GVANet: A Grouped Multiview Aggregation Network for Remote Sensing Image Segmentation

Yunsong Yang, Jinjiang Li, Zheng Chen, and Lu Ren

***Abstract*—In remote sensing image segmentation tasks, various challenges arise, including difficulties in recognizing objects due to differences in perspective, difficulty in distinguishing objects with similar colors, and challenges in segmentation caused by occlusions. To address these issues, we propose a method called the grouped multiview aggregation network (GVANet), which leverages multiview information for image analysis. This approach enables global multiview expansion and fine-grained cross-layer information interaction within the network. Within this network framework, to better utilize a wider range of multiview information to tackle challenges in remote sensing segmentation, we introduce the multiview feature aggregation block for extracting multiview information. Furthermore, to overcome the limitations of same-level shortcuts when dealing with multiview problems, we propose the channel group fusion block for cross-layer feature information interaction through a grouped fusion approach. Finally, to enhance the utilization of global features during the feature reconstruction phase, we introduce the aggregation-inhibition-activation block for feature selection and focus, which captures the key features for segmentation. Comprehensive experimental results on the Vaihingen and Potsdam datasets demonstrate that GVANet outperforms current state-of-the-art methods, achieving mIoU scores of 84.5% and 87.6%, respectively.**

***Index Terms*—Attention mechanism, multiscale fusion, remote sensing, semantic segmentation, transformer.**

I. Introduction

ADVANCEMENTS in aerospace and sensor technologies have made it increasingly convenient to access high-resolution remote sensing images. These images are characterized by abundant fine-grained details and a wealth of semantic content. Semantic segmentation, which involves predicting the semantic category or label of each pixel, serves as a fundamental pillar in the analysis of remote sensing images. In recent years, semantic segmentation has gained widespread popularity in urban scene images, driving numerous urban-related applications such as land cover mapping [1], [2], change detection [3], environmental conservation [4], road and building extraction [5], [6], and a multitude of other practical uses [7], [8].

Received 18 August 2024; accepted 10 September 2024. Date of publication 12 September 2024; date of current version 25 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61772319, Grant 62002200, Grant 62202268, and Grant 62272281, in part by the Shandong Natural Science Foundation of China under Grant ZR2020QF012 and Grant ZR2021MF068, and in part by the Yantai Science and Technology Innovation Development Plan under Grant 2022JCYJ031. *(Corresponding author: Zheng Chen.)*

The authors are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: chenzheng@sdtbu.edu.cn).

Data is available online at https://github.com/yysdck/GVANet.
Digital Object Identifier 10.1109/JSTARS.2024.3459958

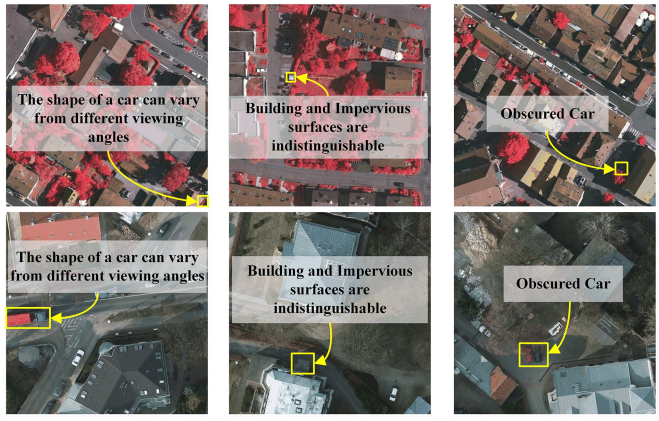


Fig. 1. Examples of various challenges in remote sensing images. The first row of images is from the Vaihingen dataset, and the second row is from the Potsdam dataset. (a) Demonstrates how cars exhibit different appearances under different viewpoints. For example, in (a), the car in the first row blends entirely with the building's shadow due to the viewpoint, making the car appear as part of the building's shadow, while the truck compartment in the second row resembles a building due to the viewpoint. (b) Shows how buildings appear similar to impervious surfaces due to lighting. In the first row, the annotated area should actually be an impervious surface but appears as a building due to lighting effects, while in the second row, the annotated building area appears like impervious surfaces due to blending with shadows. (c) Displays cars obscured by trees, making their category almost indistinguishable. These challenges can impact segmentation accuracy.

Remote sensing images, characterized by high resolution and complex scenes, present a series of challenges for segmentation, rendering traditional methods often inadequate. These challenges primarily include variations in object appearance under different lighting conditions and the difficulty in distinguishing between objects with similar colors. Additionally, occlusions where one object is partially obscured by another further complicate the accurate identification and segmentation of the occluded objects. Examples of these challenges are illustrated in Fig. 1. These factors significantly increase the complexity of the segmentation task [9]. To address these issues, researchers are continuously exploring new algorithms and techniques aimed at enhancing the accuracy and robustness of remote sensing image segmentation [10].

In recent years, significant progress has been made in remote sensing image segmentation with deep learning techniques. Compared to traditional machine learning algorithms such as support vector machine (SVM) [11], random forest [12], and conditional random field (CRF) [13], convolutional neural

network (CNN) methods based on deep learning have shown superior performance in extracting image features by learning feature representations [14], [15]. CNNs can effectively capture contextual and semantic information in images [16], [17].

For semantic segmentation, fully convolutional network (FCN) [18] was the first effective end-to-end CNN structure, but its results appeared coarse due to its simplistic design. Subsequently, more refined encoder–decoder structures were proposed [19], [20], including contracting and expanding paths, to obtain more accurate segmentation results. U-Net combines the encoder and decoder, preserving global information while capturing local details. The encoder gradually captures contextual information and abstract features from the image, while the decoder reconstructs features through deconvolution operations, gradually restoring the resolution of the feature maps to the original input image size. It also maintains the original feature information through skip connections between corresponding layers in the encoder and decoder. This architecture enables U-Net to precisely locate objects and retain details in segmentation tasks. Due to the effectiveness of U-Net, many researchers still use it as a foundational segmentation network to explore further possibilities, such as AFF-Unet [21].

Although U-Net offers numerous advantages, there are still some limitations in the context of remote sensing image segmentation. On the one hand, U-Net captures image information through a stacked convolutional approach. However, since convolutions are designed for local feature extraction, this local approach is insufficient to fully capture the diversity and complexity of objects in remote sensing images, where the appearance of objects may vary under different views. In the semantic segmentation of remote sensing images, if only local information is modeled, pixelwise classification tends to be ambiguous. Therefore, introducing global information dependencies is necessary.

To enable global modeling in remote sensing image segmentation, many studies have introduced self-attention mechanisms [22] into U-Net-based segmentation networks, creating effective long-range dependencies. Examples include Unetformer [23] and ST-Unet [24]. While self-attention mechanisms can establish global dependencies, they often require significant computation time and memory to capture global context. Subsequently, more efficient attention mechanisms have been proposed as alternatives to self-attention for extracting global context, such as dual attention [25] and CBAM [26]. In attention-based segmentation networks, although global dependencies can be effectively established, these networks usually consider only a single scale within the same layer. However, in remote sensing image segmentation, even with global information, single-scale networks struggle to adapt to the scale variations of objects, thereby affecting segmentation accuracy. To address this issue, researchers have proposed multiscale feature extraction methods, such as [27] and [28], which can extract information-rich multiscale features. However, these methods typically focus only on image resolution or scale information, neglecting the impact of different views on the model.

In remote sensing segmentation, introducing information from different views can enhance the robustness of the network. For instance, the degree of occlusion for objects varies under different views, and so does the difficulty of segmentation. Therefore, fully considering different view information can significantly improve segmentation accuracy. Although some studies have proposed methods that combine information from different views to improve classification or segmentation accuracy [29], [30], these methods usually require multiple view images of the same object. However, in practical segmentation tasks, only a single view image is usually available, creating an urgent need for a scientific method that integrates multiview information for segmentation from a single view image. To this end, we propose a multiview feature aggregation block (MVFAB). This structure first establishes dependencies between objects of the same type under different views on the same plane using the global modeling capability of the attention mechanism. It then expands these dependencies to different height views. Finally, it calculates and assigns fusion weights through postattention computation to address challenges in remote sensing segmentation.

On the other hand, the same-level skip connections in U-Net also have limitations in addressing the remote sensing segmentation challenges in remote sensing segmentation. First, same-level connections may ignore global contextual information due to the existence of local feature redundancy. In processing remote sensing images, global contextual information is crucial for understanding the relationships between objects and complex scene structures, but same-level connections may not fully utilize this information. Second, remote sensing images may contain objects of multiple scales, and same-level connections may not adapt well to this multiscale problem. Since same-level connections only connect feature maps of adjacent layers, they may not capture the features and morphology of objects at different scales accurately, leading to inaccurate segmentation results. Finally, same-level connections only connect feature maps of the same scale between the encoder and decoder, lacking cross-scale information transmission. However, objects in remote sensing images may undergo scale changes, and same-level connections may not handle this cross-scale information well. To address these limitations, we propose a channel group fusion block (CGB), which facilitates cross-level information interaction through group combination.

Based on the above considerations, we designed a grouped multiview aggregation network (GVANet) that leverages multiview information from a single image to address various challenges in remote sensing segmentation. Within this network architecture, the MVFAB enables multiview expansion of single-view information. Additionally, the CGB facilitates cross-level information interaction, allowing the model to integrate and exchange information across different scales. Finally, to enhance the efficiency of multiview information utilization, we introduced an aggregation-inhibition-activation block (AIAB) following cross-level interaction. Experimental results validate the effectiveness of our proposed segmentation method.

In summary, the contributions of this article are as follows.

1) *Proposal of the MVFAB:* This module introduces an attention mechanism to establish dependencies between features

of the same type of objects at the same height. It then expands these dependencies across different height views, and finally, it uses a postattention mechanism to calculate fusion weights and assign them to the features. This approach effectively addresses challenges in remote sensing segmentation.

2) *Proposal of the CGB:* This module facilitates cross-level information interaction between different hierarchical features through a grouped combination approach. The introduction of the CGB addresses the limitations of same-level skip connections, such as local feature redundancy, insufficient utilization of long-range dependencies, and challenges related to multiscale information, thereby improving the accuracy and robustness of the segmentation results.

3) *Proposal of GVANet:* We propose GVANet, a network designed to address various challenges in remote sensing segmentation and improve segmentation accuracy. GVANet combines MVFAB and CGB, with MVFAB enabling multiview expansion of single-view information, and CGB facilitating cross-level information interaction, allowing the model to integrate and exchange information across different scales. Additionally, to further enhance the efficiency of multiview information utilization, we designed an aggregation-inhibition-activation block (AIAB) following cross-level interaction. As an integrated network architecture, GVANet demonstrates excellent performance in remote sensing image segmentation, effectively addressing multiple remote sensing challenges and improving segmentation accuracy and robustness.

II. Related Work

Semantic segmentation of remote sensing images is a critical task in remote sensing technology. It involves assigning each pixel in a remote sensing image to its corresponding land cover category, enabling fine-grained classification and segmentation of the Earth's surface. In recent years, the rapid advancement of deep learning technology has brought about significant breakthroughs in remote sensing image semantic segmentation. In this section, we will introduce some important work related to semantic segmentation of remote sensing images and discuss their contributions to this field.

A. Semantic Segmentation of Remote Sensing Images Based on CNNs

The release of certain datasets [31], [32], along with the organization of competitions such as the IEEE Geoscience and Remote Sensing Society (IGARSS) data fusion competition, SpaceNet challenge, DeepGlobe challenge, and International Society for Photogrammetry and Remote Sensing (ISPRS) benchmarks, has played a crucial role in advancing research in semantic segmentation of remote sensing images based on convolutional neural networks (CNNs).

Fully convolutional networks (FCNs), first proposed by Long et al. in 2015, were the first effective CNN architecture for semantic segmentation. Since then, CNN-based methods have dominated the field of semantic segmentation in remote sensing, encompassing numerous research achievements [33], [34], [35]. However, FCNs suffer from low resolution segmentation results due to their simplistic decoder structure, limiting image fidelity and accuracy.

To address this segmentation issue, researchers introduced the UNet encoder–decoder network, tailored for finer semantic segmentation tasks. The UNet architecture comprises a contracting path and an expanding path, extracting and reconstructing multilevel features by progressively reducing and then restoring the spatial resolution of feature maps. This structure has become the standard model for remote sensing image segmentation, laying the foundation for subsequent research [36], [37]. Skip connections play a crucial role in encoder–decoder structures, partially bridging the semantic gap between high-level and low-level features. Some researchers have explored different skip connection strategies, such as UNet++ [38], Web-Net [39], and ResUnet-a [40]. While these improvements to skip connections have shown certain effectiveness, they still utilize same-level skip connections, resulting in redundant local features, inadequate adaptation to multiscale issues, and a lack of cross-scale information transmission, which remain limitations in addressing multiview problems in remote sensing images. In contrast, we employ a channel group fusion block (CGB) to group and combine features from different hierarchical levels, facilitating cross-layer information transmission to overcome the limitations of skip connections in addressing multiview problems.

B. Multiview Features

Multiview features represent the characteristics resulting from the fusion of features from multiple views. Researchers have noted that features extracted from images under a single view often result in insufficient or even misleading information [41]. In remote sensing image processing, utilizing multiview features can significantly improve model accuracy, as validated by several studies. For instance, methods in [29], [30], [41], [42], and [43].These methods have demonstrated that incorporating multiview information effectively enhances image classification and segmentation accuracy. However, these methods typically rely on original images from different views or simulate pseudomultiviews through techniques such as rotation. In practical applications, conditions for obtaining multiview images are limited, and usually, only single-view images are available. Furthermore, pseudomultiview methods may mislead deep networks, resulting in reduced accuracy.

To address this issue, we propose a multiview feature aggregation block (MVFAB). This approach first establishes long-range dependencies between objects of the same class from different views at the same sampling height. It then models the objects from views at different heights. Finally, it calculates weights and performs fusion to achieve effective aggregation of multiview features. This method simulates multiview characteristics within a single-view image, effectively improving the accuracy of remote sensing image segmentation.

C. Attention Mechanism

Due to the limitations of the receptive field, segmentation networks based solely on CNNs can only capture local semantic

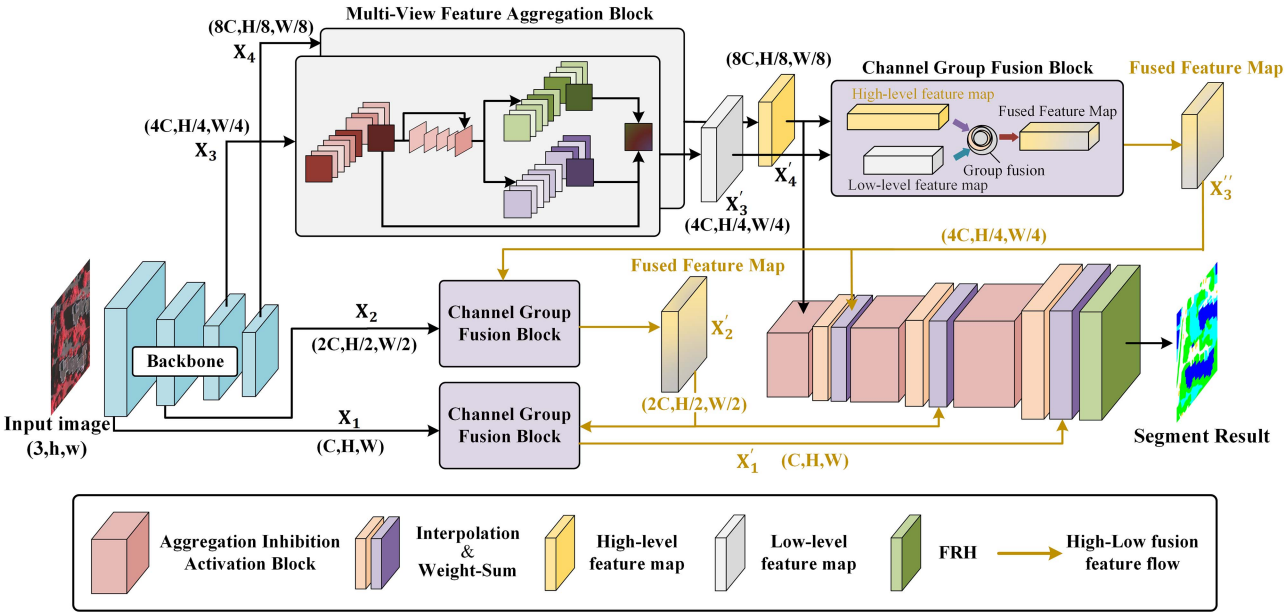


Fig. 2. GVANet structure diagram. Features extracted from the third and fourth stages of the backbone are first subjected to multiview expansion through MVFAB, followed by cross-level information interaction using CGB. Subsequently, features extracted from the first and second stages of the backbone and the output features of CGB are re-input into CGB as different level features for cross-level information interaction. This process achieves cross-level interaction between the features of the first stage and all other stages. Eventually, this process yields four different stage outputs, which serve as inputs to the feature reconstruction stage. Finally, each stage's enhanced features pass through AIAB for global information enhancement and are then processed by FRH to obtain the final segmentation result.

features, lacking the ability to model global information across the entire image. However, in high-resolution remote sensing urban scene images, many occluded objects and complex patterns, as well as frequent artificial structures, exist [44]. Relying solely on local information makes it difficult to accurately identify and interpret these complex targets.

The emergence of attention mechanisms has enabled the capture of global information to some extent. For example, DANet [45] utilizes spatial attention modules and channel attention modules to effectively extract spatial and channel dependence information. CCNet [46] designs a cross-shaped attention mechanism based on nonlocal modules. By calculating two cross-shaped dependence relationships, global dependence relationships are obtained, and the calculation steps of self-attention are optimized [47], thereby improving the algorithm's effectiveness. Zhang et al. [48] constructed a self-attention module focusing on channels and spatial locations. By multiplying the feature map by the query matrix and the key matrix, a weight map of all spatial positions and channel relationships is generated, thus obtaining global information, which is used for remote sensing image segmentation. Jha et al. [49] proposed a fusion network based on global attention (GAF-Net), providing an innovative architecture to improve remote sensing image analysis results. Wang et al. [50] designed a vector set attention module to establish relationships between channels and spatial locations for remote sensing image segmentation.

While previous researchers have effectively utilized attention mechanisms to achieve global modeling in segmentation tasks, these approaches establish dependencies between objects only based on features from the same view at the same level, which can lead to insufficient segmentation accuracy. To address this issue, we propose the MVFAB, which leverages attention mechanisms and multiview features to enable global feature expansion with multiple receptive fields. Additionally, to enhance the efficient utilization of multiview features, we have designed an AIAB based on BAM [51].

III. Method

In this section, we will commence by presenting the holistic architecture of GVANet. Subsequently, we will explore three pivotal components housed within GVANet, specifically the MVFAB, CGB, and AIAB.

A. Network Structure

The structure of GVANet is shown in Fig. 2. Our GVANet is constructed using a sampling process based on CNN and a feature reconstruction process based on attention mechanisms. Since ConvNext [50] has been proven effective and efficient in feature extraction through extensive experiments, we chose the pretrained ConvNext as the backbone, allowing us to extract deep features at a lower computational cost. ConvNext consists of four stages, with each stage downsampling the feature maps.For a given RS image $X \in R^{3\times h\times w}$, where w and h are the width and height of the input RS image, initially with three channels, X undergoes feature extraction via Convnext, resulting in four different features extracted from four Convnext stages: $X_1 \in R^{C\times H\times W}$, $X_2 \in R^{2C\times(H/2)\times(W/2)}$, $X_3 \in R^{4C\times(H/4)\times(W/4)}$, and $X_4 \in R^{8C\times(H/8)\times(W/8)}$, where W and H are the width and height after downsampling, and C is the expanded channel. The number of channels in the features extracted from the first stage is $C = 128$, and for subsequent stages,

the number of channels is twice that of the previous stage, while the height and width are halved. The features extracted from the final two stages, X_3 and X_4, first pass through an MVFAB. This stage primarily focuses on incorporating multiview features. It is worth noting that deep features become more abstract and global, so we only introduce MVFAB in the last two feature extraction stages, namely

$$\begin{cases} X_3' = F^3_{\text{MVFAB}}(X_3) \\ X_4' = F^4_{\text{MVFAB}}(X_4) \end{cases} \tag{1}$$

where F^i_{MVFAB} denotes the features passing through the MVFAB. Subsequently, to enable fine-grained cross-layer information interaction between different hierarchical features, the extracted low-level global features and high-level global features are input into a CGB for fusion, obtaining the input of the next CGB. This process derives three features with cross-layer interaction through CGB: X_3'', X_2', and X_1', with the same dimensions as X_3, X_2, and X_1, respectively

$$X_3'' = F_{\text{CGB}}(X_4', X_3') \tag{2}$$

$$X_2' = F_{\text{CGB}}(X_3'', X_2) \tag{3}$$

$$X_1' = F_{\text{CGB}}(X_2', X_1) \tag{4}$$

where $F_{\text{CGB}}(H, L)$ represents the operation of fusing high-level feature H and low-level feature L via CGB. This process yields features with fused cross-layer interaction: X_1', X_2', and X_3''. These features, along with the global feature X_4', are then aggregated via weighted sum operation with the features suppressed and activated by the AIAB in the feature reconstruction stage. The weighted sum operation selectively weights the contribution of the two features to segmentation accuracy, thereby learning more generalized fused features. The expression for the weighted sum operation can be represented as

$$\text{FF} = \alpha \cdot \text{CF} + (1 - \alpha) \cdot \text{AIF} \tag{5}$$

where FF represents the fused features, CF represents the features generated by ConvNext, and AIF represents the features produced by the AIAB, α represents a weight coefficient, where $0 \leq \alpha \leq 1$.

In particular, the feature reconstruction consists of four stages, each progressing incrementally, with output results being Y_3, Y_2, Y_1, and Y, respectively. Taking the third stage as an example, low-level features X_4' are first processed through AIAB, followed by bilinear interpolation, and then this feature is weighted and summed with high-level features X_3''. Specifically, it can be expressed as follows:

$$Y_3 = \alpha \cdot X_3'' + (1 - \alpha) \cdot \text{BI}\left(F_{\text{AIAB}}\left(X_4'\right)\right) \tag{6}$$

where $F_{\text{AIAB}}(\cdot)$ represents features processed through the AIAB block, and BI represents bilinear interpolation operations. It is important to note that in the fourth stage, there are no lower level features to input, so there is no step for weighted summation with high-level features. Additionally, the final layer is dedicated to pixel-level classification, and as such, it does not require additional reconstruction work. Therefore, the final layer does not use AIAB but instead utilizes an effective feature refinement head (FRH). The effectiveness of the feature refinement head is validated based on the detailed structure from UnetFormer [23]. Finally, the resulting features are used for pixelwise classification to obtain the segmented image.

Overall, GVANet introduces multiview information by using MVFAB in the final two stages. Additionally, GVANet employs CGB for fine-grained cross-level information interaction between different layers, addressing the issue of lack of information exchange between different levels due to same-level shortcuts in traditional networks, and uses AIAB to further enhance the efficiency of global feature utilization. In the following sections, we will provide a detailed description of these three main components of GVANet.

B. Multiview Feature Aggregation Block

In the past, most methods for remote sensing image segmentation have used single-view approaches. However, when facing the challenges of remote sensing segmentation, incorporating information from different views can enhance the network's robustness. For instance, regarding occlusion issues, a single-view network provides limited information, but different views have varying occlusion scenarios. Integrating information from views with less occlusion can effectively improve segmentation accuracy. To address this, we propose the MVFAB, which incorporates a multiview expansion module as illustrated in Fig. 3. This block first uses a preattention mechanism to establish dependencies between objects of the same type at different views but at the same sampling height [Fig. 4(a) shows instances of objects of the same type at different views with the same sampling height]. It then performs feature mapping for different sampling height views through the multiview expansion module [Fig. 4(b) shows instances of the same object at different sampling heights]. Finally, a postattention mechanism computes the fusion weights for multiview features and assigns the final feature values. The MVFAB structure consists of three parts: preattention for establishing dependencies between objects of the same type at different views but the same sampling height, multiview expansion for different sampling heights, and postattention for calculating fusion weights.

Preattention: We first establish horizontal dependencies using horizontal pooling and vertical dependencies using vertical pooling. These dependencies are then concatenated and dimensionally reduced to encode spatial information, establishing global dependencies among feature pixels. Finally, two parallel 1×1 convolutions are applied to learn channel correlations and interactions, better capturing feature differences and commonalities across different viewpoints. The expressions for this part can be described as

$$F_{\text{glb}}(x) = \text{Conv}_{1\times1}(\text{Cat}APxx, APy(x)) \tag{7}$$

$$F_{\text{pre}}(x) = x \cdot \text{Conv}_{1\times1}(F_{\text{glb}}(x)) \cdot \text{Conv}_{1\times1}(F_{\text{glb}}(x)) \tag{8}$$

where $X \in R^{C \times H \times W}$, $AP_x(\cdot)$ represents establishing horizontal dependencies, $AP_y(\cdot)$ represents establishing vertical dependencies. Cat denotes concatenation, $\text{Conv}_{x\times x}(\cdot)$ denotes

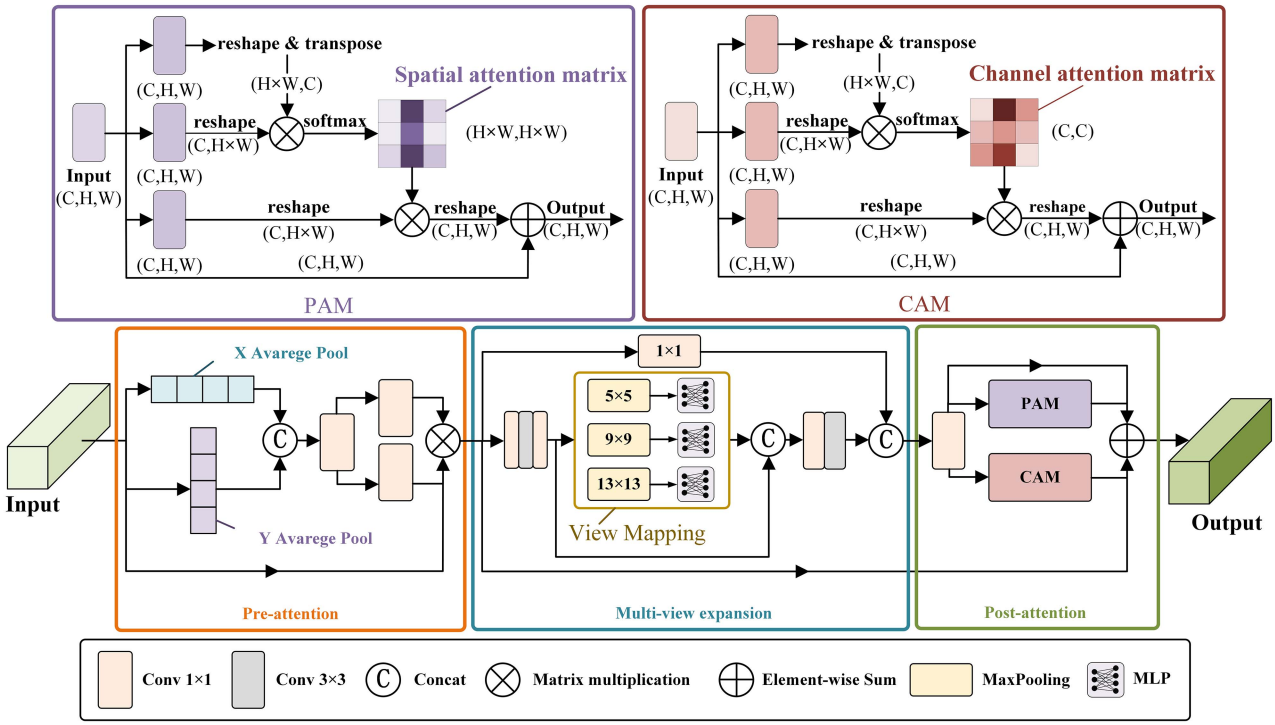


Fig. 3. MVFAB Structure Diagram. MVFAB aims to achieve multiview expansion, consisting of three key stages. The first stage is the preattention stage, which is used to establish long-range dependencies among different views at the same sampling height. Subsequently, in the multiview expansion stage, multiscale pooling is used for resolution simulation from different viewpoints and an MLP is employed to compensate for geometric deformations, thereby enabling the extraction of multiview information from different heights and perspectives. Finally, in the postattention stage, the model can selectively focus on relevant features and suppress irrelevant information. The focus of this stage lies in fusing and assigning values through concentration on feature pixels and specific channels.

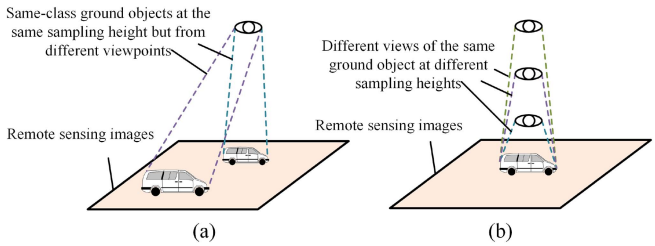


Fig. 4. Illustrates two different multiview scenarios. (a) depicts that objects of the same type at different positions have varying geometric information at the same sampling height. (b) shows that the same object has different geometric information at different sampling heights.

convolution with a kernel size of x, $F_{\text{glb}}(\cdot)$ represents establishing global dependencies, and $F_{\text{pre}}(\cdot)$ represents preattention operation.

1) Multiview Expansion: The objective of this module is to extract information from different heights and views to enhance the model's robustness to multiview inputs. Based on the perspective projection principle in image imaging (i.e., objects appear larger when closer and smaller when farther away), we approximate the original features to different view resolutions using spatial pyramid pooling (SPP) technology. Since changes in view can cause geometric deformations of objects, we mitigate these negative effects by mapping features at each resolution level using a multilayer perceptron (MLP). Specifically, this process involves two main branches: one is the view mapping branch, which first applies three convolutions to the feature map, followed by max pooling with kernels of sizes 5×5, 9×9, and 13×13. Each pooling branch uses an MLP to suppress the effects of geometric deformation. This method extends global features to multiple different height views. Finally, after two additional convolutions, the results are concatenated with the feature map processed by a single convolution to produce the output for this stage. The formulas for this stage can be described as follows:

$$F_{\text{vm}}(x) = \text{Cat}(\delta(\text{MP}_{5\times5}(x)), \delta(\text{MP}_{9\times9}(x)), \delta(\text{MP}_{13\times13}(x)), x) \quad (9)$$

$$F_{\text{Conv1}}(x) = \text{Conv}_{1\times1}\left(\text{Conv}_{3\times3}\left(\text{Conv}_{1\times1}(x)\right)\right) \quad (10)$$

$$F_{\text{Conv2}}(x) = \text{Conv}_{1\times1}\left(\text{Conv}_{3\times3}(x)\right) \quad (11)$$

$$F_{\text{mve}}(x) = \text{Cat}(F_{\text{Conv2}}(F_{Spp}(F_{\text{Conv1}}(x)), \text{Conv}_{1\times1}(x)) \quad (12)$$

where $F_{\text{vm}}(\cdot)$ denotes view mapping, $\delta(\cdot)$ represents the MLP, $F_{\text{Conv1}}(\cdot)$ and $F_{\text{Conv2}}(\cdot)$ denote two consecutive convolution operations, $\text{MP}_{x\times x}$ represents max pooling with a kernel size of x, and $F_{\text{mve}}(\cdot)$ denotes the multiview expansion operation.

2) Postattention: This part is intended to enhance the generalization ability of the multiview fusion features by adjusting the pixel and channel weights in the feature maps. Position attention and channel attention are applied to respectively improve the

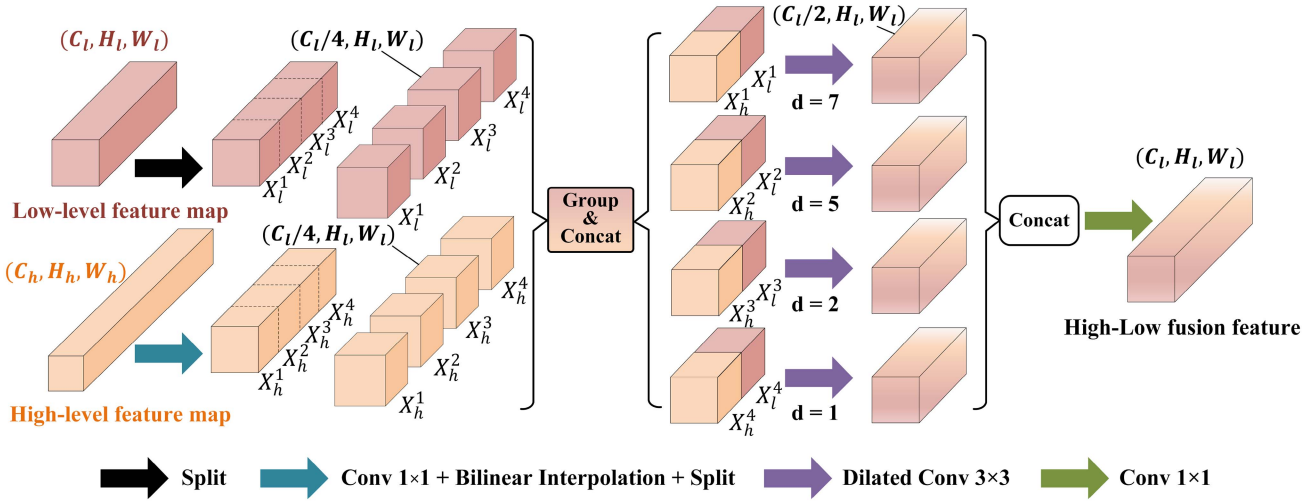


Fig. 5. CGB structure diagram. For the input high-level feature map, it is first upsampled to the size of low-level features, then divided into four groups along with the low-level features by channel. Subsequently, the features after grouping are pairwise concatenated and passed through depthwise separable convolutions. Finally, the features are mapped back to the size of low-level features for output. Grouping the features first helps better preserve and distinguish the semantic information of different-level features. Then, pairwise fusion of the grouped features enables each feature group to interact with others, facilitating finer and more comprehensive cross-level information exchange. Such interaction effectively compensates for the semantic gap between different-level features, enhancing feature representation and segmentation performance.

spatial and channel information perception capabilities of the multiscale global features. Spatial and channel attention matrices are generated to weight the original features, enhancing the perception of spatial and channel key information. Finally, to prevent network degradation, skip connections are introduced to aggregate the outputs of the two attention modules and the initial input, resulting in better feature representation for pixel-level predictions. The computational formula of MVFAB is described as follows:

$$F^i_{\text{MVFAB}} = F_{\text{pre}}(X_i) + F_{\text{post}}(F_{\text{mve}}(F_{\text{pre}}(X_i)) \tag{13}$$

where X_i represents the feature map of the ith downsampling stage, $i \in [3, 4]$, $F_{\text{post}}(x)$ denotes the postattention operation (refer to [25] for the specific formula), and F^i_{MVFAB} is the output feature map of the ith downsampling stage after passing through MVFAB.

C. Channel Group Fusion Block

In the past, networks often used intralayer skip connections to preserve original feature information, but this approach has some limitations when dealing with multiview problems. First, intralayer skip connections can only transmit feature information within the same layer, unable to effectively interact and fuse information across different layers. This results in the model's inability to fully utilize features from different layers to address multiview problems, especially those involving scenes with different scales and semantic information. Second, intralayer skip connections cannot effectively capture information about distant, blurry, or occluded objects because they only transmit features within the same layer, failing to obtain more global and comprehensive information. Therefore, when dealing with multiview problems, relying solely on intralayer skip connections may limit the model's perception and utilization of global features and multiscale information. To address these issues, we propose CGB, a module that can achieve fine-grained cross-layer fusion of information from different layers.Its structure is shown in Fig. 5, and it takes two inputs: low-level features and high-level features. First, it adjusts the size of the high-level features to match the size of the low-level features using depthwise separable convolution (DW) and bilinear interpolation. Specifically, for the two input features, high-level feature X_h and low-level feature X_l, the calculation formula is as follows:

$$f_h = \text{BI}(\text{DW}(X_h)) \tag{14}$$

where DW represents depthwise separable convolution, and BI stands for bilinear interpolation. f_h is the feature map of high-level feature X_h after being adjusted to match the size of X_l. Next, we split the two feature maps into four groups along the channel dimension and concatenate one group of low-level features with one group of high-level features, resulting in four sets of fused features. As shown in Fig. 4, this can be represented as follows:

$$\begin{cases} X^1_h, X^2_h, X^3_h, X^4_h = \text{Group}(f_h) \\ X^1_l, X^2_l, X^3_l, X^4_l = \text{Group}(X_l) \end{cases} \tag{15}$$

$$Y_i = X^i_h + X^i_l \tag{16}$$

where Group represents the grouping operation, X^i_h denotes the grouped high-level features after grouping for the ith group where $i \in [1, 2, 3, 4]$, X^i_l represents the grouped low-level features for the ith group, and Y_i is the feature obtained by combining the ith low-level feature with the ith high-level feature. Then, dilated convolutions with a kernel size of 3×3 are applied using different dilation rates $(1, 2, 5, 7)$ for the various groups to extract information at different scales. Finally, these four groups are concatenated along the channel dimension, followed by applying a regular 1×1 convolution to enable feature interaction

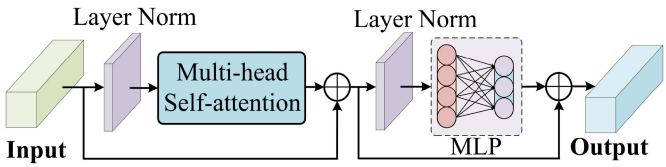


Fig. 6. Standard transformer structure.

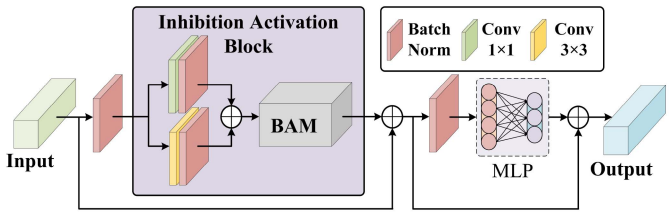


Fig. 7. AIAB structure diagram. We replaced the Transformer's multihead self-attention (MSA) mechanism with the IAB. In the inhibition activation block, we first employ a 1×1 convolution to emphasize interchannel relationships, while a 3×3 convolution focuses more on spatial features. Subsequently, the aggregated features are fed into the BAM, where important features are activated and less relevant ones are suppressed. Compared to MSA, BAM enables the model to simultaneously consider both spatial and channelwise characteristics of the features. This comprehensive approach enhances the model's capability to select and weight global features accurately.

at different scales. For the grouped features Y_i obtained from (16), the subsequent computation formulas are as follows:

$$F_{\text{CGB}} = \text{Conv}_{1\times1}\left(\sum_{i=1}^{4}\text{Conv}_{3\times3}\left(Y_i\right)\right) \tag{17}$$

where F_{CGB} represents the features fused by CGB.

D. Aggregation-Inhibition-Activation Block

To further enhance the utilization efficiency of features, we were inspired by Transformer and designed a module called AIAB, which is similar in structure to the standard Transformer as shown in Fig. 7. The structure of the standard Transformer is illustrated in Fig. 6. Compared to the standard Transformer, we replaced layer normalization with batch normalization. Additionally, we introduced an inhibition-activation block (IAB) to simplify operations, replacing the original self-attention module. Specifically, for the input features, we first perform batch normalization and then divide them into two branches: one branch undergoes a 1×1 convolution, while the other branch undergoes a 3×3 convolution. The 1×1 convolution helps the model focus more on relationships between channels, while the 3×3 convolution focuses more on spatial information in the feature maps. Subsequently, we fuse the features from both branches and further enhance the model's attention to spatial and channel information through a bottleneck attention module (BAM). The structure and implementation details of BAM can be found in reference [51]. Finally, we use an MLP for nonlinear mapping, similar to the standard Transformer block.

E. Loss Function

The loss function L we employ is a combination of a dice loss (L_{dice}) and a cross-entropy loss function (L_{ce}), which can be expressed as

$$L_{\text{ce}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \log \hat{y}_k^{(n)} \tag{18}$$

$$L_{\text{dice}} = 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\hat{y}_k^{(n)} y_k^{(n)}}{\hat{y}_k^{(n)} + y_k^{(n)}} \tag{19}$$

$$L = L_{\text{ce}} + L_{\text{dice}} \tag{20}$$

where N is the number of samples, and K is the number of classes. $y^{(n)}$ and $\hat{y}^{(n)}$ represent the one-hot encoding of the true semantic labels and their corresponding network softmax outputs, with $n \in [1, \ldots, N]$. $\hat{y}^{(k)}$ is the confidence of class k for sample n.

IV. Experiment

In this section, we will first introduce the dataset, experimental setup, and relevant metrics. Then, we will present our ablation experiments, and finally, we will discuss comparative experiments with other methods.

A. Experimental Settings

1) Datasets: We have used two commonly employed datasets in remote sensing image segmentation, namely, Vaihingen and Potsdam. The Vaihingen and Potsdam datasets are widely recognized standard datasets in the field of remote sensing image semantic segmentation, extensively used for evaluating algorithm performance to ensure the universality and comparability of research results. They offer a rich variety of land cover categories and diverse environmental conditions, including buildings, roads, trees, etc., as well as different seasons, weather, and lighting conditions, which contribute to assessing the robustness and generalization ability of models. Moreover, these datasets have been extensively utilized in research, facilitating easy comparison and benchmarking of our work with existing studies. In the following, we will introduce these two datasets.

Vaihingen dataset: The Vaihingen dataset comprises 33 high-resolution TOP image blocks, each with an average size of 2494×2064 pixels. These TOP image blocks are equipped with three multispectral bands (near-infrared, red, green), as well as a digital surface model (DSM) and a normalized digital surface model (NDSM), all with a ground sampling distance (GSD) of 9 cm. This dataset contains five primary foreground classes (impervious surfaces, buildings, low vegetation, trees, cars) and a background class (clutter). In the experiments, we selected training data according to the specific training IDs provided by the ISPRS Challenge (IDs: 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, 37). The remaining 17 images were used for testing. This selection ensures that our data are consistent with that of other researchers, facilitating comparative analysis [23],

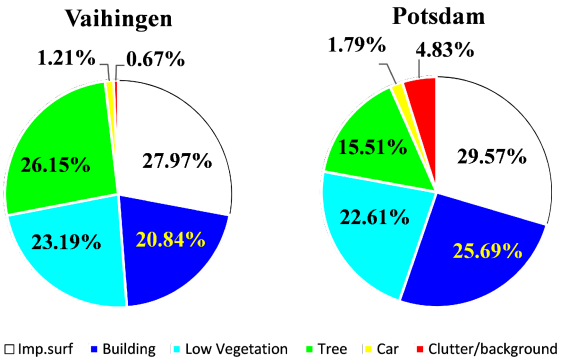


Fig. 8. Proportion of each semantic label in the two datasets.

[24]. To simplify processing and analysis, we divided the image blocks into patches measuring 1024 × 1024 pixels.

Potsdam dataset: The Potsdam dataset consists of 38 high-resolution TOP image blocks, each measuring 6000 × 6000 pixels and offering a 5-cm GSD. These images contain the same category information as the Vaihingen dataset. Moreover, they include four multispectral bands (red, green, and blue, and near-infrared), as well as DSM and NDSM. During the training process, we used specific training IDs provided by the ISPRS competition. These training IDs include images numbered 2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_8, 7_9, 7_11, and 7_12. The remaining 15 images were reserved as the test set. Similar to our approach with the Vaihingen dataset, we limited our analysis on the Potsdam dataset to use only three bands (red, green, blue). Additionally, we divided the original image blocks into 1024 × 1024 pixel patches for analysis. Notably, during our quantitative evaluations on both datasets, we excluded the "clutter/background" category.

2) Implementation Details: In our experiments, we utilized an Ubuntu 18.04 system and implemented the models using the PyTorch 1.11 framework on a single NVIDIA GeForce RTX 2080 Ti 11GB GPU to ensure efficient performance. To facilitate rapid convergence, we employed the AdamW optimizer for training all models, with a base learning rate set at 6e-4. We also incorporated a cosine learning rate scheduling strategy. For the Vaihingen and Potsdam datasets (are shown in Fig. 8), we adopted a training approach where images were randomly cropped into patches of 512 × 512 dimensions. Throughout the training process, we applied various data augmentation techniques, including random scaling factors $(0.5, 0.75, 1.0, 1.25, 1.5)$, random vertical and horizontal flipping, as well as random rotation to enhance the robustness of the models. The training process was carried out for a total of 105 epochs. During testing, we applied augmentations including multi-scale variations and random flipping.

3) Evaluation Metrics: We use common remote sensing segmentation metrics such as overall accuracy (OA), F1 score, and mean intersection over union (mIoU) as our evaluation metrics. Additionally, we use the "Parameters" metric to evaluate the model's parameter count. Before discussing these metrics, let is introduce some related metrics, such as precision and recall. We also need to understand the meaning of certain symbols: tp (true positives), fp (false positives), fn (false negatives), tn (true negatives).

Precision: Precision measures the proportion of true positive predictions among all the samples predicted as positive by the model. In other words, precision tells us the probability that a sample predicted as positive is indeed a true positive

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{21}$$

Recall: Recall refers to the proportion of true positive predictions among all samples that are actually positive. Recall measures the model's ability to discover all positive instances

$$P\text{call} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{22}$$

Overall accuracy: OA is one of the commonly used performance evaluation metrics in image classification tasks. It is the proportion of correctly classified samples to the total number of samples. However, OA may not handle class imbalance well because when the number of samples in some classes is much larger than in other classes, the model may tend to predict the class with more samples. Number of correctly classified samples: The sum of all true positives and true negatives (tp + tn). Total number of samples: The sum of all samples (tp + fp + fn + tn)

$$\text{OA} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}}. \tag{23}$$

F1 score: The F1 score is the harmonic mean of precision and recall. It combines the model's accuracy and its ability to capture positive instances. For multiclass problems, F1 scores are typically calculated for each class, and then the average of these class F1 scores is computed. For each class

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{24}$$

Average F1 score = Average of F1 Scores for all classes

Mean intersection over union: mIoU is a commonly used evaluation metric in semantic segmentation tasks to measure the accuracy of a model at the pixel level. Intersection over union (IoU) is used to assess the model's segmentation results for each class, while mIoU computes the average IoU across all classes. For each class

$$\text{IoU} = \frac{\text{tp}}{\text{tp} + \text{fp} + \text{fn}} \tag{25}$$

mIoU is the sum of IoU values for all categories divided by the number of categories.

B. Ablation Study

1) Components of GVANet: In order to separately assess the performance of the individual components of the proposed GVANet, we conducted a series of ablation experiments on the Vaihingen and Potsdam datasets. For ease of discussion, we primarily focused on mIoU and meanF1. We set up a

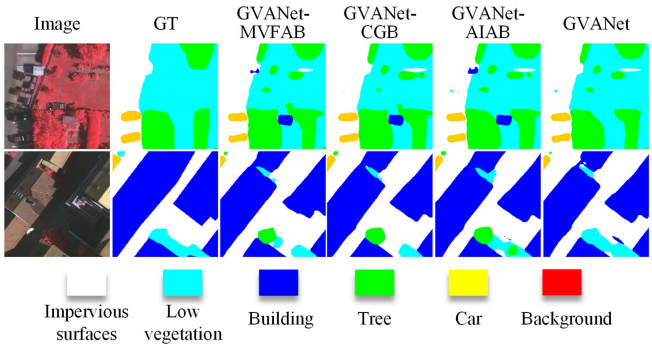


Fig. 9. Local zoom-in images of GVANet with one module removed on the Vaihingen dataset.

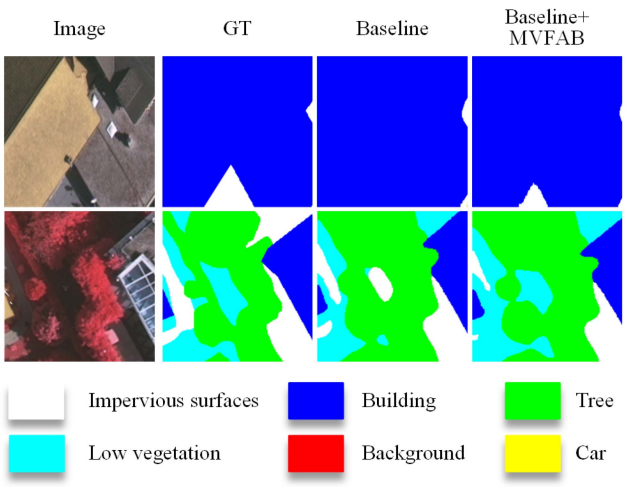


Fig. 10. Baseline + MVFAB local zoom-in view.

TABLE I
ABLATION EXPERIMENTS OF GVANET COMPONENTS

Method	Vaihingen		Potsdam	
	mIoU(%)	F1(%)	mIoU(%)	F1(%)
GVANet	84.54	91.52	87.62	93.29
GVANet−MVFAB	83.77	91.05	86.51	92.65
GVANet−CGB	83.82	91.08	86.77	92.80
GVANet−AIAB	83.90	91.12	86.80	92.82

baseline using U-Net with the ConvNext backbone and employed the feature refinement head (FRH) in the final layer.We conducted ablation experiments on the Vaihingen dataset by removing one of these components, denoted in Table I with a minus sign (−). We conducted joint ablation experiments on the three modules we introduced, and the results are presented in Table I. All experimental results are averaged from multiple trials.

From Fig. 9, it can be observed that when any module of our GVANet is removed, the performance deteriorates compared to the GVANet with all components intact. Furthermore, to demonstrate the roles and superiority of each module, we conducted individual module ablations on the Vaihingen dataset. The results are presented in Table II.

TABLE II
ABLATION EXPERIMENTS OF INDIVIDUAL MODULES IN GVANET ON THE VAIHINGEN DATASET

Method	mIoU(%)	F1(%)
Baseline	82.42	90.10
Baseline+MVFAB	83.20	90.69
Baseline+CGB	83.12	90.67
Baseline+AIAB	83.07	90.61

C. Comparison With Other Methods

1) Effect of MVFAB: From Table I, it can be observed that on the Vaihingen dataset, our model without the MVFAB module has a lower mIoU by 0.77% compared to the network with the MVFAB module. For F1, using GVANet with MVFAB results in a 0.47% increase. On the Potsdam dataset, the model without MVFAB has a 1.11% lower mIoU compared to the model with MVFAB, and using GVANet with MVFAB results in a 0.64% increase in F1. Furthermore, the network models with MVFAB are closer to our final results compared to those without MVFAB. The difference in the mIoU metric between the two models using MVFAB is not significant, as shown in Table II. On the Vaihingen dataset, the baseline model sees a 0.78% improvement in mIoU and a 0.59% improvement in F1 after adding MVFAB. Overall, using MVFAB results in at least a 0.77% improvement in mIoU and at least a 0.47% improvement in F1. From Fig. 10, it can be seen that the model with the MVFAB performs well in segmenting small instances in the first row that were not effectively addressed before, and in handling the issue of tree shadows mapped onto low vegetation in the second row that led to mixing and difficulty in segmentation. This also confirms the effectiveness of our MVFAB.

To further demonstrate the advantages of MVFAB, we provide the feature maps before and after using MVFAB in the fourth stage. The results are shown in Fig. 11, where the first column is the original image, the second column shows the feature maps without MVFAB, and the third column displays the feature maps with MVFAB. The figure reveals that after using MVFAB, the contrast in various regions of the features is reduced, indicating that the deep features are more consistent. This suggests that the model's feature changes across different regions are more stable in the deeper layers. Even with some local noise or variations in the deep features, the model is better able to identify the overall features. Additionally, the relatively lower contrast in deep features helps the model capture the overall structure and semantic information in the image more effectively. As shown in the second row of the figure, the occluded car has higher weights in the features after using MVFAB, indicating that the model has given focused attention to this region.

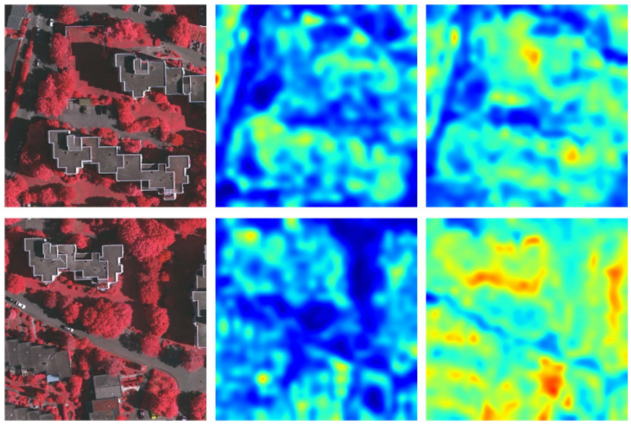


Fig. 11. Feature maps of the fourth stage of GVANet. The first column shows the original image, the second column displays the features without using MVFAB, and the third column shows the features with MVFAB. In the figure, redder colors indicate higher weights, while bluer colors indicate lower weights. The figure demonstrates that after using MVFAB, the contrast of features across different regions is reduced, indicating that the deep-layer features after applying MVFAB are more uniform and consistent. This suggests that the model pays more attention to global features rather than local ones. Additionally, it can be observed that for segmentation targets that benefit from multiview techniques, such as the occluded car in the second row of the image, the corresponding feature with MVFAB has a higher weight, indicating that the model focuses more on this region.

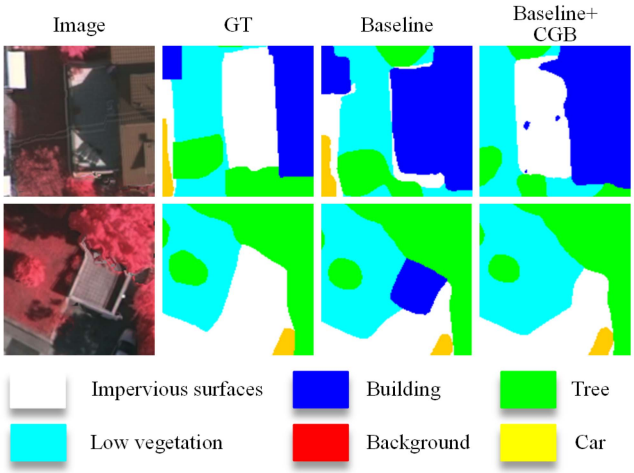


Fig. 12. Baseline + CGB local amplification.

To further investigate the advantages of MVFAB (are shown in Fig. 12), we have also dissected the MVFAB into its internal components, splitting it into preattention (Prea), multiview expansion (Mve), and postattention (Posta). We conducted ablation experiments on the Vaihingen dataset by incrementally adding individual components, as indicated by the plus sign (+) in Table III. This was done to demonstrate the effectiveness of each internal structure within our designed MVFAB. The experimental results are presented in Table III.

From Table III, it can be observed that the removal of various components leads to a reduction in performance compared to the complete MVFAB. Specifically, the MVFAB without preattention exhibits a decrease of 0.3% in mIoU compared to the

TABLE III
ABLATION OF INTERNAL COMPONENTS WITHIN MVFAB ON VAIHINGEN DATASET

Method	mIoU(%)	F1(%)
Baseline+MVFAB	83.20	90.69
Baseline+MVFAB-Prea	82.90	90.49
Baseline+MVFAB-Mve	82.93	90.51
Baseline+MVFAB-Posta	82.85	90.42
Baseline+MVFAB-Prea-Mve	82.65	90.26
Baseline+MVFAB-Prea-Posta	82.64	90.24
Baseline+MVFAB-Mve-Posta	82.69	90.35

TABLE IV
PARAMETER QUANTITY OF DIFFERENT BACKBONES AND MIOU ON THE VAIHINGEN DATASET

Method	Backbone	Parameters(M)	mIoU(%)
GVANet	ResNet50	25.56	84.16
	ResNext50	25.03	84.29
	ResNest50	27.48	84.33
	ConvNext-Tiny	28.59	84.54

MVFAB with preattention. Similarly, the MVFAB without Mve shows a 0.27% decrease in mIoU compared to the MVFAB with Mve. Moreover, the MVFAB without Posta demonstrates a 0.35% decrease in mIoU compared to the MVFAB with Posta. Furthermore, Table III indicates that the MVFAB module with both components removed performs worse than the MVFAB module with only one component removed. These experimental results affirm the effectiveness of each component within the MVFAB.

2) Effect of CGB: From Table I, it can be observed that on the Vaihingen dataset, networks utilizing CGB achieve a 0.72% higher mIoU and a 0.44% higher F1 score compared to networks that do not use CGB. On the Potsdam dataset, networks without CGB perform 0.85% lower in mIoU and 0.49% lower in F1 score compared to networks with CGB. As seen in Table II, on the Vaihingen dataset, the baseline model exhibits a 0.7% improvement in mIoU and a 0.57% improvement in F1 score after the addition of CGB. With the use of CGB, mIoU is enhanced by at least 0.7%, and F1 is improved by at least 0.44%. From Fig. 14, it can be observed that after leveraging the CGB to integrate information across different scales, the model exhibits satisfactory results in dealing with the segmentation challenges posed by occlusion or differences in viewpoint, as it effectively incorporates semantic information from deep-layer features, directing the model's focus towards the instance itself rather than individual pixels.

Furthermore, to evaluate the fusion outcomes of features using CGB, we present the feature maps from the shallowest layer, i.e., the first stage, as shown in Fig. 13. The second

TABLE V
COMPARISON OF SEGMENTATION RESULTS ON THE VAIHINGEN DATASET

Method	F1(%)					Evaluation index		
	Imp.Surf	Building	Lowveg	Tree	Car	MeanF1(%)	mIoU(%)	OA(%)
MANet[52]	84.95	88.41	78.16	88.37	70.47	82.07	70.16	84.80
ABCNet[53]	88.13	90.23	76.71	87.21	68.72	82.20	70.58	85.62
MACU-Net[54]	90.70	92.40	81.90	89.37	82.53	87.38	77.85	88.61
MAResU-Net[55]	92.91	95.26	84.95	89.94	88.33	90.28	83.30	90.86
A2-FPN[56]	92.99	95.53	84.67	90.34	87.62	90.23	82.42	91.04
DC-Swin[57]	93.46	96.00	85.32	90.03	84.88	89.94	82.01	91.29
Mask2Former[58]	92.86	**96.03**	84.15	90.50	89.30	90.57	82.99	91.10
MPCNet[59]	92.76	95.50	84.70	90.40	90.44	90.76	83.27	90.93
VMFormer[60]	93.23	95.81	85.64	**91.08**	88.65	90.88	83.48	91.54
GVANet(Ours)	**93.66**	95.99	**86.07**	90.79	**91.10**	**91.52**	**84.54**	**91.72**

The best results are indicated in bold black.

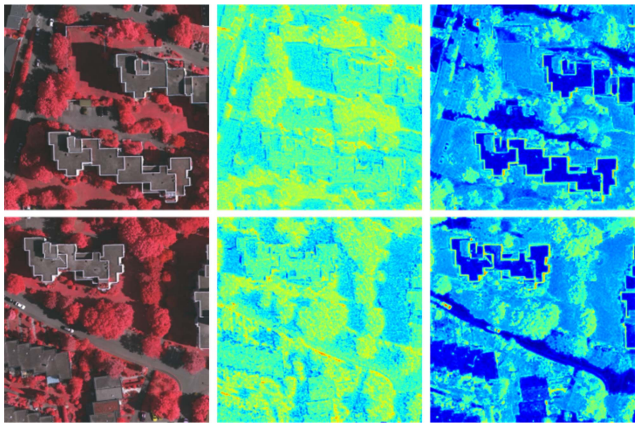


Fig. 13. Feature maps of the first stage in GVANet. The first column represents the original image, the second column represents the features without cross-level information interaction using CGB, and the third column represents the features after cross-level fusion using CGB. From the graph, it is evident that without using CGB for cross-level information interaction, the feature distribution is overly uniform, with low contrast. However, after employing CGB for cross-level fusion, the features exhibit more distinct and vivid contrasts. In segmentation networks, shallow layers require greater differences in feature positions to facilitate pixel-level classification. This helps the model better capture texture and boundary information in the image. Clearly, using CGB for cross-level fusion results in better feature representation.

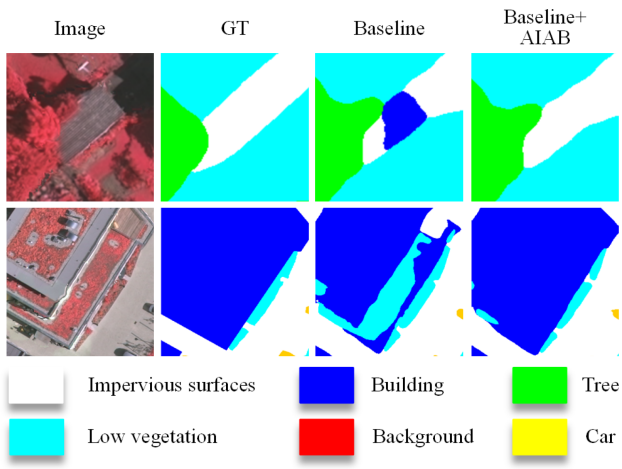


Fig. 14. Local zoom-in of baseline+AIAB on Vaihingen.

column represents features without cross-level information interaction using CGB, while the third column represents features after cross-level fusion with CGB. It is evident from the figure that without leveraging CGB for cross-level interaction, the feature distribution is excessively uniform with low contrast. Conversely, utilizing CGB for cross-level fusion results in enhanced local details and texture information in the features. Since the segmentation network's shallow layers require pixel-level classification, it is imperative for the differences in positions within the shallow features to be more pronounced, facilitating better capture of texture and boundary information in the images. Therefore, the use of CGB for cross-level fusion yields improved feature representation, further validating its effectiveness.

3) Effect of AIAB: From Table I, it can be observed that the network using AIAB on the Vaihingen dataset achieves a 0.64% higher mIoU and a 0.4% higher F1 score compared to the network without AIAB. On the Potsdam dataset, the network using AIAB outperforms the network without AIAB with a 0.82% higher mIoU and a 0.47% higher F1 score. Table II shows that on the Vaihingen dataset, the Baseline model experiences a 0.65% improvement in mIoU and a 0.51% improvement in F1 when AIAB is incorporated. From Fig. 14, it can be observed that the introduction of AIAB into the baseline model enhances the efficiency of utilizing global features in the feature reconstruction stage, allowing the model to perceive more semantic information. The segmentation results in Fig. 14 demonstrate that the model without AIAB in the feature reconstruction stage tends to focus more on individual pixels. This leads to poorer

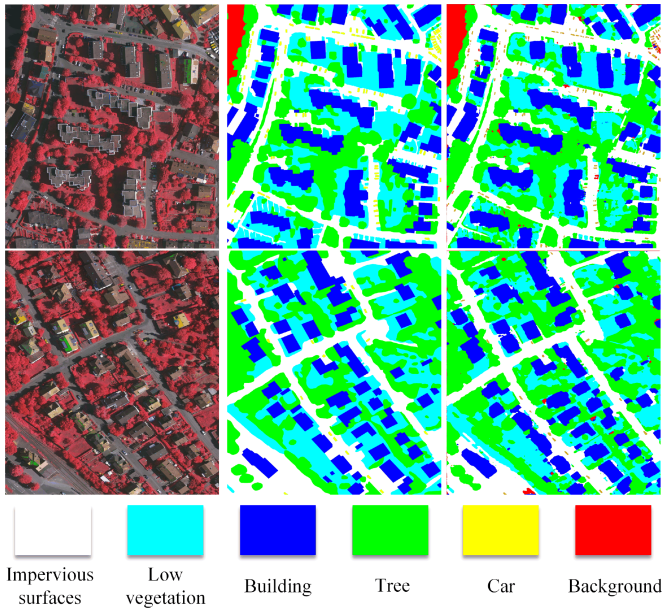


Fig. 15. Visualization results for Vaihingen Test Set ID 2 and ID 14. The first column showcases the RGB image, while the second column provides the ground truth (GT). The third column exhibits the segmentation outcomes generated by our GVANet.

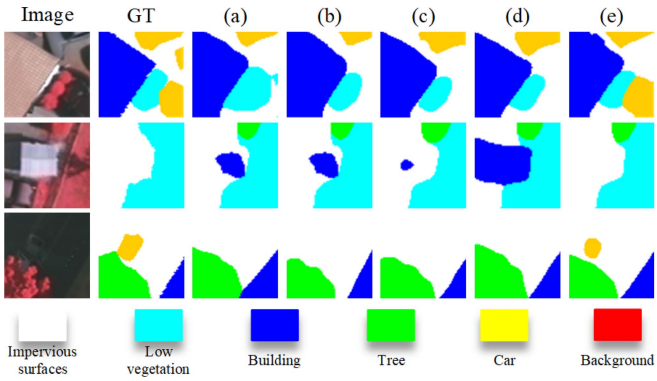


Fig. 16. Challenges in remote sensing segmentation are illustrated. The figure shows the original images along with the corresponding GT segmentation labels, as well as segmentation results from several mainstream methods and GVANet. (a) MAResU-Net, (b) A2-FPN, (c) DC-Swin, (d) MPCNet, and (e) GVANet. The first row of images demonstrates how vehicles exhibit different appearance features under varying viewpoints and lighting conditions, posing challenges for segmentation in remote sensing scenarios. The second row highlights the difficulty in distinguishing between nonpermeable surfaces and buildings with similar colors in different scenes. The third row illustrates instances where vehicles are occluded by trees, making it challenging for segmentation algorithms to correctly identify and segment occluded objects. It is evident from the images that GVANet outperforms mainstream networks in addressing common remote sensing challenges.

segmentation performance for instances with similar colors but different categories under different viewing angles. In the segmentation image of the second row in Fig. 14, Buildings with colors similar to low vegetation are erroneously segmented as low vegetation. This further confirms the effectiveness of AIAB.

From Table IV , it can be observed that even when using a ResNet backbone network, satisfactory results can still be achieved. Although the ConvNext-tiny model has a larger number of parameters, the slight increase in parameter count is entirely acceptable when compared to the improvement in final results.

The comparative models we selected are as follows: a multi-attention network with kernel attention (MANet) [52], a "Dual Path" network named ABCNet with spatial and contextual paths [53], MACU-Net based on multiscale skip connections and asymmetric convolutions [54], multistage attention residual UNet (MAResU-Net) with linear attention mechanisms [40], attention aggregation feature pyramid Network (A2-FPN) [55], DC-Swin, which incorporates a dense connection feature aggregation module (DCFAM) into the Swin-Transformer network [56], the masked Transformer network Mask2Former [57], MPCNet, a network with a multiscale prototype transformer decoder [58], and the Transformer network with variable window attention, VMFormer [59]. Our model ultimately achieves higher accuracy on the ISPRS Vaihingen and ISPRS Potsdam datasets, which are commonly used for remote sensing segmentation tasks, compared to the models mentioned above.

4) Results on the Vaihingen Dataset: Table V presents numerical results of various semantic segmentation methods compared on the Vaihingen dataset. The results indicate that our proposed GVANet achieves an average F1 of 91.52%, an mIoU of 84.54%, and an OA of 91.72%. Notably, MHLNet provides the best F1, OA, and mIoU, significantly outperforming other networks. We not only surpass the excellent lightweight convolutional network ABCNet but also outperform the DC-Swin network, which has strong global information representation capabilities.

To demonstrate the advantages of GVANet in addressing remote sensing segmentation challenges, we also provide a visual comparison with other networks. The results are shown in Fig. 16. The first row illustrates how cars exhibit different appearance features under varying viewpoints and lighting conditions in remote sensing segmentation challenges. Due to building shadows obscuring the car, the car and shadows merge, leading to segmentation errors by mainstream networks. However, GVANet successfully segments the car. The second row shows cases where impermeable surfaces and buildings with similar colors are difficult to distinguish, with GVANet correctly differentiating them. The third row depicts scenarios where cars are obstructed by trees, making it challenging for segmentation algorithms to accurately identify and segment the occluded objects. These results demonstrate the effectiveness of GVANet in addressing remote sensing segmentation challenges.

Furthermore, the prediction results for IDs 2 and 14 are shown in Fig. 15. Fig. 17 displays the prediction results of several semantic segmentation methods mentioned in Table V. From Fig. 17, it can be observed that our model outperforms current methods in some segmentation areas where the foreground and background colors are extremely similar. Most models tend to classify colors similar to the background into the same class. However, in actual images, factors such as shadows, lighting, and the object's inherent color can lead to misidentification by models that focus on local features. After incorporating multiview and global information, our model pays more attention

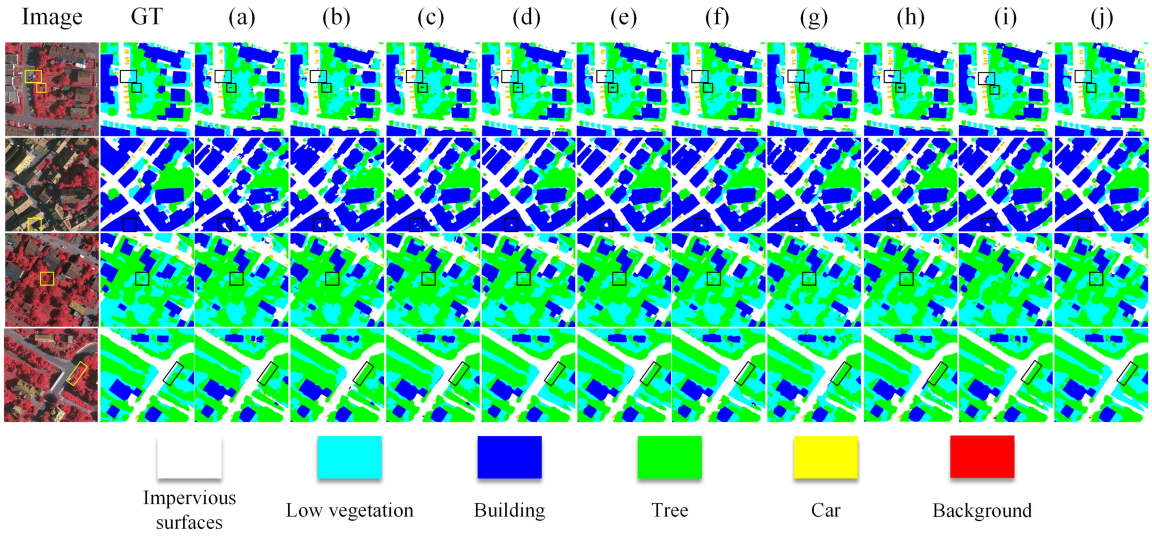


Fig. 17. Segmentation results examples of different models on Vaihingen dataset. (a) MANet. (b) ABCNet. (c) MACU-Net. (d) MAResU-Net. (e) A2-FPN. (f) DC-Swin. (g) Mask2Former. (h) MPCNet. (i) VMFormer. (j) GVANet.

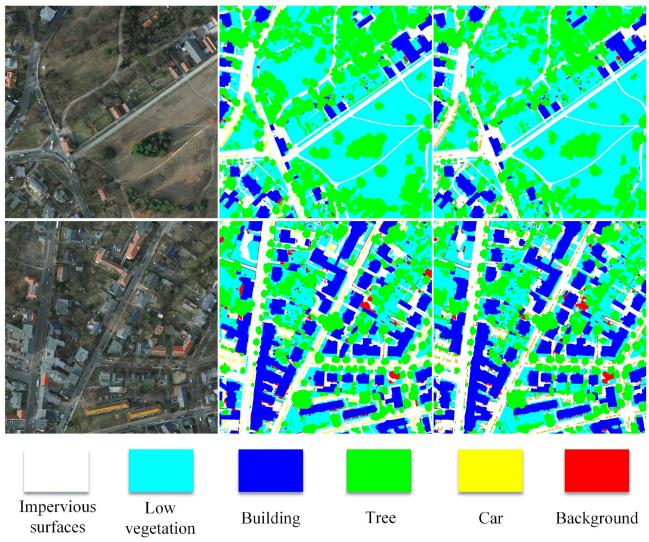


Fig. 18. Visualization of results for Potsdam test set ID 3_14 and 2_13. The first column showcases the RGB image, while the second column provides the GT. The third column exhibits the segmentation outcomes generated by our GVANet.

to the classification of the instances themselves, ignoring the influence of similar colors. This allows our model to better capture the semantic information in remote sensing images, focusing on instance classification rather than just pixel classification, resulting in improved boundary segmentation for Trees and Buildings in Fig. 17.

5) Results on the Potsdam Dataset: To provide a comprehensive evaluation, we conducted further experiments on the Potsdam dataset. As shown in Table VI, our GVANet achieved an average F1 score of 93.29% and an mIoU of 87.62% on the Potsdam test set, along with an OA score of 91.98%, all of which outperformed other methods. Due to differences in data size and data type, segmentation accuracy on the Potsdam dataset is generally higher than that on the Vaihingen dataset. From the experimental results, it is evident that the use of multireceptive fields and hybrid attention mechanisms significantly outperforms other methods.

As shown in Fig. 18, we present the segmentation results for ID 2_14 and 3_13. Additionally, we provide the prediction results for several semantic segmentation methods mentioned in Table VI. As illustrated in Fig. 19, our model performs better when segmenting instances with similar colors, even when the instance color closely resembles the background. From Fig. 19, it can be observed that our model excels in segmenting Buildings, Lowveg, and Trees.

6) Experimental Summary: The experiments demonstrate that our model outperforms current models. Figs. 17 and 19 show that after incorporating the MVFAB module to integrate multiview information, the model effectively addresses issues such as difficulty in recognizing objects due to lighting conditions, distinguishing objects with similar colors, and segmentation challenges caused by object occlusion. This indicates that our MVFAB is effective in tackling these problems. Additionally, subsequent experiments validate the effectiveness of using CGB for cross-layer information interaction and the superiority of AIAB in enhancing global feature utilization.

V. Limitations and Future Work

Although our GVANet has exhibited advantages in terms of data, it has limitations in accurately capturing object boundaries. This becomes particularly evident when segmentation results do not align perfectly with object shapes, resulting in less smooth boundaries. To address this challenge, we will delve into encoding techniques for boundary features. Furthermore, we plan to explore model compression techniques to

TABLE VI
COMPARISON OF SEGMENTATION RESULTS ON THE POTSDAM DATASET

Method	F1(%)					Evaluation index		
	Imp.Surf	Building	Lowveg	Tree	Car	MeanF1(%)	mIoU(%)	OA(%)
MANet[52]	87.46	90.66	81.96	82.81	92.76	87.13	77.45	85.05
ABCNet[53]	87.05	88.26	78.94	84.63	93.11	86.4	76.35	84.21
MACU-Net[54]	91.39	93.74	85.87	86.8	94.53	90.46	82.78	88.81
MAResU-Net[55]	92.38	95.83	86.65	88.44	96.13	91.89	85.22	90.28
A2-FPN[56]	92.85	95.84	87.17	88.76	96.13	92.15	85.66	90.68
DC-Swin[57]	93.26	96.86	87.74	88.68	95.50	92.41	86.10	91.16
Mask2Former[58]	92.48	96.41	87.53	89.37	96.08	92.37	86.03	90.83
MPCNet[59]	92.69	96.38	87.30	88.74	96.34	92.29	85.91	90.56
VMFormer[60]	93.23	95.81	85.64	**91.08**	88.65	90.88	83.48	91.54
GVANet(Ours)	**93.89**	**97.38**	**88.65**	90.01	**96.51**	**93.29**	**87.62**	**91.98**

The best results are indicated in bold black.

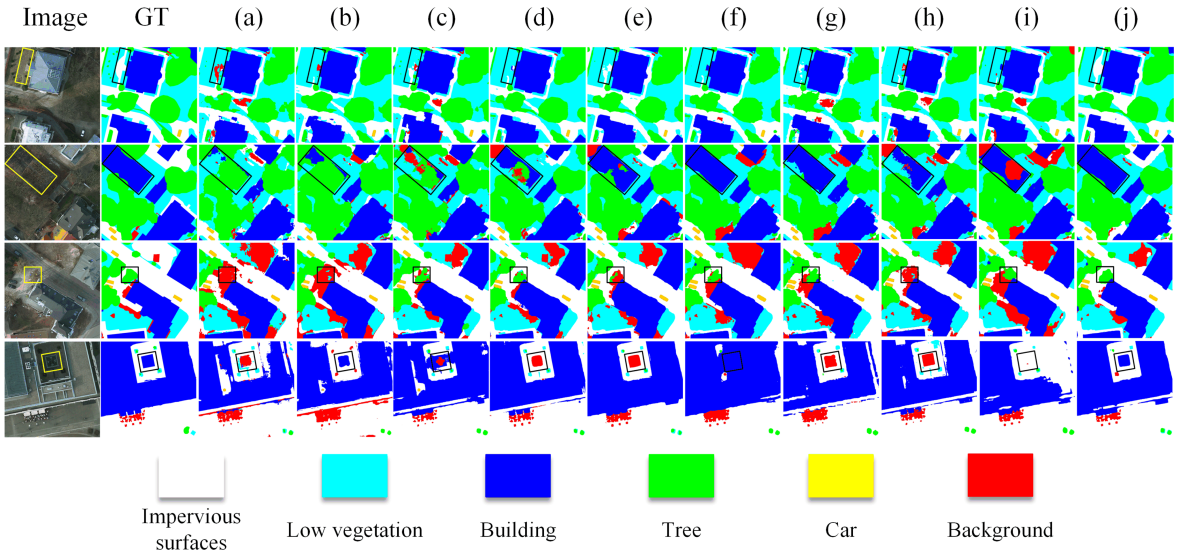


Fig. 19. Examples of segmentation results for different models on Potsdam dataset. (a) MANet. (b) ABCNet. (c) MACU-Net. (d) MAResU-Net. (e) A2-FPN. (f) DC-Swin. (g) Mask2Former. (h) MPCNet. (i) VMFormer. (j) GVANet.

enhance segmentation efficiency in our future work. Moreover, it is worth noting that our proposed method exclusively focuses on semantic segmentation of urban remote sensing images and has not yet delved into other remote sensing visual tasks such as road segmentation or parcel segmentation. In our upcoming research endeavors, we aim to develop enhanced architectures that incorporate attention mechanisms and multiview features, optimizing our network to accommodate a broader spectrum of remote sensing visual tasks.

VI. CONCLUSION

This article focuses on addressing the challenges of remote sensing segmentation by utilizing information from different views and overcoming the limitations of traditional skip connections that cannot facilitate cross-layer information interaction. We propose GVANet, a network that integrates multiview information and cross-layer scale information fusion, featuring multiview and multiinformation blending. Specifically, we first design the MVFAB to address the issue of single-view feature extraction in traditional networks. This module leverages view information at different sampling heights and expands different heights with varying views to transform single-view features into multiview features, which are then weighted and fused spatially and channelwise. To overcome the limitation of traditional same-level skip connections in handling remote sensing segmentation, we introduce the CGB module, designed for fine-grained cross-layer feature fusion to improve information interaction across different levels. Finally, to enhance the efficiency of utilizing multiview global information,

we propose the AIAB for further feature selection to extract more useful features. We have demonstrated the superiority of our network structure and the effectiveness of each module through experiments. We hope to inspire more researchers to explore the potential and applications of global multiview information and cross-layer information interaction in addressing various mainstream challenges in remote sensing segmentation.

REFERENCES

[1] R. Li, S. Zheng, C. Duan, L. Wang, and C. Zhang, "Land cover classification from remote sensing images based on multi-scale fully convolutional network," *Geo-Spatial Inf. Sci.*, vol. 25, no. 2, pp. 278–294, 2022.

[2] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[3] J. Xing, R. Sieber, and T. Caelli, "A scale-invariant change detection method for land use/cover change research," *ISPRS J. Photogrammetry Remote Sens.*, vol. 141, pp. 252–264, 2018.

[4] A. Samie et al., "Examining the impacts of future land use/land cover changes on climate in Punjab province, Pakistan: Implications for environmental sustainability and economic growth," *Environ. Sci. Pollut. Res.*, vol. 27, pp. 25415–25433, 2020.

[5] D. Griffiths and J. Boehm, "Improving public data for building segmentation from convolutional neural networks (CNNs) for fused airborne LiDAR and image data using active contours," *ISPRS J. Photogrammetry Remote Sens.*, vol. 154, pp. 70–83, 2019.

[6] 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.

[7] M. C. A. Picoli et al., "Big earth observation time series analysis for monitoring brazilian agriculture," *ISPRS J Photogrammetry Remote Sens.*, vol. 145, pp. 328–339, 2018.

[8] Y. Shen, J. Chen, L. Xiao, and D. Pan, "Optimizing multiscale segmentation with local spectral heterogeneity measure for high resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 157, pp. 13–25, 2019.

[9] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[10] D. Yu et al., "Deep convolutional neural networks with layer-wise context expansion and attention," in *Proc. Interspeech*, 2016, pp. 17–21.

[11] Y. Guo, X. Jia, and D. Paull, "Effective sequential classifier training for SVM-based multitemporal remote sensing image classification" *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3036–3048, 2018, doi: 10.1109/TIP.2018.2808767.

[12] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[13] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 582–589.

[14] J. Zhou et al., "UGIF-Net: An efficient fully guided information flow network for underwater image enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4206117.

[15] L. Ma, T. Ma, X. Xue, X. Fan, Z. Luo, and R. Liu, "Practical exposure correction: Great truths are always simple," 2022, *arXiv:2212.14245*.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[17] J. Zhou, Q. Liu, Q. Jiang, W. Ren, K.-M. Lam, and W. Zhang, "Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction," *Int. J. Comput. Vis.*, vol. 2023, pp. 1–19, 2023.

[18] 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.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Munich, Germany, 2015, pp. 234–241,.

[21] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 78–95, 2018.

[22] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

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

[24] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[25] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[28] K.-H. Liu and B.-Y. Lin, "MSCSA-Net: Multi-scale channel spatial attention network for semantic segmentation of remote sensing images," *Appl. Sci.*, vol. 13, no. 17, 2023, Art. no. 9491.

[29] G. Machado, M. B. Pereira, K. Nogueira, and J. A. Dos Santos, "Facing the void: Overcoming missing data in multi-view imagery," *IEEE Access*, vol. 11, pp. 12547–12554, 2022.

[30] Z. Qi et al., "Multi-view remote sensing image segmentation with Sam priors," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Athens, Greece, pp. 8446–8449, 2024.

[31] M. Zhang, X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang, "Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 500.

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

[33] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.

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

[35] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Semantic segmentation of multisensor remote sensing imagery with deep convnets and higher-order conditional random fields," *J. Appl. Remote Sens.*, vol. 13, no. 1, pp. 016501–016501, 2019.

[36] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention UNet for building segmentation in remote sensing images," *Sci. China Inf. Sci.*, vol. 63, pp. 1–12, 2020.

[37] Y. Hou, Z. Liu, T. Zhang, and Y. Li, "C-UNet: Complement UNet for remote sensing road extraction," *Sensors*, vol. 21, no. 6, 2021, Art. no. 2153.

[38] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet : A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support, 8th Int. Workshop*, Granada, Spain, 2018, pp. 3–11.

[39] Y. Zhang, W. Gong, J. Sun, and W. Li, "Web-Net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries," *Remote Sens.*, vol. 11, no. 16, 2019, Art. no. 1897.

[40] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResuNet for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8009205 .

[41] A. Han, L. Xing, W. Liu, and B. Liu, "MVFF: Multi-view feature fusion for few-shot remote sensing image scene classification," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2022, pp. 2452–2458.

[42] W. Zhou, Y. Shi, and X. Huang, "Multi-view scene classification based on feature integration and evidence decision fusion," *Remote Sens.*, vol. 16, no. 5, 2024, Art. no. 738.
[43] Q. Yu, X. Zhao, Y. Pang, L. Zhang, and H. Lu, "Multi-view aggregation network for dichotomous image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 3921–3930.
[44] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, 2018.
[45] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[46] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning," *Remote Sens.*, vol. 11, no. 1, 2019, Art. no. 83.
[47] Y. Ren, Y. Yu, and H. Guan, "Da-capsUNet: A dual-attention capsule U-Net for road extraction from remote sensing imagery," *Remote Sens.*, vol. 12, no. 18, 2020, Art. no. 2866.
[48] A. Jha, S. Bose, and B. Banerjee, "GAF-Net: Improving the performance of remote sensing image fusion using novel global self and cross attention learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 6354–6363.
[49] X. Wang et al., "Adaptive local cross-channel vector pooling attention module for semantic segmentation of remote sensing imagery," *Remote Sens.*, vol. 15, no. 8, 2023, Art. no. 1980.
[50] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.
[51] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.
[52] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.
[53] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 84–98, 2021.
[54] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-Net for semantic segmentation of fine-resolution remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007205.
[55] 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.
[56] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6506105.
[57] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1290–1299.
[58] Q. Wang, X. Luo, J. Feng, G. Zhang, X. Jia, and J. Yin, "Multiscale prototype contrast network for high-resolution aerial imagery semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615114.
[59] H. Yan, M. Wu, and C. Zhang, "Multi-scale representations by varying window attention for semantic segmentation," 2024, *arXiv:2404.16573*.

Yunsong Yang received the bachelor's degree in software engineering from the School of Computer Science, University of South China, Hunan, China, in 2021. He is currently working toward the master's degree in computer software and theory with the School of Computer Science and Technology, Shandong University of Finance and Economics in Yantai, Shandong, China.

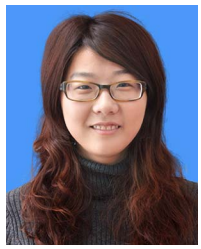
His research interests include computer graphics, computer vision, and image processing.

Jinjiang Li received the B.S. and M.S. degrees from Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shandong University, Jinan, China, in 2010, all in computer science.

From 2004 to 2006, he was an Assistant Research Fellow with the Institute of Computer Science and Technology, Peking University, Beijing, China. From 2012 to 2014, he was a Postdoctoral Fellow with Tsinghua University, Beijing, China. He is currently a Professor with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai, China. His research interests include image processing, computer graphics, computer vision, and machine learning.

Zheng Chen received the B.S. degree from Shandong Agricultural University, Tai'an, China, in 2012, the M.S. degree Shandong Normal University, Jinan, China, in 2015, and the Ph.D. degree in signal and information processing from Dalian University of Technology, Dalian, China, in 2022.

He is currently a Lecturer with Shandong Technology and Business University, Yantai, China. His research interests include computer vision, hand pose estimation, and hand shape recovery.



Lu Ren received the Ph.D. degree in signal and information processing from Dalian University of Technology, Dalian, China, in 2021.

She is currently a Lecturer with Shandong Technology and Business University, Yantai, China. Her current research interests include computer vision, sentiment analysis, and text mining.