FDGSNet: A Multimodal Gated Segmentation Network for Remote Sensing Image Based on Frequency Decomposition

Jian Cui, Jiahang Liu, *Member, IEEE*, Yue Ni, Jinjin Wang, and Manchun Li

***Abstract*—Multiple modal data fusion can provide valuable and diverse information for remote sensing image segmentation. However, the existing fusion methods often lead to feature loss during the fusion of various modal data, and the complementarity among multimodal features is insufficient. To address these problems, we propose a multimodal gated segmentation network for remote sensing images based on the frequency decomposition. Complementary information from multimodal features is extracted by establishing a long-distance correlation between the low-frequency components of different modal data. In addition, high-frequency detailed features of different modal data are preserved by residual connection. The adaptive gated fusion method is then used to control the information flow between the complementary information and each modality feature map, enabling adaptive fusion between multimodal features. These operations can effectively improve the adaptability of the proposed method in various scenarios and data changes. Extensive experiments demonstrate that the proposed method has good effectiveness, robustness, and generalization and achieved state-of-the-art performance in several remote sensing image semantic segmentation tasks.**

***Index Terms*—Frequency-domain decomposition, multimodal, remote sensing, semantic segmentation.**

I. Introduction

REMOTE sensing image (RSI) semantic segmentation is one of the primary methods for modern spatial information acquisition and has been widely applied to urban planning [1], [2], [3], smart city construction [4], and geographic information system development [5], [6].

In recent years, the increase of high-resolution remote sensing satellites has generated abundant spatial data, enhancing the interpretation and application capabilities of remote sensing information [7], [8]. Using these high-resolution images, spatial information in different modalities can be achieved for different purposes. For example, visible images (RGB bands) are mainly used to extract color features, texture features, shape features, and spatial relationship features. However, visible images are sensitive to illumination conditions and easily disturbed by environmental factors. The spectral response of the near-infrared (NIR) band is closely related to the reflection characteristics of surface coverings (such as vegetation, water, and soil). These ground cover types have significant differences in NIR band, so the differences can be used for ground cover classification and target recognition. The digital surface model (DSM) can reflect the vertical height information of ground objects, providing valuable insights into the spatial location and environmental background of buildings. This effectively reduces the influence of shadows and occlusion on the segmentation of buildings. The normalized difference water index (NDWI) is a ratio index that utilizes the green and NIR bands, which can highlight the characteristics of water bodies and reduce the impact of aquatic plants and clouds on water segmentation. The normalized difference vegetation index (NDVI) reflects the reflectance differences between vegetation in the NIR and red wavelengths, which can enhance the contrast between vegetation regions and the background, effectively reducing the impact of environmental factors, such as lighting conditions and shadows on the semantic segmentation of vegetation.

Fusing multimodal data to achieve semantic segmentation of RSIs can make full use of the complementary features of different data types to obtain more comprehensive and rich semantic information, thereby improving the accuracy and reliability of semantic segmentation. However, multimodal data generated by different sensors exhibit vastly distinct characteristics, such as cross-modal heterogeneous statistical properties and noise levels [9]. Directly performing operations, such as element addition, do not accurately extract and fuse complementary information between two data modalities and may introduce additional noise that undermines the parsing performance. Therefore, effectively extracting complementary information from different modalities to obtain richer features than those derived from a single data source has become a new technical challenge in the field of remote sensing data processing [10], [11].

An efficient feature representation serves to enhance the discriminative capacity of category information [12], [13]. Constructing the feature space of a multisource segmentation model by fully exploring the complementary and discriminative features of different data sources is key to achieving high-precision

Received 20 May 2024; revised 21 August 2024; accepted 24 September 2024. Date of publication 1 October 2024; date of current version 8 November 2024. This work was supported in part by the Innovative Talent Program of Jiangsu under Grant JSSCR2021501 and in part by the High-Level Talent Plan of NUAA, China. *(Corresponding author: Jiahang Liu.)*

Jian Cui, Jiahang Liu, Yue Ni, and Jinjin Wang are with the Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: jhliu@nuaa.edu.cn).

Manchun Li is with the School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China.

The code will be available at: https://github.com/Cuiadd/FDGSNet.

Digital Object Identifier 10.1109/JSTARS.2024.3471638

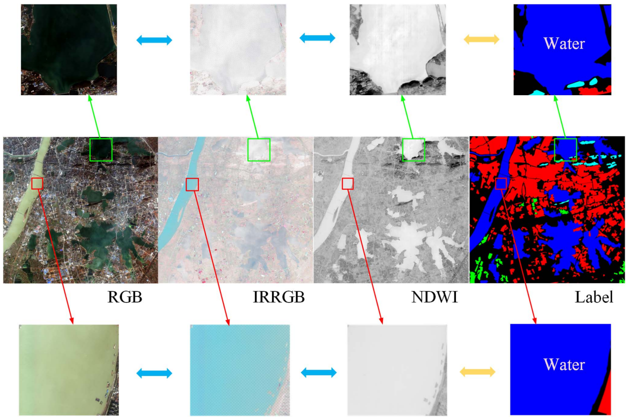


Fig. 1. Different modal data in the same scene have different distributions of features.

segmentation results of multimodal RSIs. The transformer-based mechanism is capable of extracting and integrating complementary information from the global context of the input data more effectively due to its strong ability to capture remote dependencies [14], [15].

However, as shown in Fig. 1, different modal data have different feature distributions in the same scene, which can cause information conflict and redundancy. Taking water detection as an example, water bodies in different regions have different optical properties, which may show obvious feature differences in visible light images and multispectral images. Therefore, effective information screening and feature fusion are needed to avoid the interference of redundancy and conflict in semantic segmentation. Unlike semantic segmentation tasks of ordinary scene images (e.g., cars and people), RSI targets (e.g., water bodies, vegetation, buildings, etc.) often have similar semantic features but lack consistent boundary features, and in the process of fusing features from different modalities, the underlying features of each modality are inevitably lost, resulting in spatial detail loss in the underlying features. In addition, different modal data have different recognition capabilities for different types of targets, and existing methods have not fully reflected the differences in the recognition ability of different modal data for different objects. Finally, the semantic segmentation task of RSIs requires a high generalization ability of the model. How to improve the generalization ability and interpretability of the model to adapt to different scenarios and data changes is also an important challenge.

To address the above issues, a multimodal gated segmentation network based on the frequency decomposition is proposed for RSIs. Since low-frequency components of images carry semantic information, high-frequency components contain spatial details [16]. Therefore, we aim to facilitate the extraction of modality-specific and modality-shared features by increasing and decreasing the correlation between low-frequency and high-frequency features, respectively. We initially decompose the multimodal features into high- and low-frequency components via a frequency-domain decomposition module (FDM). Then, the long-distance correlation model between the low-frequency components of different modalities was established by using the cross-modal cross-self-attention mechanism, and the correlation features between different modalities were extracted from the low-frequency components. As an independent branch, the high-frequency component is not affected by the fusion features and the residual connection retains more detailed texture features in the high-frequency features. Finally, the adaptive gated fusion (AGF) method is used to extract effective features from complementary information and adaptively fuse them with different modal data, which effectively integrates multimodal features.

FDGSNet fully utilizes the characteristics of high-frequency features and low-frequency features, extracts complementary features from low-frequency semantic features, and reduces high-frequency signal interference between different data types. This feature fusion method reduces the sensitivity of the model to data types and improves the interpretability and generalization ability of the model. The main contributions of this work can be summarized as follows.

1) We propose a generalized multimodal data fusion network (FDGSNet) for the semantic segmentation of RSIs. It can integrate different prior knowledge for different segmentation tasks to achieve the accurate segmentation results of various targets.
2) We propose an FDM to decompose the data of different modalities into high-frequency and low-frequency components, which are used to preserve the details from different modes and extract the semantic features of the target, respectively.
3) We propose a gated complementary fusion module in which the cross-modal cross-self-attention mechanism is used to establish the correlation between the low-frequency components of different modality data to extract complementary information. An AGF method is used to control the information flow between the complementary information and each modality feature map, and the cross-modal features are adaptively complementary and fused according to the differences in the recognition ability of different modal data for different objects.

The experimental results on different datasets validate the excellent feature fusion capability of our proposed FDGSNet for different models. Compared with other state-of-the-art methods, FDGSNet can efficiently integrate different prior knowledge and achieve the best accuracy in multiclass semantic segmentation tasks, as well as single-class segmentation tasks exemplified by buildings, vegetation, and water bodies.

The rest of this article is organized as follows. Section II first reviews the related works on CNN-based and transformer-based remote sensing segmentation methods. After that, Section III presents the structure of the proposed FDGSNet, whereas Section IV provides details on the extensive experiments conducted. Finally, Section V concludes this article.

II. Related Works

A. RSI Semantic Segmentation

The success of convolutional neural networks in computer vision tasks has spurred increased interest among researchers in utilizing deep learning techniques for semantic segmentation of RSIs. Zeng et al. [17] proposed a novel cross-scale feature propagation network to address the limitations of existing methods that rely on a single strategy for multiscale information capture, aiming to improve performance when processing RSIs with large-scale variance by capturing fine-grained multiscale context, embedding high-level semantic information into low-level features, and enhancing final feature representation. Zheng et al. [18] proposed a high-order semantic decoupling network to address feature distortion in RSIs caused by view angle transformation and atmospheric scattering, utilizing high-order features for semantic segmentation and decoupling to enhance feature robustness and improve segmentation performance. To solve the challenges of semantic segmentation in ultrahigh-resolution RSIs, Chen et al. [19] combined global context with spatial detail features and employed multitask learning to enhance boundary detection and segmentation, thereby achieving improved segmentation accuracy. Li et al. [20] proposed the spectrum-space collaborative network (SSCNet), which integrates spectral and spatial information to enhance the discriminative potential of representations and improve the quality of semantic segmentation in RSIs. In addition, to improve low-level feature extraction in RSI semantic segmentation and overcome traditional convolution limits, Xiao et al. [21] introduce directional convolution and large field convolution, enhancing deep learning network performance. Although these methods have demonstrated impressive performance, they rely exclusively on a single-source input, which may lack robustness across diverse scenarios and thereby hinder further improvement.

B. Modality Fusion in Remote Sensing Segmentation

Recent research has demonstrated that incorporating features from multiple sources, including HSI, thermal, NIR, and LiDAR, can significantly enhance the stability and accuracy of scene-parsing tasks. Li and Zhou [22] proposed a novel dual-modal semantic segmentation network called MASNet, which combines optical image and LiDAR point cloud features to enhance scene understanding in complex driving environments, featuring a unique MmASPP structure for multiscale contextual information capture and an adaptive synergy difference loss function to optimize cross-modal operations. To enhance scene parsing in RSIs, especially for unbalanced categories and small targets, Ma et al. [23] introduced the ABHNet, utilizing DSM and adjacent context, improving segmentation performance on benchmark datasets. To address the issue of underutilization of modality-specific characteristics and complementary information during RGB-thermal semantic segmentation, Liang et al. [24] proposed the MDBFNet, a multibranch differential bidirectional fusion network that achieves more reliable semantic scene understanding by enhancing detail and semantic information through specifically designed modules and a three-branch fusion decoder.

C. Vision Transformer

CNN-based methods are the predominant solution for semantic segmentation tasks but are limited by their local receptive fields, restricting their ability to capture long-range dependencies [25]. Transformer-based methods overcome this limitation by establishing long-distance pixel dependencies, effectively addressing complex spatial relationships. They extend the receptive field to the entire input feature map in a single pass, enabling a comprehensive pixelwise response. Therefore, it has been widely used in visual and language processing tasks due to its strong ability to acquire spatial information and establish global relationships [26], [27]. Cao et al. [28] proposed a global feature fusion network to enhance semantic segmentation accuracy in high-resolution RSIs by integrating global contextual features with local features. Li et al. [29] proposed a novel approach utilizing a multihead attention-attended module to refine the self-attention mechanism, aiming to improve the semantic segmentation accuracy of RSIs by filtering out irrelevant contexts and emphasizing informative ones. To address the challenges of limited receptive fields, insufficient global feature extraction, and inaccurate edge positioning in RSI segmentation, Cui et al. [30] proposed the global context dependency-aware network, achieving high-accuracy segmentation results through a novel dot-product attention mechanism and an edge-aware optimization module. Fan et al. [31] proposed CSTUNet, a dual-encoder model that combines CNN and Swin transformer, to address the limitations of CNNs in modeling global context for remote sensing semantic segmentation, aiming to improve segmentation accuracy by preserving details and enhancing contextual information fusion. In addition, to address the limitations of using CNNs or transformers alone in remote sensing semantic segmentation tasks, especially under resource-constrained scenarios, Dong et al. [32] proposed a novel cross-modal knowledge distillation framework named DSCTs, which harnesses the complementary advantages of both models to improve the student model's segmentation performance without adding trainable parameters.

In addition, in the field of multimodal semantic segmentation, transformer-based methods are also used for interactive fusion between different data sources. To address the limitations posed by single-modal data in land cover classification, Ren et al. [33] proposed SwinTFNet, a dual-stream deep fusion network that deeply integrates SAR and optical features, enhancing segmentation performance in clouded images and achieving superior multimodal data classification compared with other methods. Ma et al. [34] proposed a multilevel multimodal fusion approach called FTransUNet, which integrated CNN and vision transformer into a unified framework to enhance semantic segmentation accuracy by effectively fusing shallow and deep features, demonstrating superior performance on fine-resolution remote sensing datasets compared with other multimodal fusion methods. Zhang et al. [35] propose a unified fusion framework, CMX, for RGB-X semantic segmentation, achieving state-of-the-art performances on multiple datasets by effectively fusing features from different modalities and demonstrating strong generalizability across diverse sensor modalities.

While these methods greatly improve the segmentation accuracy compared with single-source inputs, they are not specifically designed for noise interference and loss of detailed features of the data in multimodal remote sensing data fusion.

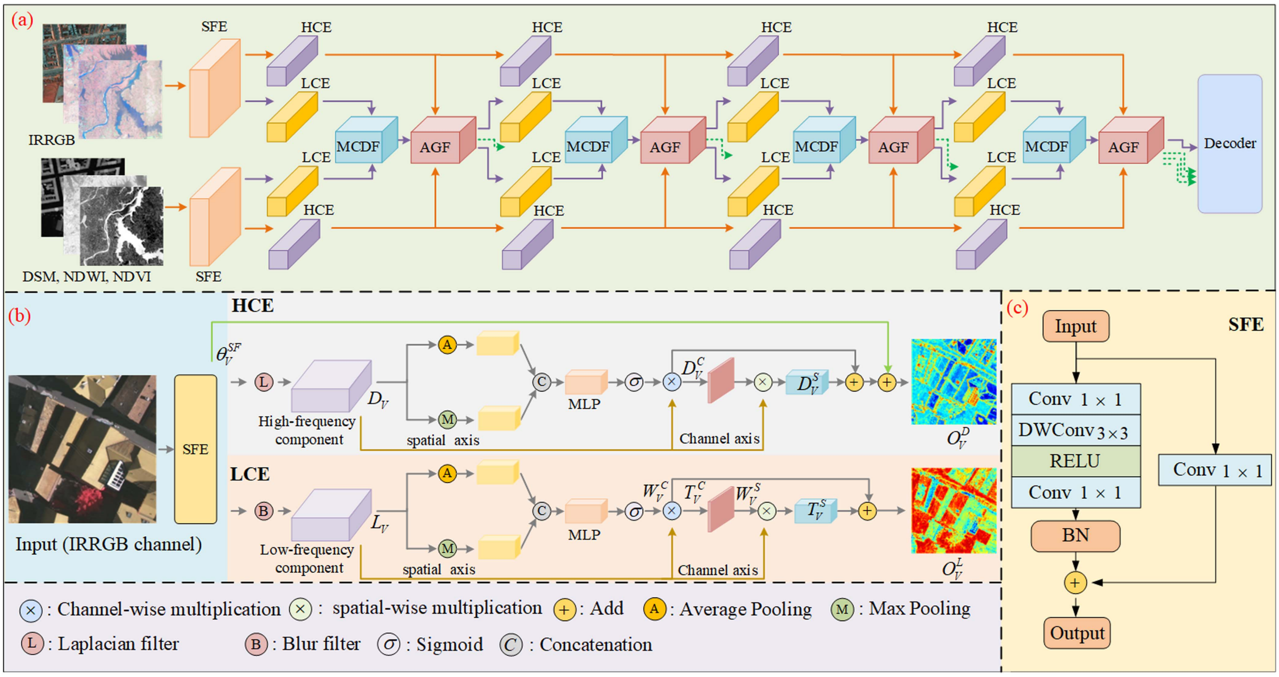


Fig. 2. Overview of FDGSNet for multimodal semantic segmentation. The inputs are an IRRGB image and another modality data (e.g., DSM, NDWI, and NDVI). (b) Detailed architecture of the FDM, including HCE and LCE. (c) Detailed architecture of the SFE.

In addition, the existing multimodal RSI fusion methods are especially designed for specific segmentation tasks and data sources, but their performance is limited to specific tasks, with a relatively narrow scope of functionality and insufficient model generalization ability. Currently, there is a lack of a general multimodal feature fusion method that can extract effective prior knowledge from different data sources and perform multiple RSI segmentation tasks. In this article, we try to address this issue.

III. Methodology

A. Framework Overview

In this work, a cross-modal fusion framework (FDGSNet) is proposed for the semantic segmentation of multimodal RSIs. The schematic diagram of FDGSNet is depicted in Fig. 2(a). This framework employs two parallel backbones for extracting features from IRRGB (NIR-RGB) images and prior knowledge inputs, encompassing IRRGB-DSM, IRRGB-NDWI, or IRRGB-NDVI. Using IRRGB-NDWI image fusion as a case study, NIR images, RGB images, and NDWI images from the same scene share common statistical characteristics in their low-frequency information, encompassing background elements and extensive environmental features. However, their high-frequency information is generally independent. Specifically, RGB images capture detailed textures and color information, while NDWI images highlight the distinct characteristics of water bodies. Therefore, we design an FDM, and the data of different modes are decomposed into high-frequency features and low-frequency features. The high-frequency component is used to retain the detailed features of the data of different modes and the low-frequency component mainly retains the main features of different objects. The complementary information between multimodal features is extracted by establishing the correlation between the low-frequency components of different modal features. In addition, we generate spatialwise gates for both correlation degree fusion feature and multimodal features and use the soft attention mechanism to control the information flow between the fusion feature and each modality feature map to realize the complementary fusion between multimodal features. The multimodal feature fusion method based on the frequency-domain decomposition effectively mitigates the noise interference and the loss of edge detail features generated during the fusion of different data sources. This enhances the generalization capability of the model and enables it to obtain valuable prior knowledge from various modal data features.

B. Frequency-Domain Decomposition Module

The FDM is used to decompose the multimodal data into high-frequency component features and low-frequency component features. First, we define some symbols for clarity in formulation. Let H, W, and C denote the height, width, and channels of an input image, respectively. The input-paired IRRGB images and multimodal images are denoted as $V \in \mathbb{R}^{H\times W\times C}$ and $I \in \mathbb{R}^{H\times W\times C}$. The share feature encoder (SFE), high-frequency component extraction (HCE), and low-frequency component extraction (LCE) are represented by $\mathrm{SF}(\cdot)$, $\mathrm{HF}(\cdot)$, and $\mathrm{LF}(\cdot)$, respectively.

SFE: SFE aims to extract shallow features $\{\theta_V^{\mathrm{SF}}, \theta_I^{\mathrm{SF}}\}$ from IRRGB multispectral and multimodal inputs $\{V, I\}$, i.e., $\theta_V^{\mathrm{SF}} = \mathrm{SF}(V), \theta_I^{\mathrm{SF}} = \mathrm{SF}(I)$.

SFE employs a bottleneck residual block; the bottleneck residual block uses a projection shortcut structure. The overview of SFE is shown in Fig. 2(c). SFE uses 1 × 1 convolution to reduce the number of input channels, which can greatly reduce the complexity and computational burden of the model. After that, 3×3 convolution and 1×1 convolution are used to increase the dimension, ensure the complexity of the network, and make full use of the input information. Batch normalization is added after each 3×3 convolution layer to normalize the input values, speed up training, and improve model performance.

LCE: The LCE is to extract low-frequency base features from the shared features. The low-frequency features of the two modalities are obtained by the blur low-pass filter, respectively. The parameter settings of the filters are based on the parameter values used in [16]. The overview of LCE is shown in Fig. 2(b). The obtained low-frequency component features $L_V \in \mathbb{R}^{H\times W\times C}$ and $L_I \in \mathbb{R}^{H\times W\times C}$ are embedded into two attention vectors along the spatial axis using global maximum pooling and global average pooling, respectively, to retain more information. And then you connect these two vectors. Then, the multilayer perceptron (MLP) and sigmoid function are used to obtain the weight $W_V^C \in \mathbb{R}^C$ and $W_I^C \in \mathbb{R}^C$, the weight value is multiplied by the low-frequency component features to obtain the channel attention activation features T_V^C and T_I^C

$$W_V^C = \sigma(f_{\text{mlp}}(\text{Pool}_{\text{Max}}(L_V) + \text{Pool}_{\text{Ave}}(L_V))) \tag{1}$$

$$W_I^C = \sigma(f_{\text{mlp}}(\text{Pool}_{\text{Max}}(L_I) + \text{Pool}_{\text{Ave}}(L_I))) \tag{2}$$

$$T_V^C = W_V^C \circledast L_V \tag{3}$$

$$T_I^C = W_I^C \circledast L_I \tag{4}$$

where σ represents the sigmoid function, $+$ denotes the concatenation, and $\circledast$ denotes the channelwise multiplication. The channel attention activation features T_V^C and T_I^C, two kinds of data, are embedded into two spatial weight plots along the channel axis using global mean pooling. The embedding operation has two 1×1 convolution layers assembled with a ReLU function. Then, the sigmoid function is applied to get two weight graphs $W_V^S \in \mathbb{R}^{H\times W}$ and $W_I^S \in \mathbb{R}^{H\times W}$. The process formula for obtaining the spatial weight graph is given as follows:

$$W_V^S = \sigma(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{1\times1}(T_V^C)))) \tag{5}$$

$$W_I^S = \sigma(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{1\times1}(T_I^C)))). \tag{6}$$

The weight value is multiplied by the low-frequency component features to obtain the spatial attention activation features T_V^S and T_I^S

$$T_V^S = W_V^S * L_V \tag{7}$$

$$T_I^S = W_I^S * L_I \tag{8}$$

where $*$ denotes the spatialwise multiplication.

Finally, short connections are used to fuse the low-frequency component features that are activated through channels and spatial dimensions to obtain the output features O_V^L and O_I^L

$$O_V^L = T_V^C + O_V^S \tag{9}$$

$$O_I^L = T_I^C + O_I^S. \tag{10}$$

HCE: The high-frequency component adopts the same structure as the low-frequency component. The overview of HCE is shown in Fig. 2(b). The difference is that the high-frequency components $D_V \in \mathbb{R}^{H\times W\times C}$ and $D_I \in \mathbb{R}^{H\times W\times C}$ are obtained by using the Laplacian high-pass filter for the shared features. The parameter settings of the filters are based on the parameter values used in [16]. Then, the high-frequency component is used to extract the features of the channel and spatial attention mechanism, and the channel attention enhancement feature $\{D_V^C, D_I^C\}$ and the spatial attention enhancement feature $\{D_V^S, D_I^S\}$ are obtained. In addition, the module adds a short connection that fuses the input primitive feature $\{\theta_V^{\text{SF}}, \theta_I^{\text{SF}}\}$ with the low-frequency component features activated through channels and spatial dimensions of attention to obtain the output features O_V^D and O_I^D

$$O_V^D = \theta_V^{\text{SF}} + D_V^C + D_V^S \tag{11}$$

$$O_I^D = \theta_I^{\text{SF}} + D_I^C + D_I^S. \tag{12}$$

As an independent branch, the high-frequency component is not affected by the fusion features and the residual connection retains more detailed texture features of multimodal features. With the forward propagation of the network, the detailed features are integrated into the gated complementary fusion features layer by layer.

The high-frequency components of the data for different modes are not correlated. During multimodal feature fusion, reducing the interference of high-frequency components is beneficial for extracting relevant semantic features from different modal data. Therefore, we improve the controllability and interpretability of feature extraction by adding correlation restrictions to the extracted features. The multimodal feature fusion method based on frequency-domain decomposition uses low-resolution features for fusion, which has a strong generalization ability for input data types. Therefore, compared with other methods, our method has better interpretability, stronger robustness to different data sources, and different semantic segmentation tasks, and can learn effective prior knowledge from different modal data features.

C. *Gated Complementary Fusion Module*

After obtaining the high-frequency and low-frequency components of each modality, we designed a two-stage gated complementary fusion module to merge the features from the two modalities and enhance cross-modal information interaction. As shown in Fig. 3, in the first stage, a multimodal feature correlation degree fusion (MCDF) module is designed to globally exchange information between the low-frequency components of both modalities, achieving complementary fusion of shared features from the two modalities and resulting in multimodal correlation fusion features. In stage 2, through the AGF module, the features of different modes can be self-adaptively selected from the MCDF features for feature enhancement. Finally, the enhanced multimodal features are fused.

MCDF: In this phase, we introduce an efficient multimodal cross-self-attention mechanism to facilitate sufficient information exchange between interactive vectors from diverse modal

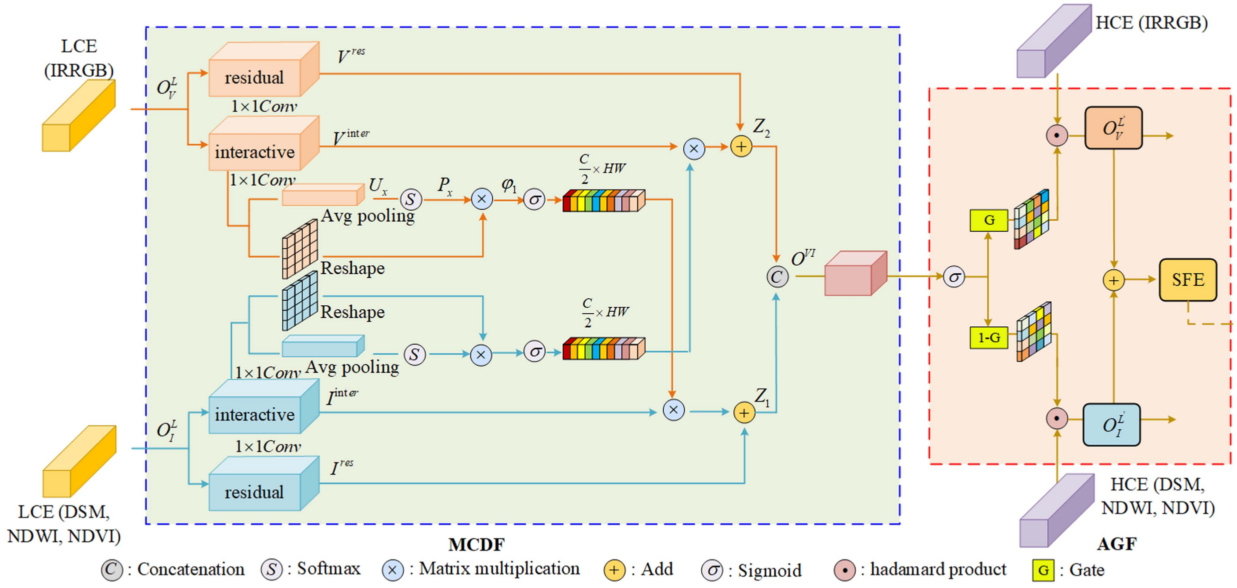


Fig. 3. Detailed architecture of gated complementary fusion module.

paths. Our multimodal cross-self-attention mechanism for enhancing cross-modal feature fusion is based on the traditional self-attention. This establishes the correlation features between different modal data from a sequence-to-sequence perspective. For brevity, we take the IRRGB multispectral image path for illustration. We first utilize two 1×1 convolutions to extract features from input features with size $O_V^L \in \mathbb{R}^{H\times W\times C}$, resulting in two feature vectors referred to as the residual vector V^{res} and the interaction vector V^{inter}.

Global average pooling is applied to the interactive vector V^{inter} across the spatial dimension to generate the feature vector U_x, facilitating the capture of global feature information for each feature layer. Subsequently, the vector P_x is obtained by employing the Softmax activation function, thereby enhancing the dynamic range of feature activation. The calculation formula is provided as follows.

$$P_x = \frac{e^{U_x}}{\sum_{i-1}^{n} e^{U_i}}. \tag{13}$$

Then, the projective function $\mu(\cdot) : \mathbb{R}^{H\times W\times (C/2)} \rightarrow \mathbb{R}^{(C/2)\times HW}$ is used to change the dimension of the interactive vector V^{inter}.

The features P_x and $\mu(V^{\text{inter}})$ are subjected to similarity calculation to obtain a similarity matrix $\varphi_1 = P_x \otimes \mu(V^{\text{inter}})$, where $\otimes$ is a matrix multiplication. Then, the sigmoid function is applied to normalize the vector φ_1, which is subsequently multiplied I^{inter} to obtain the attention-activated feature map Z_1. The calculation formula is given as follows:

$$Z_1 = I^{\text{res}} + I^{\text{inter}} \otimes \psi\left(\frac{1}{1+e^{-\varphi_1}}\right) \tag{14}$$

where Z_1 is the feature map after cross-modal cross-self-attention activation and $\psi(\cdot)$ is a projection function $\psi(\cdot) : \mathbb{R}^{(C/2)\times HW} \rightarrow \mathbb{R}^{H\times W\times (C/2)}$. In this process, I^{inter} can be called a query matrix, and P_x can be called a key matrix.

For the multimodal images path, we use the same method to calculate the feature Z_2 after cross-modal cross-self-attention activation, and add Z_1 and Z_2 to obtain the MCDF feature O^{VI}.

AGF: Different modal data have different recognition capabilities for different types of targets. To fully exploit the characteristics of different modal data, we need to aggregate spatial cross-modal features based on the difference in the representation ability of different modal data for different types of objects. To achieve this, we generate spatialwise gates for both correlation degree fusion feature O^{VI} and multimodal image features and use the soft attention mechanism to control the information flow between the fusion feature O^{VI} and each modality feature map, which is visualized in Fig. 3, and marked by the second red frame. To make the gate more precise, we use an MCDF feature O^{VI} to generate the gate. The calculation formula was given as follows:

$$O_V^{L'} = O_V^D \odot \sigma(O^{VI}) \tag{15}$$

$$O_I^{L'} = O_I^D \odot (1 - \sigma(O^{VI})) \tag{16}$$

$$O_1^{VI} = O_V^{L'} + O_I^{L'} \tag{17}$$

where σ represents the sigmoid function, and $\odot$ is a Hadamard product operator. O^{VI} contains the correlation characteristics of different modes. Sigmoid activation O^{VI} can adaptively select the information that is effective for the respective modal data from the correlation degree features and carry out feature fusion.

IV. Comparison and Discussion

In this section, we conducted semantic segmentation experiments on RSIs for water bodies, vegetation, buildings, and panoptic images, respectively, by utilizing different combinations of data sources. Then, we conducted ablation studies on each module of the proposed method. Finally, through comparisons with other methods, we further validated the effectiveness and generality of the proposed approach.

A. Dataset

Potsdam dataset: The Potsdam dataset comprises 38 high-resolution aerial RSIs and their corresponding DSMs. From these, a random selection of 24 images is designated as the training set, with the remaining images serving as the test set. Both the images and DSMs have a spatial resolution of 5 cm and dimensions of pixels. These high-resolution images include the red (R), green (G), blue (B), and NIR bands. The original labeled data encompass six major land cover classes. In this study, buildings are defined as the foreground, while all other objects are classified as the background.

GID dataset: The Gaofen image dataset (GID) is an open-source satellite remote sensing dataset. It comprises 10 finely labeled land cover images and 150 coarsely labeled land cover images. All images are captured by the GF-2 satellite, with a pixel resolution of 4 m. Each image contains the four bands of R, G, B, and NIR. We obtained NDWI and NDVI data corresponding to IRRGB images by calculating between bands. We selected 150 large-scale images with a size of 7200 × 6800 to validate the proposed method and selected 80 multispectral images (IRRGB) and the corresponding normalized index data (NDVI and NDWI) as the training set, the remaining 70 for testing. In the experiments on water body detection, we redefined the original dataset by categorizing rivers, lakes, and ponds as water bodies while assigning other categories as background. Subsequently, we combined this redefined dataset with NDWI to create the multimodal GID (water body) dataset. Similarly, in the experiments on vegetation detection, forest, grassland, and dry fields are divided into vegetation categories, while other categories are divided into background, and combined with NDVI to make the multimodal GID (vegetation) dataset.

B. Implementation Details

FDGSNet performs a downsampling operation before each FDM and increases the number of channels. After each downsampling, the network passes through one FDM and one gated complementary fusion module. Neither the FDM nor the gated complementary fusion module changes the number of channels in the input data. Therefore, each FDM has the same number of channels as the adjacent gated complementary fusion module. The model was downsampled four times in total, and the channel number C after each downsampling was successively (128, 256, 512, 1024). We take an MLP decoder with an embedding dimension of 512, introduced in SegFormer [42], select AdamW optimizer [45] with weight decay 0.01, and use cross entropy as the loss function. The original learning rate is set as $6e^{-5}$ and a poly learning rate schedule is employed.

C. Evaluation Metrics

In our conducted experiments, we utilized the Adam optimizer, configuring the learning rate to 0.0003. The evaluation of experimental outcomes was performed using metrics, including overall accuracy (OA), mean intersection over union (mIoU), and frequency-weighted intersection over union (FWIoU). The calculation formulae for these metrics are provided as follows:

$$\mathrm{OA} = \frac{\sum_{k=1}^{N} \mathrm{TP}_k}{\sum_{k=1}^{N} \mathrm{TP}_k + \mathrm{FP}_k + \mathrm{TN}_k + \mathrm{FN}_k} \tag{18}$$

$$\mathrm{MIoU} = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}. \tag{19}$$

The FWIoU metric for semantic segmentation is an improvement over the original MIoU, which assigns different weights to each class based on their frequency of occurrence

$$\mathrm{FWIoU} = \frac{\sum_{k=1}^{N} \left(\frac{\mathrm{TP}_k}{\mathrm{TP}_k+\mathrm{TN}_k+\mathrm{FN}_k} \cdot \frac{\mathrm{TP}_k+\mathrm{FN}_k}{\mathrm{TP}_k+\mathrm{FP}_k+\mathrm{TN}_k+\mathrm{FN}_k} \right)}{N+1} \tag{20}$$

where TP_k, FP_k, FN_k, and TN_k indicate the true positive, false positive, true negative, and false negatives, respectively, for object indexed as class *k*.

D. Ablation Study

Ablation studies were conducted to assess the effectiveness of key components in the proposed FDGSNet. Table I provides a detailed account of these experiments and their results. In Table I, the symbol $\surd$ indicates that the corresponding module was retained, while the absence of this symbol indicates that the module was removed. All other settings, including loss functions and optimizers, remained consistent with the complete FDGSNet.

1) FDGSNet Without FDM: The features of the output of the SFE module are directly input into the MCDF module without frequency-domain decomposition. As can be seen from Fig. 4, if the frequency decomposition of the input multimode features is not carried out, the global correlation degree fusion of the features of different modes is directly carried out, resulting in rough segmentation and failure to extract small objects.

2) FDGSNet Without MCDF: We use simple element addition instead of the MCDF module to realize feature fusion of low-frequency features of different modal data output by the FDM module. Fig. 5 shows that without the MCDF, the vegetation segmented is stuck together, and it is difficult to effectively identify regions with similar color or texture features.

To further verify the influence of selecting features of different frequencies on multimodal data fusion, we exchanged the high-frequency and low-frequency features in the network, fused the high-frequency components for cross-modal features, and used the low-frequency components as residual features, and designed the MCDF (H) module to verify the validity of the model on three different datasets. The results are shown in Table I. The experimental results show that using MCDF (H) can also improve the performance of the model compared with not using MCDF and verify the effectiveness of the cross-modal feature complementary fusion method. However, compared with the nonuse of FDM, the accuracy of the model decreased slightly. In addition, the accuracy of using the MCDF (H) module is

TABLE I
ABLATION EXPERIMENTAL RESULTS CONSIDERING THREE FDGSNET MODULES

FDM	MCDF	MCDF (H)	AGF	Postdam (Buiding)		GID (Water body)		GID (Vegetation)	
				OA(%)	mIoU(%)	OA(%)	mIoU(%)	OA(%)	mIoU(%)
	√		√	97.52	92.35	93.70	85.36	89.35	79.38
√			√	97.41	92.12	93.63	85.29	89.22	78.57
√	√			97.57	92.68	94.35	86.17	90.43	79.62
√		√	√	97.43	92.28	93.68	85.33	89.31	79.14
√	√		√	97.68	93.61	96.13	88.37	91.06	82.01

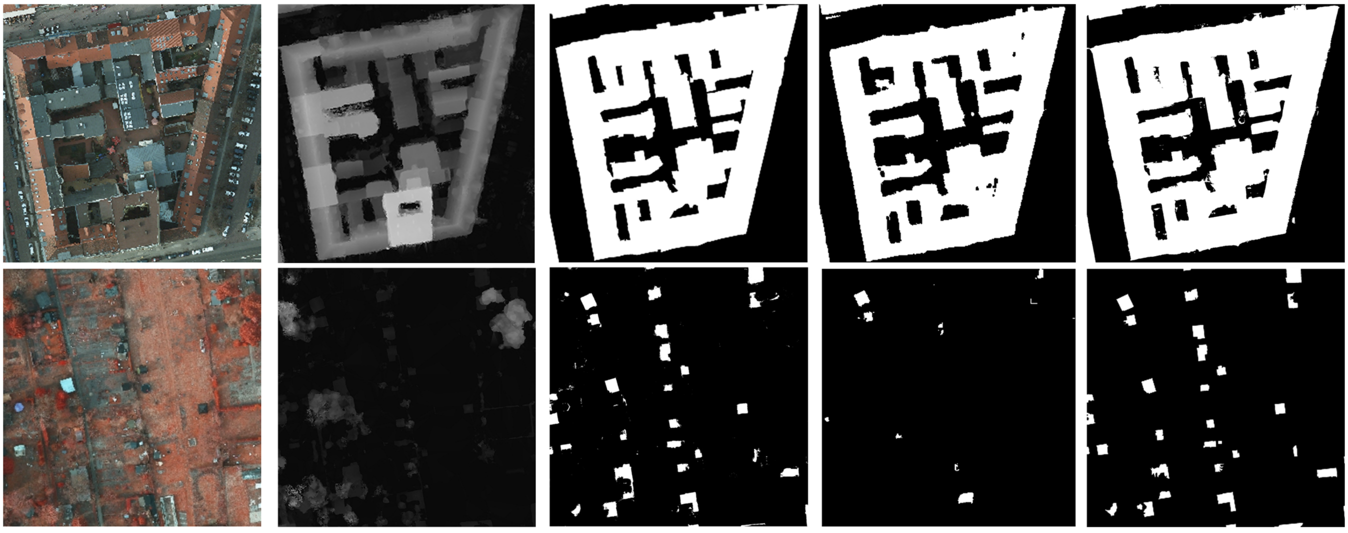


Fig. 4. Semantic segmentation comparison for FDGSNet and FDGSNet without FDM on the Postdam (Buiding) dataset.

TABLE II
COMPARISON OF ABLATION EXPERIMENTS USING INPUT DATA FROM DIFFERENT MODALITIES

IRRGB	RGB	DSM	NDWI	NDVI	Postdam (Buiding)		GID (Water body)		GID (Vegetation)	
					OA(%)	mIoU(%)	OA(%)	mIoU(%)	OA(%)	mIoU(%)
	√				96.89	91.49	88.20	80.25	86.24	76.63
√					97.63	92.41	93.82	85.58	87.58	80.39
			√		-	-	93.43	84.15	-	-
				√	-	-	-	-	67.24	50.71
√		√			97.68	93.61	-	-	-	-
√			√		-	-	96.13	88.37	-	-
√				√	-	-	-	-	91.06	82.01

significantly reduced compared with that of using the MCDF module. Therefore, we verify that when cross-mode feature fusion is carried out, selecting low-frequency features of different modes for feature fusion is conducive to improving model performance, while selecting high-frequency features of different modes for feature fusion will have a negative impact on model performance.

3) FDGSNet Without AGF: We replaced the AGF with simple 3×3 and 1×1 convolutional layers. Fig. 6 shows that without the AGF, for the parts with small interclass differences and large intraclass differences, the recognition ability is insufficient.

In addition, to verify the influence of different modal data on the segmentation accuracy, we conduct extensive ablation experiments on input data of different modalities. As shown in Table II, we used RGB and IRRGB images as the single data source, respectively, for model training and testing on Postdam (Buiding), GID (Water body), and GID (Vegetation) datasets. RGB images contain rich color and texture features, but they are easily affected by environmental factors and other factors, and the ability to recognize confusing objects is insufficient. IRRGB data increase the NIR band compared with RGB data, which improves the recognition ability of the model for different ground objects. The MIoU values of the three datasets are increased

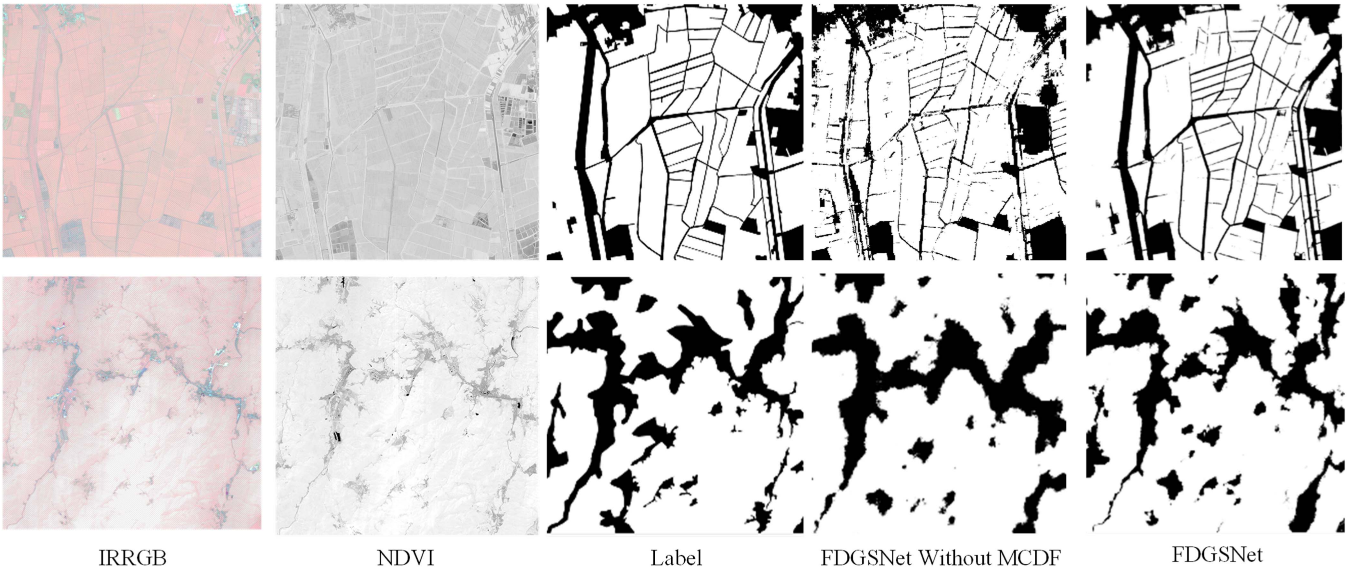


Fig. 5. Semantic segmentation comparison for FDGSNet and FDGSNet without MCDF on the GID (Vegetation) dataset.

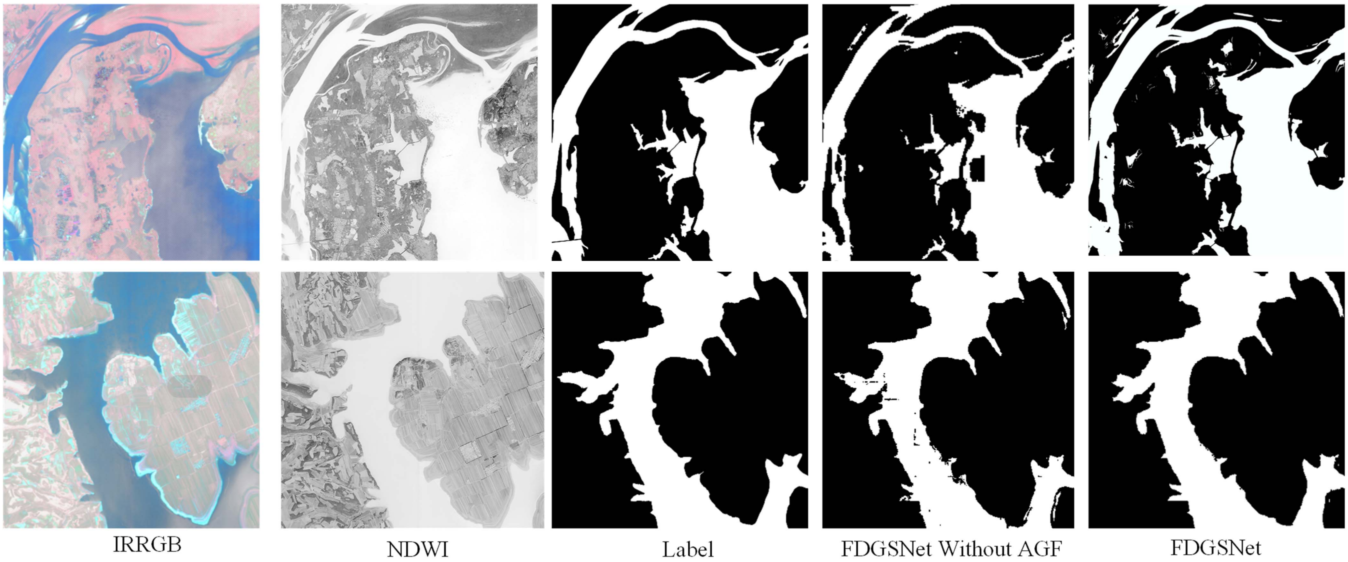


Fig. 6. Semantic segmentation comparison for FDGSNet and FDGSNet without AGF on the GID (Water body) dataset.

by 0.92%, 5.33%, and 3.76%, respectively. NDWI and NDVI data can highlight the characteristics of water and vegetation, respectively. Threshold segmentation of NDWI and NDVI data is a common detection algorithm for water and vegetation. However, using single NDWI and NDVI data cannot obtain the texture features and semantic information of the target, and it is highly dependent on the algorithm parameter settings, and the robustness is insufficient. As shown in Fig. 7, the vegetation detection method based on the threshold segmentation of NDVI data can identify smaller vegetation areas but is insufficient for detecting yellowed vegetation. The vegetation segmentation method based on RGB images cannot recognize spatial detail features, and the edge of the segmentation result is blurred. Our method improves the accuracy of semantic segmentation by combining different modal data and comprehensively utilizing their respective advantages. The MIoUs on the three datasets are 93.61%, 88.37%, and 82.01%, respectively, which achieves the most advanced segmentation accuracy.

E. Quantitative Comparison Diverse Methods

Our method was compared with state-of-the-art semantic segmentation methods and included classic multiscale context feature fusion methods, such as FCN [36], U-Net [37], and Deeplabv3+ [38], as well as advanced RSI segmentation methods that utilize transformers for global feature extraction, such as MANet [39], UNetFormer [27], BEDSN [43], CMTFNet

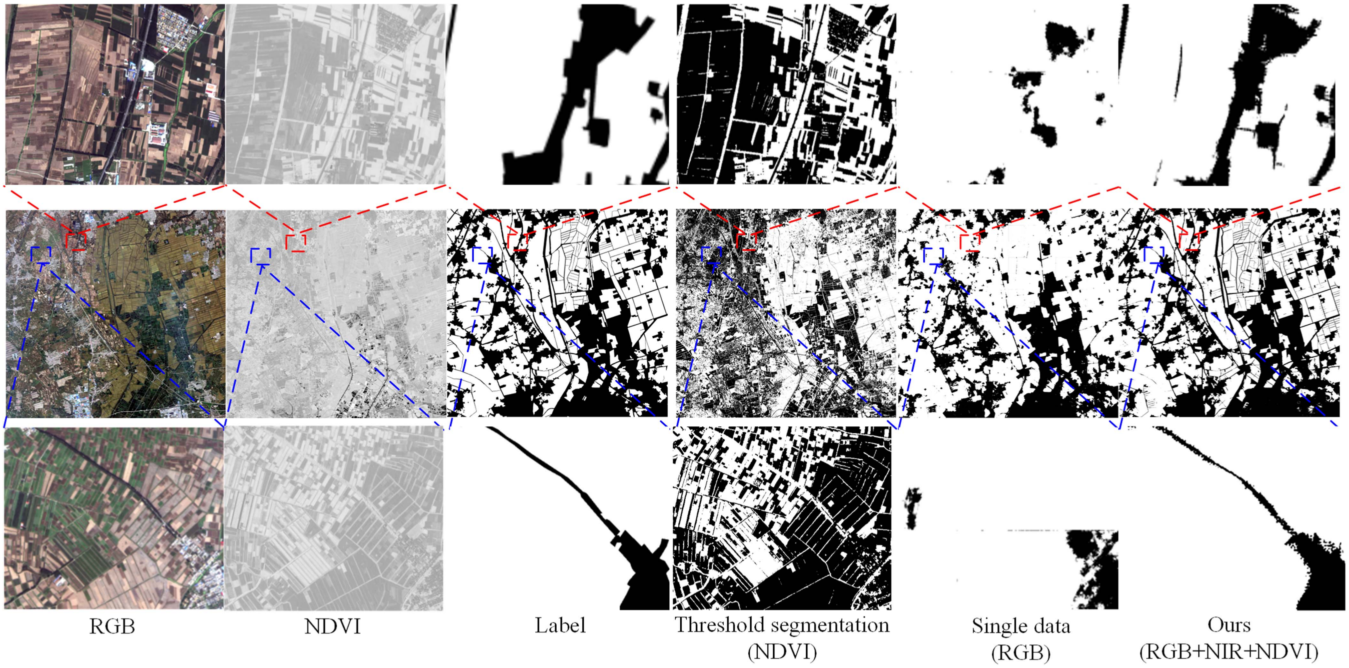


Fig. 7. Results of vegetation detection using different modal data.

TABLE III
EXPERIMENTAL RESULTS ON THE POSTDAM (BUIDING) DATASET, GID (WATER BODY) DATASET, AND GID (VEGETATION) DATASET

Methods	Postdam (Buiding)			GID (Water body)			GID (Vegetation)		
	FWIoU(%)	OA(%)	mIoU(%)	FWIoU(%)	OA(%)	mIoU(%)	FWIoU(%)	OA(%)	mIoU(%)
FCN-8s [36]	91.34	91.75	86.33	83.62	83.72	75.13	83.95	84.11	74.72
U-Net [37]	93.14	93.58	87.15	83.74	83.85	75.42	84.21	84.30	74.93
DeepLabv3+ [38]	93.85	93.91	87.62	84.69	85.27	76.91	85.18	85.49	75.06
MANet [39]	94.92	95.13	89.46	87.24	87.40	79.13	86.02	86.20	76.15
UNetFormer [27]	95.67	95.98	90.73	88.39	88.86	80.90	87.52	87.46	77.28
GCDNet [30]	96.34	96.42	91.26	90.57	90.19	81.83	87.81	87.73	77.92
PACSCNet[40]	96.82	97.14	91.95	91.89	91.16	83.42	88.28	88.35	78.54
CMGFNet [41]	97.06	97.55	92.35	92.08	92.27	84.10	88.60	88.76	78.82
FTransUNet [34]	97.20	97.31	93.14	93.23	92.84	85.05	88.74	88.92	79.63
CMX [35]	96.87	97.12	92.29	93.46	93.05	85.29	88.82	89.08	80.23
FDGSNet (Ours)	**97.45**	**97.68**	**93.61**	**96.02**	**96.13**	**88.37**	**90.91**	**91.06**	**82.01**

The bold values indicate the best results and underlined values indicate the second best results.

[44], and GCDNet [30]. These methods only consider RGB image information. These advanced single-modality methods can demonstrate the performance improvements brought by using multimodal prior data. Additionally, we compared FDGSNet with state-of-the-art multimodal feature fusion networks, including PACSCNet [40], CMGFNet [41], FTransUNet [34], and CMX [35]. The test results are shown in Table III.

MANet employs multiple efficient attention mechanisms to extract contextual dependencies, addressing the issue of underutilization of multiscale features. UNetFormer uses a lightweight ResNet18 encoder and a transformer-based decoder to establish long-range dependencies, improving the utilization of both global and local information. CMGFNet employs end-to-end cross-modal gated fusion and multilevel feature fusion techniques to improve the extraction of building footprints from VHR RSIs and DSM data. PACSCNet leverages a dual-pyramid symmetric cascade decoder and a multiscale feature extraction module to enhance segmentation accuracy by effectively harnessing multimodal contextual features.

As shown in Table III, the numeric scores for the ISPRS Postdam (Buiding), GID (Water body), and GID (Vegetation) datasets demonstrated that FDGSNet delivers high accuracy, exceeding other networks in the FWIoU, OA, and mIoU by a significant margin. The experimental results of different methods on the Potsdam (Buiding) dataset demonstrated that FDGSNet delivers the highest FWIoU of 97.45%, OA of 97.68%, and mIoU of 93.61%. The experimental results of

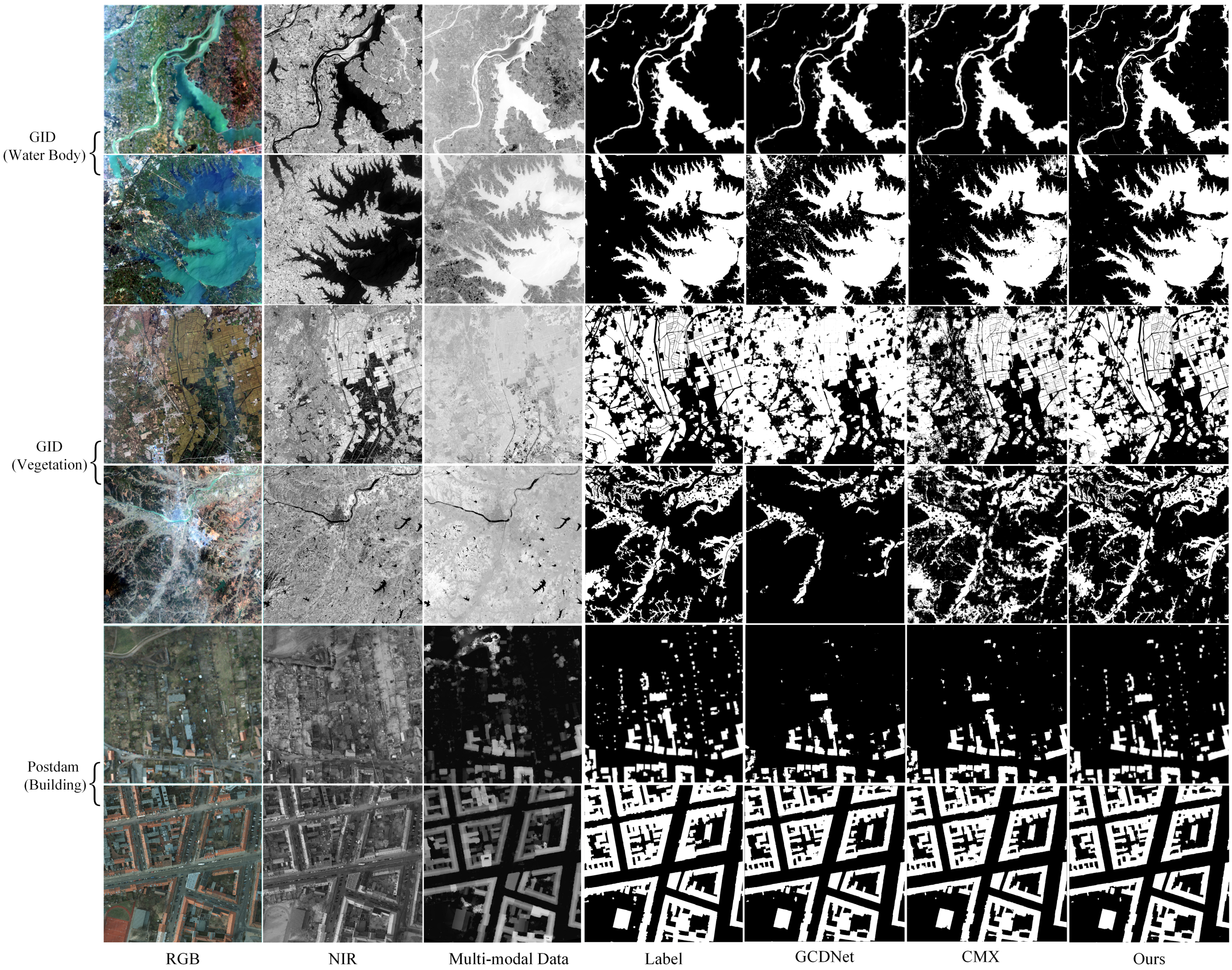


Fig. 8. Complete results of different methods. The first two rows are the test results of the GID (Water body) dataset, the multimodal data used is NDWI. The middle two rows are the test results of the GID (Vegetation) dataset, the multimodal data used is NDVI. The last two rows are the test results of the Postdam (Building) dataset, the multimodal data used is DSM.

different methods on the GID (Water body) dataset demonstrated that FDGSNet delivers the highest FWIoU of 96.02%, OA of 96.13%, and mIoU of 88.37%. The experimental results of different methods on the GID (Vegetation) dataset demonstrated that FDGSNet delivers the highest FWIoU of 90.91%, OA of 91.06%, and mIoU of 82.01%. In terms of test metrics, FDGSNet achieves the highest segmentation accuracy in different semantic segmentation tasks, indicating that our method not only can extract effective complementary information from multimodal features but also has strong generalization capabilities.

The comparison of experimental results between FDGSNet and the state-of-the-art single-modal RSI semantic segmentation network (GCDNet) and the state-of-the-art multimodal RSI semantic segmentation network (CMX) is shown in Fig. 8. In terms of segmentation performance, FDGSNet has a strong recognition capability for confusing objects, such as paddy fields and dry fields, and rivers with different colors. It is effective at segmenting small objects, such as tiny buildings hidden in forests. In addition, it achieves excellent segmentation results for complex edge contours of objects.

The existing multimodal RSI fusion methods are especially designed for specific segmentation tasks and data sources, but their performance is limited to specific tasks, with a relatively narrow scope of functionality and insufficient model generalization ability. For example, FTransUNet has achieved higher detection accuracy than CMX on ISPRS Postdam (Buiding), while CMX performs better on GID (Water body) and GID (Vegetation). FDGSNet uses the frequency-domain decomposition method to decompose the input multimodal data into low-frequency components containing common semantic features and high-frequency components containing their unique fine-grained features. By establishing the correlation between the low-frequency components of different modal features, the effective complementary information between multimodal features is extracted, and the residual connection is used to retain

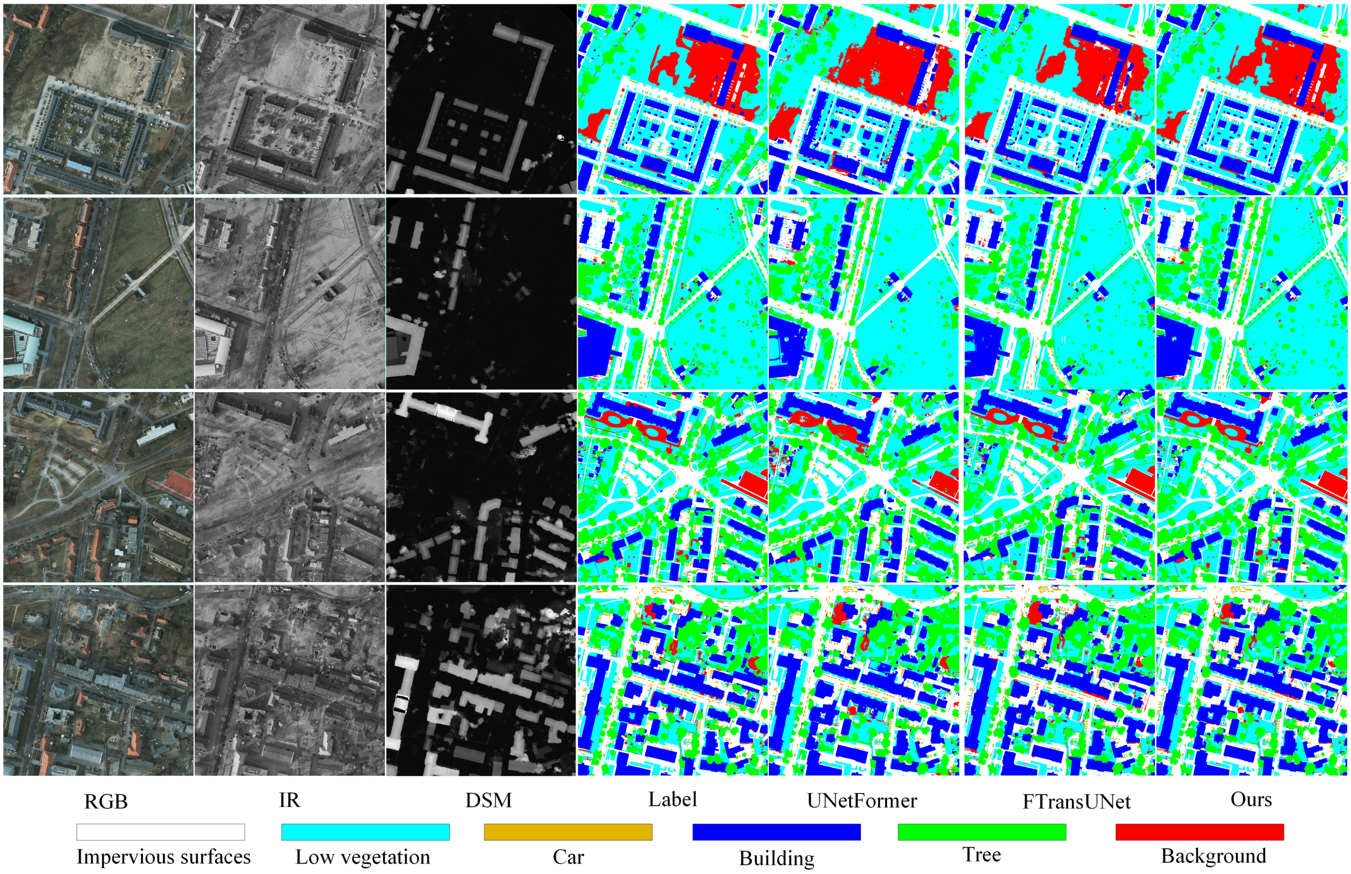


Fig. 9. Complete results of different methods based on the ISPRS Potsdam datasets.

TABLE IV
EXPERIMENTAL RESULTS ON THE ISPRS POSTDAM DATASETS

Methods	Imp.surf.	Building	Low veg.	Tree	Car	AF(%)	mACC(%)	mIoU(%)	OA(%)
Deeplabv3+	87.24	92.15	83.72	86.58	86.42	87.22	88.85	84.36	87.72
UNetFormer	88.45	92.71	85.25	86.43	88.32	88.23	88.94	84.53	89.17
PACSCNet	91.45	94.38	89.96	86.72	91.29	90.76	90.01	84.92	90.23
BEDSN	91.62	95.41	88.64	86.47	94.23	91.27	90.75	84.86	90.02
CMTFNet	**92.37**	96.48	87.38	87.55	92.91	91.34	91.06	85.04	90.44
CMX	91.28	96.19	91.62	87.24	95.79	92.42	91.68	85.76	91.03
FTransUNet	91.94	96.34	91.56	86.93	96.28	92.61	91.75	85.91	91.12
FDGSNet (Ours)	91.34	**96.62**	**92.05**	**87.80**	**96.52**	**92.87**	**92.02**	**86.34**	**91.67**

The bold values indicate the best results and underlined values indicate the second best results.

the fine-grained features of different modal data. The multimodal feature fusion method based on the frequency-domain decomposition uses shared semantic features at low frequencies for fusion, reducing the mutual interference problem in the fusion process between different modalities and reducing the sensitivity of the model to different data sources. Compared with other methods, our method has better interpretability and strong generalization ability for different data sources and different segmentation tasks. To validate the exceptional performance and generalization capability of FDGSNet in multimodal RSI semantic segmentation tasks, we evaluated the model's panoramic semantic segmentation performance on the ISPRS Potsdam dataset. We utilized multispectral data (IRRGB) and DSM as inputs to accomplish the semantic segmentation tasks for various object categories within the Potsdam dataset.

FDGSNet was compared with state-of-the-art semantic segmentation methods on the ISPRS Potsdam datasets. The test results are depicted in Fig. 9. As illustrated in Table IV, in the task of panoramic RSI semantic segmentation, FDGSNet achieved the optimal $F1$-score for most detected objects and demonstrated the highest OA of 91.67%, mIoU of 86.34%, and AF of 92.87%.

Experimental results demonstrate that FDGSNet outperforms other models in terms of its ability to effectively extract intricate image features, accurately discern target locations and boundaries, and exhibit superior generalization capabilities. It successfully mitigates the challenges posed by noise interference

TABLE V
COMPUTATIONAL COMPLEXITY ANALYSIS FOR MEASURING 512 × 512 IMAGES ON A SINGLE NVIDIA GEFORCE RTX 3090TI GPU

Methods	Multimodal	FLOPs (G) ↓	Parameter (M) ↓	Memory (MB) ↓	Speed (FPS) ↑	mIoU (%) ↑
Deeplabv3+	N	138.63	40.35	4336	35.59	84.36
UNetFormer	N	46.04	24.20	2483	43.06	84.53
CMTFNet	N	32.85	30.07	3472	40.27	85.04
PACSCNet	Y	276.54	359.64	9847	15.96	84.92
CMX	Y	136.62	141.46	7359	20.83	85.76
FTransUNet	Y	45.21	160.88	3463	29.85	85.91
FDGSNet (Ours)	Y	202.58	228.73	8851	18.46	86.34

mIoU value is the result of the ISPRS Potsdam datasets.

and the loss of detailed features of different modal data, which often leads to inaccurate target localization and boundary pixel segmentation. Moreover, it addresses the issue of inadequate model robustness in semantic segmentation tasks involving multimodal RSIs. Our approach exhibits excellent performance on diverse datasets and achieves superior segmentation outcomes compared with alternative networks.

F. Model Parameters and Computation Complexity Analysis

We evaluate the computational complexity of the proposed FDGSNet using the following evaluation metrics: the floating point operation count (FLOP), the number of model parameters, the memory footprint, and the frames per second (FPS). We fed input data of size 512×512 into all models and also evaluated the computational complexity and parameter sizes of various methods in the same runtime environment, and the results are shown in Table V. From Table V, it can be seen that the computational complexity of multimodal models is generally higher than that of single-modal models. The number of model parameters of FDGSNet is at a medium level compared with other methods, but the computational complexity of the models is higher. This is mainly due to the introduction of the transformer, which is more computationally intensive. Overall, although FDGSNet is in the middle of the range in terms of computational complexity, it shows superior performance.

V. CONCLUSION

In this article, a multimodal gated segmentation network (FDGSNet) for RSIs based on frequency decomposition is proposed. First, the multimodal features are decomposed into high-frequency and low-frequency components. The high-frequency component is used to preserve the detailed information of different modality data, while the low-frequency component mainly retains the semantic features of different objects. Then, complementary information between multimodal features can be extracted by establishing the correlation between the low-frequency components of different modality features. Finally, the AGF module is used to achieve adaptive fusion between multimodal features. Frequency-domain decomposition-based multimodal feature fusion methods effectively mitigate noise interference and edge detail loss during the fusion of diverse data sources. This multimodal feature fusion method has a strong generalization capability, which enables it to obtain valuable prior knowledge from various modal data features. The experimental results show that FDGSNet is well on the ISPRS Postdam, ISPRS Postdam (Buiding), GID (Water body), and GID (Vegetation) datasets, and achieves better segmentation results than other networks. In future research, we will further explore how to reduce the computational cost of the multimodal data fusion module for wide applications.

REFERENCES

[1] Y. Ni, J. Liu, W. Chi, X. Wang, and D. Li, “CGGLNet: Semantic segmentation network for remote sensing images based on category-guided global-local feature interaction,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Mar. 2024, Art. no. 5615617.

[2] X. Yuan, J. Shi, and L. Gu, “A review of deep learning methods for semantic segmentation of remote sensing imagery,” *Expert Syst. Appl.*, vol. 169, 2020, Art. no. 114417.

[3] X. Guo et al., “GAN-based virtual-to-real image translation for urban scene semantic segmentation,” *Neurocomputing*, vol. 394, pp. 127–135, 2020.

[4] W. Boonpook, Y. Tan, and B. Xu, “Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry,” *Int. J. Remote Sens.*, vol. 42, pp. 1–19, 2021.

[5] Y. Wang, X. Zeng, and X. Liao, “B-FGC-net: A building extraction network from high resolution remote sensing imagery,” *Remote Sens.*, vol. 14, no. 2, 2022, Art. no. 269.

[6] G. Grekousis, “Local fuzzy geographically weighted clustering: A new method for geodemographic segmentation,” *Int. J. Geograph. Inf. Sci.*, vol. 35, pp. 152–174, 2021.

[7] P. Shamsolmoali, M. Zareapoor, H. Zhou, R. Wang, and J. Yang, “Road segmentation for remote sensing images using adversarial spatial pyramid networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4673–4688, Jun. 2021.

[8] S. Ji, D. Wang, and M. Luo, “Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3816–3828, May 2021.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[10] R. Shen, A. Huang, B. Li, and J. Guo, “Construction of a drought monitoring model using deep learning based on multi-source remote sensing data,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 79, pp. 48–57, 2019.

[11] N. Venugopal, “Automatic semantic segmentation with deeplab dilated learning network for change detection in remote sensing images,” *Neural Process. Lett.*, vol. 51, pp. 2355–2377, 2020.

[12] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[13] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, “Multistage attention ResU-net for semantic segmentation of fine-resolution remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 800920505.

[14] X. Meng, Y. Yang, L. Wang, T. Wang, R. Li, and C. Zhang, “Class-guided swin transformer for semantic segmentation of remote sensing imagery,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2022, Art. no. 6517505.

[15] J. Liu, W. Zhou, Y. Cui, L. Yu, and T. Luo, "GCNet: Grid-like context-aware network for RGB-thermal semantic segmentation," *Neurocomputing*, vol. 506, pp. 60–67, 2022.

[16] Z. Zhao et al., "CDDfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5906–5916.

[17] Q. Zeng, J. Zhou, and X. Niu, "Cross-scale feature propagation network for semantic segmentation of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Aug. 2023, Art. no. 6008305.

[18] C. Zheng, J. Nie, Z. Wang, N. Song, J. Wang, and Z. Wei, "High-order semantic decoupling network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5401415.

[19] Y. Chen, Y. Wang, S. Xiong, X. Lu, X. X. Zhu, and L. Mou, "Integrating detailed features and global contexts for semantic segmentation in ultrahigh-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Apr. 2024, Art. no. 4703914.

[20] X. Li et al., "SSCNet: A spectrum-space collaborative network for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 15, no. 23, 2023, Art. no. 5610.

[21] R. Xiao, C. Zhong, W. Zeng, M. Cheng, and C. Wang, "Novel convolutions for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5907313.

[22] X. Li and J. Zhou, "MASNet: Road semantic segmentation based on multiscale modality fusion perception," *IEEE Trans. Instrum. Meas.*, vol. 73, Nov. 2024, Art. no. 2500713.

[23] J. Ma, W. Zhou, J. Lei, and L. Yu, "Adjacent bi-hierarchical network for scene parsing of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Feb. 2023, Art. no. 3000705.

[24] W. Liang, C. Shan, Y. Yang, and J. Han, "Multi-branch differential bidirectional fusion network for RGB-T semantic segmentation," *IEEE Trans. Intell. Veh.*, to be published, doi: 10.1109/TIV.2024.3374793.

[25] R. Li, S. Zheng, C. Zhang, C. Duan, and L. Wang, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, 2021.

[26] R. Li, L. Wang, C. Zhang, C. Duan, and S. Zheng, "A2-FPN for semantic segmentation of fine-resolution remotely sensed images," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 1131–1155, 2022.

[27] 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, Sep. 2022.

[28] Y. Cao, C. Huo, S. Xiang, and C. Pan, "GFFNet: Global feature fusion network for semantic segmentation of large-scale remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4222–4234, Jan. 2024.

[29] X. Li et al., "AAFormer: Attention-attended transformer for semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, May 2024, Art. no. 5002805.

[30] J. Cui, J. Liu, J. Wang, and Y. Ni, "Global context dependencies aware network for efficient semantic segmentation of fine-resolution remoted sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, Sep. 2023, Art. no. 2505205.

[31] L. Fan, Y. Zhou, H. Liu, Y. Li, and D. Cao, "Combining swin transformer with UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5530111.

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

[33] B. Ren, B. Liu, B. Hou, Z. Wang, C. Yang, and L. Jiao, "SwinTFNet: Dual-stream transformer with cross attention fusion for land cover classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, Jan. 2024, Art. no. 2501505.

[34] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, Mar. 2024, Art. no. 5403215.

[35] J. Zhang, H. Liu, K. Yang, X. Hu, and R. Stiefelhagen, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14679–14694, Dec. 2023.

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

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

[38] L. 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.

[39] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5607713.

[40] X. Fan et al., "Progressive adjacent-layer coordination symmetric cascade network for semantic segmentation of multimodal remote sensing images," *Expert Syst. Appl.*, vol. 238, 2024, Art. no. 121999.

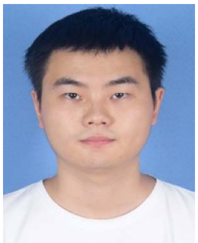
[41] H. Hosseinpour, F. Samadzadegan, and F D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 96–115, 2022.

[42] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[43] X. Li et al., "Boundary-enhanced dual-stream network for semantic segmentation of high-resolution remote sensing images," *GISci. Remote Sens.*, vol. 61, no. 1, 2024, Art. no. 2356355.

[44] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 2004612.

[45] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, Dec. 2014, pp. 1–13.



Jian Cui received the B.S. degree in communication engineering and the M.S. degree in control engineering from the Zhongyuan University of Technology, Zhengzhou, China, in 2017 and 2020, respectively. He is currently working toward the Ph.D. degree in optical engineering with the Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His research interests include instance segmentation, object detection, and edge detection.

Jiahang Liu (Member, IEEE) received the B.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2000, the M.S. degree in tectonics from the Institute of Geology, China Earthquake Administration, Beijing, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University, Shanghai, China, in 2011.

He has been a Full Professor with the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests include remote sensing, image processing, computer vision, and digital twin.

Yue Ni received the B.E. degree in information engineering from the Nanjing University of Information Science and Technology, Nanjing, China, in 2019, and the M.S. degree in communication and information engineering in 2022 from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, where he is currently working toward the Ph.D. degree in optical engineering with the College of Astronautics.

His current research interests include semantic segmentation of remote sensing images, remote sensing image processing, and deep learning.

Jinjin Wang received the B.S. degree in communication engineering and the M.S. degree in information and communication engineering from the Zhongyuan University of Technology, Zhengzhou, China, in 2017 and 2020, respectively. She is currently working toward the Ph.D. degree in optical engineering with the Nanjing University of Aeronautics and Astronautics, Nanjing, China.

Her research interests include image processing, hyperspectral dimensionality reduction, and deep learning.

Manchun Li received the Ph.D. degree in cartography from Nanjing University, Nanjing, China, in 1992.

He is currently a Professor with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University. His research interests include GIS and remote sensing applications.