Multimodal Fusion Remote Sensing Semantic Segmentation Based on CNN and Transformer

Youwei Cheng[1], Lixin Chen[1], Runxin Dai[1], Youbiao Wang[2], Bin Xie[2]
23020241154383,23020241154341,23020241154385,36920241153251,36920241153262

[1]Information School Class
[2]AI Class

Abstract

With the rapid advancements in remote sensing technology, remote sensing semantic segmentation has found widespread application in areas such as land cover mapping and urban change detection. Compared to traditional single-modal segmentation techniques, multi-modal fusion-based segmentation models have demonstrated superior performance and garnered considerable attention in recent years. However, many of these models rely on Convolutional Neural Networks (CNNs) or Vision Transformer (ViT) for fusion operations, leading to limited capabilities in modeling and representing local-global context. In this study, we propose a multi-modal fusion method that integrates CNNs and ViT within a unified framework, offering an efficient solution for remote sensing semantic segmentation. First, shallow features are extracted and fused using convolutional layers and shallow feature fusion (SFF) modules. Next, deep features representing semantic information and spatial relationships are extracted through a specially designed deep feature fusion (DFF) module. The DFF module comprises the self-attention (SA) layers and mutually boosted attention (MBA) layers, where MBA computes SA and cross-attention (CA) in parallel, enhancing both intra-modal and cross-modal contextual information while directing attention to semantically relevant regions. Therefore, the proposed method is capable of fusing shallow and deep features at multiple layers, fully leveraging CNNs to accurately represent local details and transformers to capture global semantics. Extensive experiments conducted on publicly available high-resolution remote sensing dataset validate the effectiveness and superiority of the proposed method.

1.Introduction

Image semantic segmentation is a classic task in computer vision, aiming to divide an image into multiple regions and assign a semantic label to each pixel to identify the category to which the pixel belongs. With the continuous development of technology, semantic segmentation has been widely applied in the field of remote sensing imagery processing. The semantic segmentation of urban scene images has driven various city-related applications, including land cover mapping(Li et al. 2022), change detection(Xing, Sieber, and Caelli 2018) and road and building extraction(Griffiths and Boehm 2019).

Although significant progress has been made in semantic segmentation technology for remote sensing images, multiple challenges remain, such as multi-scale and resolution variations, complex backgrounds and diversity of ground objects, and the impact of noise and occlusion on image quality. Traditional semantic segmentation algorithms, such as K-nearest neighbors (KNN)(Larose and Larose 2014) and random forests(Rigatti 2017), have achieved certain results in early computer vision tasks but are limited in performance when dealing with complex scenes. In recent years, the development of deep learning has greatly advanced the field of image semantic segmentation, with convolutional neural networks (CNNs) making significant progress in particular. However, CNNs rely on local receptive fields for convolution operations, which limits their ability to capture global contextual information. The introduction of the Vision Transformer (ViT)(Han et al. 2022) has improved this issue. ViT utilizes attention mechanisms, allowing the model to capture relationships between pixels at arbitrary positions within the image, making it more effective in handling global information and achieving remarkable results.

Compared to single-modal data provided by a single sensor, multi-modal data processing(Roche et al. 2021) demonstrates clear advantages in many computer vision applications. By combining information from different modalities, models can obtain richer and more diverse features, thereby improving overall performance. However, the differences and incompatibility between different modalities pose challenges for multi-modal data fusion. Typically, three strategies are employed: early fusion, mid-level fusion, and late fusion. Early fusion relies on precise alignment of multi-modal data but is less robust in handling task-irrelevant information. Late fusion processes different modalities independently before merging them, but it falls short in capturing the cross-modal relationships of the data. In contrast, mid-level fusion can capture cross-modal dependencies at intermediate feature levels, making it more effective in representation learning tasks. However, most existing research adopts simple summation or concatenation-based single-layer feature fusion methods, neglecting the modeling of long-range cross-modal dependencies across different feature levels, which limits the potential of multi-modal fusion.

To address the aforementioned challenges, we propose a

cross-modal fusion network that integrates CNN and ViT. First, shallow features are extracted and fused using convolutional layers and shallow feature fusion (SFF) modules. Next, deep features representing semantic information and spatial relationships are extracted through a specially designed deep feature fusion module (DFF). Finally, the obtained shallow and deep features were sent to a decoder for feature fusion and upsampling, restoring the input image size.

2.Related Work

2.1.CNN-based Semantic Segmentation Methods

The Fully Convolutional Network (FCN) was the first effective CNN architecture to address semantic segmentation in an end-to-end manner. Since then, CNN-based methods have dominated semantic segmentation tasks in the field of remote sensing(Kotaridis and Lazaridou 2021). However, due to the overly simplified decoder in FCN, segmentation resolution is limited, impacting both image fidelity and accuracy.

To address this issue, an encoder-decoder network called UNet was proposed for semantic segmentation, featuring two symmetrical paths known as the contracting path and the expanding path(Ronneberger, Fischer, and Brox 2015). Building on the encoder-decoder structure, researchers have designed various skip connections to capture richer context(Diakogiannis et al. 2020). However, the limited receptive field of CNN-based segmentation networks restricts their ability to capture local semantic features, lacking the capacity to model global information across the entire image. Given that remote sensing image scenes are more complex, identifying these intricate targets poses a significant challenge.

2.2.Transformer-based Semantic Segmentation Methods

The Vision Transformer (ViT)(Alexey 2020)is an innovative model capable of effectively extracting global feature information, particularly beneficial in the field of remote sensing image semantic segmentation. Researchers have extensively explored methods for applying ViT to semantic segmentation tasks. For instance, ViT-Adapter(Chen et al. 2022) is a flexible framework that incorporates image-related prior knowledge into the pre-trained ViT network, enabling more effective handling of complex scenes. SETR(Zheng et al. 2021), another Transformer-based segmentation model, treats semantic segmentation as a sequence-to-sequence problem, achieving efficient segmentation.

However, ViT also presents certain limitations in semantic segmentation tasks. Firstly, when processing high-resolution images, ViT's computational complexity grows quadratically with the increase in input sequence length, which restricts its feasibility in practical applications. Secondly, ViT is heavily dependent on large-scale datasets, which can lead to reduced performance in tasks with limited data. Additionally, ViT's global receptive field may overlook local detail information.

2.3.Multimodal Semantic Segmentation Based on CNN and Transformer

With the application of advanced remote sensing technologies, the field of remote sensing can now comprehensively acquire a broad spectrum of multimodal data, including hyperspectral imaging (HSI), visible light imaging (VIS), and LiDAR. Multimodal semantic segmentation networks, by integrating the distinct information from various modalities, have significantly enhanced segmentation accuracy. The Sigma network(Wan et al.)utilizes a Selective Structured State-Space Model (Mamba), achieving global receptive field coverage with linear complexity. LET-Net(Ta et al. 2023) effectively combines U-shaped CNNs with Transformers, using capsule embeddings to address each model's shortcomings.

Although these approaches perform well, the optimal integration of CNNs and Transformers for feature extraction and fusion has not been fully explored, limiting the network's ability to capture both local details and global context, thus affecting the quality of feature representation. To address this, this paper explores the integration of Transformers and CNNs with multimodal remote sensing data to achieve more robust semantic segmentation.

3.Proposed Solution

Figure 1 illustrates the structure of the proposed framework. Specifically, the CNN backbone incorporates the SFF module for feature fusion, while the MBA layer in DFF enables deep feature fusion. More specifically, our method extracts features from VIS and DSM using two separate ResNet branches. The shallow features at varying scales are enhanced by the SFF modules following each CNN block. These features are then flattened into sequences and passed through DFF, where deep-level feature fusion is performed via MBA to capture additional contextual information. The fused outputs from DFF are summed and reshaped before being sent to a cascaded upsampler decoder, which helps restore spatial details with higher precision. In the following sections, we will provide a detailed explanation of the key components of our method.

3.1.Shallow Feature Fusion

We denote VIS images and DSM images by $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ and $\mathbf{Y} \in \mathbb{R}^{H\times W\times 1}$, where H and W represent the height and width of the images, respectively. Our method employs a dual-branch encoder to extract multi-scale features from each modality. Each branch consists of four convolutional layers for multi-scale feature extraction, producing downsampled feature maps of size $\left(\frac{H}{2^{i-1}}\right) \times \left(\frac{W}{2^{i-1}}\right) \times C_i$,where C_i represents the number of channels at the i-th layer of the encoder. These shallow features are then fused through the SFF module, where the features from the auxiliary DSM modality are integrated into the VIS modality before being passed to the next branch of the VIS encoder. Additionally, skip connections are utilized, directly feeding the outputs of the SFF module into the corresponding decoder layers to recover local and contextual information.

As shown in Figure 2, the SFF module first aggregates global information by applying Global Average Pooling (AvgPool) separately on the VIS and DSM branches. Subsequently, the squeeze and excitation process is carried out using AvgPool followed by two 1×1 convolutional layers, with ReLU and Sigmoid activation functions applied. Finally, the features from VIS and DSM are weighted and element-wise summed, resulting in the final fused shallow-level features.

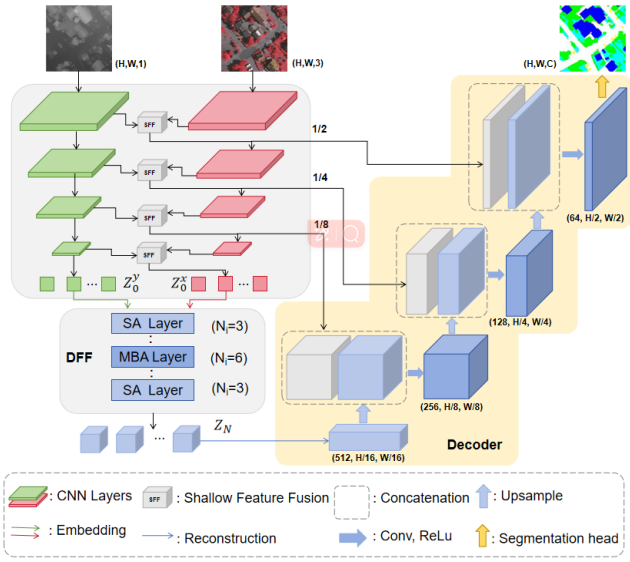

Figure 1: Overview of our method

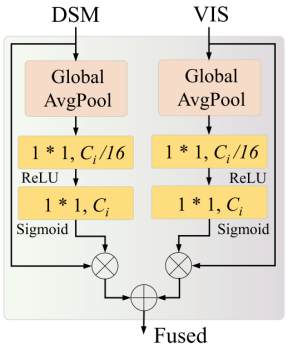

Figure 2: Proposed SFF module for SFF in the CNN blocks.

3.2.Deep Feature Fusion

$\mathbf{x}_I$ and $\mathbf{y}_I$ are the feature maps of VIS and DSM, respectively, obtained from a certain layer of the convolutional neural network (CNN). The dimensions of these feature maps are $\left(\frac{H}{2^{I-1}}\right) \times \left(\frac{W}{2^{I-1}}\right) \times C_I$, where I is the layer index, and C_I is the number of output channels from the last layer of the CNN.The feature maps $\mathbf{x}_I$ and $\mathbf{y}_I$ are first passed through two embedding layers to change the channel size from C_I to C_{hid} for further processing. These embedded feature maps are then flattened into two 2D patch sequences of length L, denoted as z_0^x and z_0^y. At this point, the feature map size becomes $C_{\text{hid}} \times L$, where $L = \frac{H \times W}{2^{(I-1)} \times 2^{(I-1)}}$.To retain the position information of the image patches in the original image, position embeddings are added to z_0^x and z_0^y. Finally, the position-encoded and embedded sequences z_0^x and z_0^y are passed as inputs to the DFF module.

The input to the DFF module sequentially undergoes three processes: the SA layers for enhancing deep-level feature representations, the MBA layers for deep-level feature fusion, and the SA layers for further enhancement of the fused feature representations. The number of layers for each process is N_1, N_2, and N_3, respectively. In each layer, the features from the VIS and DSM branches are processed simultaneously, with z_n^x and z_n^y denoting the hidden features at the n-th layer, where $n \in \{1, 2, \ldots, N_1 + N_2 + N_3\}$. Throughout the entire process, the DFF module maintains the feature map dimensions as $C_{\text{hid}} \times L$. The SA layers consist of multiple modules, including two SA modules, two multilayer perceptron (MLP) modules, and Layer Normalization (LN) operations,as shown in Figure 3(a)

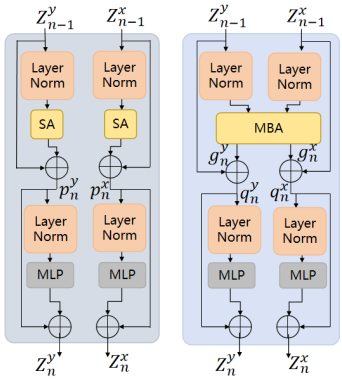

Figure 3: (a) Proposed SA layer in DFF. (b) Proposed MBA layer in DFF

Given the multimodal feature inputs $z_{(n-1)}^x$ and $z_{(n-1)}^y$, the global relationships within the features are learned using the multihead self-attention mechanism. The output of the n-th layer SA can be written as:

$$p_n^x = \text{SA}(\text{LN}(z_{(n-1)}^x)) + z_{(n-1)}^x$$

$$p_n^y = \text{SA}(\text{LN}(z_{(n-1)}^y)) + z_{(n-1)}^y$$

$$z_n^x = \text{MLP}(\text{LN}(p_n^x)) + p_n^x$$

$$z_n^y = \text{MLP}(\text{LN}(p_n^y)) + p_n^y$$

After the feature enhancement through the SA layers, the MBA layers further fuse the multimodal features in an abstract semantic space and leverage rich contextual information,as shown in Figure 3(b) .During the deep feature fusion stage, the MBA module computes both CA and SA information simultaneously to learn the correlation between the two modalities. The output of the MBA layer can be written as:

$$(g_n^x, g_n^y) = \text{MBA}(\text{LN}(z_{(n-1)}^x), \text{LN}(z_{(n-1)}^y))$$

Let $q_n^x = g_n^x + z_{(n-1)}^x$, $q_n^y = g_n^y + z_{(n-1)}^y$, the fused output at the n-th layer is then:

$$z_n^x = \text{MLP}(\text{LN}(q_n^x)) + q_n^x$$

$$z_n^y = \text{MLP}(\text{LN}(q_n^y)) + q_n^y$$

With the support of multihead design, the MBA module divides the multimodal feature inputs $z_{(n-1)}^x, z_{(n-1)}^y$ into H equal segments, denoted as $z_{(n-1,h)}^x, z_{(n-1,h)}^y$, where $h = 1, 2, \ldots, H$. The projection matrices for the multimodal information, $\{Q_x, K_x, V_x\}$ and $\{Q_y, K_y, V_y\}$, are obtained through linear projections. Then, SA information (sa_x, sa_y) and CA information (ca_x, ca_y) are derived for both modalities.Figure 4 illustrates the structure of the proposed MBA module. Specifically, the SA module uses $\{Q_x, K_x, V_x\}$ and $\{Q_y, K_y, V_y\}$ to compute intra-modal features, while the CA module computes cross-modal associations.

Next, an adaptive fusion mechanism is proposed to fuse the SA and CA features:

$$g_n^x = \lambda_{\text{sa}}^x \, \text{sa}_x + \lambda_{\text{ca}}^x \, \text{ca}_x, \quad g_n^y = \lambda_{\text{sa}}^y \, \text{sa}_y + \lambda_{\text{ca}}^y \, \text{ca}_y$$

where λ_{sa}^x, λ_{ca}^x, λ_{sa}^y, and λ_{ca}^y are learnable weighting coefficients used to balance the contributions from SA and CA. Finally, the fused features are further enhanced through the SA layers. The final output of DFF, denoted as $z_N \in \mathbb{R}^{C_{\text{hid}} \times L}$, is the feature map derived from the last SA layer.

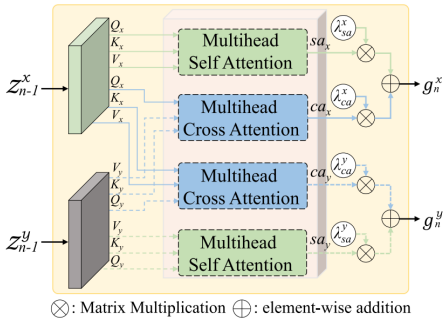

Figure 4: Schematic of the MBA module

3.3.Cascaded Decoder

The Cascaded Decoder uses multiple upsampling modules to recover the fused features from the hidden layers, thereby completing the final segmentation process. The decoder first uses a reconstruction module to convert the 2D input sequence z_N into a 3D tensor of size $C_{\text{dec}} \times \frac{H}{2^{(I-1)}} \times \frac{W}{2^{(I-1)}}$, where C_{dec} is the number of input channels for the first block of the decoder. Then, multiple cascaded decoder blocks gradually restore the spatial resolution to $H \times W$ by concatenating skip connections from the corresponding CNN backbone layers. Each decoder block consists of an upsampling operation, a convolution layer, and a ReLU layer. Finally, the segmentation head performs the final semantic prediction.

4.Experiments

Dataset. The Vaihingen dataset consists of 16 very high-resolution orthophotos, with an average image size of 2500 × 2000 pixels. Each orthophoto has three channels—near-infrared, red, and green (NIRRG)—along with a normalized digital surface model (DSM). The ground sampling distance (GSD) is 9 cm. The dataset includes five foreground classes: buildings (Bui.), trees (Tre.), low vegetation (Low.), cars (Car), impermeable surfaces (Imp.), and one background class (Clutter). The 16 orthophotos are split into a training set with 12 patches and a test set with 4 patches. Specifically, the training set includes images numbered 1, 3, 23, 26, 7, 11, 13, 28, 17, 32, 34, and 37, while the test set includes images numbered 5, 21, 15, and 30.

Evaluation metrics. To evaluate segmentation results of multimodal remote sensing data, overall accuracy (OA), mean F1 score (mF1), and mean Intersection over Union (mIoU) are used. These metrics provide a fair basis for comparing our method with other state-of-the-art approaches.

Implementation details. All experiments were implemented using PyTorch on a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. Stochastic gradient descent (SGD) was used for model training with a learning rate of 0.01, momentum of 0.9, weight decay of 0.0005, and batch size of 10. After collecting samples using sliding windows, simple data augmentation techniques such as random rotation and flipping were applied.The method we proposed uses a CNN backbone composed of two ResNet50 models, each with four convolutional layers (I = 4) and ahidden layer size of $C_{ild} = 768$. The DFF module contains $N_1 + N_2 + N_3 = 12$ transformer layers, distributed as $N_1 = 3, N_2 = 6$, and $N_3 = 3$. Each layer has $H = 12$ attention heads, with a channel size of $C_{dec} = 512$.Finally, all transformer backbones and ResNet50 were pre-trained on ImageNet for better initialization.

4.1.Performance Comparison

We benchmarked the performance of our proposed method against five representative state-of-the-art approaches, including PSPNet(Zhao et al. 2017), ESANet(Seichter et al. 2021), CMGFNet(Hosseinpour, Samadzadegan, and Javan 2022), TransUNet(Chen et al. 2021), and MFTransNet(He et al. 2023).

As shown in Table 1, our proposed method significantly outperforms the baseline TransUNet in terms of OA, mF1, and mIoU, confirming that our method successfully fused shallow and deep features by extracting complementary information from the assisting modality, resulting in robust representations. Compared to other methods, our proposed method outperforms the others in two categories: Low vegetation and Car. Specifically, our method outperforms the existing MFTransNet by 2.03% in the Low vegetation category. Additionally, the classification accuracy for the Car category is improved by 4.19% over the existing CMGFNet. These results can be attributed to our method's more effective extraction and fusion of multi-level multimodal features through the sequential use of CNN and Transformer. Overall, our method achieved 92.22% OA, 89.00% mF1 score, and 82.11% mIoU, reflecting improvements of 1.22%, 2.89%, and 5.41%, respectively, compared to the baseline TransUNet. These results validate that our method achieves superior generalization performance.

Figure 5 presents the visualization results of all six methods in the experiment. It is evident that remote sensing images are more complex than natural images, for example: 1) Object edges are distinct, but shapes vary, requiring accurate edge segmentation; 2) Small target objects are more

Table 1: Segmentation Results Comparison

Type	Method	OA(%)						mF1(%)	mIoU(%)
		Bui.	Tre.	Low.	Car	Imp.	Total		
CNN-based	PSPNet	**98.39**	94.26	64.97	55.58	90.92	89.45	82.32	71.87
	ESANet	98.34	94.08	75.53	68.19	87.92	90.41	85.52	75.87
	CMGFNet	92.50	93.67	76.23	78.84	92.50	91.75	88.50	80.03
Transformer-based	TransUNet	91.95	**95.11**	72.99	62.63	91.95	91.00	86.11	76.70
	MFTransNet	95.20	93.22	79.57	76.07	**93.23**	91.56	87.36	78.47
	our	97.94	91.60	**81.60**	**83.03**	93.15	**92.22**	**89.00**	**82.11**

difficult to segment. Clearly, our method excels at identifying complex edges, producing smoother results, and providing complete and connected object segmentation with fewer isolated points. Specifically, the SFF module helps preserve details of objects with various scales and shapes, resulting in accurate edges for Building objects. Additionally, the DFF module can more accurately capture complex long-range semantic information, aiding in the identification of complete objects with fewer scattered points. These advantages allow our method to achieve more accurate classification than other approaches. In Figures 5(d)–(i), two red boxes are highlighted. In the upper box, our method clearly delineates the edges of the building, yielding cleaner and more complete building segmentation results. In the lower box, our method efficiently identifies small target objects.So, our method was able to achieve significant overall performance improvement.

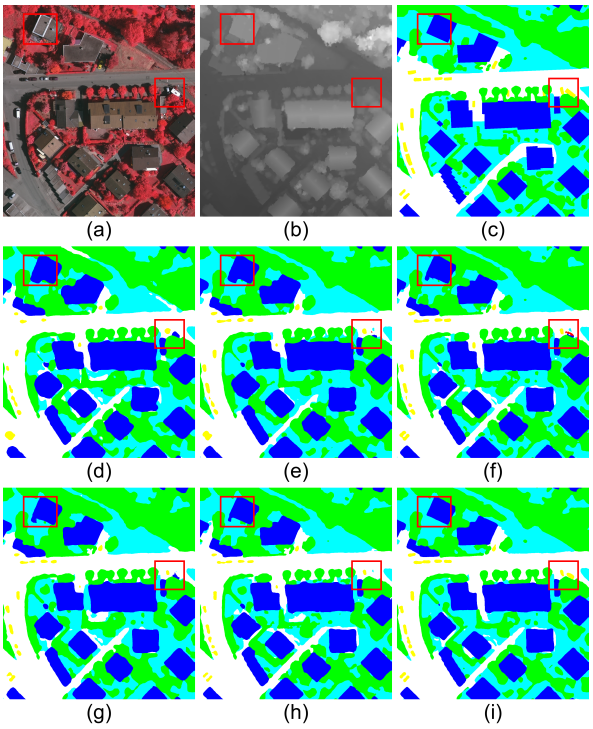

Figure 5: : Qualitative performance comparisons. (a) NIRRG images, (b) DSM, (c) Ground truth, (d)PSPNet, (e) ESANet, (f) CMGFNet, (g) TransUNet, (h) MFTransNet, (i) Our. Two red boxes are added to all subfigures to highlight the differences.

4.2.Ablation Study

To validate the effectiveness of each component in our proposed method, we conducted ablation experiments by systematically removing specific components while maintaining the dual-branch framework. As presented in Table 2, we designed two ablation experiments based on our fusion scheme. In the first experiment, the proposed DFF module was decoupled into two single-modal ViT modules, i.e., two independent self-attention-based transformers, while the SFF module within the CNN block remained unchanged. In contrast, the second experiment removed the SFF module from CNN, with the shallow features of the two modalities being independently extracted by two separate branches.

Table 2 illustrates that both the shallow and deep feature fusion modules are critical for the enhanced performance of our proposed method. Specifically, the SFF module is capable of learning and providing robust representations of the fundamental features of ground objects, such as shape, boundaries, color, and texture, irrespective of scale variations. Furthermore, the DFFF module helps to distinguish complex remote sensing scenes by leveraging the semantic information extracted by the SFF module.

Table 2: Ablation study of the proposed approach.

SFF	DFF	OA(%)	mF1(%)	mIoU(%)
✓		91.91	88.36	81.25
	✓	91.59	88.41	81.32
✓	✓	**92.22**	**89.00**	**82.11**

4.3.Model Complexity Analysis

We assess the computational complexity of our proposed method using the following metrics: floating-point operations (FLOPs), the number of model parameters, memory usage, and frames per second (FPS). FLOPs are employed to assess model complexity, while the number of model parameters and memory usage serve to evaluate memory requirements. Finally, FPS is intended to evaluate execution speed.

Table 3 presents the results of the complexity analysis for all the compared methods. Compared to PSPNet, our proposed method exhibits lower FLOPs, despite having a larger number of parameters. Compared with other methods, although the approach we proposed achieves optimal performance, its complexity is higher. While it outperforms other models, it also has certain limitations. In the future, we will

Table 3: Computational complexity analysis measured on a single Nvidia Geforce RTX 3090 GPU.

Method	Multimodal	FLOPS (G)	Parameter (M)	Memory (MB)	Speed (FPS)	MIoU(%)
PSPNet	N	49.03	46.72	3124	**66.01**	71.87
ESANet	Y	**7.73**	**34.03**	1914	10.42	75.87
TransUNet	Y	32.27	93.23	3028	10.81	76.70
CMGFNet	Y	19.51	64.20	2463	11.61	80.03
MFTransNet	Y	8.44	43.77	**1549**	14.88	78.47
our	Y	45.21	160.88	3463	9.74	**82.11**

further optimize this method to develop a more lightweight version.

5.Conclusion

In this work, we integrate CNN and ViT into a unified framework and propose a multi-layer multi-modal fusion method for remote sensing semantic segmentation. Specifically, we design a CNN-based SFF module to extract and fuse detailed shallow features across multiple scales, followed by the DFF module for deep semantic feature extraction and fusion. The proposed DFF utilizes an MBA module integrated with SA and CA to extract deep features and guide the multi-modal deep feature fusion, enabling effective segmentation of complex remote sensing images. Experiments on the ISPRS Vaihingen dataset show that our method outperforms several other semantic segmentation approaches in multiple metrics. However, there are some limitations in our method, such as suboptimal performance on certain categories, and the model has a large number of parameters and high computational complexity. In the future, we will further optimize this method to achieve higher segmentation accuracy and develop a more lightweight version.

References

Alexey, D. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Chen, J.; Gao, Y.; Yu, L.; Li, K.; Cai, J.; and Heng, P.-A. 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.

Diakogiannis, F. I.; Waldner, F.; Caccetta, P.; and Wu, C. 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162: 94–114.

Griffiths, D.; and Boehm, J. 2019. Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154: 70–83.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

He, S.; Yang, H.; Zhang, X.; and Li, X. 2023. MFTransNet: A Multi-Modal Fusion with CNN-Transformer Network for Semantic Segmentation of HSR Remote Sensing Images. *Mathematics*, 11(3): 722.

Hosseinpour, H.; Samadzadegan, F.; and Javan, F. D. 2022. CMGFNet: A Deep Cross-Modal Gated Fusion Network for Building Extraction from Very High-Resolution Remote Sensing Images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184: 96–115.

Kotaridis, I.; and Lazaridou, M. 2021. Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173: 309–322.

Larose, D. T.; and Larose, C. D. 2014. k-nearest neighbor algorithm.

Li, R.; Zheng, S.; Duan, C.; Wang, L.; and Zhang, C. 2022. Land cover classification from remote sensing images based on multi-scale fully convolutional network. *Geo-spatial information science*, 25(2): 278–294.

Rigatti, S. J. 2017. Random forest. *Journal of Insurance Medicine*, 47(1): 31–39.

Roche, J.; De-Silva, V.; Hook, J.; Moencks, M.; and Kondoz, A. 2021. A multimodal data processing system for LiDAR-based human activity recognition. *IEEE Transactions on Cybernetics*, 52(10): 10027–10040.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Seichter, D.; Köhler, M.; Lewandowski, B.; Wengefeld, T.; and Gross, H.-M. 2021. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 13525–13531.

Ta, N.; Chen, H.; Liu, X.; and Jin, N. 2023. LET-Net: locally enhanced transformer network for medical image segmentation. *Multimedia Systems*, 29(6): 3847–3861.

Wan, Z.; Wang, Y.; Yong, S.; Zhang, P.; Stepputtis, S.; Sycara, K.; and Xie, Y. ???? Sigma: Siamese Mamba Network for Multi-Modal Semantic Segmentation. arXiv 2024. *arXiv preprint arXiv:2404.04256*.

Xing, J.; Sieber, R.; and Caelli, T. 2018. A scale-invariant change detection method for land use/cover change research. *ISPRS Journal of Photogrammetry and Remote Sensing*, 141: 252–264.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid Scene Parsing Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2881–2890.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.