Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 10, No. 3, 2024

Multi-Scale Adaptive Road Extraction Network for High-Resolution Remote Sensing Images

Jiajia Liu *, Chidong Huang

College of Aviation Electronics and Electrical, Civil Aviation Flight University of China, Jianyang, Sichuan, 641400, China
*** Corresponding author:** Jiajia Liu (Email: cafucljj@163.com)

Abstract: Road extraction from high-resolution remote sensing images faces challenges such as discontinuities caused by occlusions from trees and buildings, as well as computational complexity arising from varying road orientations and widths. In order to solve these problems, this paper proposes a Multi-Scale Adaptive Road Extraction Network (MARENet), which takes ConvNeXt as the backbone and designs the Atrous Pyramid Convolution (APC) module, which can accurately extract the roads in the image, and the Frequency Adaptive Transformer (FAT) module, which can focus on the important features in the image and avoid the problem of incomplete extraction. The network can effectively solve the above two key problems and achieve accurate road extraction. EARENet was evaluated experimentally on a commonly used and challenging DeepGlobe dataset. The proposed method is superior to the existing methods in terms of accuracy, precision, recall, F1-score and IoU, and the results show that the proposed network has robustness and superiority in processing high-resolution road remote sensing images.

Keywords: Remote Sensing Image; Road Extraction; Semantic Segmentation; Atrous Pyramid Convolution; Attention Mechanisms.

1. Introduction

Remote sensing images have critical applications in urban planning, navigation, disaster management, and autonomous driving[1]. Road extraction from remote sensing images has thus become a prominent research focus in both the remote sensing and computer vision domains[2]. However, challenges such as occlusions from vegetation, shadows cast by buildings, and the diverse appearance of roads across different environments make this task highly challenging[3]. To improve the effectiveness of road extraction from remote sensing images, numerous methods have been developed, ranging from traditional image processing techniques to modern deep learning approaches.

With the advancement of deep learning, road extraction tasks have seen significant improvements. Convolutional neural network (CNN) models, such as Fully Convolutional Network (FCN) and U-Net, marked a paradigm shift from pixel-based methods to learning-based methods, enabling the automatic extraction of hierarchical features from images[4]. Following the success of basic CNN models, advanced architectures like D-LinkNet[5] introduced additional techniques such as dilated convolutions, which expand the receptive field while maintaining computational efficiency. By combining the LinkNet encoder-decoder architecture with dilated convolutions, D-LinkNet achieved state-of-the-art performance on several road extraction tasks, including the DeepGlobe Challenge.

Transformer models, initially developed for natural language processing, have been adapted for computer vision tasks, including road extraction[6] For example, Seg-Road[7] leverages the self-attention mechanism to capture long-range dependencies and global contextual information, which are critical for understanding complex road networks. Similarly, the Swin Transformer[8] demonstrated that hierarchical attention mechanisms can further enhance road extraction accuracy by processing images at multiple resolutions[9].

The integration of CNNs and Transformer models has proven particularly effective for road extraction[10]. The Next-ViT model[11] combines the strengths of both architectures, using CNNs for efficient local feature extraction and Transformers for capturing global relationships. This hybrid approach enhances the model's ability to process multi-frequency signals, enabling the capture of high-level structural features and fine-grained road details[12].

Despite the success of many network-based approaches in road extraction, challenges remain. These include continuity issues caused by occlusions from trees and buildings, as well as difficulties arising from varying road orientations and widths. To address these challenges, this study proposes MARENet, which combines the strengths of CNNs and Transformers. The backbone of the proposed network is ConvNeXt[13], and it incorporates two key modules: the APC and the FAT. The APC module enables multi-scale feature representation, facilitating efficient multi-scale feature aggregation and enhancing the network's ability to handle varying road orientations and widths. Meanwhile, the FAT module helps the model focus on the most relevant features in the image, effectively capturing both low-frequency and high-frequency information. This addresses road discontinuities caused by occlusions and improves segmentation accuracy by enhancing road continuity.

The proposed network was evaluated on the DeepGlobe dataset, which is widely used for road extraction tasks in remote sensing images. Extensive experiments demonstrated that MARENet achieved advanced performance on the DeepGlobe dataset, surpassing most existing methods across evaluation metrics.

2. Method

2.1. Overall Network Architecture

This study aims to develop a semantic segmentation method for road extraction in high-resolution remote sensing images. The structure of the proposed MARENet is shown in

Figure 1. The backbone of MARENet for feature extraction is ConvNeXt, and the network integrates the designed APC module and FAT module. Additionally, a decoder structure based on D-LinkNet is employed. The goal is to address the challenges of road extraction in complex environments by fusing multi-scale contextual information and multi-frequency signals, achieving efficient and accurate segmentation performance.

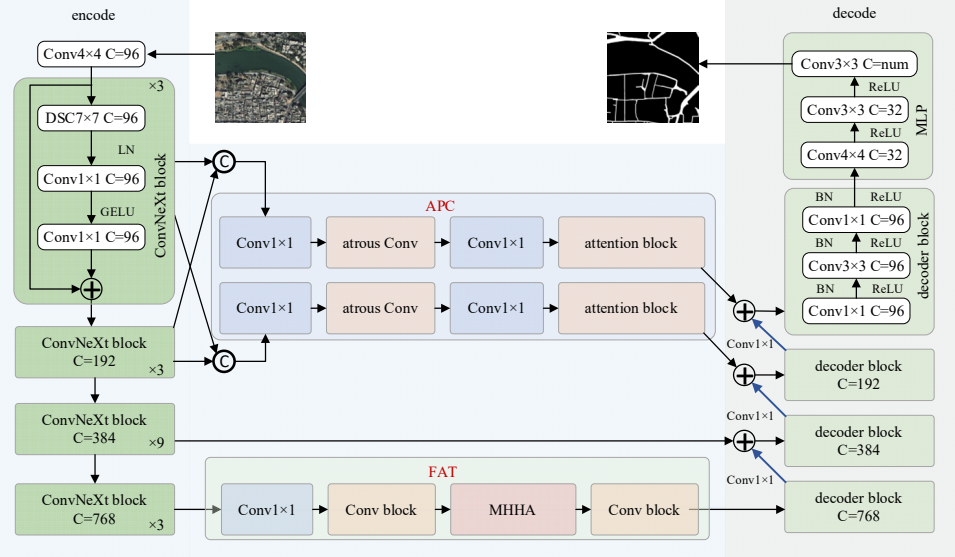

Fig 1. Multi-Scale Adaptive Road Extraction Network

MARENet employs ConvNeXt as the encoder, which consists of a 4×4 convolution followed by four ConvNeXt blocks. Each ConvNeXt block contains one 7×7 depthwise separable convolution (DSC) and two 1×1 convolutions, arranged with a residual connection structure. ConvNeXt blocks of different dimensions are designed with either 3 layers or 9 layers, as annotated in the figure.To further enhance multi-scale feature representation, the network incorporates the APC and FAT modules. The APC module consists of two 1×1 convolutions, dilated convolutions with varying dilation rates, and an attention mechanism module. The FAT module is composed of a 1×1 convolution, a convolution block (CB), and a hybrid multi-head attention mechanism (HMHA). In the decoder, decoder blocks progressively perform scale transformations on feature maps of different dimensions. These transformations are performed using 1×1 convolutions to adjust the scales, and the feature maps are added step by step. Finally, an MLP composed of convolution layers is employed to complete the road extraction task. This design ensures the effective integration of features from different scales and dimensions, leading to accurate and robust road extraction.

2.2. Atrous Pyramid Convolution Module

In this work, the structure of the proposed APC module is illustrated in Figure 2. The main design of the APC module includes dilated convolutions with varying dilation rates and an attention mechanism. In road extraction tasks, dilated convolutions enable the model to capture large-scale contextual information while preserving fine details, making it well-suited for handling complex road structures. The attention mechanism adaptively focuses on the critical feature regions within the image, enhancing the resolution of key areas and thereby allowing for more accurate identification of road boundaries. The combination of these two components improves the precision and robustness of road extraction.

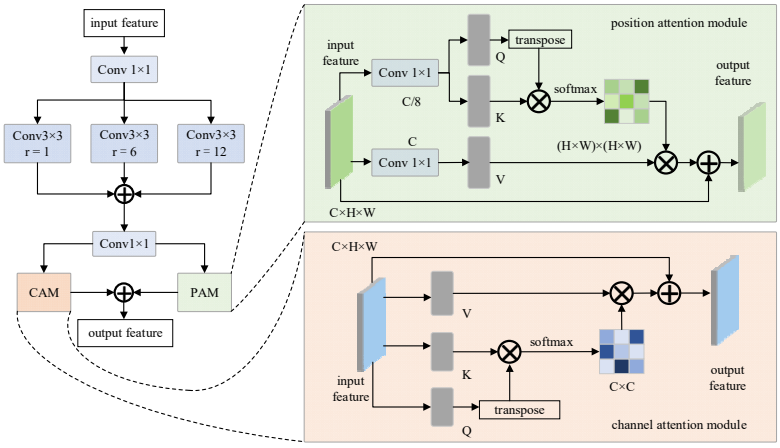

Fig 2. Atrous Pyramid Convolution (APC) Module

The APC module first applies a 1×1 convolution to perform dimensional transformation on the input features, followed by three parallel 3×3 dilated convolutions for feature extraction, with dilation rates set to 1, 6, and 12, respectively. This configuration allows the model to capture feature representations at multiple scales without significantly

increasing the number of parameters. The APC module achieves computational efficiency while capturing broader spatial dependencies. The output of the APC module is calculated as follows:

$$F_{APC} = \sum_{i=1}^{1,6,12} Conv_{3\times3}^{rate(i)}(F) \tag{1}$$

Here F represents the input feature map, and F_{APC} represents the output feature map. $Conv_{3\times3}^{rate(i)}$ represents the 3×3 convolution operation with a dilation rate of i, This design enables the module to simultaneously capture details at different levels across the entire feature map, thereby improving the performance of road extraction.

2.2.1. Channel Attention Mechanism

The core idea of the channel attention mechanism (CAM) is to enhance feature representation by capturing the global dependencies between different channels[14]. First, the input features are flattened into a two-dimensional matrix, where the channels are represented as one-dimensional vectors, and the spatial dimensions are flattened into the other dimension. The key matrix (K) is obtained through a transpose operation, enabling it to perform matrix multiplication with the query matrix (Q), resulting in a channel correlation matrix. This matrix has dimensions C×C, representing the inter-dependencies between channels. Next, the attention matrix is normalized using the softmax function to ensure the numerical stability of the generated attention weights. Then, the attention weights are applied to the value matrix (V) through a weighted summation to obtain enhanced channel features. Finally, the enhanced features are reshaped back to their original dimensions of C×H×W.

This process captures long-range dependencies between channels, significantly improving the network's ability to model global features. The CAM demonstrates notable effectiveness in enhancing the robustness of semantic feature representation.

2.2.2. Positional Attention Mechanism

In addition to CAM, the ACP module introduces a positional attention mechanism (PAM) to highlight important spatial regions in the feature map. First, the input feature map is processed through two 1×1 convolutions, reducing the number of channels to C/8 to generate the Q and K, while another convolution maintains the original number of channels C to generate the V. These convolution operations preserve the spatial dimensions of the feature map. Next, Q and K are flattened into matrices with a channel dimension of C/8. Their similarity is calculated through a matrix dot product to produce an attention matrix, which represents the correlation between each position and all other positions. To stabilize training, the similarity results are normalized using the softmax function. The attention matrix is then applied to the feature map to optimize the model's ability to focus on target regions. The attention matrix and weighted calculation formula are as follows:

$$A(i,j) = soft\max\left(\frac{Q(i)\bullet K(j)}{\sqrt{d_k}}\right) \tag{2}$$

Where $A(i,j)$ represents the attention weight between the i position and the j position, and $\sqrt{d_k}$ is the scaling factor used to stabilize training.

$$O(i) = \sum_j A(i,j)\bullet V(j) \tag{3}$$

The attention weights $A(i,j)$ are used to perform a weighted summation over $V(j)$, resulting in the position-enhanced features $O(i)$.

This method allows the network to adaptively emphasize spatial regions that are critical for the segmentation task, thereby improving the accuracy of boundary delineation and object recognition. The final output of the APC module is obtained by combining the channel and positional attention feature maps. This fusion of multi-scale, channel, and positional attention information enables the model to more effectively capture complex relationships and contextual connections within the data. The APC module provides a comprehensive feature aggregation strategy, enhancing the network's ability to perform accurate and robust semantic segmentation, especially in tasks involving varying object scales and spatial structures.

2.3. Frequency Adaptive Transformer module

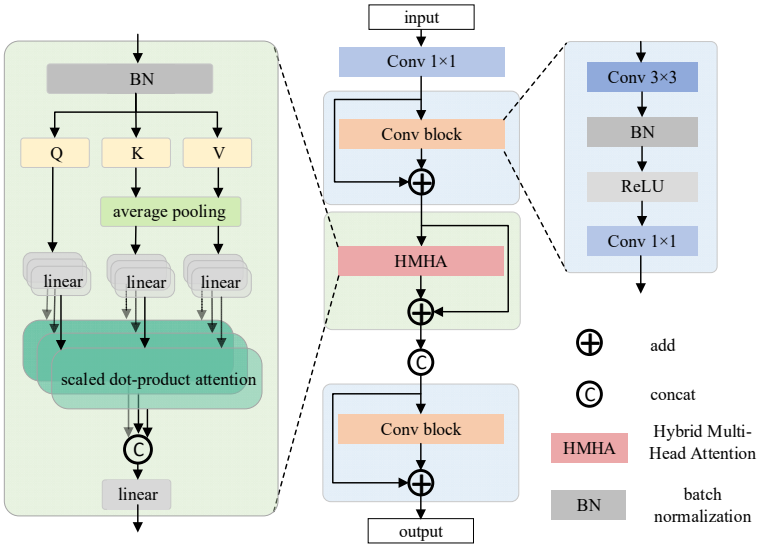

Fig 3. Frequency Adaptive Transformer (FAT) Module

Transformer modules are particularly effective at capturing low-frequency signals, which provide contextual information and help in understanding the overall shape of roads[15]. However, previous studies have shown that Transformer blocks might weaken high-frequency signals, such as local detailed textures. Since both high-frequency and low-frequency signals are critical for human vision, it is necessary to design an efficient combination module to capture comprehensive image features.

Based on these considerations, the study modified the Transformer module to design the FAT module, as shown in Figure 3, which can effectively handle information from multiple frequency bands. The FAT module begins with a 1×1 convolution for scale transformation. The CB module, designed in a residual connection manner, processes high-frequency information to preserve finer local details. In addition to the CB module, FAT integrates a HMHA module, which is inserted between two CB modules. In the HMHA module, batch normalization is first applied, followed by the generation of Q, K, and V features. The K and V features are enhanced for global information through average pooling before being combined with the Q features and input into a linear transformation layer. These three components then interact and aggregate features through a scaled dot-product attention mechanism to extract important information and achieve adaptive information fusion. The CB and HMHA modules work collaboratively to capture signals of different frequencies. After the outputs of the two modules are fused, an MLP layer is employed to refine the fused features and enhance key representations. In summary, FAT effectively combines multi-frequency signals, significantly improving the model's performance.

3. Experiments

3.1. Implementation Details

The experiments were conducted using randomly cropped 512×512 resolution windows from the DeepGlobe dataset for network training. Data augmentation techniques such as image rotation, flipping, and mirroring were applied. The Adam optimizer was used, with a batch size of 4 and an initial learning rate of 2e-4. A total of 90 epochs were trained. The training of the MARENet model was implemented in PyTorch and conducted on a single NVIDIA 2080 SUPER GPU for all experiments.

During the prediction phase, test-time augmentation was employed, which included horizontal flipping, vertical flipping, and diagonal flipping of images. The outputs were then restored to match the original image orientations. The predicted probabilities from each augmented output were averaged, and a threshold of 0.5 was applied to generate binary outputs. In the final binary output, roads were labeled as foreground, while other objects were labeled as background.

3.2. Dataset

The DeepGlobe dataset contains 6,226 road remote sensing images, with images originating from three different regions. The ground resolution is 50 cm/pixel, and the pixel resolution is 1024×1024. The dataset was divided into 4,980 training images, 996 validation images, and 250 test images for the experiments. This dataset is formulated as a binary segmentation problem, where roads are labeled as the foreground and other objects are labeled as the background. It is one of the most widely used datasets for road extraction.

3.3. Method Comparison

In this section, the road segmentation performance of MARENet is quantitatively and qualitatively compared with existing methods. The performance is compared against A2FPNet [16], ABCNet[17], BASNet[18], D-LinkNet[5], DSNet [19], FFNet[20], and UNet-former[21].

3.3.1. Quantitative Results

To evaluate the performance of the proposed model, the experiment adopted five key evaluation metrics: Accuracy, Precision, Recall, F1-score, and Intersection over Union (IoU). Accuracy measures the proportion of pixels in the entire image that are correctly classified by the model. Precision quantifies the proportion of false positives in the results, while Recall measures the proportion of false negatives. F1-score is the harmonic mean of Precision and Recall, balancing the trade-off between the two. Additionally, IoU measures the overlap between the segmentation results and the ground truth, making it one of the core metrics for evaluating the performance of semantic segmentation models. The quantitative data is summarized in Table 1.

Table 1. Comparison results between MARENet and other methods

method	Accuracy	Precision	Recall	F1-score	IoU
A2FPNet[16]	98.01	75.77	76.01	75.89	61.05
ABCNet[17]	98.02	76.94	74.20	75.55	60.55
BASNet[18]	98.37	80.36	78.78	79.56	66.05
D-LinkNet[5]	98.16	77.89	75.98	76.92	62.37
DSNet[19]	98.16	77.36	77.46	77.41	63.02
FFNet[20]	98.04	75.92	75.68	75.79	60.89
UNetformer[21]	97.97	73.51	78.14	75.75	60.95
MARENet	**98.46**	**81.84**	**81.66**	**81.75**	**69.10**

From Table 1, it can be seen that MARENet achieves optimal results in terms of precision, accuracy, recall, F1-score, and IoU on the DeepGlobe dataset. These quantitative results confirm that MARENet is capable of capturing precise road information, enabling accurate road extraction from remote sensing images.

3.3.2. Qualitative Results

For qualitative analysis, the study employed recently proposed networks for model training and selected various types of images from the **DeepGlobe dataset** for segmentation result comparisons. These images were categorized into two groups based on **occlusion conditions** and **other complex road scenarios** to facilitate comparison

of the segmentation results.

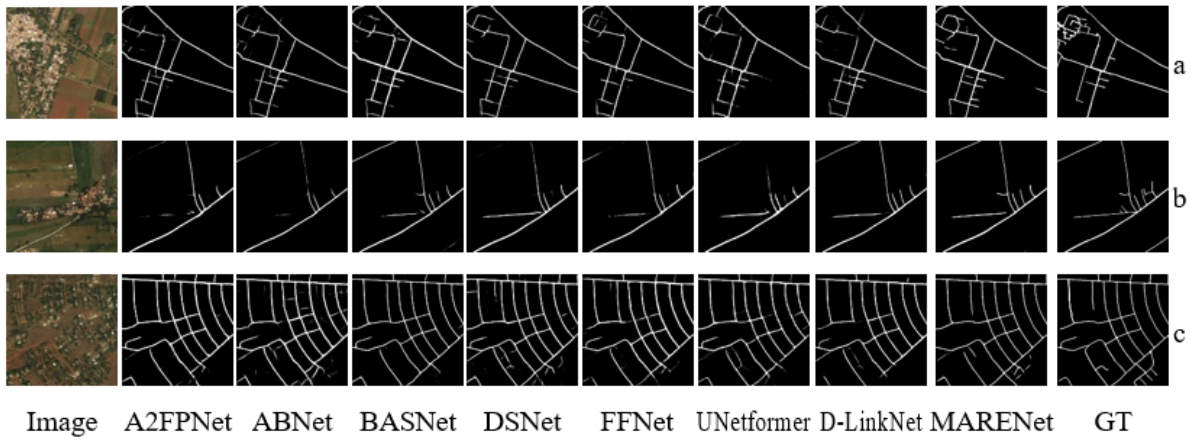

Fig 4. Comparison of road extraction effects when the road is occluded

From Figure 4, in groups a, b, and c, it can be observed that MARENet outperforms other methods in road extraction under conditions where roads are occluded. It demonstrates superior detail preservation and continuity in road extraction. When roads are occluded by trees and buildings, MARENet leverages global contextual information to infer road conditions, ensuring continuity in road extraction. This set of comparative experiments highlights that MARENet effectively addresses the issue of continuity caused by occlusions from trees and buildings, showcasing its robustness in handling such challenges.

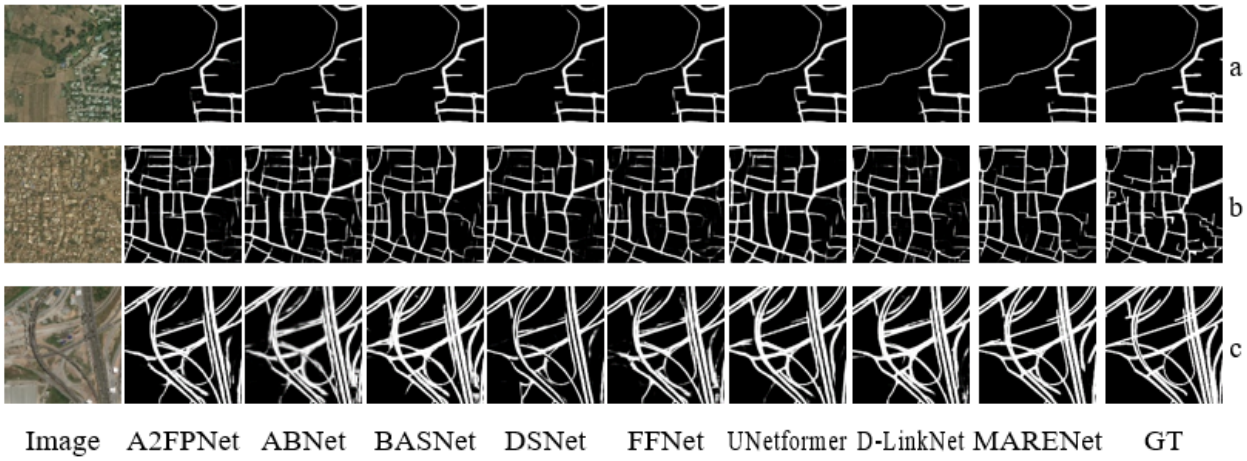

Fig 5. Comparison of the extraction effect of roads with different directions and widths

From Figure 5, it is evident that the complexity of the roads is high. In groups a, b, and c, there are cases where roads differ in orientation and width, as well as situations with frequent road intersections, which place high demands on road extraction networks. Comparing the results from groups a, b, and c, it can be observed that the road extraction results of MARENet are very similar to the ground truth (GT). The road extraction performance of MARENet outperforms other methods, and its advantages are visually apparent in the figure. The comparison of road extraction results demonstrates that MARENet performs well in handling complex road scenarios, adapting to different road conditions, and addressing challenges caused by varying road orientations and widths.

3.4. Ablation Study

An ablation study was conducted to demonstrate the impact of the FAT module and APC module on road segmentation. As shown in Table 2, combining the FAT and APC modules individually with the ConvNeXt backbone significantly improves road extraction performance. Furthermore, integrating both FAT and APC modules into the network results in superior performance compared to using a single module. In terms of global context processing, the FAT module demonstrates strong global feature extraction capabilities, which benefit road segmentation continuity. In terms of local details, the APC module excels at extracting fine-grained features, effectively distinguishing road boundary information and handling roads with different orientations.

The experimental results confirm that the proposed MARENet has the ability to extract both local features and global contextual features, leading to significantly improved segmentation performance.

Table 2. Ablation experiments

method	F1	IoU
ConvNext only	80.45	67.17
ConvNext + FAT	81.49	68.64
ConvNext + APC	81.56	68.78
ConvNext + FAT APC	81.75	69.10

4. Conclusion

In this paper, a novel MARENet is proposed to address the

complex challenges in road extraction tasks. The architecture employs ConvNeXt as the backbone for feature extraction, enabling efficient capture of both local and global information in images, thereby providing a strong foundation for subsequent modules. To enhance the model's performance, two key modules were designed: the APC and FAT modules. The APC module, through dilated convolutions with varying rates, significantly improves the capture of multi-scale contextual information, allowing the model to handle diverse road shapes and scales. The FAT module combines the global feature extraction capabilities of Transformers with the local detail extraction strengths of CNNs, enabling the model to preserve road details while capturing long-range dependencies, thus reducing the impact of road occlusions and ensuring the continuity of road extraction.

By optimizing the handling of different feature signals, MARENet achieves a good balance between global contextual understanding and local feature precision, effectively overcoming the over-smoothing issues observed in previous Transformer-based models. Evaluation results on benchmark datasets such as DeepGlobe demonstrate that the network achieves state-of-the-art performance in terms of accuracy, precision, recall, F1-score, and IoU. These results highlight its robustness in diverse and complex environments and its potential for real-world applications in fields such as autonomous driving, disaster management, and urban planning.

References

[1] Li Y H, Wang M, Su X P, et al. Road extraction from remote sensing images combining attention and context fusion[J/OL]. Journal of Jilin University(Engineering and Technology Edition), 1-10[2024-10-07].https://doi.org/ 10.13229/j. cnki. jdxbgxb. 20240442.

[2] Mo S, Shi Y, Yuan Q, et al. A Survey of Deep Learning Road Extraction Algorithms Using High-Resolution Remote Sensing Images[J]. Sensors, 2024, 24(5): 1708.

[3] Zhao L, Guo D D, Wang Q Q, et al. Deep learning based road extraction from remote sensing images[J]. Modern Electronics Technique, 2023,46(23):48-54. DOI:10.16652/j.issn.1004-373x. 2023. 23.009.

[4] Li K, Tan M, Xiao D, et al. Research on road extraction from high-resolution remote sensing images based on improved UNet++[J]. IEEE Access, 2024.

[5] Zhou L, Zhang C, Wu M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 182-186.

[6] Liu X, Wu Y, Liang W, et al. High resolution SAR image classification using global-local network structure based on vision transformer and CNN[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[7] Tao J, Chen Z, Sun Z, et al. SEG-Road: a segmentation network for road extraction based on transformer and CNN with connectivity structures[J]. Remote Sensing, 2023, 15(6): 1602.

[8] Liu Z, Hu H, Lin Y, et al. Swin transformer v2: Scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12009-12019.

[9] Zhu X, Huang X, Cao W, et al. Road Extraction from Remote Sensing Imagery with Spatial Attention Based on Swin Transformer[J]. Remote Sensing, 2024, 16(7): 1183.

[10] Wang J, Zeng Z, Sharma P K, et al. Dual-path network combining CNN and transformer for pavement crack segmentation [J]. Automation in Construction, 2024, 158: 105217.

[11] Li J, Xia X, Li W, et al. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios[J]. arXiv preprint arXiv:2207.05501, 2022.

[12] Hu X, Zhong B, Liang Q, et al. Transformer tracking via frequency fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(2): 1020-1031.

[13] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[14] Zhang W, Zhao W, Li J, et al. CVANet: Cascaded visual attention network for single image super-resolution[J]. Neural Networks, 2024, 170: 622-634.

[15] Huan H, Zhang B. FDAENet: frequency domain attention encoder-decoder network for road extraction of remote sensing images[J]. Journal of Applied Remote Sensing, 2024, 18(2): 024510-024510.

[16] Li R, Wang L, Zhang C, et al. A2-FPN for semantic segmentation of fine-resolution remotely sensed images[J]. International journal of remote sensing, 2022, 43(3): 1131-1155.

[17] Li R, Zheng S, Zhang C, et al. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery[J]. ISPRS journal of photogrammetry and remote sensing, 2021, 181: 84-98.

[18] Bo W, Liu J, Fan X, et al. BASNet: Burned area segmentation network for real-time detection of damage maps in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-13.

[19] Guo Z, Bian L, Huang X, et al. DSNet: A Novel Way to Use Atrous Convolutions in Semantic Segmentation[J]. arXiv preprint arXiv:2406.03702, 2024.

[20] Mehta D, Skliar A, Ben Yahia H, et al. Simple and efficient architectures for semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2628-2636.

[21] Wang L, Li R, Zhang C, et al. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196-214.