Hybrid Attention Fusion Embedded in Transformer for Remote Sensing Image Semantic Segmentation

Yan Chen, Quan Dong, Xiaofeng Wang, Qianchuan Zhang, Menglei Kang, Wenxiang Jiang, Mengyuan Wang, Lixiang Xu, and Chen Zhang

***Abstract*—In the context of fast progress in deep learning, convolutional neural networks have been extensively applied to the semantic segmentation of remote sensing images and have achieved significant progress. However, certain limitations exist in capturing global contextual information due to the characteristics of convolutional local properties. Recently, Transformer has become a focus of research in computer vision and has shown great potential in extracting global contextual information, further promoting the development of semantic segmentation tasks. In this article, we use ResNet50 as an encoder, embed the hybrid attention mechanism into Transformer, and propose a Transformer-based decoder. The Channel-Spatial Transformer Block further aggregates features by integrating the local feature maps extracted by the encoder with their associated global dependencies. At the same time, an adaptive approach is employed to reweight the interdependent channel maps to enhance the feature fusion. The global cross-fusion module combines the extracted complementary features to obtain more comprehensive semantic information. Extensive comparative experiments were conducted on the ISPRS Potsdam and Vaihingen datasets, where mIoU reached 78.06% and 76.37%, respectively. The outcomes of multiple ablation experiments also validate the effectiveness of the proposed method.**

***Index Terms*—Global cross fusion, hybrid attention, remote sensing image, semantic segmentation, Transformer.**

I. Introduction

With the continuous development of satellite and remote sensing technologies, numerous high-resolution images can be easily acquired [1]. Semantic segmentation plays a crucial role in various remote sensing applications, including urban construction and planning, land surveying, environmental monitoring, and disaster assessment, to name a few. In the classical paradigm of geographic object analysis based on image analysis, it is expected first to use unsupervised segmentation methods to segment the image and then classify the segmented regions [2]. However, semantic segmentation employs a pixel-level supervised style and assigns each pixel with a predesigned label. Although remote sensing images contain a wealth of detailed ground object information, the distribution of ground object categories is often imbalanced. In addition, due to variations in shape, color, texture, and other features, the ground objects exhibit significant intra-class variance and slight interclass variance in the imaging process. Consequently, the semantic segmentation task for remote-sensing images poses substantial challenges [3].

Methods based on hand-crafted feature extraction were the primary approach in the early stages of addressing semantic segmentation on remotely sensed images [4]. However, as remote sensing image resolution continues to improve, conventional methods face significant challenges in extracting ground object features and achieving accurate semantic segmentation [5]. These conventional methods often rely heavily on domain knowledge and expertise, requiring manual feature design and the selection of suitable machine learning algorithms. However, high-resolution remote sensing images contain abundant details and intricate scenes, making it difficult for these methods to perform refined feature extraction and segmentation. In addition, traditional methods have poor adaptability, as each task may require parameter adjustments or method redesign, which increases manual intervention and development costs and limits the scalability and generality of the algorithms [6].

Deep learning techniques, particularly convolutional neural networks (CNNs), have emerged as the dominant approach for various semantic segmentation tasks in recent years [7], [8]. Unlike conventional methods, deep learning models do not require manual feature design. They automatically learn feature representations from data, thereby increasing automation and possessing strong nonlinear modeling capabilities [9]. By employing deep learning architectures such as CNNs, it becomes possible to capture more fine-grained local contextual information, effectively extracting the complex advanced features of objects on the ground and achieving more accurate semantic segmentation. Inspired by the end-to-end fully CNN (FCN) framework [10], many semantic segmentation networks have been developed. For example, SegNet [11] employs an encoder–decoder architecture for feature extraction and upsampling, thereby generating efficient pixel-level segmentation results. DeepLab v3+ [12] introduced the atrous spatial pyramid pooling (ASPP) module to obtain spatial contextual feature

Manuscript received 21 November 2023; revised 3 January 2024; accepted 22 January 2024. Date of publication 29 January 2024; date of current version 12 February 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176085, in part by the Key Scientific Research Foundation of the Education Department of Province Anhui under Grant KJ2020A0658, in part by the University Natural Sciences Research Project of Province Anhui under Grant KJ2021ZD0118, in part by the Hefei University Talent Research Funding under Grant 20RC13, in part by the Hefei University Scientific Research Development Funding under Grant 20ZR03ZDA, in part by the Program for Scientific Research Innovation Team in Colleges and Universities of Anhui Province under Grant 2022AH010095, and in part by the Hefei Specially Recruited Foreign Expert support. *(Corresponding author: Quan Dong.)*

The authors are with the School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China (e-mail: chenyan@hfuu.edu.cn; dq2112774778@163.com; xfwang@hfuu.edu.cn; qianczhang@163.com; kangml@stu.hfuu.edu.cn; jiangwx@stu.hfuu.edu.cn; wangmy@stu.hfuu.edu.cn; xulixiang@hfuu.edu.cn; zhangchen@hfuu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3358851

information. Regular convolutions are replaced by atrous convolutions to expand the receptive field and achieve multiscale feature extraction. PSPNet [13] achieves multiscale receptive fields and fine-grained feature fusion through a pyramid pooling structure to enhance the understanding of semantic information in images. Both BiSeNet V1 [14] and V2 [15] adopt a concise and efficient dual-path encoder structure, which respectively extracts spatial detail information and high-level semantic information, achieving high accuracy and fast speed while enhancing discriminative ability. MPSegNet [16] incorporates a scale guidance module, enabling the subnetwork to focus on specific scale objects with large-scale variations in the image. CGFDN [17] designs a feature decoupling module to encode the co-occurrence relationship into the convolutional features, thereby decoupling the most discriminative features. SLA-NET [18] proposes a spatial-logic aggregation network based on morphological transformations, where morphological operators effectively embed trainable structural elements to form unique morphological representations. However, compared to natural images, remote sensing images contain more abundant information. The convolutional operations used for feature extraction also introduce significant noise interference. Spectral data provides valuable information for scene understanding. SpectralGPT [19], ExViT [20], and GSANet [21], among others, all integrate complementary information from different modalities to achieve more comprehensive and accurate results. Other works specifically design attention mechanisms targeting specific problems and challenges to effectively alleviate the impact of noise on segmentation [22].

Subsequently, attention mechanisms aroused a wave of enthusiasm in the field of deep learning and gained widespread application in semantic segmentation tasks. For example, the convolutional block attention module (CBAM) [23] combines mechanisms of channel attention and spatial attention, enabling the network to adaptively adjust the performance of feature maps in the channel dimension and spatial dimension, thereby enhancing the model's ability to focus on essential features and suppress noise. MACU-Net [24] adopts asymmetric convolutional blocks to enhance the feature representation capability, replacing the standard convolutional layers. It also designs multiscale skip connections combined with channel attention to combine and refine the semantically generated features at multiple levels. MANet [25] proposes a multiscale strategy based on kernel and channel attention to aggregate relevant contextual features at different levels. MsanlfNet [26] uses multiscale attention and fast Fourier transform to obtain fine multiscale spatial features and global contextual information, effectively balancing performance and computational complexity. The introduction of these efficient attention mechanisms has effectively alleviated the weakness of CNNs in handling global dependence relationships, resulting in significant improvements in semantic segmentation accuracy. However, there is still room for improvement in modeling global context. For example, as shown in Fig. 1(a), the lack of global context information in MANet leads to an inaccurate understanding of the complete shape and boundaries of the target object, resulting in erroneous segmentation results. In addition, as shown in Fig. 1(b), trees and low-vegetation areas may be influenced by clouds, shadows, occlusions, and other interferences, causing them to exhibit similar texture features. This can lead to category confusion when MsanlfNet encounters these regions.

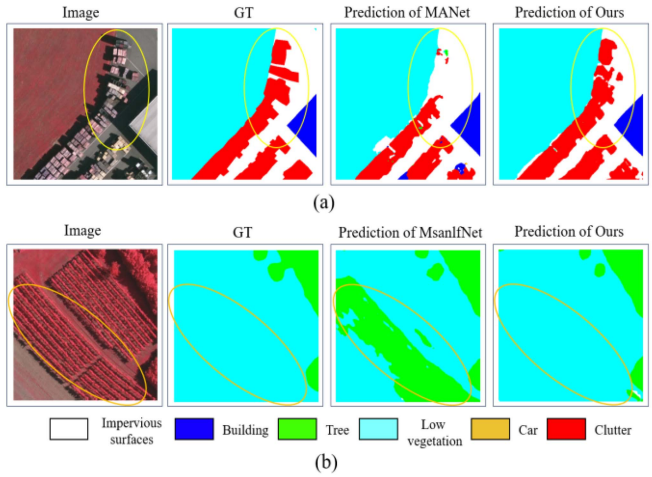

Fig. 1. Examples of predictions on the ISPRS Vaihingen dataset. (a) MANet incorrectly segmented the complete boundaries and shape of the target object. (b) MsanlfNet encountered category confusion and misclassified low-vegetation as trees.

Many studies have also started to address the challenges posed by remote-sensing images. For example, HighDAN [27] proposes a high-resolution domain adaptation network, by embedding the adversarial learning-based DA's idea into HR-Net and utilizing the Dice Loss to mitigate the effects of class imbalance. MMT [28] proposes a mixed-mask attention mechanism that assists the network in learning more explicit intraclass and interclass correlations by capturing long-range interdependent representations. The study in [29] highlights the significant progress and breakthroughs achieved by introducing Transformer methods into semantic segmentation tasks. Drawing from the recent vital advances of Transformers in computer vision, we propose a novel Transformer architecture to address the aforementioned issues. HAFNet primarily combines a CNN-based encoder with a specially crafted Transformer decoder, forming a hybrid architecture. This design fully leverages the strengths of both CNN and Transformer, enabling the network to efficiently extract and process complex contextual information. The main contributions of this article are outlined as follows.

1) A novel hybrid attention-based Transformer architecture, Channel-Spatial Transformer Block (CSTB), is intended to aggregate local detail, channel, and global information at different levels to obtain more comprehensive semantic information.
2) GCFM employs an interactive fusion strategy that promotes synergy between different branching features to establish effective correlations between features so that the network can capture critical information from the image as a whole.
3) A novel decoder architecture constructed by CSTB and GCFM based on the encoder of ResNet50 has been proposed for improving representation capacity. Comparisons between the different methods are carried out on the ISPRS Potsdam and Vaihingen datasets, and the experimental

results indicate that our proposed approach exhibits higher efficiency compared to other advanced approaches.

The subsequent sections of this article are organized in the following manner. Section II reviews some related work. Section III provides a thorough description of the proposed approach. Section IV introduces the datasets and evaluation metrics used in the experiments and presents a detailed analysis of the experimental results. Section V concludes this article.

II. Related Work

This section concisely reviewed the semantic segmentation methods related to our study and analyzed their constraints.

A. Encoder–Decoder-Based Architectures

The FCN is a classic end-to-end method for semantic segmentation. It uses convolution and deconvolution operations to restore feature maps to the original image resolution, achieving pixel-level classification. However, the overly simplified encoder–decoder structure in FCN leads to coarse segmentation results, thereby reducing segmentation accuracy. U-Net [30] effectively solves this issue by employing a symmetric structure that connects the encoder and decoder, known as the contraction-expansion path. This architecture preserves abundant contextual information and high-resolution features, leading to more accurate image segmentation. Specifically, the encoder progressively decreases the feature map size and captures semantic features, whereas the decoder restores details and spatial detail information through a combination of upsampling and skip connections, achieving accurate segmentation while preserving details. The encoder–decoder architecture has subsequently achieved excellent performance and wide application in remote sensing image semantic segmentation tasks [31]. In related studies [32], [33], various improvements have been made at the decoder stage to extract rich semantic information.

While the CNN-based encoder–decoder architecture has achieved remarkable performance, it faces certain limitations regarding the receptive field. If the emphasis is solely placed on extracting local semantic features, the network may have difficulty effectively capturing the comprehensive image information, especially in high-resolution remote sensing urban scene images with rich features. This limitation can pose considerable challenges in accurately identifying complex target objects, leading to erroneous segmentation results.

B. Attention Mechanism

Combining deep learning and attention mechanisms has seen widespread use in many fields. To address the problem of CNN focusing too much on local patterns, numerous attempts have been made to model global information, with the widely favored approach being to introduce attention mechanisms into the network. For example, nonlocal neural networks [34] introduce nonlocal modules to capture global dependencies by computing similarities between pixels, overcoming the limitations of conventional CNNs in handling long-range dependencies. DANet [35] proposes combining positional attention with channel attention to better capture dependencies and contextual information between diverse pixel positions. OCRNet [36] constructs soft object regions for each category in advance and enhances the feature representation ability of pixels by learning the relationship between pixels and object regions, thus fusing local feature information with global contextual information.

With the development and broad application of numerous attention mechanisms further validating their great potential, attention mechanisms have also contributed significantly to the progress in the semantic segmentation of remotely sensed images. LSCNet [37] introduces a large kernel sparse ConvNet weighted by multifrequency attention, utilizing two parallel rectangular convolutional kernels and employing an adaptive sparse optimization strategy to dynamically optimize the fixed neuron connections between different convolutional layers. In [38], a spectral attention subnetwork and a spatial attention subnetwork are constructed to focus on more discriminative information in the spectral domain and spatial domain, respectively. In [39], a multimodal attention-aware convolutional network is proposed, where designed cascaded blocks facilitate multistage information exchange. LANet [40] proposes a local attention embedding method, which allows the network to focus on leveraging the connection between local detail features and global features to capture comprehensive contextual information. AFNet [41] designs an attention fusion network that allows the network to retain more detailed information while adaptively performing multiscale feature fusion to capture more integrated semantic information. CANet [42] leverages the techniques of multiscale residual concatenation and spatial pyramid pooling to aggregate rich contextual information at different levels. Despite these advantages, the convolution operation based on limited receptive fields primarily focuses on extracting local feature information. At the same time, these attention modules heavily rely on convolutional operations, making it challenging to efficiently extract global context dependencies and acquire long-range dependencies. As a result, it can easily lead to ambiguity in classifying certain pixels in remote sensing images [43], [44]. Finally, if only a single attention module is used at the decoder, the network lacks the ability to globally model multilevel semantic features.

C. Transformer-Based Semantic Segmentation Methods

With its excellent sequence-to-sequence modeling capability, the Transformer model has shown outstanding performance in extracting global contextual information compared to the models mentioned above that only use the regular attention mechanism. Subsequently, many researchers began to explore applying the Transformer model to the semantic segmentation task of remotely sensed images [45]. The main advantage of the Transformer model in dealing with fine high-resolution remote sensing imagery of urban scenes is its capacity to efficiently establish long-range dependence relationships, thereby improving the accuracy and robustness of semantic segmentation.

Currently, most Transformer models used for semantic segmentation still adopt an encoder–decoder structure. Based on different combinations of encoders and decoders, Transformer models for semantic segmentation can be classified into two types. The first type is models that fully utilize the Transformer structure. For example, SegFormer [46], SwinUnet [47], and

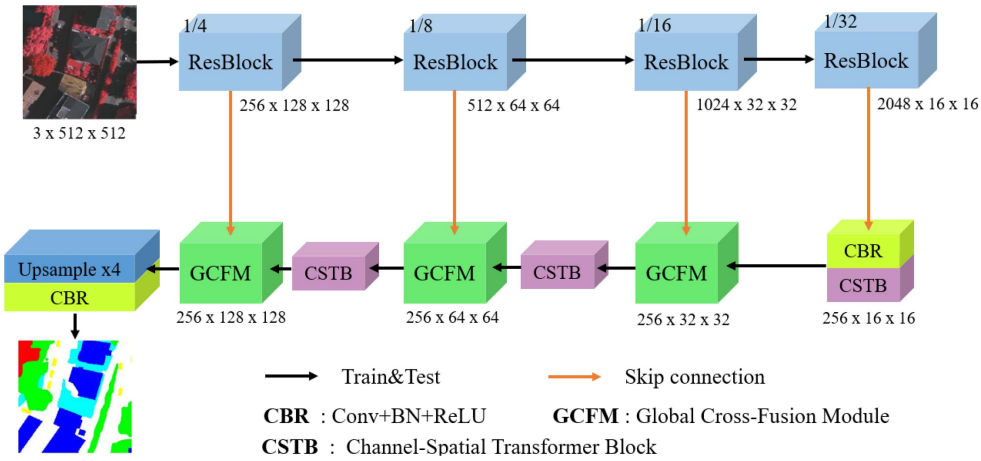

Fig. 2. Schematic of the HAFNet.

Segmenter [48] leverage the advantages of Transformers to better consider global semantic information, leading to significant improvements in segmentation performance and efficiency. The second type involves using the Transformer in the encoder while employing CNN in the decoder. For example, GMHF [49] proposes a bibranch global^{+} multiscale hybrid network that extracts both multiscale global and local features. These features are further integrated through correlation modules to provide the network with more comprehensive semantic information. In DC-Swin [50], the Swin Transformer [51] is used in the encoder to enhance the network's ability to handle long-range dependencies. Conversely, the decoder incorporates a densely connected feature aggregation module based on convolution operations. In the research by [52], SwinTF-U-Net, SwinTF-PSP, and SwinTF-FPN also utilize the Swin Transformer as the encoder, whereas the decoder is based on multiscale CNN architectures, such as U-Net, PSP, and FPN. Although the second type of Transformer combined with CNN has achieved impressive results in remote sensing image semantic segmentation tasks, its computational cost is much higher than that of a CNN-based encoder [53]. Therefore, the proposed network architecture in this article adopts CNN as the encoder and utilizes the Transformer in the decoder part.

III. Methodology

This section presents the structure of HAFNet and provides an overview of the model's framework, as shown in Fig. 2. The construction of the model involves the utilization of a CNN-based encoder and a Transformer-based decoder. A comprehensive description of every component follows.

A. CNN-Based Encoder

The feasibility of ResNet50 in remote sensing image semantic segmentation tasks has been verified in previous studies [25], [26]. We adopt a pretrained ResNet50 as the encoder, which consists of four residual block layers. Each residual block layer performs a down-sampling of the feature maps by a factor of 2. The deep structure enhances the network's generalization ability and representation power. Multiscale skip connections ensure that our model efficiently handles both fine-grained and coarse-grained semantic information in the images. To extract multilevel semantic features while maintaining computational efficiency, we apply channel compression to the end of the backbone network, reducing the final output channels to 256 and ensuring consistency throughout the decoder with 256 channels. This design aims to balance performance and computational cost.

B. Transformer-Based Decoder

The mainstream methods for capturing global contextual information can be divided into two types. The first approach is to add attention modules at the end of the encoder [34]. However, this design may make it difficult for the network to capture multiscale global semantic features. The second approach is to directly use a Transformer model in the encoder [54]. However, this not only increases the computational burden and number of parameters but may also result in the loss of spatial detail feature information. In contrast, HAFNet introduces the CSTB and GCFM in our decoder. These components work together to extract global context correlation without sacrificing spatial details. Different from the traditional multihead self-attention in the regular Transformer, the proposed CSTB mainly includes a channel attention module and a spatial attention module to optimize the network's utilization of channel information and capture global contextual information. In addition, the residual branch in the CSTB effectively aggregates local details, channels, and global contextual information by leveraging the local feature information extracted by ResNet50, which helps the network capture more comprehensive semantic information.

C. Channel Adaptive Module (CAM)

Channel attention is to learn channel weights to weigh different channels of the input feature map, which enables the model to dynamically select and adjust the importance of channels. CAM adopts a two-branch structure compared to previous classical channel attention modules [55], [56]. Specifically, as shown in

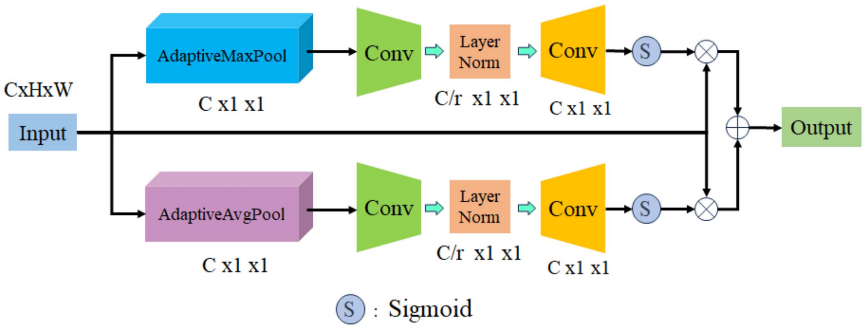

Fig. 3. Schematic of the channel adaptive module.

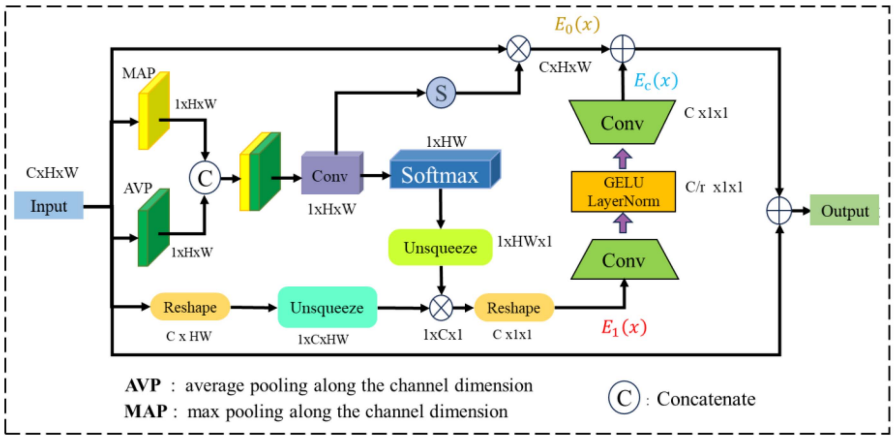

Fig. 4. Schematic of the spatial context module.

Fig. 3., for an input feature map of size $X \in \mathrm{R}^{\mathrm{CxHxW}}$, first, maximum pooling and average pooling operations are performed on it to aggregate the global semantic information of the feature map, resulting in two channel maps with the same dimensions, i.e., the maximum pooling feature $X_{Mc} \in \mathrm{R}^{\mathrm{Cx1x1}}$ and the average pooling feature $X_{Ac} \in \mathrm{R}^{\mathrm{Cx1x1}}$. In the subsequent bottleneck structure, the channel dimension is reduced by a 4x factor through convolution. Next, LayerNorm is introduced to enhance the model's robustness. Subsequently, the channel dimension is restored to its original value through convolution. Finally, we use the Sigmoid activation function to acquire weight coefficients and multiply the input feature map with these two channel weights to enhance the effective features. The two corresponding results are then fused together using a summation operation. The relevant formulas are as follows:

$$\begin{aligned} M_C(X) &= \sigma(W_{1x1}(L_N(W_{1x1}(\mathrm{AvgPool}(x))))) \cdot X \\ &\quad + \sigma(W_{1x1}(L_N(W_{1x1}(\mathrm{MaxPool}(x))))) \cdot X \\ &= \sigma(W_{1x1}(L_N(W_{1x1}(X_{Mc})))) \\ &\quad + \sigma(W_{1x1}(L_N(W_{1x1}(X_{Ac})))) \end{aligned} \tag{1}$$

where AvgPool represents the average pooling operation, MaxPool denotes the maximum pooling operation, L_N denotes the Layernorm operation, σ represents the Sigmoid activation function, and W_{1x1} represents the 1×1 convolution operation.

D. Spatial Context Module (SCM)

SCM can focus the attention on the global scope of the whole image or feature map, enhancing the model's ability to perceive global features, achieving a more comprehensive and integrated grasp of the overall information of the image, and better interpreting the relationship between different regions. As shown in Fig. 4., the primary emphasis lies in merging two distinct branches. First, in the upper part, the input feature map $X \in \mathrm{R}^{\mathrm{CxHxW}}$ is subjected to average pooling and max pooling operations along the channel dimension separately. The two results are then concatenated, and a convolutional operation is performed to generate the spatial feature map $(\in)$. Finally, by applying the Sigmoid activation function, the resulting map is multiplied with the input feature map to acquire the spatial attention map $E_0(x)$, capturing the dependence relationships between different positions. The formula is as follows:

$$E_0(X) = \sigma(W_{1x1}[\mathrm{MAP};\mathrm{AVP}]) \cdot X = \sigma(\in) \cdot X \tag{2}$$

where MAP represents max pooling across the channel dimension, whereas AVP represents average pooling along the channel dimension. In the following part, the spatial feature

map $(\in)$is dimensionally transformed, and the Softmax activation function is applied to the dimensionally adjusted result (i.e., H∗W dimension). The dimension is added to the output to obtain $X_S \in \mathrm{R}^{1\mathrm{xHWx1}}$. Furthermore, the input feature map is reshaped to obtain $X_R \in \mathrm{R}^{\mathrm{CxHW}}$. Subsequently, the result of the unsqueeze transformation, denoted as $X_U \in \mathrm{R}^{1\mathrm{xCxHW}}$, undergoes matrix multiplication with $X_S \in \mathrm{R}^{1\mathrm{xHWx1}}$. Finally, the channel map $E_1(x)$ with dimensions of Cx1x1 is obtained through the reshape operation. This part is the distillation of spatial information for global perceptual convergence. The specific steps are formulated as follows.

$$\begin{aligned} E_1\ (X) &= R(U_n\,(S_{d2}\,(R\,(\in)))\ \cdot\ U_n\,(R\,(X))) \\ &= R\,(X_S \cdot\ U_n\,(X_R)) \\ &= R\,(X_S \cdot X_U) \end{aligned} \tag{3}$$

where R denotes the Reshape operation, S_{d2} denotes the Softmax activation function acting on the second dimension and U_n signifies the unsqueeze operation that expands on the data dimensions. The subsequent bottleneck structure is used to enhance the interchannel dependencies. First, the channels are compressed to *C*/*r* by a convolutional layer, where *r* is the reduction factor. In addition, LayerNorm is inserted in the middle, allowing the normalization operation to be performed before the nonlinear activation function GELU. This facilitates the preservation of the data's dynamic range and improves the model's generalization ability. After that, the result $E_C \in \mathrm{R}^{\mathrm{Cx1x1}}$, which aggregates the global feature information, is superimposed on the spatial feature $Eo(x)$. Finally, the fused result is summed with the input feature map to obtain $E(X)$ as follows:

$$\begin{aligned} E\ (X) &= W_{3x3}(\delta(L_N(W_{3x3}\,(E_1\,(X))))) + E_0\,(X) + X \\ &= E_C\ + E_0\,(X) + X \end{aligned} \tag{4}$$

where W_{3x3} represents the 3x3 convolution operation, and δ denotes the GELU activation function.

E. Redesigned MLP Layer in Transformer

Although the FFN layer in the previous Transformer Block has strong nonlinear modeling ability, its local perceptiveness is inadequate as it only focuses on the feature information of the current position. Therefore, further improvements have been made to the FFN in the Transformer Block. As shown in Fig. 5, the improved FFN layer first utilizes a 7 × 7 depthwise convolution to enhance the model's local perceptiveness.

Next, in the subsequent fully connected layer, we introduce the simple and efficient global response normalization (GRN) from ConvNext v2 [57]. It aggregates global features, normalizes features, and calibrates features, thereby helping the network suppress feature collapse and enhance channel contrast and selectivity. The three specific steps of GRN are as follows. First, the spatial features X_i are aggregated into a vector gx using the global function $\mathcal{G}(\cdot)$ as follows:

$$\mathcal{G}\,(X) :=\ X \in R^{HxWxC} \rightarrow gx \in R^{C}. \tag{5}$$

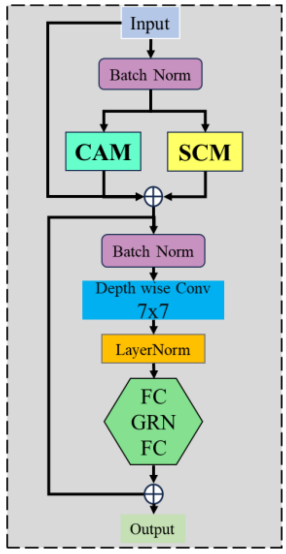

Fig. 5. Schematic of the Channel-Spatial Transformer Block.

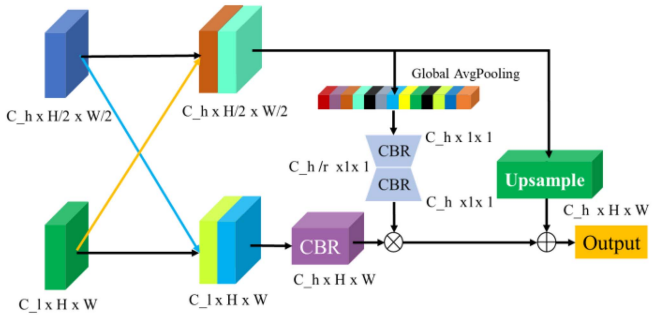

Fig. 6. Schematic of the global cross-fusion module.

This can be seen as an elementary pooling layer. Next, we apply the response normalization function $\mathcal{N}(\cdot)$ to the aggregated value. Specifically, we use the following standard division normalization:

$$\mathcal{N}\,(\parallel X_i \parallel) :=\parallel X_i \parallel \in R \rightarrow \frac{\parallel X_i \parallel}{\sum_{j\ =\ 1,\ldots C} \parallel X_j \parallel} \in R \tag{6}$$

where $\parallel X_i \parallel$ represents the L2-norm of the *i*th channel. Intuitively, (6) quantifies the importance of the *i*th channel relative to all other channels in the entire feature map. Similar to other normalization methods [58], this step aims to induce competition and mutual inhibition among different channels, thereby enhancing the expressive power of features. Ultimately, the original input response is adjusted using the computed feature normalization scores, which is formulated as follows:

$$X_i = X_i * \mathcal{N}(\mathcal{G}\,(X)_i)\ \in R^{HxW}. \tag{7}$$

F. Global Cross-Fusion Module

In the decoder, conventional feature fusion methods typically involve using fixed interpolation or convolution operations to upsample the image and then directly fuse it with the features

obtained from skip connections. However, this often leads to loss of detailed information and blurriness. Thus, we propose a global cross-fusion module (GCFM) to obtain more comprehensive and global semantic information, enabling more accurate restoration of image details. As shown in Fig. 6, the information interaction fusion of different branches compensates for the information loss caused by the reduction in the number of channels and enhances the spatial and semantic information. Secondly, a global average pooling layer is utilized to generate the attention map $X_C \in \mathrm{R}^{\mathrm{Cx1x1}}$. Then, the channel dimension C is reduced by four times, expanded to the original size, and multiplied with the low-level features to obtain the attention feature map $Y_C \in \mathrm{R}^{\mathrm{CxHxW}}$. This suppresses unnecessary noise and selectively focuses on essential parts of the image. Finally, the upsampling operation is performed on the high-level features, and the obtained result is fused with the attention feature map using summation. The formula is as follows:

$$\begin{aligned} G(X) &= \mathrm{CBR}(C_r\ BR(\mathrm{AvgPool}(X \cdot Y))) \cdot \mathrm{CBR}(Y \cdot X) \\ &\quad + U_p(X) \\ &= \mathrm{CBR}(C_r\ BR(X_C)) \cdot \mathrm{CBR}(Y \cdot X) + U_p(X) \\ &= Y_C + U_p(X) \end{aligned} \tag{8}$$

where *X* represents the deep semantic features, *Y* represents the shallow semantic features, $X \cdot Y$ denotes the feature fusion by summation of *Y* after upsampling and dimension adjustment with *X*, and $Y \cdot X$ follows the same logic. CBR represents convolution, batch normalization, and ReLU, whereas U_p represents bilinear interpolation upsampling.

IV. Experimental Results and Analysis

In this section, we will outline the datasets used, the evaluation metrics employed, and provide details of the conducted experiments. We will analyze and discuss the experimental results on both datasets.

A. Dataset Description

1) Potsdam: This dataset, provided by ISPRS, was collected from aerial imagery of the German city of Potsdam, with a total of 38 high-resolution images. Each image has an average size of 6000 × 6000 pixels and a ground sampling distance of 5 cm. This dataset includes six categories: Impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background. Specifically, we use images with IDs 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_10, 7_11, and 7_12 for training. We use the images with IDs 2_10 for validation and the rest for testing. To reduce the amount of computation, a digital surface model (DSM) and a normalized DSM are not used in our experiments. In addition, red, green, and blue bands are used in our experiments. We also crop the original images into patches of 512 × 512 size and perform data enhancement by randomly rotating, scaling the size, vertically flipping, horizontally flipping, and adding random Gaussian noise.

TABLE I
Ablation Study of Each Component of the HAFNet

Dataset	Method	OA	mF1	mIoU
Potsdam	Baseline	89.62	84.86	76.19
	Baseline + GCFM	89.73	85.59	76.88
	Baseline + CAM	90.01	85.44	76.99
	Baseline + SAM	89.70	85.47	76.75
	Baseline + CS	89.81	85.89	77.16
	Baseline + CSTB	90.44	86.02	77.66
	Baseline + CSTB + GCFM	**90.45**	**86.47**	**78.06**
Vaihingen	Baseline	89.89	84.13	74.25
	Baseline + GCFM	90.04	84.70	74.84
	Baseline + CAM	90.22	84.50	74.77
	Baseline + SAM	89.88	84.85	75.00
	Baseline + CS	90.08	85.01	75.18
	Baseline + CSTB	90.14	85.23	75.45
	Baseline + CSTB + GCFM	**90.29**	**85.93**	**76.37**

The best values in the column are emphasized in bold.

2) Vaihingen: This dataset, provided by ISPRS, was collected from aerial imagery of the German city of Vaihingen and consists of 33 high-resolution images. Each image has an average size of 2994 × 2064 pixels and a ground sampling distance of 5 cm. The ground reality comprises six categories identical to those in the ISPRS Potsdam benchmark. The training set includes 16 images, whereas the test set consists of 17 images. For training, we use images with IDs 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 32, 34, and 37. Image with ID 30 was reserved for validation, whereas the rest 17 images were used for testing. The dataset is processed in the same manner as Potsdam.

B. Evaluation Metrics

The experimental results are evaluated based on three commonly used metrics: overall accuracy (OA), mean F1 score (mF1), and mean intersection over union (mIoU). Their calculations are as follows:

$$\mathrm{OA} = \frac{\sum_{i=0}^{C} K_{ii}}{\sum_{i=0}^{C}\sum_{j=0}^{C} K_{ij}} \tag{9}$$

$$\mathrm{F1} = \frac{1}{C+1}\sum_{i=0}^{C}\frac{K_{ii}}{K_{ii} + \frac{1}{2}\sum_{j=0}^{C}(K_{ij}+K_{ji})} \tag{10}$$

$$\mathrm{mIou} = \frac{1}{C+1}\sum_{i=0}^{C}\frac{K_{ii}}{\sum_{j=0}^{C} K_{ij} + \sum_{j=0}^{C}(K_{ji}-K_{ii})} \tag{11}$$

where K_{ii} expresses the number of correctly classified pixels for class *i*. K_{ij} expresses the number of pixels incorrectly classified as class *j*. K_{ji} denotes the number of pixels from class *j* that are incorrectly classified as class *i*. *C* characterizes the total number of categories.

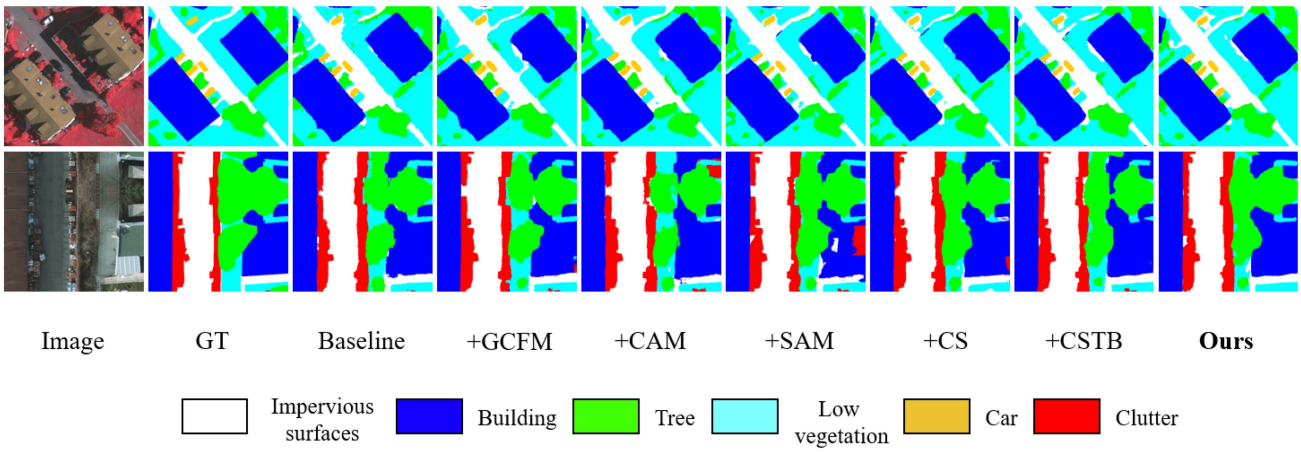

Fig. 7. Visualization of ablation experiments on two datasets.

TABLE II
RESULTS FROM EXPERIMENTS CONDUCTED ON THE POTSDAM TEST SET

Method	Per-class IoU (%)						Evaluate Metrics (%)		
	Imp.surf	Building	Low.veg.	Tree	Car	Clutter	mF1	OA	mIoU
PSPNet	82.77	90.01	74.45	76.97	90.48	31.69	83.46	88.65	74.39
BiSeNeV1	82.95	89.63	73.92	74.36	88.22	35.45	83.60	88.25	74.09
BiSeNeV2	80.91	87.18	71.95	73.16	81.94	34.66	82.05	86.95	71.63
MAResU-Net	84.34	91.76	75.40	76.73	90.70	40.90	85.53	89.54	76.64
MsanlfNet	83.98	90.24	74.02	74.98	88.70	37.57	84.26	88.69	74.92
ABCNet	84.78	92.02	76.52	78.49	91.70	36.39	85.21	89.92	76.65
MANet	84.49	91.48	76.36	78.12	91.03	41.17	85.85	89.81	77.11
BANet	83.35	89.14	73.55	74.56	87.99	33.88	83.26	88.17	73.74
PVT	83.35	91.36	73.33	75.03	81.41	37.24	85.79	88.91	77.32
UnetFormer	84.81	91.40	74.34	77.54	91.77	36.51	84.85	89.38	76.06
CMTFNet	85.63	**92.65**	76.18	78.36	91.60	40.52	86.01	90.17	77.49
BuildFormer	83.04	89.02	72.84	74.8	88.87	37.56	83.84	88.00	74.23
MCCANet	85.84	92.51	76.60	79.05	**92.16**	39.01	85.93	90.31	77.53
GCDNet	85.67	92.40	75.65	78.22	91.26	38.32	85.51	89.99	76.92
Ours	**85.94**	92.54	**76.89**	**79.32**	91.58	**42.12**	**86.47**	**90.45**	**78.06**

The best values are highlighted in bold.

C. Implementation Details

All experiments were conducted using PyTorch on a single server equipped with an NVIDIA GeForce RTX 3090 GPU, which has 24 GB of memory. In all experiments, we utilized the AdamW optimizer to accelerate convergence. The baseline learning rate was set to 6e-4, and a cosine strategy was employed to update the learning rate. The batch size was set to 8, and the maximum number of epochs for training was set to 105. Moreover, during the training process, we combined the soft cross entropy loss and dice loss using a weighted sum to form the final joint loss function. Lastly, all our experimental results are individual results, and the experimental environment is PyTorch 2.0.1 and CUDA 11.8.

D. Ablation Experiments

To comprehensively evaluate the performance of different modules, Table I lists the conducted ablation experiments on two datasets under different configurations. The baseline model uses ResNet50 as the backbone network and models only local contextual information at the decoder. Baseline + GCFM represents the use of only the GCFM, whereas Baseline + CSTB represents the use of only the Channel-Spatial Transformer Block. To

TABLE III
RESULTS FROM EXPERIMENTS CONDUCTED ON THE VAIHINGEN TEST SET

Method	Per-class IoU (%)						Evaluate Metrics (%)		
	Imp.surf	Building	Low.veg.	Tree	Car	Clutter	mF1	OA	mIoU
PSPNet	84.76	89.25	70.56	80.13	72.30	44.95	83.95	87.88	73.66
BiSeNetV1	83.94	87.75	68.73	78.96	69.60	42.09	82.63	88.77	71.84
BiSeNetV2	81.14	85.56	66.65	78.10	60.27	39.90	80.29	87.52	68.60
MAResU-Net	84.92	90.26	70.31	79.84	79.58	39.67	83.92	89.75	74.10
MsanlfNet	84.20	88.85	68.12	78.89	71.31	38.02	82.19	88.93	71.57
ABCNet	84.86	89.72	69.14	79.85	76.80	45.81	84.44	89.49	74.37
MANet	85.37	90.50	70.18	80.12	**80.23**	41.49	84.37	89.90	74.65
BANet	82.82	86.61	67.42	78.19	65.67	39.36	81.25	88.09	70.01
PVT	80.77	85.31	67.71	75.77	57.12	36.60	82.05	88.87	71.79
UnetFormer	84.71	89.22	69.59	80.22	75.44	42.90	83.86	89.49	73.68
CMTFNet	85.89	90.91	71.44	80.29	79.66	42.84	84.78	90.23	75.17
BuildFormer	83.44	87.83	68.50	79.15	72.99	41.54	82.87	88.75	72.24
MCCANet	**85.98**	90.95	71.07	80.79	80.05	44.96	85.19	90.31	75.63
GCDNet	85.39	90.40	70.24	80.11	79.24	44.48	84.76	89.87	74.98
Ours	85.35	**90.96**	**71.57**	**80.88**	79.01	**50.44**	**85.93**	**90.29**	**76.37**

The best values are highlighted in bold.

demonstrate the contribution of the improved MLP in CSTB, we also construct a simple variant Baseline + CS by removing MLP and using the attention mechanism only.

1) Baseline + GCFM: At the decoder, the interactive fusion of features from different branches can be accomplished by using GCFM to establish the correlation between the features and enable the network to better extract the global contextual information. mIoU is improved by 0.69% and 0.59% on the Potsdam and Vaihingen datasets, respectively, and this improvement proves the effectiveness of the GCFM module.
2) Baseline + CAM: CAM can learn channel weights to weigh different channels of the input feature map, enabling the model to dynamically select and adjust the importance of channels. This leads to significant improvements over the Baseline on two datasets.
3) Baseline + SAM: SAM models the relationships between different positions to enhance the model's perception ability and better understand the semantic information in the image. It improves mIoU by at least 0.56% on two datasets.
4) Baseline + CS: By effectively aggregating complementary features using the channel-spatial attention module, the network gains more discriminative features. This leads to significant improvements of 1.03% in mFl and 0.97% in mIoU on the Potsdam dataset, showcasing the effectiveness of CS.
5) Baseline + CSTB: CSTB is applied at different positions in the encoder. It fuses local details, channels, and global information, enabling the network to capture more comprehensive semantic information. On both datasets, the performance of mFl and mIoU has improved by at least 1.1% and 1.47%, respectively. Furthermore, the mIoU of "Baseline + CSTB" is higher than that of "Baseline + CS" by 0.5%/0.27%, highlighting the necessity of CSTB.
6) Baseline + CSTB + GCFM: This efficient fusion approach utilizes three CSTBs and three GCFMs, enabling the network to understand and analyze images from different perspectives. It significantly suppresses the impact

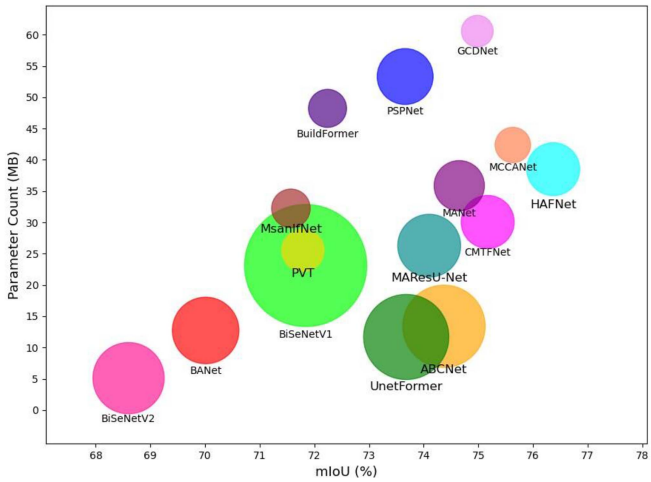

Fig. 8. Visualization of network performance evaluation.

TABLE IV
COMPARISON OF COMPLEXITIES BETWEEN OUR METHOD AND OTHER METHODS

Method	mF1(%)	mIoU(%)	Params(M)	Flops(G)
PSPNet	83.95	73.66	53.32	201.59
BiSeNetV1	82.63	71.84	23.11	40.90
BiSeNetV2	80.29	68.60	**5.10**	**11.19**
MAResU-Net	83.92	74.10	26.77	35.11
MsanlfNet	82.19	71.57	32.24	27.81
ABCNet	84.44	74.37	11.68	15.63
MANet	84.37	74.65	35.86	77.76
BANet	81.25	70.01	28.58	58.10
PVT	82.05	71.79	25.52	-
UnetFormer	83.86	73.68	11.68	11.74
CMTFNet	84.78	75.17	30.68	33.62
BuildFormer	82.87	72.24	48.11	126.76
MCCANet	85.19	75.63	42.38	102.53
GCDNet	84.76	74.98	60.56	281.13
Ours	**85.93**	**76.37**	38.51	114.64

The best values are highlighted in bold.

TABLE V
ABLATION STUDIES OF DIVERSE ATTENTION MECHANISMS ON THE VAIHINGEN TEST SET

Attention Mechanism	Per-class IoU (%)						Evaluate Metrics (%)		
	Imp.surf	Building	Low.veg.	Tree	Car	Clutter	mF1	OA	mIoU
+SCSE	85.18	90.26	71.02	80.43	76.51	45.55	84.73	89.93	74.83
+CBAM	85.30	90.49	70.68	81.11	77.95	46.57	85.10	90.10	75.35
+DA	**85.82**	90.82	**71.71**	**81.27**	78.99	43.98	85.02	**90.39**	75.43
+KAM&CAM	85.72	90.94	71.31	81.08	78.91	45.48	85.18	90.31	75.57
+EGLA	85.65	90.59	71.38	80.77	**79.54**	45.30	85.16	90.22	75.54
+M2SA	85.33	90.84	71.13	81.05	77.93	48.77	85.52	90.23	75.84
+SEAA	85.59	90.67	71.27	81.15	78.69	44.95	85.04	90.24	75.39
Ours	85.35	**90.96**	71.57	80.88	79.01	**50.44**	**85.93**	90.29	**76.37**

The best values are highlighted in bold.

of redundant feature information and achieves the highest accuracy on both datasets. This also further validates the feasibility of our designed decoder.

In summary, the experimental results not only show the effectiveness of GCFM but also demonstrate the essential role of CGTB. Finally, our method not only achieves the best values on the three metrics but also brings an improvement of at least 1.87% compared to the baseline mIoU on the two public datasets. In addition, we visualize the results of the ablation experiments, as shown in Fig. 7. The segmentation performance on both datasets gradually improves.

E. Comparing With Existing Works

We have also conducted extensive experiments on ISPRS Potsdam and Vaihingen datasets. Also, to ensure a fair comparison, all experiments were performed under the same training and testing setup. We compare our method with PSPNet [13], BiSeNet V1 [14], BiSeNet V2 [15], MAResU-Net [59], MsanlfNet [26], MANet [25], ABCNet [60], BANet [45], UnetFormer [61], PVT [62], CMTFNet [63], BuildFormer [64], MCCANet [65], and GCDNet [66]. As shown in Table II, the proposed HAFNet achieves the best F1, OA, and mIoU metrics

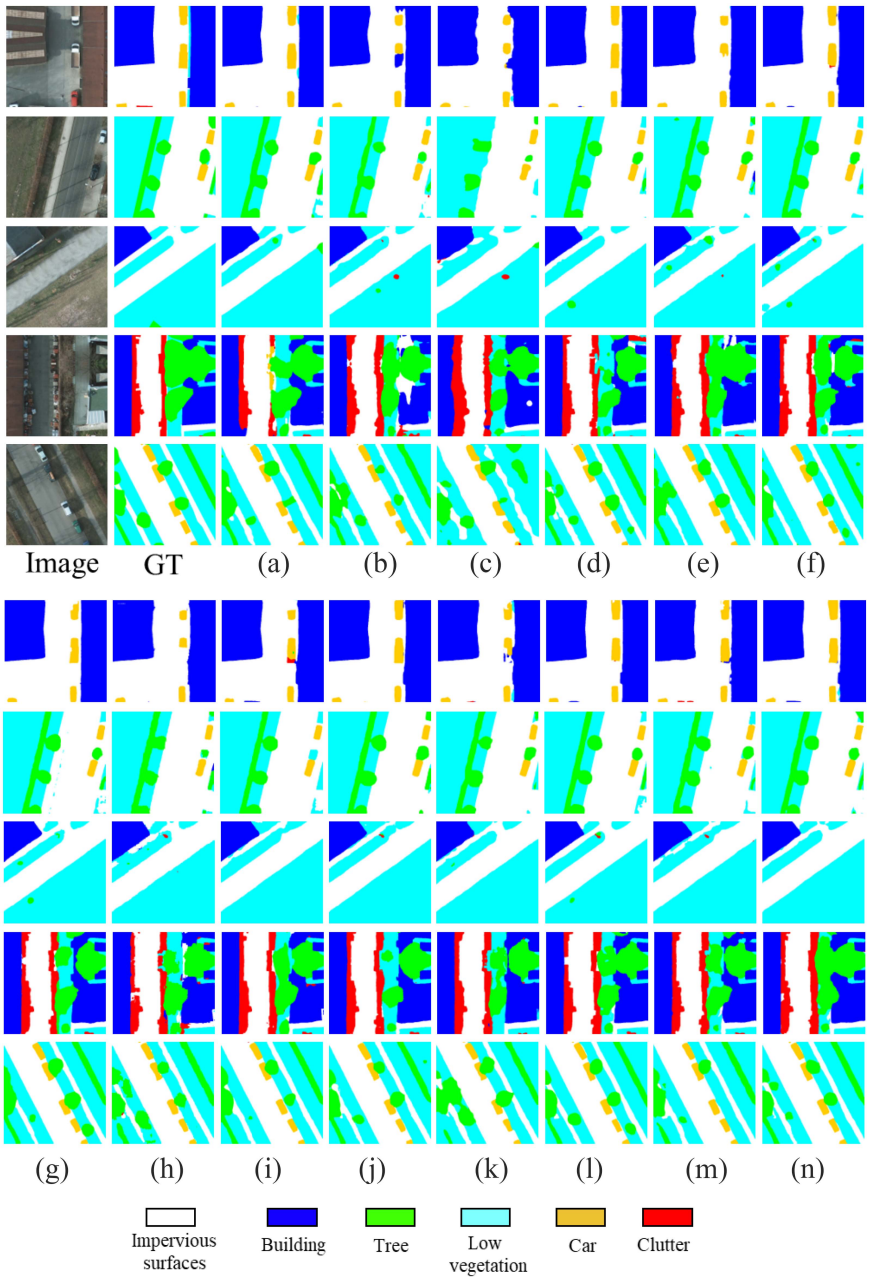

Fig. 9. Visualization comparisons on the Potsdam test set. (a)-(n) : PSPNet, BiSeNetV1, BiSeNetV1, MAResU-Net, MsanlfNet, ABCNet, MANet, BANet, UnetFormer, CMTFNet, BuildFormer, MCCANet, GCDNet, HAFNet.

on the Potsdam dataset, significantly outperforming other CNN and Transformer based networks. MCCANet captures channel attention at various scales through multiscale channelwise cross attention, allowing dynamic and adaptive feature fusion in a context-scale aware manner, thus focusing on both large and small objects distributed throughout the input. It achieves the best results on tasks involving small objects such as cars. CMTFNet combines the advantages of CNN and Transformer by using attention to learn multiscale feature representations and efficiently aggregate deep and shallow features, achieving the best results on the category of buildings. It is worth noting that our method's IoU on car and building categories is slightly lower than the best value by around 0.35%, whereas the remaining four categories achieve the best results.

To further illustrate the performance of our proposed approach, we conducted the same comparative experiments on the Vaihingen dataset. In Table III, our HAFNet achieved mF1 of 85.93% and mIoU of 76.37% on the Vaihingen dataset. It is worth noting that because of the unequal class distribution in the Vaihingen dataset, the Clutter class often poses significant challenges during prediction. However, our method performs exceptionally well in the Clutter class, with an IoU surpassing other networks by more than 4.63%. In addition, especially in the categories of Low Vegetation and Trees, which have complex

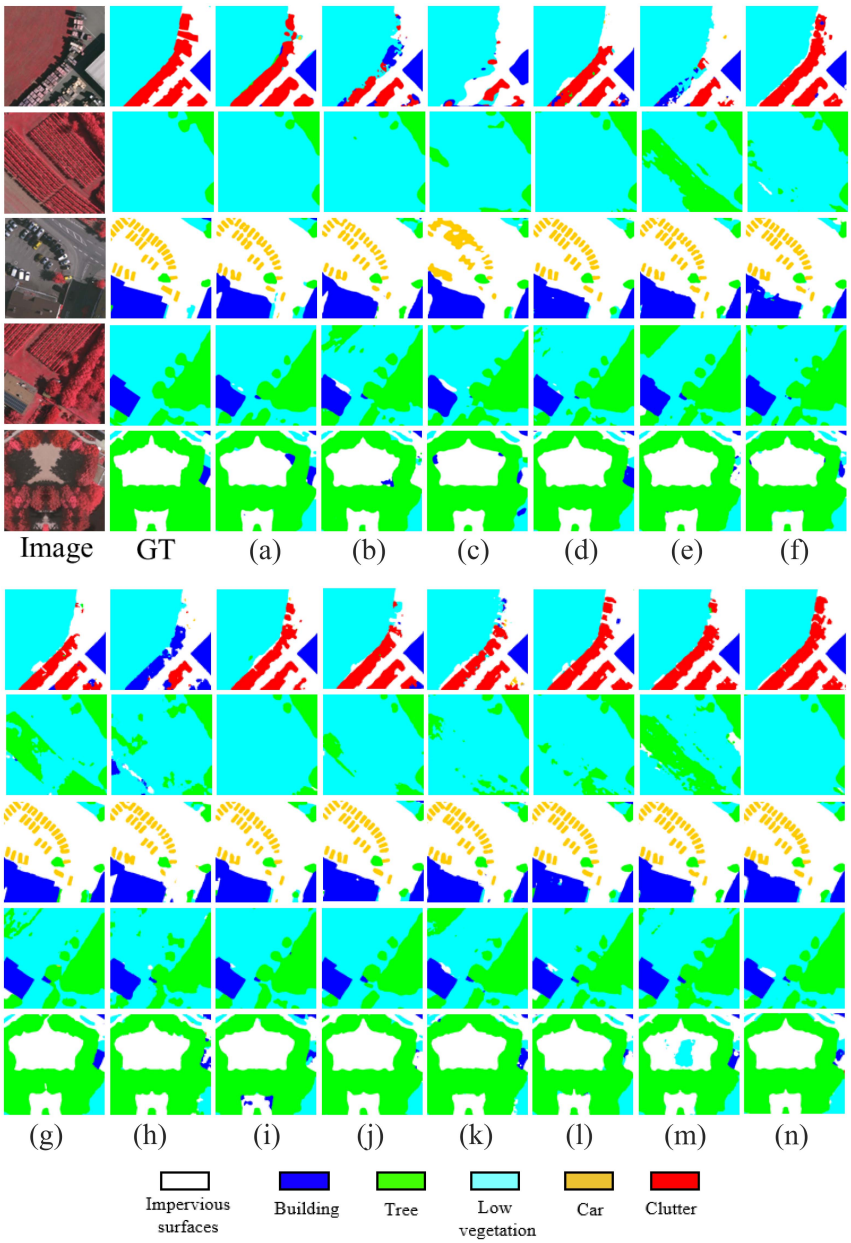

Fig. 10. Visualization comparisons on the Vaihingen test set. (a)-(n) : PSPNet, BiSeNetV1, BiSeNetV1, MAResU-Net, MsanlfNet, ABCNet, MANet, BANet, UnetFormer, CMTFNet, BuildFormer, MCCANet, GCDNet, HAFNet.

features, our approach yields the best results on both datasets. This further proves that HAFNet can alleviate interference and error propagation among categories, thereby improving the network's response capability to differences between samples. Furthermore, we also conducted a simple performance evaluation of HAFNet, as shown in Fig. 8. The circles in the figure denote FPS, and the larger the circle, the higher the FPS value. We also provide the model's parameter quantity and computational complexity in Table IV. Based on the comprehensive analysis of the presented figures and table, although our model has achieved a favorable balance between performance and efficiency, BiSeNetV2 is a highly efficient network for lightweight real-time semantic segmentation. Although its precision is not as high as our proposed method, it exhibits significant advantages in terms of Params, FPS, and Flops. This is precisely where the limitation of our method lies.

To verify the effectiveness of the proposed channel-spatial attention, we conducted ablation experiments by replacing it with several other cutting-edge attention mechanisms while keeping other modules unchanged. These include concurrent spatial and channel squeeze and excitation (SCSE) [67], CBAM [23], dual attention (DA) [35], kernel attention mechanism and channel attention mechanism (KAM&CAM) [25], efficient global-local attention (EGLA) [61], multiscale multihead self-attention (M2SA) [63], and squeeze-enhanced axial attention (SEAA) [68]. Due to the efficient coupling of the channel-spatial attention module in HAFNet, which enhances the network's understanding of images, our method consistently achieves the

best mF1 and mIoU on the Vaihingen dataset, as shown in Table V. Although our channel-spatial attention did not reach the highest IoU values in the four categories, it was only lower by about 0.38% from the best value. Significantly, our method still outperforms other attention mechanisms by over 1.67% in the clutter category, which further demonstrates the strong learning ability of our approach, as well as its superiority and robustness in handling tasks with few samples. This also provides valuable references and inspiration for applying our approach to other fields.

F. Qualitative Analysis of the Segmentation Results

As shown in Figs. 9 and 10, we have provided visualization results on two public remote sensing datasets, denoted as a–n in the following order: PSPNet, BiSeNetV1, BiSeNetV1, MAResU-Net, MsanlfNet, ABCNet, MANet, BANet, UnetFormer, CMTFNet, BuildFormer, MCCANet, GCDNet, HAFNet. First, the segmentation results on the Potsdam test set are shown in Fig. 9, where CSTB leverages the contextual information and interdependencies within feature maps from different layers to support the network obtain more comprehensive semantic information, which enables the segmentation results to better preserve the geometric details and complex contours. For regular round and regular shaped objects, the segmentation results from our method are clearer and have smoother edges than other methods. From the segmentation results of the first three figures, our method not only suppresses the interference of background features well but also predicts outcomes very close to ground truth. Next, the segmentation outcomes on the Vaihingen test set are shown in Fig. 10, where GCFM fuses semantic information at different scales, effectively mitigates semantic gaps between features, and precisely controls features with fine and coarse-grained, and demonstrates a high degree of granularity in the segmentation results of the first two graphs, allowing HAFNet to outperform other methods in dealing with the problems of missegmentation and category confusion. Especially in the first figure, when dealing with the "clutter category", our method is better at extracting the objects' full contours than other methods. In addition, for objects with complex texture features such as trees and low vegetation, the final two segmentation result images in Figs. 9 and 10 vividly showcase the outstanding performance of our approach.

V. Conclusion

This article proposes an inimitable decoder based on the Transformer architecture with ResNet50 as the encoder. Our designed CSTB can effectively integrate local details, channels, and global information, reducing the interference of redundant feature information and enhancing the significant feature representation of objects. The GCFM performs an interactive fusion of features from different branches, establishing correlations between various elements and improving the model's understanding of contextual information to acquire a more comprehensive understanding of semantic information. The experimental results on the ISPRS Potsdam and Vaihingen datasets demonstrate that HAFNet outperforms the compared baselines. HAFNet also performs exceptionally well in handling complex objects with similar texture features, such as low vegetation and trees, effectively addressing the issue of class confusion. In addition, it also effectively alleviates the problem of class imbalance in remote sensing images. Finally, comprehensive ablation studies have provided evidence of the effectiveness of every component in the proposed approach. In addition, our research only focuses on how to improve segmentation accuracy, whereas there are still deficiencies in the number of model parameters and computational complexity. Therefore, in future research, we will continue to explore how to fully utilize the respective advantages of CNN and Transformer to further improve the model efficiency and performance by addressing this aspect of the problem.

References

[1] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019, doi: 10.1016/j.isprsjprs.2019.04.015.

[2] P. Wang, C. Huang, J. C. Tilton, B. Tan, and E. C. B. de Colstoun, "HOTEX: An approach for global mapping of human built-up and settlement extent," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 1562–1565, doi: 10.1109/IGARSS.2017.8127268.

[3] G. Chen, Q. Weng, G. J. Hay, and Y. He, "Geographic object-based image analysis (GEOBIA): Emerging trends and future opportunities," *GISci. Remote Sens.*, vol. 55, no. 2, pp. 159–182, Mar. 2018, doi: 10.1080/15481603.2018.1426092.

[4] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 277–293, Aug. 2017, doi: 10.1016/j.isprsjprs.2017.06.001.

[5] D. Phiri and J. Morgenroth, "Developments in Landsat land cover classification methods: A review," *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 967, [Online]. Available: https://www.mdpi.com/2072-4292/9/9/967

[6] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114417, doi: 10.1016/j.eswa.2020.114417.

[7] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3376–3385.

[8] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2650–2658.

[9] J. Lv, Q. Shen, M. Lv, Y. Li, L. Shi, and P. Zhang, "Deep learning-based semantic segmentation of remote sensing images: A review," *Front. Ecol. Evol.*, vol. 11, Jul. 2023, Art. no. 1201125, doi: 10.3389/fevo.2023.1201125.

[10] 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.

[11] 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, doi: 10.1109/TPAMI.2016.2644615.

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

[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[14] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

[15] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021, doi: 10.1007/s11263-021-01515-2.

[16] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012, doi: 10.1109/tgrs.2022.3207551.

[17] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2245–2255, Mar. 2021, doi: 10.1109/tgrs.2020.3006872.

[18] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212, doi: 10.1109/tgrs.2022.3233847.

[19] D. Hong et al., "SpectralGPT: Spectral Foundation model," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.07113

[20] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal Deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415, doi: 10.1109/tgrs.2023.3284671.

[21] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5526914, doi: 10.1109/tgrs.2022.3166252.

[22] M. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, Sep. 2022, doi: 10.1007/s41095-022-0271-y.

[23] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, vol. 11211, pp. 3–19.

[24] 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, doi: 10.1109/lgrs.2021.3052886.

[25] 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, doi: 10.1109/TGRS.2021.3093977.

[26] L. Bai, X. Lin, Z. Ye, D. Xue, C. Yao, and M. Hui, "MsanlfNet: Semantic segmentation network with multi-scale attention and nonlocal filters for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6512405, doi: 10.1109/lgrs.2022.3185641.

[27] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113856, doi: 10.1016/j.rse.2023.113856.

[28] Z. Xu, J. Geng, and W. Jiang, "MMT: Mixed-mask transformer for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613415, doi: 10.1109/tgrs.2023.3289408.

[29] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[31] Y. Sun, Y. Tian, and Y. Xu, "Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning," *Neurocomputing*, vol. 330, pp. 297–304, Feb. 2019, doi: 10.1016/j.neucom.2018.11.051.

[32] 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. Photogramm. Remote Sens.*, vol. 145, pp. 78–95, Nov. 2018, doi: 10.1016/j.isprsjprs.2017.12.007.

[33] Y. Shen, J. Chen, L. Xiao, and D. Pan, "Optimizing multi-scale segmentation with local spectral heterogeneity measure for high resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 157, pp. 13–25, Nov. 2019, doi: 10.1016/j.isprsjprs.2019.08.014.

[34] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

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

[36] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 173–190, doi: 10.1007/978-3-030-58539-6_11.

[37] J. Wang, W. Li, M. Zhang, and J. Chanussot, "Large kernel sparse ConvNet weighted by multi-frequency attention for remote sensing scene understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5626112, doi: 10.1109/TGRS.2023.3333401.

[38] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021, doi: 10.1109/tgrs.2020.3007921.

[39] H. Zhang, J. Yao, L. Ni, L. Gao, and M. Huang, "Multimodal attention-aware convolutional neural networks for classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3635–3644, 2023, doi: 10.1109/jstars.2022.3187730.

[40] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021, doi: 10.1109/tgrs.2020.2994150.

[41] R. Liu, L. Mi, and Z. Chen, "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7871–7886, Sep. 2021, doi: 10.1109/tgrs.2020.3034123.

[42] W. Cheng, W. Yang, M. Wang, G. Wang, and J. Chen, "Context aggregation network for semantic labeling in aerial images," *Remote Sens.*, vol. 11, no. 10, May 2019, Art. no. 1158, doi: 10.3390/rs11101158.

[43] M. Yang, S. Kumaar, Y. Lyu, and F. Nex, "Real-time semantic segmentation with context aggregation network," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 124–134, Aug. 2021, doi: 10.1016/j.isprsjprs.2021.06.006.

[44] Z. Chen, J. Zhao, and H. Deng, "Global multi-attention UResNeXt for semantic segmentation of high-resolution remote sensing images," *Remote Sens.*, vol. 15, no. 7, Apr. 2023, Art. no. 1836, doi: 10.3390/rs15071836.

[45] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, Aug. 2021, Art. no. 3065, doi: 10.3390/rs13163065.

[46] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12077–12090, 2021.

[47] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2022, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.

[48] 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.

[49] A. Zhao, C. Wang, and X. Li, "A global$^+$ multiscale hybrid network for hyperspectral image classification," *Remote Sens. Lett.*, vol. 14, no. 9, pp. 1002–1010, Sep. 2023, doi: 10.1080/2150704x.2023.2258467.

[50] 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, doi: 10.1109/lgrs.2022.3143368.

[51] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[52] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Transformer-based decoder designs for semantic segmentation on remotely sensed images," *Remote Sens.*, vol. 13, no. 24, Dec. 2021, Art. no. 5100, doi: 10.3390/rs13245100.

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

[54] J. Chen et al., "TransUnet: Transformers make strong encoders for medical image segmentation," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2102.04306

[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Pro. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[56] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11534–11542.

[57] S. Woo et al., "ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16133–16142.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[59] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8009205, doi: 10.1109/lgrs.2021.3063381.

[60] 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. Photogramm. Remote Sens.*, vol. 181, pp. 84–98, Nov. 2021, doi: 10.1016/j.isprsjprs.2021.09.005.

[61] L. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022, doi: 10.1016/j.isprsjprs.2022.06.008.

[62] S. Du and M. Liu, "Class-guidance network based on the Pyramid Vision transformer for efficient semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5578–5589, 2023, doi: 10.1109/jstars.2023.3285632.
[63] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multi-scale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art, no. 2004612, doi: 10.1109/tgrs.2023.3314641.
[64] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711, doi: 10.1109/tgrs.2022.3186634.
[65] J. Zheng, A. Shao, Y. Yan, J. Wu, and M. Zhang, "Remote sensing semantic segmentation via boundary supervision-aided Multiscale Channelwise Cross attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4405814, doi: 10.1109/tgrs.2023.3292112.
[66] J. Cui, J. Liu, J. Wang, and Y. Ni, "Global context dependencies Aware network for efficient semantic segmentation of fine-resolution remoted sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2505205, doi: 10.1109/lgrs.2023.3318348.
[67] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Intervention*, 2018, pp. 421–429, doi: 10.1007/978-3-030-00928-1_48.
[68] Q. Wan, Z. Huang, J. Lu, G. Yu, and L. Zhang, "Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation," in *Proc. 11th Int. Conf. Learn. Representations*, 2023.

Yan Chen received the M.Sc. degree in cartography and geographical information system from the China University of Mining & Technology, Xuzhou, China, in 2014, and the Ph.D. degree in spatial information management and modeling from TU Dortmund University, Dortmund, Germany, in 2019.

At TU Dortmund University, he studied spatial information management and modeling as a member of the Spatial Information Management and Modelling Department of the School of Spatial Planning. In 2022, he joined the Collaborative Innovation Centre for Computer Vision and Pattern Recognition within the School of Artificial Intelligence and Big Data, Hefei University. His research interests include remote sensing image analysis, computer vision and pattern recognition, and deep learning and optimization algorithms.

Quan Dong received the B.Sc. degree in computer science and technology from Nanjing Normal University Zhongbei College, Nanjing, China, in 2022. He is currently working toward the M.Sc. degree in electronic information with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China.

His current research focuses on the semantic segmentation of remote-sensing images.

Xiaofeng Wang received the B.Sc. degree in software engineering from Anhui University, Hefei, China, in 1999, the M.Sc. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2005, and the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2009, both in pattern recognition and intelligent system.

He is currently a Professor with the School of Artificial Intelligence and Big Data, Hefei University and a provincial academic and technical leader reserve candidate. His research interests include computer vision and pattern recognition and image processing.

Qianchuan Zhang received the B.Sc. degree in computer science and technology from Sichuan Minzu College, Kangding, China, in 2020. He is currently working toward the M.Sc. degree in electronic information with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China.

His current research focuses on the intelligent parsing of remote sensing images and deep learning.

Menglei Kang received the B.Sc. degree in computer and information engineering from Chuzhou University, Chuzou, China, in 2021. He is currently working toward the M.Sc. degree in electronic information with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China.

His current research focuses on the semantic segmentation of remote-sensing images.

Wenxiang Jiang received the B.Sc. degree in computer and information engineering from the Anhui University of Technology, Ma'anshan, China, in 2021. He is currently working toward the M.Sc. degree in electronic information with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China.

His research interests include computer vision in deep learning and remote sensing image analysis.

Mengyuan Wang received the B.Sc. degree in Internet of Things engineering from Applied Technology College of Soochow University, Suzhou, China, in 2020. She is currently working toward the M.Sc. degree in electronic information with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China.

Her research interests include semantic segmentation of high-resolution remote sensing images and deep learning.

Lixiang Xu received the B.Sc. degree from Fuyang Normal University, Fuyang, China, in 2005, the M.Sc. degree from the Harbin University of Science and Technology, Harbin, China, in 2008, both in applied mathematics, and the Ph.D. degree in computer application technology from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2017.

In 2008, he was with Huawei Technologies Co., Ltd., before joining Hefei University in the following year. From 2015 to 2017, he was awarded a scholarship to pursue his studies in Germany as a joint Ph.D. student. He is currently a postdoctoral researcher with the University of Science and Technology of China, Hefei, China. His research interests include structural pattern recognition, machine learning, graph spectral analysis, image and graph matching, especially in kernel methods, and complexity analysis on graphs and networks.

Chen Zhang was born in Anhui, China. She received the M.S. degree in computational mathematics from Anhui University, Hefei, China, in 2011, and the Ph.D. degree in information management and systems from Hefei University of Technology, Hefei, China, in 2016.

She is currently an Associate Professor with the School of Artificial Intelligence and Big Data, Hefei University, Hefei, China. Her research interests include machine learning and artificial intelligence.