Multielement-Feature-Based Hierarchical Context Integration Network for Remote Sensing Image Segmentation

Yunsong Yang, Genji Yuan, and Jinjiang Li

Abstract—**In the current remote sensing segmentation tasks, we identify issues of insufficient accuracy in segmenting objects and types with similar colors, along with a lack of adequate smoothness and coherence in edge segmentation. To address these challenges, we propose a network framework called the multielement-feature-based hierarchical context integration network (MHCINet). This framework achieves deep integration of global information, local information, multiscale information, and edge information. First, we introduce an Edge and Levels Grouped Aggregator to fuse shallow features, deep features, and edge information, enhancing foreground saliency. Finally, to better identify instances with similar colors during the feature reconstruction stage, we design a constant multivariate feature integrator to fully exploit multiscale information and global context, thereby improving the segmentation model's performance. Comprehensive experimental results on the Vaihingen and Potsdam datasets demonstrate that MHCINet outperforms existing state-of-the-art methods, achieving mean intersection over union of 84.8% and 87.6% on the Vaihingen and Potsdam datasets, respectively.**

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

I. Introduction

As SENSOR and aerospace technologies continue to advance, high-resolution satellite and aerospace remote sensing images can be easily obtained. These images provide high-resolution observations of diverse landscapes on Earth, covering various scenes from urban areas to farmlands, forests to lakes. Remote sensing image segmentation is a crucial technology aimed at partitioning remote sensing images of the Earth into different objects or land cover categories. This is vital for geographic information systems, resource management, environmental monitoring, and crisis management. The following are key applications utilizing remote sensing image segmentation, such as land cover mapping [1], [2], change detection [3], [4], environmental protection [5], [6], road and building extraction [7], [8], and many other practical applications [9], [10].

In recent years, deep-learning-based remote sensing techniques for image processing have seen rapid development. In comparison to traditional machine learning algorithms such as Random Forest (RF) [11], conditional random field (CRF) [12], and support vector machine (SVM) [13], deep learning can automatically learn representations from raw data, alleviating the burden of feature engineering. Additionally, deep learning models with multiple layers of feature extraction can capture different levels of features, from low-level textures to high-level semantics, aiding in understanding the complex structure and semantics of the data. Various deep-learning-based methods have been proposed in the field of remote sensing to analyze different types of data [14], [15], including hyperspectral images (HSI), optical images, LiDAR data, and integrated multisensor data. Convolutional neural network (CNN) methods have proven effective in classifying and segmenting each pixel in a given image into semantic labels [16].

For semantic segmentation, the fully convolutional network (FCN) [17] is an architecture that transforms a CNN into a suitable structure for semantic segmentation. However, its design is relatively coarse. Subsequently, more refined encoder–decoder structures [18] have been proposed. U-Net [19] combines an encoder and a decoder, and its uniqueness lies in having skip connections that link the encoder and decoder parts. These connections help combine high-level semantic information with low-level feature information, thereby enhancing segmentation accuracy. This architecture allows the network to maintain detailed information about resolution. While U-Net has shown good performance on remote sensing images, subsequent researchers have made improvements to adapt U-Net for segmentation tasks, such as U-Net++ [20] and AFF-UNet [21].

While the aforementioned CNN networks are encouraging in terms of overall accuracy (OA), in remote sensing image segmentation, challenges arise due to factors such as lighting, shadows, and seasonal variations, which result in color similarity among classes and consequently fragmented segmentation results [22]. Addressing color similarity is crucial for improving accuracy, ensuring precise differentiation of similar-colored

Manuscript received 20 December 2023; revised 27 January 2024 and 23 February 2024; accepted 13 March 2024. Date of publication 18 March 2024; date of current version 12 April 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 ZR2023MF026 and Grant ZR2021MF068, and in part by Yantai Science and Technology Innovation Development Plan under Grant 2022JCYJ031 and Grant 2023JCYJ040. *(Corresponding author: Genji Yuan.)*

The authors are with the School of Computer Science and Technology, Shandong Technology and Business University, Yantai 264005, China (e-mail: yuangenji@outlook.com).

The code is available online at https://github.com/yysdck/MHCINet
Digital Object Identifier 10.1109/JSTARS.2024.3378301

objects, and reducing fragmentation to achieve more coherent results. Considering color similarity also aids in capturing semantic information, enhancing the overall semantic consistency of segmentation results.

The root cause of this problem lies in the fact that modeling based on local information typically leads to ambiguous segmentation results. Therefore, for remote sensing image semantic segmentation, more advanced and specialized methods are needed to fully utilize global context information. The transformer's self-attention mechanism [23] has the capability to capture global information in images, effectively creating long-range dependencies, but it often requires substantial computational time and memory. To address this issue, scholars have proposed more efficient alternative attention mechanisms, such as dual attention [24] specifically designed for segmentation and BAM [25], which enhances the ability to obtain global information.

Combining convolutional networks with attention mechanisms has become a preferred choice for current remote sensing image segmentation, such as Unetformer with GlobalLocalAttention [26]. In addition, multiscale analysis helps enhance the contextual understanding of color information in images. At larger scales, the overall color distribution of target objects can be captured while at smaller scales, local color variations can be observed more finely.

The following networks utilize multiscale information: Deeplabv3+ [27], which achieves good segmentation results on multiple datasets by integrating the ASPP module into the encoder–decoder structure; CTMFNet [28], which proposes a method to fuse global, local, and multiscale information; and MSCSA-Net [29], which designs local channel spatial attention and multiscale attention to effectively extract information-rich multiperspective features. Although the aforementioned methods utilize multiscale information, they do not address the color similarity issue present in remote sensing images.

To address the color similarity issue in remote sensing segmentation, we propose a method of integrating and analyzing global information at different scales of original features through multiscale information fusion attention. This approach enables the model to effectively pay attention to color distribution information based on various scales.

Furthermore, for remote sensing semantic segmentation, the segmentation of some objects still tends to have inaccurate edges. This is because the process of downsampling the original image and then reconstructing features may lead to the loss of boundary information, making segmentation tasks more challenging [30]. Some scholars have introduced edge information based on deep learning to address the problem of boundary information loss. Yuan et al. [31] proposed a method that integrates cloud segmentation and cloud edge detection, focusing on more accurately detecting cloud edges to achieve high-precision cloud detection. Cheng et al. [32] introduced an edge-aware convolutional network for the segmentation task of remote sensing harbor images. However, although these studies utilize edge information for segmentation, they focus only on the accuracy of edge feature extraction, neglecting the importance of the way in which edge information is integrated. Simple integration of edge information into the network may introduce information redundancy or confusion, and the inconsistent fusion of edge information with other features in the network can lead to feature representation inconsistency. These factors can have a negative impact on the final segmentation results. We adopt a group-based edge fusion method, which can more effectively integrate edge information into the segmentation method.

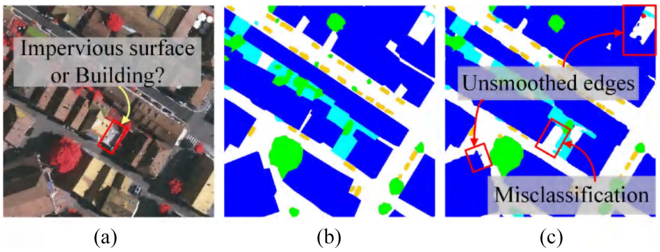


Fig. 1. Current challenges in remote sensing image segmentation are as follows. (a) Original image, where it is difficult to distinguish impervious surfaces and buildings due to color similarity. (b) Ground truth (GT). (c) Segmentation result of MPCNet, which does not utilize foreground saliency and is challenged by color similarity, resulting in edges that are not smooth enough.

Among the segmentation methods mentioned before, there are still some limitations, such as a lack of global information, insufficient depth of edge information fusion, and a lack of multiscale information. These issues restrict the general ability to address the challenges of remote sensing image segmentation tasks and consequently limit the improvement of accuracy in remote sensing segmentation images. Considering these challenges (as seen in Fig. 1), we propose a network called Multielement-Feature-based Hierarchical Context Integration Network (MHCINet), which not only performs better in terms of edges but also exhibits the ability to address challenges related to color similarity.

In summary, the contributions of our work are mainly reflected in the following aspects:

1) *Introduction of the Edge Level Group Aggregator (ELGA):* To overcome the problem of insufficiently deep feature fusion that traditional skip connections may cause, we introduce ELGA, which deeply fuses high-level features, low-level features, and edge features in the feature extraction stage through a grouping method. This helps efficiently utilize edge information in remote sensing semantic segmentation.
2) *Introduction of the constant multivariate feature integrator (CMFI):* In remote sensing images, the segmentation of objects or types with similar colors may face challenges due to color similarity. To address this issue, we introduce CMFI, which integrates multiscale information and then performs spatial and channel analysis. This enhances the model's ability to capture color transformation information inherent in different scale features, thereby improving the segmentation accuracy of objects with similar colors.
3) *Construction of MHCINet and multiloss joint-constrained training:* We constructed a two-stage semantic segmentation model, MHCINet, which improves the final semantic segmentation results by effectively guiding with

edge features. Additionally, we designed a multiloss joint-constrained training, including multidimensional boundary loss and segmentation loss. Experimental results on the Vaihingen Dataset and the Potsdam Dataset demonstrate the significant improvement in semantic segmentation accuracy achieved by MHCINet.

II. Related Work

Remote sensing semantic segmentation is a highly specialized and complex field that involves handling various types of remote sensing image data, such as high-resolution satellite images, multispectral images, and synthetic aperture radar images. In this field, researchers and scientists continually strive to find better methods and technologies to accurately segment and classify different objects in images. When addressing the challenges in this field, researchers need to consider the characteristics of the data to ensure the accuracy and robustness of segmentation results. The success of remote sensing semantic segmentation is crucial for urban planning, agricultural monitoring, natural disaster management, and other areas. Consequently, relevant research has been evolving to meet the demands of these applications. This section will introduce some important work related to remote sensing image semantic segmentation and their contributions to the field.

A. CNN-Based Remote Sensing Image Semantic Segmentation

In the field of remote sensing image semantic segmentation, traditional methods often focus on designing more robust features that combine spectral information and local image textures [33], [34]. For example, Huang et al. [35] proposed considering environmental and spectral information to effectively represent objects such as buildings. With the continuous advancement of remote sensing technology, current research tends to rely on high-resolution datasets. Although these datasets have clear geometric information and fine textures [36], the relationships between foreground and background in these datasets are more complex, posing greater challenges for more accurate segmentation.

In recent years, deep learning methods have been widely used in the field of remote sensing segmentation. The FCN [17], first proposed by Long et al. in 2015, was the first CNN structure to effectively address semantic segmentation problems. Subsequently, methods based on CNNs have dominated the field of semantic segmentation in remote sensing, covering many research achievements [37], [38]. However, despite being a pioneer, FCNs decoder structure is too simple, resulting in lower resolution of segmentation results, thereby limiting the fidelity and accuracy of images.

To overcome this issue, researchers have proposed a encoder–decoder network called UNet [19], focusing on more fine-grained semantic segmentation tasks. The structure of UNet exhibits symmetry and consists of two key components: 1) the encoder and 2) the decoder. The encoder extracts multilevel features by progressively downsampling feature maps' spatial resolutions. The decoder is used for feature reconstruction, gradually restoring the spatial resolutions of feature maps. These two components are interconnected through skip connections.

This encoder–decoder network structure has become the standard model for remote sensing image segmentation, laying the foundation for subsequent research [27]. Subsequent research based on this structure includes Unet++ [20], which incorporates dense convolutional blocks to bridge semantic gaps. MAResUnet-a [39] combines residual concatenation, subattribute convolution, pyramid-style scene understanding, and multitask inference to establish conditional relationships between tasks and improve segmentation accuracy. Ma et al. [40] proposed Factseg, a foreground-activated small target semantic segmentation network.

However, these methods essentially overlook the importance of other foreground saliency. Our approach enhances foreground saliency by integrating edge information from downsampling features of different sizes.

B. Semantic Segmentation Based on Edge Constraints

To address the issue of insufficiently coherent and smooth edges in image segmentation results, scholars often employ edge constraints in segmentation. The commonly used approach is to apply edge constraints to segmentation, with the initial consideration being postprocessing for classification, such as using CRF [12] for edge optimization of classification results. Later, with the rapid development of deep learning, attention turned to integrating edge information with deep learning models to improve the accuracy of the segmented edges. For example, Michieli and Zanuttigh [41] proposed an edge-aware graph matching network, GMENet, for segmentation. Chen et al. [42] introduced an edge-aware convolutional kernel that effectively utilizes geometric information embedded in deep channels to enhance the quality of feature mapping for RGB-D images, significantly improving semantic segmentation accuracy. Kuang et al. [43] presented a novel body and edge-aware network to enhance the accuracy of medical image segmentation.

In the field of remote sensing, edge constraints are also frequently applied in semantic segmentation. For instance, Jung et al. [22] proposed an approach applicable to various semantic segmentation networks, including encoder–decoder structures. This method combines holistic nested edge detection with a boundary enhancement module. Zheng et al. [44] introduced an optimization algorithm based on Markov random fields (MRFs) for multiscale edge-preserving in remote sensing segmentation. Sui et al. [45] presented a segmentation network structure with blockwise edge detection. Unfortunately, the aforementioned methods mostly focus on the extraction of finer edges and overlook the importance of the fusion of edge information in the segmentation model for effective edge utilization.

This article adopts a joint training approach based on the combination of edge and segmentation networks, with a specific focus on the constraint methods of integrating edge information with the segmentation network. This approach aims to enhance segmentation accuracy.

C. Multiscale Features

In the case of remote sensing images containing instances of various scales and details, challenges such as recognizing too small objects, obscured targets, and increased noise in high-resolution images arise. To address these issues, researchers tend to use multiscale features. Zhang et al. [46] drew inspiration from HRNet's multibranch parallel convolution structure and creatively generated multiscale feature maps. They introduced an adaptive spatial pooling module to better aggregate local context information. DeeplabV3 [47] optimized image feature information using the atrous spatial pyramid pooling (ASPP) algorithm. ASPP employs atrous convolutions with four different rates to extract feature maps. AFFPN [48] constructs multiscale feature maps on high-level feature maps using atrous convolution and adaptive global context. A2-FPN [49] encodes semantic features at multiple scales using a feature pyramid and enhances multiscale feature learning with an attention aggregation module. Tian et al. [50] designed a multiscale background information aggregation network to automatically extract highland lake areas. The multikernel pyramid pooling module in this network aggregates background information from different lakes globally. MSLANet [51] is a multiscale position network with a dual-branch multiscale aggregation unit that achieves multiscale feature aggregation without increasing computational parameters.

While the mentioned algorithms enhance the model's robustness to changes in object or region scale, they primarily focus on improving the adaptability of the model to scale changes. They overlook the potential of multiscale features in addressing color similarity issues in remote sensing images. To leverage this characteristic, this article proposes a method to extract multiscale features from various features during the feature reconstruction stage. This approach aims to capture color transformation information at different scales, providing more robust features for subsequent analysis.

D. Global Context Modeling

Global context modeling holds a crucial position in the field of computer vision, aiming to deeply understand and effectively utilize the global information within images or videos to enhance the performance of various image processing tasks. Global context information encompasses the overall background, context, and intricate relationships between objects in an image.

To capture the long-range dependencies between features in images, the introduction of attention mechanisms has become one of the popular methods. This mechanism allows the network to selectively focus on specific areas of the image, thereby better capturing the global contextual information. This mechanism includes self-attention mechanisms in the transformer [23], spatial attention models [25], etc. By introducing attention mechanisms, researchers can more effectively model the correlations between different regions, significantly improving the performance of image segmentation tasks.

Inspired by the remarkable capabilities of the transformer in sequence-to-sequence modeling, many researchers in the field of remote sensing image segmentation have begun to introduce transformer into remote sensing image processing. Some models using pure transformer structures, such as Segmenter [52] and SwinUNet [53], as well as models that combine transformer and CNN, such as GLOTS [54] and EMRT [55], have achieved significant success in this field. However, many researchers still prefer convolution-based attention mechanisms with stronger generalization abilities, better handling of information loss or noise, and better balancing of the importance of local and global information.

For example, Chen et al. [56] introduced local aggregation with graph convolution and global attention blocks to more fully capture contextual information. Additionally, Li et al. [39] proposed a linear attention mechanism that reduces computational complexity while maintaining performance. Nevertheless, a single attention module struggles to fully capture the global information of multilayer semantic features. Therefore, this article adopts the approach of combining mixed CNN attention with standard transformer blocks to more comprehensively capture the global contextual information of multilayer semantic features.

III. Method

In this section, we will first introduce the structure of MHCINet. Following that, we will present two crucial modules of MHCINet, namely ELGA and CMFI. Finally, we will describe the loss functions employed in our study.

A. MHCINet Structure

The overall structure of MHCINet is shown in Fig. 2. The input image undergoes downsampling using a CNN-based method, with Convnext serving as the backbone for four downsampling stages, resulting in four distinct feature maps: X_1, X_2, X_3, and X_4. Specifically, X_1, X_2, and X_3 pass through a traditional edge extraction network, as illustrated in Fig. 3. The traditional edge extraction network produces edge features e_1, e_2, and e_3. The edges predicted by the Canny algorithm constrain the loss on the predicted edges. The input to the segmentation network comprises $X_1, X_2, X_3, X_4, e_1, e_2$, and e_3. The features are fused through ELGA in a grouped manner, following these steps:

$$X_3' = \text{ELGA}(X_4, X_3, e_3) \tag{1}$$

$$X_2' = \text{ELGA}(X_3', X_2, e_2) \tag{2}$$

$$X_1' = \text{ELGA}(X_2', X_1, e_1). \tag{3}$$

Here, $\text{ELGA}(X, Y, Z)$ denotes the fusion of high-level feature X, low-level feature Y, and edge feature Z, combining low-level texture information, high-level semantic information, and edge information.

During the feature reconstruction phase, to address color similarity issues more effectively, we introduced a method named CMFI. The goal of CMFI is to model more effective multiscale information in the global context to tackle color similarity problems. We employ CMFI in all subsequent feature reconstruction stages, taking X_4 as an example, where it undergoes CMFI

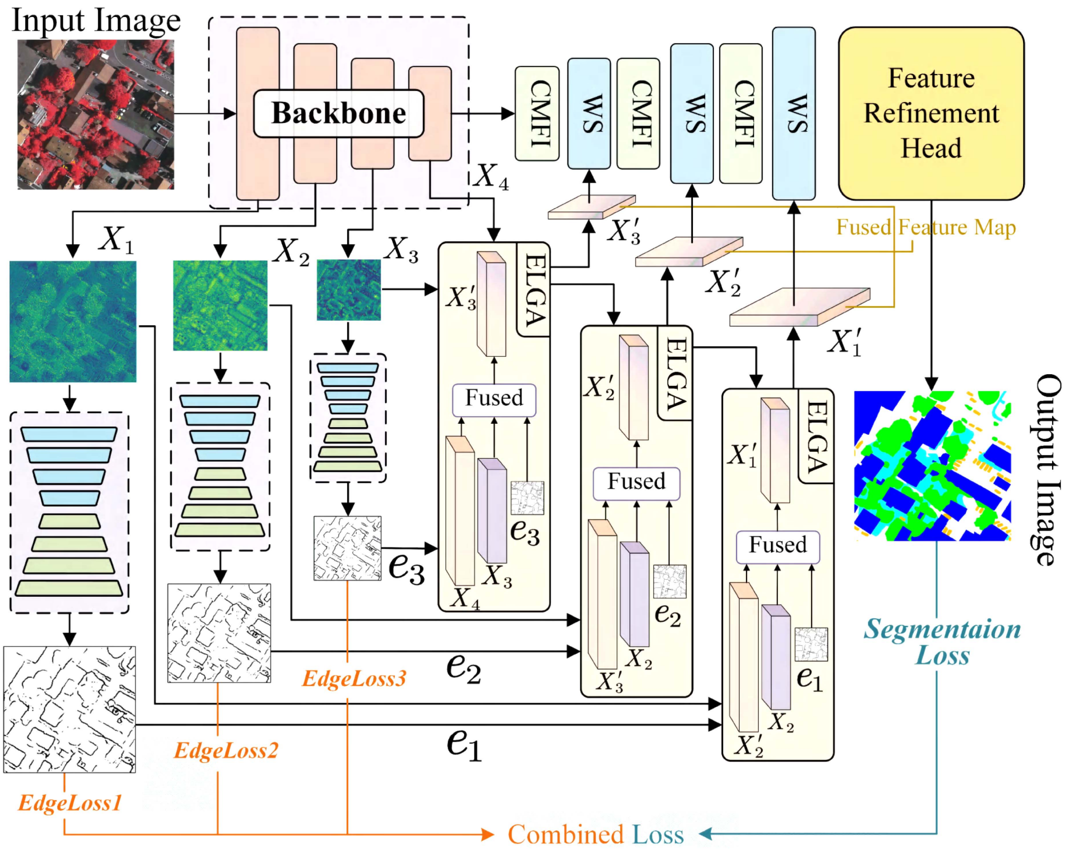


Fig. 2. Overall structure diagram of MHCINet.

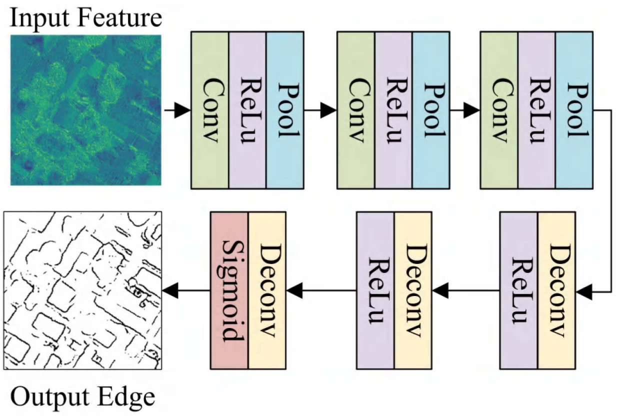


Fig. 3. Structure of the traditional edge feature extraction network consists of three standard convolutional blocks and three layers of deconvolutional blocks.

processing. Through this approach, we obtain X_4' enriched with global information, which is then fused with the mixed features X_3' obtained through ELGA. Finally, a weighted fusion strategy is applied to balance global and mixed information, as specified by the following formula:

$$Y = \mathrm{ReLu}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(a \cdot \mathrm{CMFIF} + (1-a) \cdot \mathrm{ELGAF}))). \tag{4}$$

Here, CMFIF represents the global fusion feature obtained through CMFI, ELGAF represents the mixed feature obtained through ELGA, and $\mathrm{Conv}_{1\times1}$, BN, and ReLu denote the pointwise convolution, batch normalization, and ReLu processes, respectively. The parameter a is a learnable weight. After completing three reconstruction stages, the final segmentation result is obtained through the feature refinement head [26].

Edge information is typically more pronounced in low-level features as it reflects the local structure and details of the image, primarily caused by changes in brightness or color. In contrast, as features progress through multiple layers of convolutional networks, high-level features extract more abstract information. During this process, the original detailed information, such as edges, gradually becomes abstracted and integrated into higher-level feature representations. Consequently, extracting edge information from high-level features becomes relatively challenging and less critical. Considering this, MHCINet utilizes edge extraction only in the first three relatively low-level features.

B. *Edge Level Group Aggregator*

The structure of ELGA can be referenced in Fig. 4. The deep features undergo convolution followed by bilinear interpolation to obtain features consistent in size with shallow features. Edge features are combined with both shallow and deep features and fused through grouped aggregation to obtain the final mixed

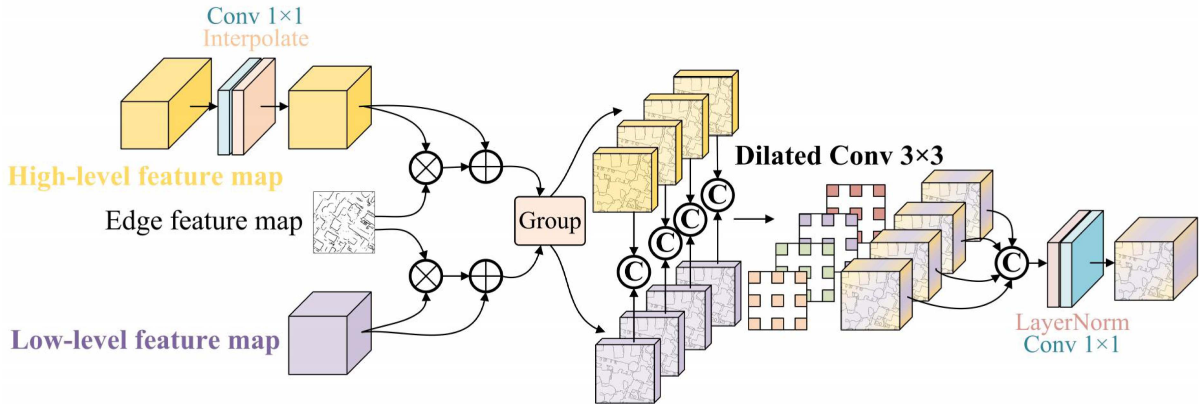


Fig. 4. ELGA structure diagram.

features. Specifically, the expressions for enhancing edge information in deep and shallow features are as follows:

$$\mathrm{LE} = \mathrm{EFM} \cdot \mathrm{LFM} + \mathrm{LFM} \tag{5}$$

$$\mathrm{HE} = \mathrm{EFM} \cdot \mathrm{HFM} + \mathrm{HFM} \tag{6}$$

where EFM represents the input edge features, LFM represents shallow features, and HFM represents deep features after resizing, LE represents the shallow features enhanced by edge information, and HE represents the deep features enhanced by edge information. Subsequently, LE and HE are grouped along the channel dimension into four chunks each

$$\begin{cases} X_h^1, X_h^2, X_h^3, X_h^4 = \mathrm{Group}\,(\mathrm{HE}) \\ X_l^1, X_l^2, X_l^3, X_l^4 = \mathrm{Group}\,(\mathrm{LE}) \end{cases} \tag{7}$$

$$Y_i = \mathrm{Concat}(X_l^i, X_h^i) \tag{8}$$

where Group denotes the grouping operation, X_h^i represents the grouped high-level features for the ith group, $i \in [1, 2, 3, 4]$, X_l^i represents the ith group of grouped shallow features, and Y_i represents the aggregated features of the ith shallow feature and the ith deep feature. Subsequently, dilated convolutions with kernel size 3×3 and dilation rates $(1, 2, 5, 7)$ are applied to obtain information with different receptive fields. Finally, the four groups are concatenated along the channel dimension, followed by a regular convolution with a kernel size of 1, allowing better interaction between different groups of features. This facilitates enhanced integration of deep, shallow, and edge features, enabling mutual supplementation and reinforcement to achieve interaction between features of different scales. The expression is as follows:

$$\begin{aligned} F_{\text{out}} = {} & \mathrm{Conv}_{1\times1}(\mathrm{Concat}(\mathrm{Conv}_{3\times3}^{dr=1}(Y_1), \\ & \mathrm{Conv}_{3\times3}^{dr=2}(Y_2), \mathrm{Conv}_{3\times3}^{dr=5}(Y_3), \\ & \mathrm{Conv}_{3\times3}^{dr=7}(Y_4))) \end{aligned} \tag{9}$$

where $\mathrm{Conv}_{k\times k}^{dr=j}$ denotes convolution with a kernel size of k and dilation rate j.Concat represents the concatenation operation. F_{out} represents the final output features.

The deep features of remote sensing images contain semantic information about objects such as buildings, roads, and vegetation, which is crucial for land cover classification and target detection. Shallow features, on the other hand, contain detailed information such as texture, edges, and colors, which help improve the accuracy of image analysis. Additionally, edge information typically includes clear boundaries and contours between objects or foreground objects and the background. Enhancing foreground saliency enables more precise segmentation of foreground and background, particularly aiding in extracting target objects from complex backgrounds. We enhance foreground saliency by integrating semantic, detail, and edge information using a grouping fusion approach. The ELGA grouping fusion structure provides superior feature inputs for subsequent feature reconstruction compared to simple feature addition or concatenation methods, thereby enhancing the performance of the segmentation model.

C. *Constant Multivariate Feature Integrator*

For the feature reconstruction stage of MHCINet, CMFI was employed in this study. Its core idea is to combine global information and enhance multiscale information. By emphasizing key regions of synthesized information at different scales, the segmentation performance of the decoder in addressing challenges related to image color similarity is improved. The structure of CMFI is illustrated in Fig. 5, where we integrate the standard transformer encoder structure [23].

In CMFI, for input features, a batch normalization operation is first applied. Then, two branches are used to process the features. One branch employs a 1×1 convolution to transform the features' dimensions. The other branch initially utilizes three Conv layers for key feature extraction. Subsequently, spatial pyramid pooling is performed with different scales ($5 \times 5, 9 \times 9, 13 \times 13$) to process the multiscale features with two convolution operations to match the original features. Finally, the features from this branch are concatenated with those from the other branch to obtain a multiscale synthesized feature.

Following this, the feature is passed through the bottleneck attention module (BAM) [25] after the multiscale synthesized feature. BAM can focus attention on key regions in the features, enhancing the model's ability to distinguish color-similar objects by emphasizing significant color differences. Unlike the generic

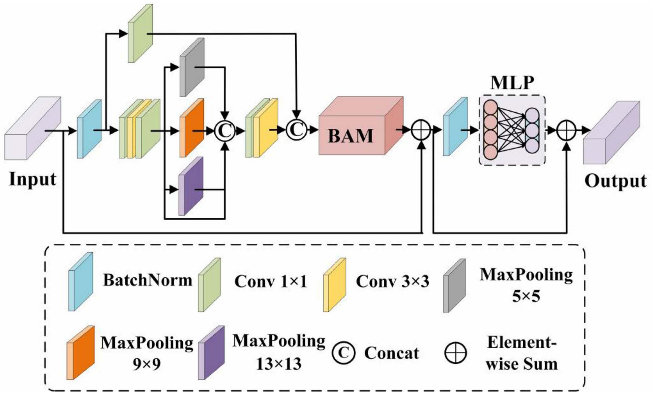


Fig. 5. CMFI structure diagram. In CMFI, input features are standardized using BatchNorm, followed by processing through two branches: One for dimension transformation and the other for extracting key features followed by spatial pyramid pooling to obtain multiscale integrated features. BAM is introduced to focus on key regions, enhancing the ability to distinguish color-similar objects. Finally, through residual connections and MLP, complex patterns are captured to obtain the final feature output.

global self-attention module, this method can be adjusted based on the nature of the input data and the type of task.

Subsequently, the original features are residual-connected and batch-normalized again, before passing through an MLP and being residual-connected again to obtain the final feature output. The final MLP aims to capture complex patterns in the feature maps, aiding in more accurate segmentation.

In remote sensing images, due to factors such as lighting and artificial structures, two different types of objects may appear with similar colors, making precise segmentation difficult. To address this issue, we introduce multiscale features, allowing the model to observe the shape and texture features of objects at different scales, providing more contextual information to better distinguish objects with similar colors but different scales. By analyzing images at multiple scales, the model can capture the diversity of objects and relationships between different scales, thereby achieving more accurate segmentation of color-similar objects. When facing the challenge of color similarity, attention mechanisms become crucial. These mechanisms enable the model to focus attention on key areas, helping to capture regions where color differences are more significant. By introducing attention mechanisms, the model can better identify objects with similar colors but different semantics, thereby improving segmentation accuracy. This strategy of combining multiscale features and attention mechanisms helps overcome the challenge of color similarity and enhances the performance of segmentation models in complex scenarios.

D. Loss Function

The schematic diagram of the joint loss structure is shown in Fig. 6. The entire network needs to effectively detect the boundary mask and segmentation mask to obtain a well-regularized segmentation mask. For this purpose, we design two types of loss functions for each output: One is the edge loss (denoted as L_e) for boundary mask detection, and the other is the segmentation loss (denoted as L_s) for segmentation mask detection. Joint training with these two losses is performed, and the joint loss (L_{com}) is expressed as

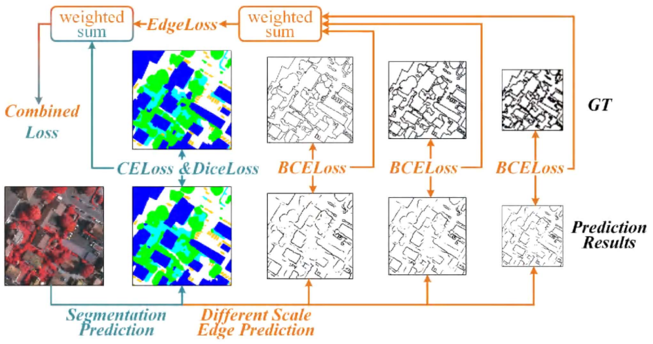


Fig. 6. Loss function schematic diagram. In the diagram, the blue lines represent processes related to segmentation loss while the orange lines represent processes related to edge loss. The three edge prediction results in the figure represent the edge predictions for the first, second, and third layers of features extracted from the backbone of the main network.

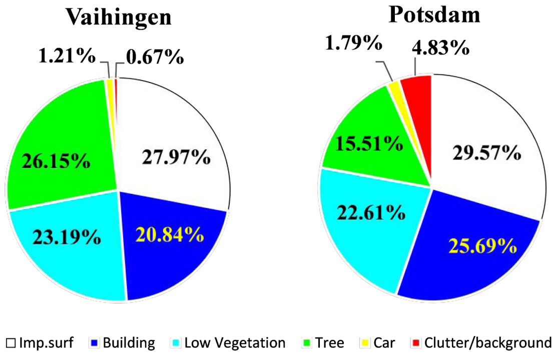


Fig. 7. This figure illustrates the proportion of each semantic label in the two datasets.

$$L_{\text{com}} = k_1 L_e + k_2 L_s. \tag{10}$$

Here, k_1 and k_2 represent the parameters for L_e and L_s, respectively. The behavior and learning emphasis of the model can be controlled through weighted operations. Balancing between the tasks can be achieved by assigning appropriate weights to each loss function, ensuring a balanced model across all losses. As the primary task in this article is segmentation and edge prediction serves as an auxiliary task to enhance segmentation results, we set k_1 to 0.7 and k_2 to 0.3 based on this consideration.

Edge Loss: In the proposed method, there are three different scales of edges. Therefore, for edge loss, the method treats edge prediction as a binary prediction. The joint loss for edge loss is defined using the binary cross-entropy loss (BCELoss) for the three scales, denoted as L_e^n, where $n \in [1, 2, ..N]$ and here $N = 3$. The overall edge loss is denoted as L_e and calculated as

$$L_e^n = -\sum_{i=1}^{H}\sum_{j=1}^{W}[G^n(i,j)\log P^n(i,j) + (1 - G^n(i,j))\log(1 - P^n(i,j))] \tag{11}$$

$$L_e = \frac{1}{N}\sum_{n=1}^{N} L_e^n \tag{12}$$

where $P^n(i,j)$ represents the predicted value at the ith row and jth column of the edge feature map for the nth scale, and $G^n(i,j)$ represents the true value at the ith row and jth column of the edge feature map for the nth scale.

Segmentation Loss: Cross-entropy loss is commonly used for image segmentation tasks, encouraging the model to make segmentation results closer to the true labels. However, in regions with similar colors, cross-entropy loss may make it challenging for the model to distinguish object boundaries, leading to poor segmentation results in areas with high color similarity. To address the challenge of color similarity, dice loss may be better suited since it focuses on contour matching rather than color. In our proposed method, the segmentation loss (L_s) for the segmentation model is a combination of dice loss (L_{dice}) and cross-entropy loss (L_{ce}), expressed as

$$L_{\text{ce}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \log \hat{y}_k^{(n)} \tag{13}$$

$$L_{\text{dice}} = 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\hat{y}_k^{(n)} y_k^{(n)}}{\hat{y}_k^{(n)} + y_k^{(n)}} \tag{14}$$

$$L_s = L_{\text{ce}} + L_{\text{dice}}. \tag{15}$$

Here, N is the number of samples, K is the number of classes, $y^{(n)}$ and $\hat{y}^{(n)}$ represent the one-hot encoding of the true semantic label and its corresponding network Softmax output for sample n ($n \in [1,\ldots,N]$). Additionally, $\hat{y}_k^{(n)}$ represents 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: The ISPRS Potsdam and ISPRS Vaihingen datasets consist of high-resolution remote sensing images captured in both urban and rural environments (as seen in Fig. 7). This diversity reflects real-world scenarios, making these datasets suitable for evaluating model performance across various backgrounds.

Vaihingen: This dataset is primarily based on remote sensing images from the Vaihingen region in Germany. Vaihingen is characterized by numerous independent buildings and small multistory structures. The dataset comprises 33 different-sized high spatial resolution true orthophoto (TOP) image blocks, with an average size of 2494 × 2064 pixels each. Each image block consists of TOP and a digital surface model (DSM) extracted from a larger TOP mosaic, along with a normalized DSM (NDSM). Each TOP image block includes three multispectral bands (near-infrared, red, and green). The dataset encompasses five foreground land cover classes (impervious surfaces, buildings, low vegetation, trees, and cars) and one background land cover class (clutter). In our experiments, only TOP image blocks were utilized, excluding DSM and NDSM. The training was conducted on images with IDs 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, and 37 while the remaining 17 images were reserved for testing. Image blocks were cropped into 1024 × 1024 pixel patches for processing and analysis.

Potsdam: This dataset utilizes aerial images from Potsdam, Germany, containing 38 high spatial resolution TOP image blocks with a ground sampling distance of 5 cm, and each image block has a size of 6000 × 6000 pixels. Similar to Vaihingen, each image block is composed of true TOP and DSM extracted from a larger TOP, along with an NDSM. The land cover classes in this dataset are identical to Vaihingen, including impervious surfaces, buildings, low vegetation, trees, cars as foreground land cover classes, and clutter as the background land cover class. The dataset provides four multispectral bands (red, green, blue, and near-infrared), along with DSM and NDSM. For training, images with IDs 2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_8, 7_9,7_11, and 7_12 were selected (excluding image 7_10 with annotation errors) for training while the remaining 15 images were used for testing. Similar to Vaihingen, only three spectral bands (red, green, and blue) were used, and the original image blocks were cropped into 1024 × 1024 pixel patches for analysis. In the quantitative evaluation of these two datasets, the "clutter/background" category was disregarded.

2) Implementation Details: In this experiment, Ubuntu 18.04 operating system was chosen, and all models were deployed on a single NVIDIA GeForce RTX 2080 Ti 11-GB GPU using the PyTorch 1.11 framework. To achieve faster model convergence, the experiment employed the AdamW optimizer with a base learning rate set to 6e-4 and utilized a cosine learning rate schedule for adjustment. For the Vaihingen and Potsdam datasets, the following data preprocessing steps were applied. First, images were randomly cropped into patches of size 512 × 512. During the training phase, multiple data augmentation techniques were introduced, including random scaling ([0.5, 0.75, 1.0, 1.25, 1.5]), random vertical and horizontal flipping, and random rotation. The training process comprised 105 epochs. In the testing phase, multiscale evaluation and random flipping augmentation techniques were adopted.

3) Evaluation Metrics: Commonly used remote sensing segmentation metrics, including OA, F1 score, and Mean Intersection over Union (mIoU), were used as evaluation metrics in this experiment. In addition, the experiment employed the number of parameters as an evaluation metric. Before introducing these metrics, other related metrics such as precision and recall will be discussed, along with the meanings of some symbols: tp (true positive), fp (false positive), fn (false negative), and tn (true negative).

Precision: Precision measures the proportion of true positives among the samples predicted as positive by the model. In other words, precision informs us of the probability that a sample predicted as positive is truly positive

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}}. \tag{16}$$

Recall: Recall is the proportion of true positives among all samples truly positive. Recall measures the model's ability to discover all positives.

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}. \tag{17}$$

Overall Accuracy: OA is a commonly used performance evaluation metric in image classification tasks. It is the ratio of correctly classified samples to the total number of samples. However, OA may not handle class imbalance well, as the model may bias toward predicting the class with more samples

$$\text{OA} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fn} + \text{tn}}. \tag{18}$$

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

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}. \tag{19}$$

Overall F1 score: Average of F1 scores for all classes.

Mean Intersection over Union: mIoU is a commonly used evaluation metric in semantic segmentation tasks, measuring the model's accuracy in pixel-level segmentation. Intersection over Union (IoU) is used to assess the model's segmentation results for each class, and mIoU calculates the average of IoUs for all classes.

$$\text{IoU} = \frac{\text{tp}}{\text{tp} + \text{fp} + \text{fn}}. \tag{20}$$

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

B. Ablation Experiment

1) Components of MHCINet: To assess the performance of each component of MHCINet, a series of ablation experiments were conducted on the Vaihingen and Potsdam datasets. For ease of discussion, the main focus is on mIoU and meanF1. The metrics presented in this experiment are the averaged results of multiple experiments. Table I provides the results of removing individual modules from MHCINet and training with different losses. Here, Diceloss represents the dice loss in the segmentation loss, Celoss represents the cross-entropy loss in the segmentation loss, and Seloss represents the combined segmentation loss composed of dice loss and cross-entropy loss. The removed modules are indicated by (-). Fig. 8 illustrates the segmentation results after removing a single module from MHCINet.

We conducted additional experiments on the Vaihingen dataset by adding a single module to the baseline. The experimental setup involved using U-Net and ConvNext as the backbone networks, augmented with FRH as our baseline. The

TABLE I
RESULTS OF MHCINET AFTER REMOVING MODULES

Method	Vaihingen		Potsdam	
	mIoU (%)	F1 (%)	mIoU (%)	F1 (%)
MHCINet	84.80	91.67	87.64	93.31
MHCINet-ELGA	83.79	91.06	86.29	92.51
MHCINet-CMFI	83.89	91.12	86.42	92.60
MHCINet-Diceloss	84.48	91.48	87.02	92.94
MHCINet-Celoss	84.47	91.47	87.07	93.00

The performance will deteriorate after removing any module.

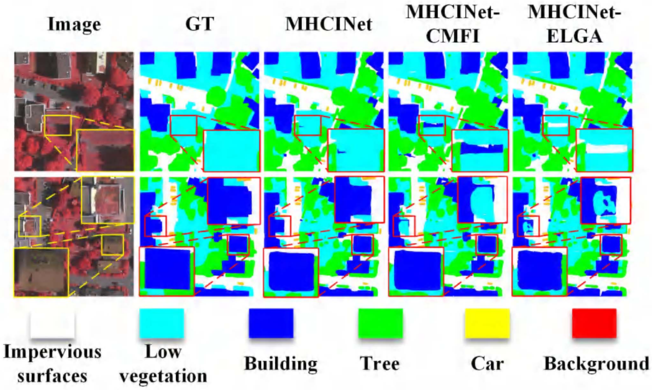


Fig. 8. Effect diagrams of MHCINet with individual modules removed. No matter which structure is removed from the network, the final segmentation performance will degrade.

TABLE II
RESULTS AFTER ADDING ANY MODULE TO THE BASELINE ON VAIHINGEN DATASET

Method	mIoU (%)	F1 (%)
Baseline	82.07	90.01
Baseline+ELGA	83.63	90.96
Baseline+CMFI	83.49	90.88
Baseline+Seloss	82.36	90.18

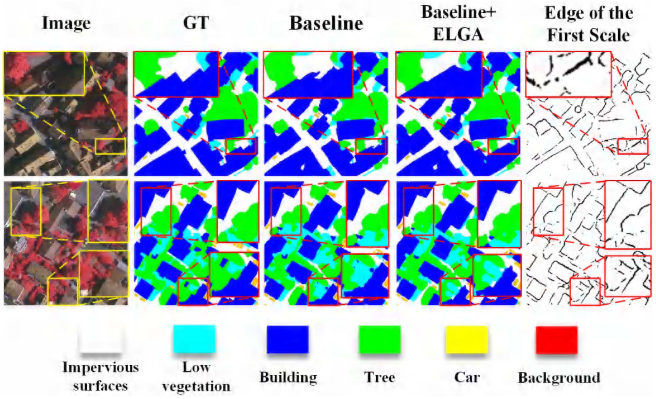


Fig. 9. Illustrates the effect of incorporating ELGA into the Baseline. The addition of ELGA results in smoother and more coherent edges.

TABLE III
COMPARISON OF RESULTS BETWEEN ELGA METHOD AND OTHER EDGE FUSION METHODS

Method	Vaihingen		Potsdam	
	mIoU (%)	F1 (%)	mIoU (%)	F1 (%)
MHCINet	84.80	91.67	87.64	93.31
MHCINet-ELGA	83.79	91.06	86.29	92.51
MHCINet-ELGA+ra	83.43	90.84	85.69	92.16
MHCINet-ELGA+cat	83.50	90.87	85.72	90.74

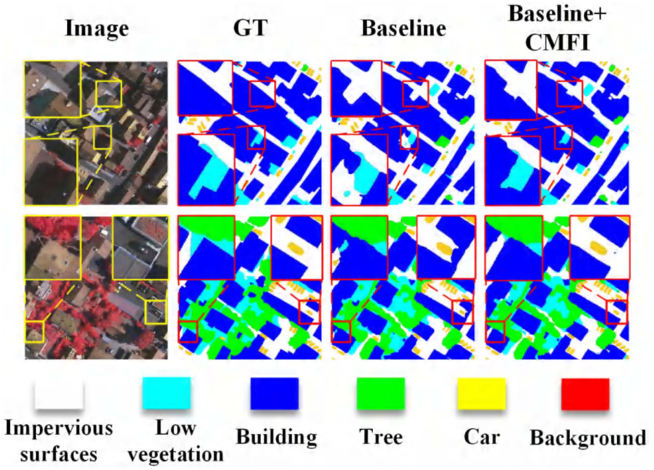


Fig. 10. Incorporating the effect of CMFI into the baseline: Results and analysis.

TABLE IV
COMPARISON OF RESULTS BETWEEN ELGA METHOD AND OTHER EDGE FUSION METHODS

Method	Lowveg		Tree	
	mIoU (%)	F1 (%)	mIoU (%)	F1 (%)
Baseline	72.52	84.07	82.80	90.59
Baseline+CMFI	75.07	85.76	83.55	91.04

training of the baseline utilized cross-entropy loss (celoss) as the loss function. The modules added to the baseline are indicated with a plus sign (+). Table II presents the results of adding a single module to the baseline and training with Seloss as the loss function.

2) *Effect of ELGA:* From Table I, it can be observed that for MHCINet, removing the ELGA module results in a decrease in mIoU and F1 scores by 1.01% and 0.61%, respectively, on the Vaihingen dataset. Similarly, on the Potsdam dataset, the gaps in mIoU and F1 scores are 1.35% and 0.8%, respectively, when compared to the configuration with ELGA. Table II provides a more intuitive view, indicating that the utilization of ELGA in the baseline enhances mIoU and F1 scores by 1.56% and 0.95%, respectively, on the Vaihingen dataset. It can be concluded that the use of ELGA brings at least a 1.01% improvement in mIoU and a 0.61% improvement in F1 to the model.

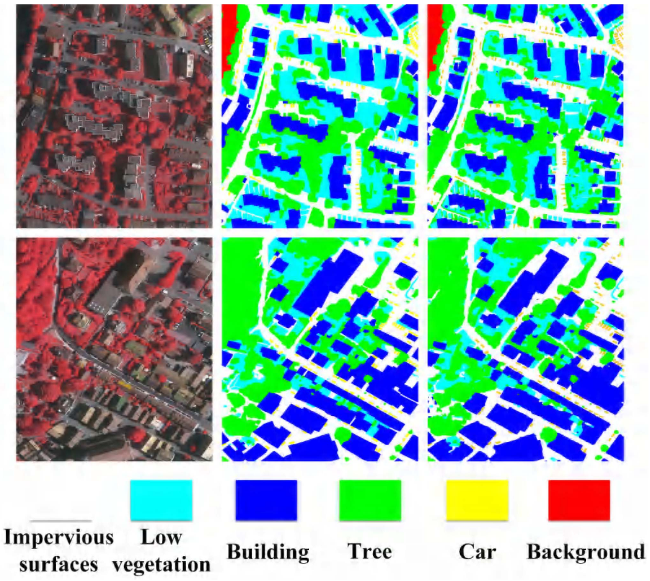


Fig. 11. Segmentation results of MHCINet on the Vaihingen dataset for IDs 2 and 6 are shown in the images.

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

Method	Backbone	Parameters (M)	mIoU (%)
MHCINet	ResNet50	25.56	84.44
	ResNext50	25.03	84.48
	ResNest50	27.48	84.54
	ConvNext-Tiny	28.59	84.80

From the second row of Fig. 8, it can be observed that, without ELGA, MHCINet lacks smoothness and continuity in the segmentation of building edges. This situation is more apparent in Fig. 9, where the addition of ELGA to the baseline leads to more accurate edge detection. The first row of Fig. 9 illustrates that incorporating ELGA enhances the coherence and smoothness of the image edges.

To further demonstrate the advantages of ELGA, we conducted experiments using two common feature fusion methods, regular addition (ra) and concat (cat), as alternatives to ELGA. Specifically, the features of the three edges were added or concatenated with the corresponding downsampling features at the same layer, and the fusion results were directly outputted. As shown in Table III, when using regular addition (83.43% mIoU and 90.84% F1 on the Vaihingen dataset) and concatenation (83.50% mIoU and 90.84% F1 on the Vaihingen dataset), the differences in performance were marginal, and both were lower than using ELGA (84.80% mIoU and 91.67% F1 on the Vaihingen dataset). It is worth noting that MHCINet without ELGA but with ra and cat outperforms Baseline+CMFI because MHCINet-ELGA+ra includes Baseline+CMFI, utilizes seloss training, and incorporates edge features. However, the results using ra and cat in MHCINet are comparable to those of Baseline+CMFI (83.49% mIoU and 90.88% F1 on the Vaihingen dataset). This phenomenon suggests that inappropriate edge

TABLE VI
COMPARISON OF SEGMENTATION RESULTS ON THE VAIHINGEN DATASET

Method	F1 (%)					Evaluation index		
	Imp.Surf	Building	Lowveg	Tree	Car	MeanF1 (%)	mIoU (%)	OA (%)
MANet[58]	84.95	88.41	78.16	88.37	70.47	82.07	70.16	84.80
ABCNet[59]	88.13	90.23	76.71	87.21	68.72	82.20	70.58	85.62
MACU-Net[60]	90.70	92.40	81.90	89.37	82.53	87.38	77.85	88.61
MAResU-Net[40]	92.91	95.26	84.95	89.94	88.33	90.28	83.30	90.86
A2-FPN[50]	92.99	95.53	84.67	90.34	87.62	90.23	82.42	91.04
UnetFormer[26]	92.72	95.72	84.36	89.93	88.55	90.26	82.47	90.76
DC-Swin[61]	93.46	96.00	85.32	90.03	84.88	89.94	82.01	91.29
MPCNet[62]	92.76	95.50	84.70	90.40	90.44	90.76	83.27	90.93
MHCINet(Ours)	**93.70**	**96.21**	**86.09**	**91.33**	**91.04**	**91.67**	**84.80**	**91.95**

The best results are indicated in bold.

information fusion methods lead to redundancy or confusion in network information and inconsistent fusion of edge information with other features in the network, resulting in inconsistent feature representations. This not only emphasizes the importance of an appropriate edge feature fusion method but also validates the effectiveness of ELGA.

3) Effect of CMFI: CMFI, proposed to address the issue of color similarity, deeply integrates multiscale and synthesizes global and local information. From experimental data, on the Vaihingen dataset, MHCINet experienced a decrease of 0.91% in mIoU and 0.55% in F1 scores after removing CMFI. On the Potsdam dataset, the removal of CMFI resulted in a decrease of 1.22% in mIoU and 0.71% in F1 scores. Intuitively, the addition of CMFI to the baseline is more evident, as seen in Table II, where the mIoU and F1 scores improved by 1.42% and 0.87%, respectively, after incorporating CMFI. The use of CMFI leads to at least a 0.91% increase in mIoU and at least a 0.55% increase in F1.

Considering the Vaihingen and Potsdam datasets, the classes “lowveg” and “tree” face significant color similarity issues in the original images due to lighting, seasons, etc. Adding CMFI to the model proves effective in addressing color similarity issues, particularly for “lowveg” and “tree” classes. Experimental results in Table IV show that, after adding CMFI to the baseline, the IoU and F1 for “lowveg” improved by 2.55% and 1.69%, respectively, and for “tree,” the IoU and F1 scores increased by 0.73% and 0.45%. The effectiveness of CMFI in resolving color similarity issues is evident. In Fig. 8, the sensitivity of MHCINet to global information diminishes after removing CMFI, resulting in misclassification of buildings as “lowveg” due to the red color. Fig. 10 illustrates the segmentation results with and without CMFI added to the baseline, emphasizing the impact of CMFI on improving accuracy in classification, particularly when global information is considered.

4) Effect of Loss Function: To demonstrate the effectiveness of the joint loss, experiments were conducted using only individual losses. The experimental results, as shown in Table I, indicate that when employing only celoss and diceloss training strategies, the results are consistently lower than those achieved with the joint loss. Additionally, in the baseline, we employed a joint training strategy with celoss and diceloss. According to the experimental results in Table II, using the joint loss (seloss) improves mIoU and F1 by 0.29% and 0.17%, respectively, compared to the single celoss.

5) Effect of Backbone: To eliminate the influence of the backbone, we conducted experiments involving the replacement of the backbone. The selected backbones include ConvNext-Tiny, used in our network, as well as commonly used backbones such as ResNet50, ResNext50, and ResNest50. The results are presented in Table V, revealing that ConvNext-Tiny has a larger parameter count, yet MHCINet achieves the best performance when using ConvNext-Tiny as the backbone. Therefore, we recommend using ConvNext-Tiny as the backbone.

C. Comparative Experiments

The selected models for comparison in this experimental study are as follows: the multiattention network (MANet) [57] with kernel attention, the “bilateral network” ABCNet [58] incorporating spatial and context pathways, MACU-Net [59] based on multiscale skip connections and asymmetric convolutions, the multistage attention residual UNet (MAResU-Net) [39] featuring a linear attention mechanism, A2-FPN [49] with attention-aggregated feature pyramid network, UnetFormer [26] is a U-shaped neural network with integrated global and local feature reconstruction, DC-Swin [60] employing a dense connection feature aggregation module, and MPCNet [61], a network with a multiscale prototype transformer decoder. Our proposed model ultimately achieved higher accuracy on the widely used ISPRS Vaihingen and ISPRS Potsdam datasets for remote sensing segmentation tasks compared to the aforementioned models.

Results on the Vaihingen Dataset: Table VI presents the numerical results of various semantic segmentation methods on the Vaihingen dataset for comparison. The results indicate that our proposed MHCINet achieved an average F1 of 91.67%,

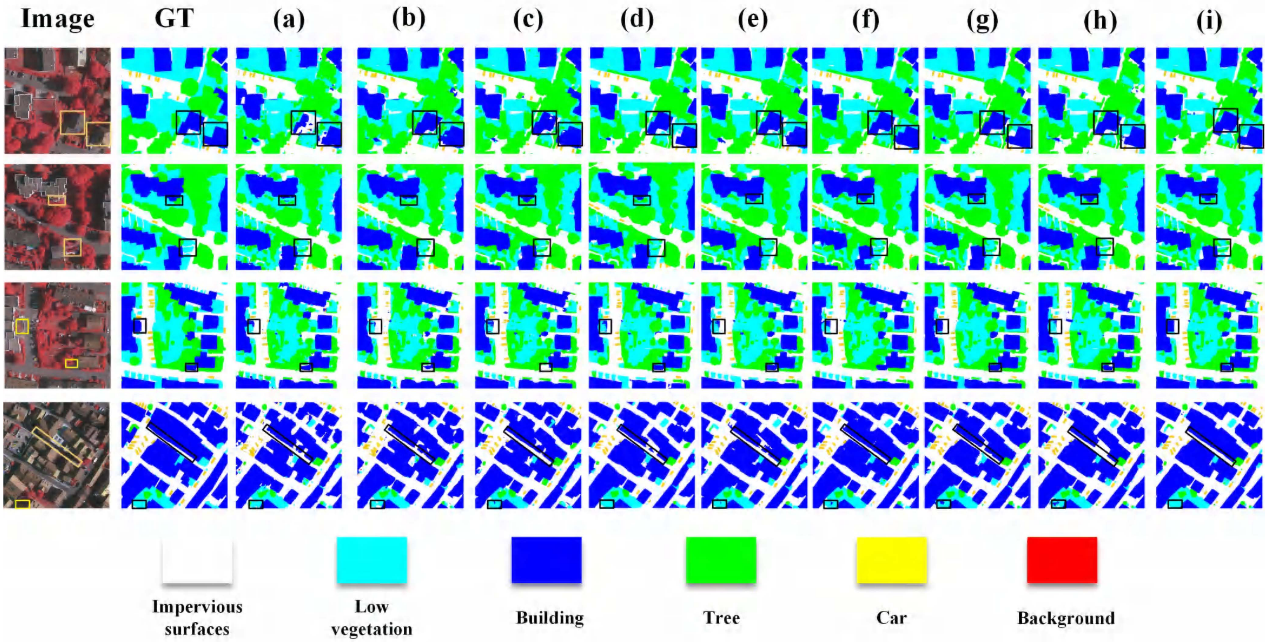


Fig. 12. Examples of segmentation results for different models on the Vaihingen dataset. (a) MANet. (b) ABCNet. (c) MACU-Net. (d) MAResU-Res. (e) A2-FPN. (f) UnetFormer. (g) DC-Swin. (h) MPCNet. (i) MHCINet. In areas with significant differences, regions are outlined with yellow boxes in the original image, and with black boxes in the GT and segmentation network's prediction results. As highlighted in the second row for building areas and the fourth row for tree and low vegetation areas, our model demonstrates superior results in addressing color similarity challenges compared to other state-of-the-art methods. The first and fourth rows emphasize the more coherent and smooth segmentation edges achieved by our model.

mIoU of 84.80%, and OA of 91.95%. MHCINet outperformed other networks with the best F1, OA, and mIoU. Notably, MHCINet surpassed the excellent convolutional lightweight network ABCNet and outperformed the DC-Swin network, which has strong global information representation capabilities. It is worth mentioning that our proposed MHCINet method achieved F1 scores of 86.09% and 91.33% for the "Lowveg" and "Tree" classes, respectively, surpassing other state-of-the-art networks by more than 0.77% and 0.93%. We believe that most lowveg and tree instances exhibit certain color similarities. Experimental results demonstrate that our model can address color similarity challenges to some extent.

Fig. 11 illustrates the segmentation results of MHCINet on images with ID 2 and 6 from the Vaihingen dataset. Fig. 12 compares the segmentation results of MHCINet with other state-of-the-art networks on the Vaihingen dataset. From Fig. 12, the third row shows that buildings and impervious surfaces exhibit color similarities in the original images due to factors such as lighting. MHCINet with CMFI successfully segmented buildings and impervious surfaces precisely. Compared to networks that do not consider global and multiscale information, the model fused with CMFI analyzes and understands segmentation instances from a global perspective by combining information from various scales. This allows the model to observe the overall segmentation instance as much as possible, enabling it to perform pixelwise classification without solely considering individual pixel values. This is why MHCINet with CMFI excels in addressing color similarity challenges compared to other networks. From Fig. 12, the fourth row shows that MHCINet, with the use of CMFI, not only excels in addressing color similarity challenges but also, with the incorporation of ELGA, achieves smoother and more coherent edges in the segmentation results compared to models without edge information. Through various experiments, we demonstrate that the proposed ELGA is effective.

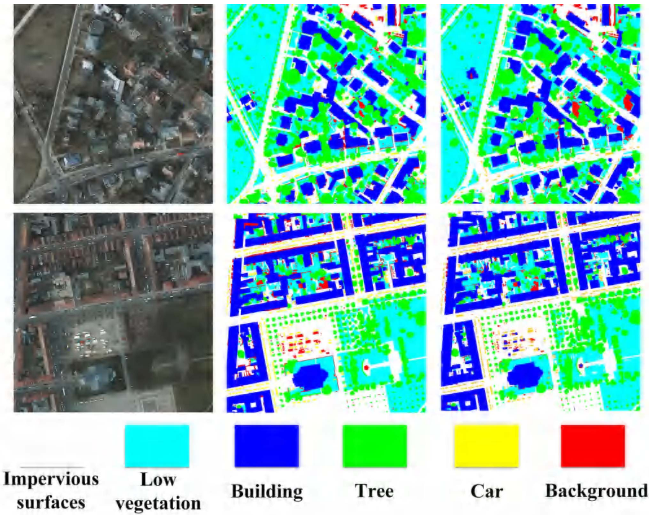


Fig. 13. Segmentation results of MHCINet on images with IDs 2_13 and 5_13 from the Potsdam dataset are illustrated.

Results on the Potsdam Dataset: To comprehensively evaluate the network performance, we conducted further experiments on the Potsdam dataset. The experimental results are presented in Table VII, where MHCINet achieved an average F1 score of 93.31%, an mIoU of 87.64%, and an OA index of 91.90% on the Potsdam test set, outperforming other methods. Due to differences in data size and types, the segmentation accuracy on

TABLE VII
COMPARISON OF SEGMENTATION RESULTS ON THE POTSDAM DATASET

Method	F1 (%)					Evaluation index		
	Imp.Surf	Building	Lowveg	Tree	Car	MeanF1 (%)	mIoU (%)	OA (%)
MANet[58]	87.46	90.66	81.96	82.81	92.76	87.13	77.45	85.05
ABCNet[59]	87.05	88.26	78.94	84.63	93.11	86.4	76.35	84.21
MACU-Net[60]	91.39	93.74	85.87	86.8	94.53	90.46	82.78	88.81
MAResU-Net[40]	92.38	95.83	86.65	88.44	96.13	91.89	85.22	90.28
A2-FPN[50]	92.85	95.84	87.17	88.76	96.13	92.15	85.66	90.68
UnetFormer[26]	92.37	96.21	86.87	88.24	95.31	91.8	85.06	90.39
DC-Swin[61]	93.26	96.86	87.74	88.68	95.50	92.41	86.10	91.16
MPCNet[62]	92.69	96.38	87.30	88.74	96.34	92.29	85.91	90.56
MHCINet(Ours)	**93.76**	**97.21**	**88.79**	**90.12**	**96.64**	**93.31**	**87.64**	**91.90**

The best results are indicated in bold.

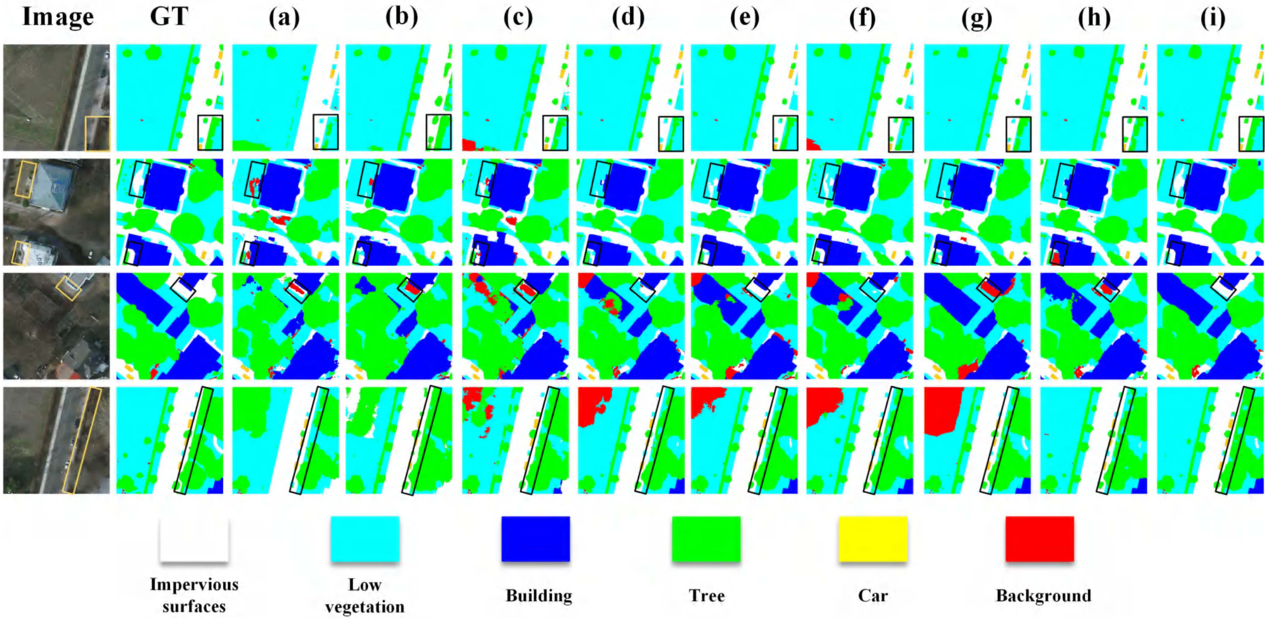


Fig. 14. Examples of segmentation results for different models on the Potsdam dataset. (a) MANet. (b) ABCNet. (c) MACU-Net. (d) MAResU-Res. (e) A2-FPN. (f) UnetFormer. (g) DC-Swin. (h) MPCNet. (i) MHCINet. The regions with significant differences are highlighted with yellow bounding boxes in the original image, and with black bounding boxes in both the GT and the prediction results from the segmentation network. Compared to other state-of-the-art networks, we achieve superior segmentation results.

the Potsdam dataset is generally higher than that on the Vaihingen dataset. Similar to the results on the Vaihingen dataset, the Lowveg and Tree classes on the Potsdam dataset also surpassed other state-of-the-art networks by 1.05% and 1.38%, respectively.

As illustrated in Fig. 13, we provide overall segmentation images for ID2_13 and 5_13, and Fig. 14 showcases the segmentation results of the network models involved in Table VI on the Potsdam dataset. Similar to its performance on the Vaihingen dataset, MHCINet exhibited superior performance in addressing the challenge of color similarity compared to other networks. For instance, in Fig. 14, the second and third rows demonstrate that networks without CMFI tend to misclassify imp.surf and lowveg classes that exhibit similar colors. However, MHCINet with CMFI effectively leveraged global information to correctly classify these instances. For the first and fourth rows in Fig. 14, models using ELGA showed increased sensitivity to edges, enhancing the accuracy of edge segmentation.

V. Limitations and Future Prospects

While our MHCINet demonstrates superior performance in terms of data and partially addresses issues related to edge fusion and color similarity, there are several limitations to consider. The proposed model in this article focuses solely on semantic segmentation of urban scenes in remote sensing imagery and has not explored other remote sensing visual tasks, such as road segmentation and parcel segmentation. In our future work, we intend to delve further into developing more optimal network architectures that incorporate foreground saliency and refine our model to cater to a broader range of remote sensing visual tasks. Furthermore, we may explore segmentation tasks that go beyond pixelwise classification, potentially utilizing approaches beyond the conventional pixel-level labeling. Additionally, our model relies heavily on labeled data for both segmentation and edge detection. In real-world scenarios, obtaining accurate and comprehensive labels can be challenging. Therefore, we plan to enhance our segmentation network by incorporating improvements from unsupervised models, making it more adaptable to practical remote sensing segmentation challenges. Finally, our focus will extend to model compression as we continue our research endeavors.

VI. Conclusion

This article aimed to enhance the sensitivity of segmentation models to edges by incorporating edge information and addressing the challenge of color similarity through the integration of global and multiscale information. These directions were pursued to improve the segmentation accuracy of remote sensing images. The proposed method, MHCINet, is based on foreground saliency and incorporates cross-scale and global information. Specifically, we introduced ELGA, a module designed to fuse low-level and high-level information enhanced by edge features, mitigating edge discontinuity and smoothness issues. In comparison with common fusion methods, ELGA demonstrated a deeper integration of the mentioned information. Additionally, CMFI was introduced to tackle color similarity challenges. This module utilized the encoder part of a standard transformer, replacing layer normalization with batch normalization, and incorporating attention mechanisms after multiscale features to ensure effective integration of multiscale and global information. To adapt the model to our task, a joint loss was designed for training. Experimental results validated the superiority of our network architecture and the effectiveness of individual modules. The findings of this study pave the way for further exploration of fusion models in the remote sensing domain, utilizing foreground saliency, global information, and multiscale features. The potential and applications of such models are substantial. We encourage researchers to delve into the possibilities of foreground saliency-based fusion models, contributing to advancements and applications in remote sensing.

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] 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.

[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] I. de Gélis, S. Lefèvre, and T. Corpetti, "Siamese KPConv: 3D multiple change detection from raw point clouds using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 274–291, 2023.

[5] 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.

[6] F. Chen, H. Balzter, F. Zhou, P. Ren, and H. Zhou, "DGNet: Distribution guided efficient learning for oil spill image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 4201317, pp. 1–17, Jan. 2023.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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, Jun. 2018.

[14] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5506812, pp. 1–12, Jul. 2021.

[15] X. Liu et al., "Deep multiview union learning network for multisource image classification," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 4534–4546, Jun. 2022.

[16] 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.

[17] 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.

[18] 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.

[19] 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. Intervention*, 2015, pp. 234–241.

[20] Z. Zhou, M. M. R. 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., 8th Int. Workshop Multimodal Learn. Clin. Decis. Support*, 2018, pp. 3–11.

[21] X. Wang, Z. Hu, S. Shi, M. Hou, L. Xu, and X. Zhang, "A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved UNet," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 7600.

[22] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5215512, pp. 1–12, Sep. 2022.

[23] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, Art. no. 124.

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

[25] J. Park, "BAM: Bottleneck attention module," in *Proc. Brit. Mach. Vis. Conf.*, 2018, Art. no. 124.

[26] 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.

[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] P. Song, J. Li, Z. An, H. Fan, and L. Fan, "CTMFNet: CNN and transformer multiscale fusion network of remote sensing urban scene imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5900314, pp. 1–14, Dec. 2023.

[29] 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.

[30] 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.

[31] K. Yuan, G. Meng, D. Cheng, J. Bai, S. Xiang, and C. Pan, "Efficient cloud detection in remote sensing images using edge-aware segmentation network and easy-to-hard training strategy," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 61–65.

[32] D. Cheng, G. Meng, S. Xiang, and C. Pan, "FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 12, pp. 5769–5783, Dec. 2017.

[33] K. Wang and D. Ming, "Road extraction from high-resolution remote sensing images based on spectral and shape features," *Proc. SPIE*, vol. 7495, pp. 968–973, 2009.

[34] D. Li, G. Zhang, Z. Wu, and L. Yi, "An edge embedded marker-based watershed algorithm for high spatial resolution remote sensing image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 10, pp. 2781–2787, Oct. 2010.

[35] X. Huang, W. Yuan, J. Li, and L. Zhang, "A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 654–668, Feb. 2017.

[36] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, Art. no. 156.

[37] 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.

[38] 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.

[39] 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, pp. 1–5, 2022.

[40] A. Ma, J. Wang, Y. Zhong, and Z. Zheng, "FacSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5606216, pp. 1–16, Jul. 2022.

[41] U. Michieli and P. Zanuttigh, "Edge-aware graph matching network for part-based semantic segmentation," *Int. J. Comput. Vis.*, vol. 130, no. 11, pp. 2797–2821, 2022.

[42] R. Chen, F.-L. Zhang, and T. Rhee, "Edge-aware convolution for RGB-D image segmentation," in *Proc. 35th Int. Conf. Image Vis. Comput. New Zealand*, 2020, pp. 1–6.

[43] H. Kuang, Y. Liang, N. Liu, J. Liu, and J. Wang, "BEA-SegNet: Body and edge aware network for medical image segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2021, pp. 939–944.

[44] C. Zheng, Y. Chen, J. Shao, and L. Wang, "An MRF-based multigranularity edge-preservation optimization for semantic segmentation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, no. 8008205, pp. 1–5, Feb. 2022.

[45] B. Sui, Y. Cao, X. Bai, S. Zhang, and R. Wu, "BIBED-Seg: Block-in-block edge detection network for guiding semantic segmentation task of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1531–1549, Jan. 2023.

[46] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sens.*, vol. 12, no. 4, 2020, Art. no. 701.

[47] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[48] Z. Zuo et al., "AFFPN: Attention fusion feature pyramid network for small infrared target detection," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3412.

[49] 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.

[50] Z. Tian, X. Guo, X. He, P. Li, X. Cheng, and G. Zhou, "MSCANet: Multiscale context information aggregation network for Tibetan Plateau Lake extraction from remote sensing images," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 1–30, 2023.

[51] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multi-scale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5609519, pp. 1–19, May 2023.

[52] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[53] H. Cao et al., "Swin-UNet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 205–218.

[54] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5617515, pp. 1–15, Aug. 2023.

[55] T. Xiao, Y. Liu, Y. Huang, M. Li, and G. Yang, "Enhancing multiscale representations with transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5605116, pp. 1–16, Mar. 2023.

[56] Z. Chen, G. Wu, H. Gao, Y. Ding, D. Hong, and B. Zhang, "Local aggregation and global attention network for hyperspectral image classification with spectral-induced aligned superpixel segmentation," *Expert Syst. Appl.*, vol. 232, 2023, Art. no. 120828.

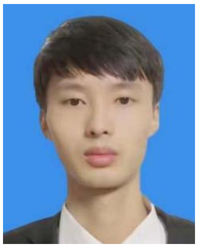
[57] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 5607713, pp. 1–13, Jul. 2022.

[58] 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.

[59] R. Li, C. Duan, and S. Zheng, "MACU-Net semantic segmentation from high-resolution remote sensing images," 2020, *arXiv:2007.13083*.

[60] 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, no. 6506105, pp. 1–5, Jan. 2022.

[61] Q. Wang, X. Luo, J. Feng, G. Zhang, X. Jia, and J. Yin, "Multi-scale prototype contrast network for high-resolution aerial imagery semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5615114, pp. 1–14, Jul. 2023.



Yunsong Yang received the bachelor's degree in computer science and technology 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 science and technology with the School of Computer Science and Technology, Shandong University of Finance and Economics, Yantai, China.

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

Genji Yuan received the bachelor's degree in electronic science and technology from Shandong Technology and Business University, Yantai, China, in 2017, and the master's and Ph.D. degrees in computer science and technology from Qingdao University, Qingdao, China, in 2020 and 2024, respectively.

His research interests include computer vision, pattern recognition, machine learning, data analysis, and intelligent transportation systems.

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

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. He is currently a Professor with the School of Computer Science and Technology, Shandong Technology and Business University. His research interests include image processing, computer graphics, computer vision, and machine learning.