})\boldsymbol{W}'_{\mathrm{FC2}} \tag{6}$$

where σ_{G} denotes the GELU activation function, and $\boldsymbol{W}'_{\mathrm{FC1}} \in \mathbb{R}^{768\times3072}$ and $\boldsymbol{W}'_{\mathrm{FC2}} \in \mathbb{R}^{3072\times768}$ are the MLP weight matrices trained via D²LoRA. By applying the proposed D²LoRA to perform PEFT on all linear layers of the MTP encoder, the degradation of the multimodal encoder's feature extraction capability is alleviated, enabling it to more effectively extract features from MRSI.

3.3. GCFF

Due to the inherent heterogeneity of MRSI, significant discrepancies exist between feature distributions of different modalities. Consequently, conventional fusion operations, such as element-wise addition or concatenation, fail to achieve effective cross-modal integration. Among current approaches, CNN-based fusion is constrained by limited receptive fields and thus cannot capture long-range dependencies across modalities, whereas transformer-based fusion suffers from high computational complexity that slows inference and increases resource consumption. To address these limitations, we propose a graph-convolutional feature fusion method (FGCN), illustrated in Fig. 5. By modeling cross-modal feature interactions with a graph structure, FGCN preserves global context modeling capability while effectively fusing different modalities with lower computational complexity than transformer-based methods. In the proposed FGCN, the input feature maps comprise the optical modality feature map $\boldsymbol{X}_{\mathrm{V4}} \in \mathbb{R}^{h/16\times w/16\times768}$ and the DSM (SAR) modality feature map $\boldsymbol{X}_{\mathrm{D4}} \in \mathbb{R}^{h/16\times w/16\times1024}$. Each is first passed through a convolutional layer to produce $\boldsymbol{X}'_{\mathrm{V4}} \in \mathbb{R}^{h/32\times w/32\times512}$ and $\boldsymbol{X}'_{\mathrm{D4}} \in \mathbb{R}^{h/32\times w/32\times512}$, thereby aligning both spatial resolution and channel dimensionality. Feature alignment between these representations is then used to construct the graph Laplacian. Assuming fusion of the DSM modality into the optical modality, let $\boldsymbol{X}_A \in \mathbb{R}^{h/32\times w/32\times512}$ denote the optical features and $\boldsymbol{X}_B \in \mathbb{R}^{h/32\times w/32\times512}$ the DSM (SAR) features. The graph Laplacian is computed as follows:

$$\boldsymbol{L}_{\mathrm{G}} = \sigma_{\mathrm{R}}(\boldsymbol{X}_A\boldsymbol{W}_A)\mathrm{Diag}(\sigma_{\mathrm{S}}(\mathrm{GP}(\boldsymbol{X}_B\boldsymbol{W}_B)))\sigma_{\mathrm{R}}(\boldsymbol{X}_A\boldsymbol{W}_A)^\top \tag{7}$$

$$d_{ii} = \sum_j a_{ij} \tag{8}$$

$$\widetilde{\boldsymbol{L}}_{\mathrm{G}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{L}_{\mathrm{G}}\boldsymbol{D}^{-1/2} \tag{9}$$

where matrices $\boldsymbol{W}_A \in \mathbb{R}^{512\times K}$ and $\boldsymbol{W}_B \in \mathbb{R}^{512\times K}$ are linear layers initialized with Kaiming uniform initialization and Diag(·) denotes a diagonal matrix with the input vector as its diagonal elements. $\boldsymbol{I} \in \mathbb{R}^{hw/1024\times hw/1024}$ is the identity matrix used to introduce residual connections. $\boldsymbol{D} \in \mathbb{R}^{hw/1024\times hw/1024}$ is a diagonal matrix whose diagonal elements d_{ii} represent the sum of the ith row or column, and $\boldsymbol{D}^{-1/2} \in \mathbb{R}^{hw/1024\times hw/1024}$ is a normalization matrix, where $(\cdot)^{-1/2}$ denotes element-wise square root after matrix inversion. In this study, global average pooling is applied to compress the feature representation of $\boldsymbol{X}_B$, and $\boldsymbol{W}_B$ is used to reduce its channel dimension to K. Subsequently, channel-wise correlation between the two modalities is computed to enable efficient alignment and fusion of their features. After computing the normalized graph Laplacian $\widetilde{\boldsymbol{L}}_{\mathrm{G}} \in \mathbb{R}^{hw/1024\times hw/1024}$, the feature $\boldsymbol{X}_A$ is reshaped into graph nodes $\boldsymbol{X}_{\mathrm{G}} \in \mathbb{R}^{hw/1024\times512}$, which are then processed through a graph convolution operation as formulated below to effectively integrate multimodal feature information:

$$\boldsymbol{X}_{\mathrm{F}} = \sigma_{\mathrm{R}}(\widetilde{\boldsymbol{L}}_{\mathrm{G}}\boldsymbol{X}_{\mathrm{G}}\boldsymbol{W}_{\mathrm{G}}) \tag{10}$$

where $\boldsymbol{W}_{\mathrm{G}} \in \mathbb{R}^{512\times512}$ denotes the linear transformation matrix for graph node features and is initialized using Kaiming uniform initialization. Compared with transformer-based fusion methods, FGCN maintains low computational complexity by employing a small channel dimension K, while dynamically modulating the graph Laplacian of modality A through inner products with the node features of modality B. This dynamic neighborhood-weighted aggregation mechanism enables substantially greater spatial-dimensional fusion of features across modalities under a low computational budget. To enable deeper fusion of complex remote sensing objects across significantly different modalities, we adopt the process illustrated in Fig. 6. Specifically, features from the two modalities are first independently enhanced using separate FGCN modules to strengthen their semantic representations. Subsequently, two additional FGCN modules are employed to perform cross-modal feature fusion. After this enhancement–fusion–enhancement process for MRSI features, the fused representations are integrated through element-wise addition.

Table 1
Comparative experimental results on the Potsdam dataset.

Methods	IoU (%)					mIoU (%)	mF1 (%)	OA (%)
	ImSur.	Buil.	LowVe.	Tree	Car			
ABCNet [45]	84.86	88.69	71.52	73.87	88.04	81.40	89.56	87.94
PSPNet [43]	86.96	92.90	75.72	78.04	92.41	85.21	91.85	90.16
TransUNet [47]	87.02	90.74	73.97	77.14	90.09	83.79	91.02	89.54
UNetFormer [49]	87.03	92.21	75.33	78.66	92.18	85.09	91.79	90.19
FuseNet [19]	86.42	93.65	73.76	75.65	87.22	83.34	90.73	89.52
vFuseNet [18]	86.68	93.99	74.80	77.87	89.40	84.55	91.46	89.94
ESANet [16]	87.24	92.94	75.35	78.30	90.31	84.83	91.64	90.27
SA-GATE [17]	85.59	92.90	73.43	78.02	91.02	84.19	91.24	89.34
CMFNet [21]	87.32	93.00	75.09	78.20	90.47	84.81	91.63	90.33
FTransUNet [23]	88.20	94.07	75.73	78.21	90.82	85.41	91.97	90.59
MFNet [37]	88.64	94.07	75.77	79.08	90.41	85.59	92.08	90.66
Mul-VMamba [52]	87.54	94.03	74.93	77.35	89.64	84.70	91.54	90.17
DLGCNet (Ours)	**89.50**	**95.03**	**76.62**	**79.72**	**93.69**	**86.91**	**92.83**	**91.47**

4. Experiments result

4.1. Datasets

To evaluate the effectiveness and generalization capability of DLGCNet, experiments are conducted on the Potsdam, Vaihingen [64], and WHU-OPT-SAR [65] datasets. The Potsdam and Vaihingen datasets comprise the following classes: Impervious Surface (ImSur.), Building (Buil.), Low Vegetation (LowVe.), Tree, Car, and Clutter. The WHU-OPT-SAR dataset includes: Farmland (Farml.), City, Village (Villa.), Water, Forest (Fores.), Road, and Other.

The **Potsdam** dataset comprises high-resolution remote sensing imagery acquired over the Potsdam, Germany. It consists of 38 orthorectified images of size 6000 × 6000 pixels, each accompanied by a normalized DSM (nDSM), all at a spatial resolution of 5 cm. In this study, each pair of RGB optical image and its corresponding nDSM (SAR) was cropped into tiles of 512 × 512 pixels. Of the 38 multimodal image pairs, 32 were used for training and the remaining 6 for testing.

The **Vaihingen** dataset comprises high-resolution remote sensing imagery acquired over Vaihingen, Germany. It contains 33 orthorectified images with an average size of approximately 2000 × 2500 pixels, each accompanied by a DSM at a spatial resolution of 9 cm. Of the 33 multimodal image pairs, 28 were used for training and the remaining 5 for testing.

The **WHU-OPT-SAR** dataset's optical and SAR images are provided by the Gaofen Hub Center of Hubei Province. It contains 100 image pairs of approximately 5500 × 3700 pixels. The optical images were acquired by the GF-1 satellite, and the SAR images by the GF-3 satellite. Of the 100 multimodal image pairs, 80 were used for training and 20 for testing.

4.2. Evaluation metrics

This study employs several metrics to evaluate the performance of MRSI semantic segmentation, namely Intersection over Union (IoU), mean IoU (mIoU), mean F1 score (mF1), and Overall Accuracy (OA). The computation of each metric follows the definitions in [23,46]. OA is computed as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{11}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. For each non-background class n in datasets, the class-specific $\mathrm{Precision}_n$ and Recall_n are defined as follows:

$$\mathrm{Precision}_n = \frac{\mathrm{TP}_n}{\mathrm{TP}_n + \mathrm{FP}_n} \tag{12}$$

$$\mathrm{Recall}_n = \frac{\mathrm{TP}_n}{\mathrm{TP}_n + \mathrm{FN}_n} \tag{13}$$

IoU_n computes the intersection over union for a single class, more directly reflecting its segmentation quality, whereas $\mathrm{F1}_n$ serves as a comprehensive measure combining $\mathrm{Precision}_n$ and Recall_n.

$$\mathrm{IoU}_n = \frac{\mathrm{TP}_n}{\mathrm{TP}_n + \mathrm{FP}_n + \mathrm{FN}_n} \tag{14}$$

$$\mathrm{F1}_n = 2 \times \frac{\mathrm{Precision}_n \times \mathrm{Recall}_n}{\mathrm{Precision}_n + \mathrm{Recall}_n} \tag{15}$$

The corresponding mIoU and mF1 are defined as the mean IoU and mean F1, respectively, calculated over all classes excluding the background.

4.3. Implementation details

The experiments were conducted on a workstation equipped with an NVIDIA RTX 3060 GPU, an Intel Core i7-12700F CPU, and 32 GB of RAM, using PyTorch 2.2.1 and CUDA 11.8. Data augmentation techniques included random flipping, random brightness adjustment, and random contrast adjustment. Both DLGCNet and all comparison models were trained with the AdamW optimizer, an initial learning rate of 1×10^{-4}, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay of 2×10^{-5}. Training proceeded for 50 epochs under a cosine annealing learning rate schedule. The rank of D^2LoRA followed the configuration in [33], with $r = 4$ and $\alpha = 1$, while the GCFF module employed $K = 64$ channels. The loss function was the sum of cross-entropy and Dice losses. Both training and testing batch sizes were set to 4. More detailed experimental settings follow those described in [49].

4.4. Comparative experiments

4.4.1. MRSI segmentation comparative analysis of the Potsdam dataset

Table 1 presents the quantitative results on the Potsdam dataset, and Fig. 7 shows the corresponding qualitative segmentation maps. The first four rows report the performance of single-modality (optical) baseline models, while the remaining entries correspond to multimodal methods using MRSI. Table 1 shows that the proposed DLGCNet outperforms the SOTA across overall evaluation metrics: specifically, DLGCNet achieves a mIoU that is 1.32% higher than SOTA, a mF1 that is 0.75% higher, and an OA that is 0.81% higher. Additionally, DLGCNet attains the highest IoU for every class. Notably, it delivers exceptional results on the highly sensitive Building, Tree, and Car categories—thanks to the DSM's height information, the strong feature extraction capability of the multimodal encoder, and the efficient MRSI feature fusion provided by GCFF. In summary, DLGCNet enhances both feature extraction and fusion on the Potsdam dataset, yielding superior overall performance across all classes. The visualizations in Fig. 7 further confirm that DLGCNet's predictions align closely with the ground truth (GT). Taken

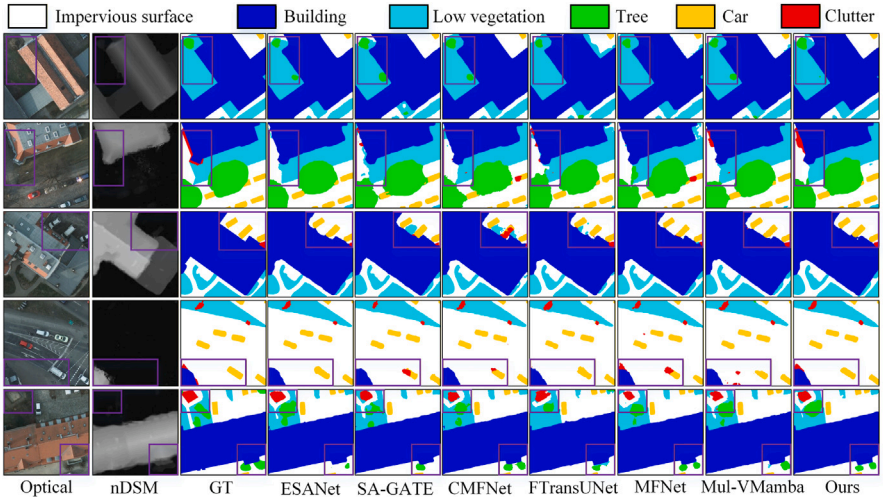

Fig. 7. Visualization of comparative experimental results on the Potsdam dataset. Purple boxes highlight the differences among the subfigures.

together, Table 1 and Fig. 7 clearly show that networks leveraging MRSI outperform those relying solely on optical imagery.

Although most multimodal models in the experiments employ dual ResNet backbones for feature extraction, their limited adaptability to new data hinders segmentation performance improvement. In contrast, DLGCNet uses a multimodal encoder with D^{2}LoRA, into which the MFI injects complementary modality features to leverage the VFM's prior knowledge for MRSI feature extraction. This design enables the model to capture complex contextual dependencies within MRSI. Moreover, owing to the feature-alignment fusion mechanism of GCFF, the modality-specific features extracted by the encoder can efficiently establish long-range inter-modal dependencies, yielding fully integrated outputs and superior segmentation performance.

4.4.2. MRSI segmentation comparative analysis of the Vaihingen dataset

Table 2 presents the comparative evaluation results on the Vaihingen dataset, and Fig. 8 shows the corresponding visual segmentation outputs. As shown in Table 2, the proposed DLGCNet outperforms most competing models overall: its mIoU is 1.10% higher, its mF1 is 0.64% higher, and its OA is 0.17% higher than those of the best-performing model. Some class-level IoU scores do not reach their maxima; this is likely because the backbone MTP was pre-trained on RGB optical imagery, whereas the Vaihingen dataset uses IRRG optical images, yielding a larger domain gap that makes adaptation to IRRG more challenging. Even so, DLGCNet still achieves strong overall performance, delivering competitive IoU across all classes. The segmentation visualizations in Fig. 8 further demonstrate that DLGCNet's predictions align closely with the GT. Taken together, the results in Table 2 and Fig. 8 are consistent with the findings from our earlier comparative experiments, yielding similar conclusions regarding the effectiveness of MRSI segmentation.

4.4.3. MRSI segmentation comparative analysis of the WHU-OPT-SAR dataset

Table 3 presents the comparative evaluation results on the WHU-OPT-SAR dataset, while Fig. 9 shows the corresponding visual segmentation outputs. As shown in Table 3, the proposed DLGCNet outperforms SOTA overall: its mIoU is 2.11% higher, its mF1 is 1.95% higher, and its OA is 0.60% higher than SOTA. DLGCNet achieves the best performance across all classes and exhibits particularly marked improvements on the Village, Water, and Road categories. This superior performance on these classes is attributable to their distinct SAR backscatter characteristics, which make them more readily distinguishable in SAR imagery. The segmentation visualizations in Fig. 9 further confirm that DLGCNet's predictions closely match the GT. Taken together, the results in Table 3 and Fig. 9 are consistent with our previous comparative experiments, yielding similar conclusions regarding the efficacy of DLGCNet for MRSI segmentation.

4.5. Ablation experiments

4.5.1. Ablation study of different pre-trained backbones

To investigate the impact of different encoders on DLGCNet's segmentation performance, we design a corresponding ablation study. In this experiment, three alternative encoders are compared against the MTP: ResNet [24], ViT [26], and SegFormer [48]. Specifically, we employ ResNet-101, ViT-Base, and MiT-B2 as backbones. Furthermore, to demonstrate D^{2}LoRA's effectiveness in enhancing encoder adaptability, D^{2}LoRA is additionally applied for PEFT to ViT and MTP; the remaining backbones are trained with full-parameter fine-tuning. The segmentation results presented in Table 4 reveal that ResNet-101 achieves relatively strong performance; however, its limited receptive field impairs its ability to model global contextual information, resulting in inferior segmentation accuracy compared to the encoders with D^{2}LoRA. The results of the ViT-Base, MiT-B2, and MTP-Base without D^{2}LoRA are inferior to those of the fully fine-tuned ResNet-101, indicating that transformer-based encoders exhibit limited adaptability under full-parameter fine-tuning, which consequently leads to inferior segmentation performance. By contrast, DLGCNet achieves superior performance when employing the MTP encoder with D^{2}LoRA. This is because the MTP encoder with D^{2}LoRA freezes the backbone parameters while introducing trainable low-rank matrices and diagonal matrices, thereby enhancing the adaptability of the VFM. As a result, the proposed model attains SOTA accuracy in the MRSI semantic segmentation task. Furthermore, comparing ViT-Base and MTP-Base with and without D^{2}LoRA confirms that D^{2}LoRA substantially enhances each encoder's adaptability to downstream data, leading to improved segmentation accuracy and demonstrating its potential applicability across a wide range of VFMs. In summary, this ablation study shows that the D^{2}LoRA-trained MTP encoder used in DLGCNet outperforms other mainstream backbones and that D^{2}LoRA effectively increases encoder adaptability to complex land cover data, further boosting DLGCNet's segmentation performance on MRSI.

4.5.2. Effect of the dual diagonal matrices in D^{2}LoRA

To investigate the influence of the dual diagonal matrices on D^{2}LoRA's adaptation capability, we conduct experiments with five configurations to explore the relationship between the diagonal matrices

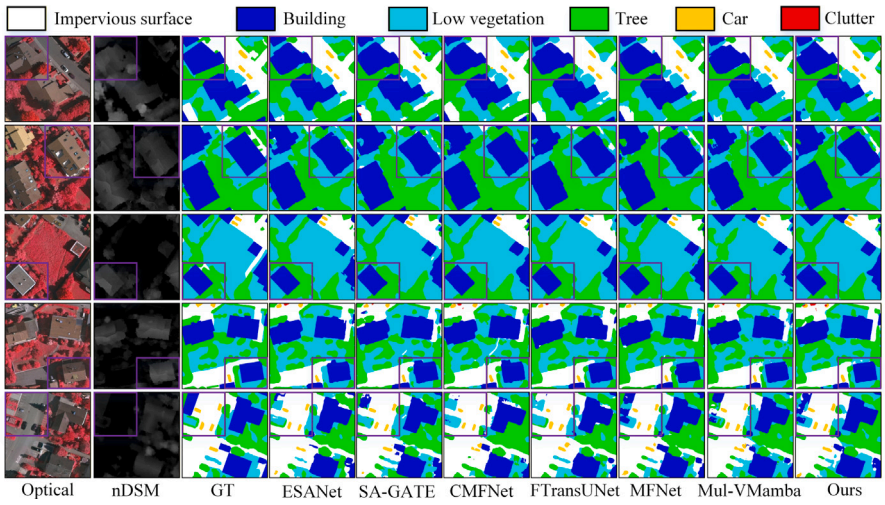

Fig. 8. Visualization of comparative experimental results on the Vaihingen dataset. Purple boxes highlight the differences among the subfigures.

Table 2
Comparative experimental results on the Vaihingen dataset.

Methods	IoU (%)					mIoU (%)	mF1 (%)	OA (%)
	ImSur.	Buil.	LowVe.	Tree	Car			
ABCNet [45]	86.44	92.54	72.61	81.33	69.11	80.41	88.88	91.13
PSPNet [43]	86.73	93.81	72.97	81.99	80.86	83.27	90.72	91.58
TransUNet [47]	87.17	92.71	74.07	82.48	78.58	83.00	90.57	91.74
UNetFormer [49]	86.38	92.72	72.34	82.04	**81.71**	83.04	90.59	91.30
FuseNet [19]	86.63	92.83	73.13	83.05	54.51	78.03	86.98	91.43
vFuseNet [18]	87.00	93.85	72.86	81.45	73.41	81.71	89.72	91.52
ESANet [16]	87.39	93.57	73.94	83.37	78.21	83.30	90.74	92.01
SA-GATE [17]	86.97	93.78	73.51	82.40	81.00	83.53	90.88	91.77
CMFNet [21]	87.37	93.76	73.69	82.65	78.87	83.27	90.72	91.87
FTransUNet [23]	87.95	94.96	74.83	83.01	74.56	83.06	90.55	92.35
MFNet [37]	88.07	94.31	75.12	**83.85**	75.29	83.33	90.73	92.42
Mul-VMamba [52]	86.70	93.244	72.65	82.32	74.00	81.78	89.78	91.45
DLGCNet (Ours)	**88.37**	**95.35**	**75.15**	83.50	80.79	**84.63**	**91.52**	**92.59**

Table 3
Comparative experimental results on the WHU-OPT-SAR dataset.

Methods	IoU (%)						mIoU (%)	mF1 (%)	OA (%)
	Farml.	City	Villa.	Water	Fores.	Road			
ABCNet [45]	65.07	55.32	44.47	61.99	81.89	33.96	57.12	71.49	82.49
PSPNet [43]	65.18	58.13	46.22	62.68	82.66	34.33	58.20	72.39	82.99
TransUNet [47]	63.91	58.31	46.70	61.61	82.16	30.12	57.14	71.34	82.60
UNetFormer [49]	65.18	58.63	46.77	62.89	82.54	35.69	58.61	72.81	83.02
FuseNet [19]	65.55	59.68	47.80	62.45	82.41	31.44	58.22	72.28	82.99
vFuseNet [18]	65.33	57.55	45.44	62.25	82.48	35.03	58.02	72.27	82.93
ESANet [16]	65.36	59.61	46.52	64.38	82.53	33.97	58.73	72.79	83.06
SA-GATE [17]	64.85	58.19	45.34	63.30	82.27	33.92	57.98	72.18	82.99
CMFNet [21]	64.97	57.87	45.48	60.47	82.26	31.22	57.04	71.30	82.74
FTransUNet [23]	63.43	59.58	45.32	62.30	81.09	32.33	57.34	71.64	81.92
MFNet [37]	65.83	58.94	47.96	65.11	82.26	32.29	58.73	72.72	83.24
Mul-VMamba [52]	64.75	58.61	47.24	64.47	81.67	33.94	58.44	72.61	82.51
DLGCNet (Ours)	**66.08**	**61.40**	**48.55**	**65.65**	**83.18**	**40.20**	**60.84**	**74.74**	**83.84**

and performance. Specifically, the settings were: LoRA (only the low-rank matrix $\boldsymbol{BA}$), Row (only $\boldsymbol{W}_1$ and $\boldsymbol{BA}$), Column (only $\boldsymbol{W}_2$ and $\boldsymbol{BA}$), Shared (two diagonal matrices shared across layers plus $\boldsymbol{BA}$), and D^{2}LoRA (separate $\boldsymbol{W}_1$, $\boldsymbol{W}_2$, and $\boldsymbol{BA}$). The results in Table 5 show that LoRA performs worst; single-direction transforms (Row or Column) improve over LoRA but remain inferior to D^{2}LoRA. The cross-layer shared variant also yields suboptimal performance, indicating that each pair of diagonal matrices should be learned independently for the low-rank-adapted matrix in each layer. These findings demonstrate that applying two diagonal matrices to perform row- and column-wise transformations of the low-rank-adapted matrix enhances adaptation, and that independently learned dual diagonals per layer provide the greatest gains in VFM feature extraction ability and segmentation performance.

4.5.3. Ablation study of fusion methods

To investigate the impact of different fusion methods on DLGC-Net's segmentation performance, we design ablation experiments and evaluated fusion complexity using floating point operations (FLOPs), model parameter count, and inference speed. FLOPs measure computational complexity, model parameter count reflects memory requirements, and inference speed assesses runtime efficiency. In this

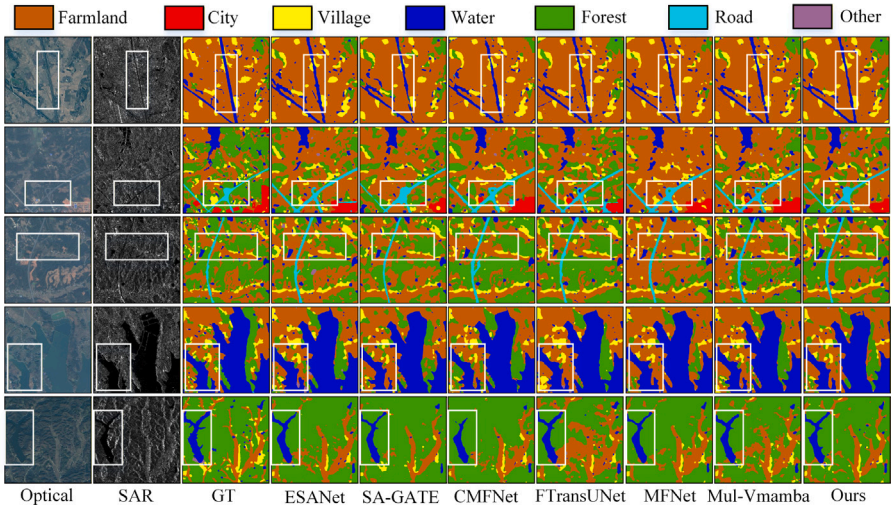

Fig. 9. Visualization of comparative experimental results on the WHU-OPT-SAR dataset. White boxes highlight the differences among the subfigures.

Table 4
Results of ablation experiments for different pre-trained backbones. Red values denote the performance gains achieved by encoders with D^{2}LoRA.

Backbones	mIoU (%)	mF1 (%)	OA (%)
ResNet-101 [24]	85.87	91.88	90.62
MiT-B2 [48]	85.63	92.08	90.78
ViT-Base [26]	85.57	91.57	90.39
ViT-Base (D^{2}LoRA)	86.58(+1.01)	92.66(+1.09)	**91.54**(+1.15)
MTP-Base [30]	84.91	91.64	90.28
MTP-Base (D^{2}LoRA) (Ours)	**86.91**(+1.62)	**92.83**(+0.97)	91.47(+1.05)

Table 5
Effect of the dual diagonal matrices in D^{2}LoRA. LoRA: $\boldsymbol{BA}$ only; Row: $\boldsymbol{W}_1$ with $\boldsymbol{BA}$; Column: $\boldsymbol{W}_2$ with $\boldsymbol{BA}$; Shared: layer-shared diagonals with $\boldsymbol{BA}$; D^{2}LoRA: $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ with $\boldsymbol{BA}$.

Methods	mIoU (%)	mF1 (%)	OA (%)
Full	84.91	91.64	90.28
LoRA	85.73	91.88	90.35
Row	86.60	92.66	91.15
Column	86.69	92.71	91.31
Shared	86.2	92.43	91.10
D^{2}LoRA(Ours)	**86.91**	**92.83**	**91.47**

Table 6
Results of ablation experiments for different fusion methods. GCFF variants from Eq. (7) are also evaluated: GCFF without global average pooling and GCFF without modality B fusion.

Methods	mIoU (%)	mF1 (%)	OA (%)	FLOPs(G)	Parameter(M)	Speed(FPS)
RGBDF [16]	85.87	91.87	89.60	**98.45**	**92.72**	**16.38**
FViT [23]	86.08	92.04	89.91	144.13	263.35	11.69
GCFF w/o B	85.97	91.15	90.68	99.65	97.38	15.74
GCFF w/o GP	86.26	92.38	90.91	99.80	97.90	15.67
GCFF(Ours)	**86.91**	**92.83**	**91.47**	99.66	97.90	15.70

experiment, two alternative fusion schemes are compared against our GCFF: the convolution-based RGB-D Fusion (RGBDF) [16] and the transformer-based FViT [23]. We also ablate GCFF variants in Eq. (7) – one without global average pooling and one without the modality B branch – to assess the impact of pooling on GCFF's fusion performance. Table 6 shows that RGB-D Fusion yields poorer segmentation results, likely because its limited receptive field weakens its ability to model inter-modal dependencies, although it has the lowest computational complexity and the fastest speed. FViT performs better than RGB-D Fusion, indicating stronger fusion capability; however, its transformer-based design has the highest computational complexity and the slowest speed, making efficient multimodal fusion difficult. In addition, the GCFF variant without modality B performs worse than the proposed GCFF, because no modality fusion is applied to the graph Laplacian, making it difficult for modality A to effectively integrate information from modality B. In contrast, GCFF with global average pooling achieves the best segmentation performance, indicating that the method effectively captures long-range inter-modal dependencies by computing correlations between different modalities from global features. Moreover, global average pooling and a small channel reduction factor K keep GCFF's computation and parameter count low, resulting in lower computational complexity and slightly faster speed than transformer-based fusion methods.

Table 7
Results of ablation experiments for the PEFT methods, with the mean and standard deviation from five random runs.

Methods	mIoU (%)	mF1 (%)	OA (%)
SAM-Adapter [31]	85.93 ± 0.38	92.24 ± 0.32	91.03 ± 0.23
LoRA [33]	85.73 ± 0.37	91.88 ± 0.44	90.35 ± 0.49
AdaLoRA [34]	85.64 ± 0.51	92.08 ± 0.36	90.81 ± 0.45
DoRA [35]	86.21 ± 0.29	92.32 ± 0.25	91.08 ± 0.17
LoRA-Dash [36]	86.12 ± 0.34	92.22 ± 0.18	90.97 ± 0.21
D^{2}LoRA (Ours)	**86.91 ± 0.31**	**92.83 ± 0.24**	**91.47 ± 0.20**

4.5.4. Ablation study of the PEFT methods

This section presents an evaluation of the impact of D^{2}LoRA on the segmentation performance of the proposed DLGCNet in comparison with other PEFT methods, through a set of ablation experiments. In these experiments, different PEFT techniques are applied to the multimodal encoder of DLGCNet to assess the effectiveness of the proposed approach. Specifically, five ablation variants of DLGCNet are constructed, each employing a distinct PEFT method: SAM-Adapter [31], LoRA [33], AdaLoRA [34], DoRA [35], and LoRA-Dash [36]. According to the results in Table 7, SAM-Adapter, LoRA, and AdaLoRA improve VFM adaptability to some extent, but they perform worse than D^{2}LoRA because they do not further adjust the low-rank-adapted matrix. DoRA and LoRA-Dash further enhance adaptability by adjusting the low-rank-adapted matrix through trainable vectors and task-specific directions, respectively. D^{2}LoRA achieves the best adaptability because it introduces two diagonal matrices that perform row- and column-wise transformations of the low-rank-adapted matrix. The row transformation matrix $\boldsymbol{W}_1$ enhances adaptation to spatial semantic information,

Table 8
Ablation experiment results for different modality branches across three datasets.

Datasets	Modalities	IoU (%)					mIoU (%)	mF1 (%)	OA (%)
		ImSur.	Buil.	LowVe.	Tree	Car			
Potsdam	RGB	87.02	90.17	75.37	78.19	93.21	84.79	91.62	89.95
	nDSM	47.01	89.12	28.52	56.21	44.49	53.07	67.22	66.23
	RGB-nDSM	**89.50**	**95.03**	**76.62**	**79.72**	**93.69**	**86.91**	**92.83**	**91.47**
Vaihingen	IRRG	88.01	94.19	**79.41**	82.05	**81.24**	83.78	91.02	91.93
	nDSM	58.43	86.01	33.45	69.13	14.41	52.29	64.66	76.60
	IRRG-nDSM	**88.37**	**95.35**	75.15	**83.50**	80.79	**84.63**	**91.52**	**92.59**
	Modalities	IoU (%)					mIoU (%)	mF1 (%)	OA (%)
		Farml.	City	Villa.	Water	Fores.	Road			
WHU-OPT-SAR	RGB	63.62	59.79	46.12	62.57	80.78	34.50	57.90	72.23	81.90
	SAR	55.07	51.93	30.46	51.85	72.71	15.84	46.31	60.99	75.14
	RGB-SAR	**66.08**	**61.40**	**48.55**	**65.65**	**83.18**	**40.20**	**60.84**	**74.74**	**83.84**

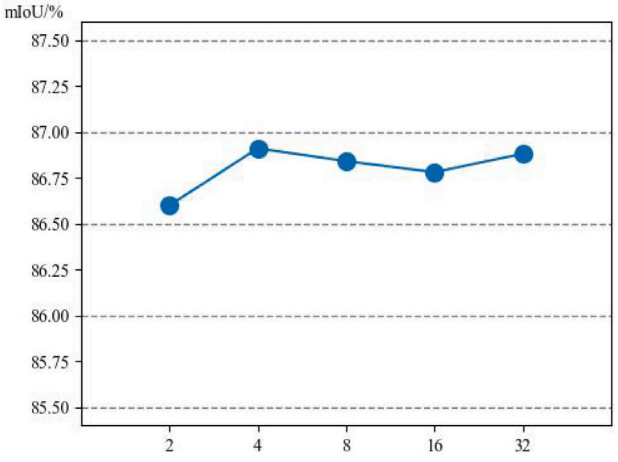

Fig. 10. Effect of different r on the adaptation capability of the pre-trained backbone.

while the column transformation matrix $\boldsymbol{W}_2$ improves adaptation to feature salience. Since MRSI exhibits large variations in spatial scale, positional relationships, and feature diversity, VFM adaptation becomes more difficult; the two diagonal matrices in D^{2}LoRA help address these characteristics and thus yield the best feature representation performance. In summary, this ablation study shows that the proposed D^{2}LoRA effectively improves VFM adaptability to different data distributions compared with other PEFT methods. Table 7 also demonstrates the importance of modifying the low-rank-adapted matrix, especially through separate row- and column-wise transformations, which better enhance VFM feature representation for new data distributions. Therefore, D^{2}LoRA more effectively improves DLGCNet's feature extraction capability for MRSI than the other PEFT methods.

4.5.5. Ablation study on modality-specific feature extraction branches across multiple datasets

To investigate the influence of individual modality branches on DLGCNet's segmentation performance, we design a series of ablation experiments. We compare a single-modality branch network against the dual-branch DLGCNet across all three datasets. For each dataset, two ablation variants are constructed: one employing only the optical modality branch, and the other using only the DSM (SAR) modality branch. The results in Table 8 for all three datasets show that segmentation using a single-modality branch performs poorly, demonstrating the effectiveness of our dual-branch multimodal encoder. On the Potsdam and Vaihingen datasets, both the DSM-only and the DSM-optical multimodal networks achieve superior results on the Building and Tree classes. This is because the DSM primarily encodes height information, making vertically prominent objects more distinguishable; On the WHU-OPT-SAR dataset, both the SAR-only and the SAR-optical multimodal networks excel on the Farmland, Water, and Forest classes, since SAR imagery highlights differences in surface scattering properties, allowing classes with distinct backscatter characteristics to be more readily separated. Consequently, this ablation experiment confirms that the dual-branch multimodal encoder in DLGCNet outperforms single-branch, single-modality encoders, and that incorporating multiple sensor modalities enhances DLGCNet's segmentation accuracy.

4.6. Analysis

4.6.1. Effect of rank r on the adaptation capability of the pre-trained backbone

To investigate how the rank r affects a pre-trained backbone's adaptability, we conduct fine-tuning experiments using D^{2}LoRA with ranks $r = 2, 4, 8, 16, 32$ to explore the relationship between the number of trainable parameters and model performance. The results, presented in Fig. 10, indicate that $r = 4$ yields the best segmentation performance, while overall the choice of rank has a relatively modest effect on the adaptation capability of the pre-trained backbone. This indicates that D^{2}LoRA is not highly sensitive to changes in r, and that a relatively low rank is sufficient for effective adaptation with fewer trainable parameters.

4.6.2. Heatmap analysis

To visualize and analyze the segmentation results of DLGCNet, Fig. 11 presents output layer heatmaps from the baseline FTransUNet and DLGCNet. The heatmaps show each model's attention to multiple land cover classes; higher activations are rendered in warmer colors (closer to red), whereas lower activations are rendered in cooler colors (closer to blue). We also compute the IoU between each thresholding heatmap and the GT; values closer to 1 indicate better agreement between the heatmap and the GT for that class. For k_2 and k_4 in all subfigures ($k = \mathrm{a}, \mathrm{b}, \ldots, \mathrm{f}$), DLGCNet exhibits more concentrated high-confidence activations at the segmentation head and higher correlation coefficients, indicating greater confidence and more accurate, reliable segmentation across classes compared with FTransUNet.

4.6.3. Complexity analysis

To assess the complexity of multimodal fusion models, we evaluate three metrics: FLOPs, model parameter count, and inference speed. Fig. 12 visually presents the complexity analysis of mainstream multimodal data fusion methods evaluated in this work. The results indicate that, despite a relatively larger parameter count, the proposed DLGCNet achieves comparable FLOPs and parameters to transformer-based fusion methods such as CMFNet and FTransUNet while exhibiting lower inference latency. This efficiency stems from DLGCNet's use of

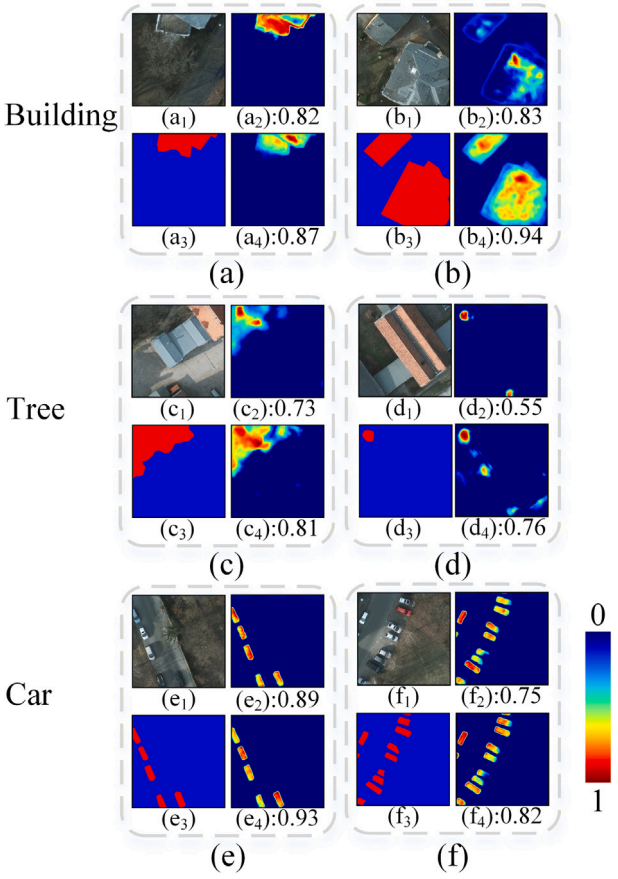

Fig. 11. Heatmap visualizations. In (a), (a_1)–(a_4) show the optical image, the FTransUNet output layer heatmap, the GT, and the DLGCNet output layer heatmap, respectively; the IoU of the thresholding heatmaps is also reported with a threshold of 0.5. (b)–(f) follow the same layout.

a lightweight DSC block for DSM (or SAR) feature extraction and from the small channel-reduction factor K used in the GCFF module's graph Laplacian, which together enable efficient feature fusion. Moreover, compared with ESANet and Mul-VMamba – lightweight fusion networks – DLGCNet yields substantially improved segmentation performance with only a modest increase in model complexity. Relative to the existing best-performing model MFNet, DLGCNet attains superior segmentation accuracy under a more favorable complexity profile. In summary, compared with transformer-based multimodal fusion approaches, DLGCNet performs efficient MRSI feature extraction and fusion at lower computational cost while delivering the SOTA segmentation performance.

5. Conclusion and discussions

In this study, we develop a multimodal encoder by employing the proposed D^2LoRA for PEFT of the VFM and integrating multimodal information via the MFI. This approach effectively preserves the prior knowledge acquired by the VFM during pre-training and provides a more comprehensive feature representation for MRSI. The proposed GCFF employs effective feature-alignment fusion to overcome the limitations of existing fusion approaches: conventional CNN-based fusion methods struggle to capture long-range inter-modal dependencies, and transformer-based fusion approaches incur high computational complexity. In contrast, GCFF achieves effective cross-modal fusion with reduced complexity. Comparative experiments demonstrate that DLGCNet, leveraging the synergistic interplay between feature extraction by

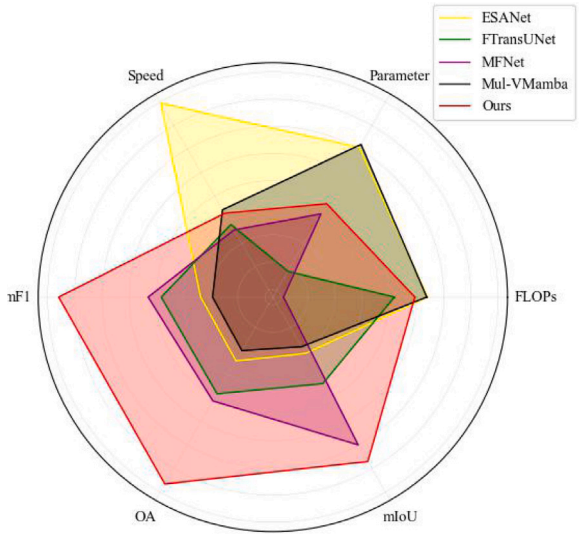

Fig. 12. Complexity results of multimodal data fusion methods. In the radar chart, superior metric values are represented closer to the outer edge.

the multimodal encoder and GCFF's effective feature fusion, achieves significantly superior performance on the MRSI semantic segmentation task. Ablation studies further validate the factors contributing to this performance improvement, and underscore the fundamental principles and overall effectiveness of DLGCNet.

CRediT authorship contribution statement

Junying Zeng: Methodology, Funding acquisition, Conceptualization. **Jiahua Xu:** Writing – original draft, Software, Methodology. **Xudong Jia:** Validation, Resources, Data curation. **Bin Deng:** Writing – review & editing, Supervision. **Yikui Zhai:** Supervision, Funding acquisition. **Chuanbo Qin:** Supervision, Funding acquisition. **Pasquale Coscia:** Supervision, Project administration. **Angelo Genovese:** Supervision, Project administration. **Xiaolin Tian:** Project administration, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the 2025 Key Research Platform Project for Ordinary Universities in Guangdong Province and 2024 Key Research Platform Project for Ordinary Universities in Guangdong Province (Nos. 2025ZDZX1039 and 2024ZDZX1008), the National Natural Science Foundation of China (No. 62273109), the Guangdong Province Key Disciplines Scientific Research Capacity Enhancement Project (No. 2024ZDJS034), and the Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19).

Data availability

The authors do not have permission to share data.

References

[1] Hu Zhu, Shiming Liu, Lizhen Deng, Yansheng Li, Fu Xiao, Infrared small target detection via low-rank tensor completion with top-hat regularization, IEEE Trans. Geosci. Remote Sens. 58 (2) (2020) 1004–1016.

[2] Xuechen Bai, Xinghua Li, Jianhao Miao, Huanfeng Shen, A front-back view fusion strategy and a novel dataset for super tiny object detection in remote sensing imagery, Knowl.-Based Syst. 326 (2025) 114051.

[3] Xin Li, et al., Frequency-guided denoising network for semantic segmentation of remote sensing images, IEEE Trans. Geosci. Remote Sens. 64 (2026) 1–17.

[4] Xin Li, Feng Xu, Anzhu Yu, Xin Lyu, Hongmin Gao, Jun Zhou, A frequency decoupling network for semantic segmentation of remote sensing images, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–21.

[5] Chengqiang Zhao, Shijie Chen, Jiashu Zhang, Xuanmei Fan, Mingzhe Liu, CRLMDG-LM: Causal representation learning-guided multi-target domain generalization network for fine-grained landslide mapping from high-resolution remote sensing images, Knowl.-Based Syst. 330 (2025) 114796.

[6] Xin Li, et al., Position-aware differential denoising transformer for semantic segmentation of remote sensing images, IEEE Geosci. Remote. Sens. Lett. 23 (2026) 1–5.

[7] Wujie Zhou, Chuanming Ji, Weiwei Qiu, Yuanyuan Liu, Adaptive adjustment multiteacher distillation network for visible–Infrared transmission line detection, IEEE Trans. Geosci. Remote Sens. 64 (2026) 1–10.

[8] Han-Sol Ryu, Jeong-Eun Park, Jaehoon Jeong, Sungwook Hong, Generation of hypothetical radiances for missing green and red bands in geostationary environment monitoring spectrometer, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16 (2023) 9025–9037.

[9] Jin Xie, Wujie Zhou, Caie Xu, Yuanyuan Liu, Fangfang Qiang, HCL-net: Heterogeneous collaborative learning for lightweight remote sensing image segmentation, IEEE Trans. Geosci. Remote Sens. 64 (2026) 1–14.

[10] Xuan Hou, Yunpeng Bai, Yefan Xie, Ying Li, Changjing Shang, Qiang Shen, Language-guided change detection for high-resolution remote sensing imagery with limited labelled data, Knowl.-Based Syst. 326 (2025) 113994.

[11] Yu Shen, Jianyu Chen, Liang Xiao, Delu Pan, Optimizing multiscale segmentation with local spectral heterogeneity measure for high resolution remote sensing images, ISPRS J. Photogramm. Remote Sens. 157 (2019) 13–25.

[12] Shunping Zhou, et al., DSM-assisted unsupervised domain adaptive network for semantic segmentation of remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–16.

[13] Wujie Zhou, Jin Xie, Caie Xu, Semantic prompt and graph-convolution-structure distillation framework for semantic segmentation of remote sensing images, IEEE Trans. Geosci. Remote Sens. (2026) 1–14.

[14] Chao Zhan, Kui Yang, WCMamba: Enhancing high-resolution remote sensing image semantic segmentation with pyramid wavelet convolution and SS2d, Knowl.-Based Syst. 324 (2025) 113877.

[15] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, Chen Wu, ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data, ISPRS J. Photogramm. Remote Sens. 162 (2020) 94–114.

[16] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, Horst-Michael Gross, Efficient RGB-D semantic segmentation for indoor scene analysis, in: Proc. IEEE Int. Conf. Robot. Autom., ICRA, 2021, pp. 13525–13531.

[17] Xiaokang Chen, et al., Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2020, pp. 561–577.

[18] Nicolas Audebert, Bertrand Le Saux, Sébastien Lefèvre, Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks, ISPRS J. Photogramm. Remote Sens. 140 (2018) 20–32.

[19] Caner Hazirbas, Lingni Ma, Csaba Domokos, Daniel Cremers, FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture, in: Proc. Asian Conf. Comput. Vis., ACCV, 2017, pp. 213–228.

[20] Jinglin Zhang, et al., GLCANet: Global–local context aggregation network for cropland segmentation from multi-source remote sensing images, Remote. Sens. 16 (24) (2024) 4627.

[21] Xianping Ma, Xiaokang Zhang, Man-On Pun, A crossmodal multiscale fusion network for semantic segmentation of remote sensing data, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 16 (2022) 3463–3474.

[22] Xiaoliang Meng, Yuechi Yang, Libo Wang, Teng Wang, Rui Li, Ce Zhang, Class-guided swin transformer for semantic segmentation of remote sensing imagery, IEEE Geosci. Remote. Sens. Lett. 19 (2022) 1–5.

[23] Xianping Ma, Xiaokang Zhang, Man-On Pun, Ming Liu, A multilevel multimodal fusion transformer for remote sensing semantic segmentationy, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–15.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2016, pp. 770–778.

[25] Wujie Zhou, Beibei Tang, Meixin Fang, Yuanyuan Liu, Dependency then compression: Global dependency network with three-stage knowledge transfer for visible-infrared transmission line detection, IEEE Trans. Geosci. Remote Sens. 64 (2026) 1–13.

[26] Alexey Dosovitskiy, et al., An image is worth 16×16 words: Transformers for image recognition at scale, in: Proc. Int. Conf. Learn. Represent., ICLR, 2021, pp. 1–21.

[27] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen, Towards geospatial foundation models via continual pretraining, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2023, pp. 16760–16770.

[28] Xian Sun, et al., Ringmo: A remote sensing foundation model with masked image modeling, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–22.

[29] Danfeng Hong, et al., SpectralGPT: Spectral remote sensing foundation model, IEEE Trans. Pattern Anal. Mach. Intell. 46 (8) (2024) 5227–5244.

[30] Di Wang, et al., MTP: Advancing remote sensing foundation model via multitask pretraining, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 17 (2024) 11632–11654.

[31] Tianrun Chen, et al., SAM-adapter: Adapting segment anything in underperformed scenes, in: Proc. IEEE/CVF Conf. Comput. Vis. Workshops, ICCVW, 2023, pp. 3359–3367.

[32] Chenze Shao, Yang Feng, Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation, in: Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, ACL, 2022, pp. 2023–2036.

[33] Edward J. Hu, et al., Lora: Low-rank adaptation of large language models, in: Proc. Int. Conf. Learn. Represent., ICLR, 2022, pp. 1–13.

[34] Qingru Zhang, et al., Adaptive budget allocation for parameter-efficient fine-tuning, in: Proc. Int. Conf. Learn. Represent., ICLR, 2023, pp. 1–17.

[35] Shih-Yang Liu, et al., Dora: Weight-decomposed low-rank adaptation, in: Proc. Int. Conf. Mach. Learn., ICML, 2024, pp. 32100–32121.

[36] Chongjie Si, Zhiyi Shi, Shifan Zhang, Xiaokang Yang, Hanspeter Pfister, Wei Shen, Unleashing the power of task-specific directions in parameter efficient fine-tuning, in: Proc. Int. Conf. Learn. Represent., ICLR, 2025.

[37] Xianping Ma, Xiaokang Zhang, Man-On Pun, Bo Huang, A unified framework with multimodal fine-tuning for remote sensing semantic segmentation, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–15.

[38] Alexander Kirillov, et al., Segment anything, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV, 2023, pp. 3992–4003.

[39] Keyan Chen, et al., Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–17.

[40] Wujie Zhou, Xiaomin Fan, Weiqing Yan, Shengdao Shan, Qiuping Jiang, Jenq-Neng Hwang, Graph attention guidance network with knowledge distillation for semantic segmentation of remote sensing images, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–15.

[41] Evan Shelhamer, Jonathan Long, Trevor Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.

[42] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, MICCAI, 2015, pp. 234–241.

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, Pyramid scene parsing network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 6230–6239.

[44] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.

[45] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, Peter M. Atkinson, ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery, ISPRS J. Photogramm. Remote Sens. 181 (2021) 84–98.

[46] Xin Li, et al., A synergistical attention model for semantic segmentation of remote sensing images, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–16.

[47] Jieneng Chen, et al., TransUNet: Transformers make strong encoders for medical image segmentation, 2021, arXiv:2102.04306.

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, in: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 34, 2021, pp. 12077–12090.

[49] Libo Wang, et al., UNetFormer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, ISPRS J. Photogramm. Remote Sens. 190 (2022) 196–214.

[50] Haiming Zhang, Guorui Ma, Hongyang Fan, Hongyu Gong, Di Wang, Yongxian Zhang, SDCINet: A novel cross-task integration network for segmentation and detection of damaged/changed building targets with optical remote sensing imagery, ISPRS J. Photogramm. Remote Sens. 218 (2024) 422–446.

[51] Li Yan, Jianming Huang, Hong Xie, Pengcheng Wei, Zhao Gao, Efficient depth fusion transformer for aerial image semantic segmentation, Remote. Sens. 14 (5) (2022) 1294.

[52] Rongrong Ni, et al., Mul-vmamba: Multimodal semantic segmentation using selection-fusion-based vision-mamba, Knowl.-Based Syst. 334 (2026) 115119.

[53] Hamidreza Hosseinpour, Farhad Samadzadegan, Farzaneh Dadrass Javan, CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images, ISPRS J. Photogramm. Remote Sens. 184 (2022) 96–115.

[54] Wujie Zhou, Jin Xie, Caie Xu, Yuanyuan Liu, Yunchao Wang, Adapt, generate, and supervise: Geometry-aware diffusion-guided SAM framework for remote sensing semantic segmentation, IEEE Trans. Geosci. Remote Sens. 64 (2026) 1–14.
[55] Xinyang Pu, Feng Xu, Low-rank adaption on transformer-based oriented object detector for satellite onboard processing of remote sensing images, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–13.
[56] Xiaoyan Lu, Qihao Weng, Multi-LoRA fine-tuned segment anything model for urban man-made object extraction, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–19.
[57] Xichuan Zhou, et al., MeSAM: Multiscale enhanced segment anything model for optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–15.
[58] Binbin Song, Hui Yang, Yanlan Wu, Peng Zhang, Biao Wang, Guichao Han, A multispectral remote sensing crop segmentation method based on segment anything model using multistage adaptation fine-tuning, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–18.
[59] Liye Mei, et al., SCD-SAM: Adapting segment anything model for semantic change detection in remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–13.
[60] Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, Lorenzo Bruzzone, Adapting segment anything model for change detection in VHR remote sensing images, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–11.
[61] Ashish Vaswani, et al., Attention is all you need, in: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 1–11.
[62] Andrew G. Howard, et al., MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv:1704.04861.
[63] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2018, pp. 7132–7141.
[64] F. Rottensteiner, et al., The ISPRS benchmark on urban object classification and 3D building reconstruction, ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. 1 (1) (2012) 293–298.
[65] Xue Li, et al., MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification, Int. J. Appl. Earth Obs. Geoinf. 106 (2022) 102638.