DBARCT: Road Extraction Based on Double Branch Architecture and Random Block Coding Transformer

Ziyi Chen, Senior *Member, IEEE*, Yucai Chen, Lujuan Gao, Dilong Li, *Member, IEEE,* Linlin Xu, *Member, IEEE*, Jonathan Li, *Fellow, IEEE*, Cheng Wang, *Senior Member, IEEE*, and Yewang Chen*, Senior *Member, IEEE*

***Abstract*—Although Transformer models are main network architectures for the delineation of roads from remote sensing imagery, they have critical limitations due to their regular patch mechanism and inefficiency in local information learning. To address these limitations for enhanced road extraction, this paper presents a novel Double Branch Architecture and Random block Coding Transformer (DBARCT), with the following contributions. First, to improve local spatial details learning, we integrate transformer with convolutional neural network (CNN) into a novel dual-branch encoder-decoder architecture, such that the resulting model is efficient at learning both the local edge information and the global context information that are highly complementary for accurate road extraction. Second, to additionally augment the learning of global contextual information, we integrate the regular patching approach in traditional transformer models with a new irregular patching approach, such that it can better capture the global spatial information correlations that might be ignored by the regular patching approach. Third, an array of tests was carried out to meticulously scrutinize the efficacy of the fundamental elements of the suggested model. The empirical findings reveal that the Intersection over Union (IoU) metric attained by the proposed methodology on the LRSNY dataset stands at 88.53%, thereby corroborating the efficacy and preeminence of our approach in tasks related to road extraction.**

***Index Terms*—Remote sensing images, transformer, road extraction, convolutional neural network, random block coding.**

I. Introduction

PATHWAY delineation from aerial imagery stands as a pivotal investigative domain within the realms of remote sensing and geospatial information systems. This technology involves analyzing images acquired by satellites or aircraft to identify and extract road information [1], which has significant application value in urban planning [2], traffic management [3], navigation system construction [4], disaster response [5]. The road information in the high precision optical remote sensing image is the important method to construct the intelligent traffic system, improve the efficiency of the city management and deal with the emergency [6].

In recent times, CNNs and Transformers have catalyzed profound advancements in the realms of machine learning and image processing, emerging as central subjects of scholarly inquiry within these disciplines. CNNs play a vital role in remote sensing image object detection and image segmentation due to their excellent ability to extract local spatial features. They efficiently extract features of targets, such as roads from remote sensing images, demonstrating strong feature extraction capabilities when dealing with image data [7]. Recently, several pioneering CNN-based methods have been introduced to derive data from remote sensing imagery. For example, the Segmentation Depth-wise Separable Graph Convolutional Network was designed in [8], which adjusts the convolutions to improve the road features. Tan et al.[9] proposed a holistic end-to-end methodology for road segmentation to efficiently extract information from different layers in a convolutional network , addressing the contrast between the complexity of a network's layers and the detail within the image it processes. The Coordinate-Dense-Global(CDG) model was proposed in [10], which combines coordinate transformation modules, dense convolutional networks and global attention modules, significantly improving the model's category classification ability. Given the robust ability of CNN models to efficiently extract local features, the precision of road delineation using remote sensing imagery has seen considerable enhancement.

However, compared to transformer models, CNNs fall short in representing global features among pixels [11]. To tackle this problem, researchers have proposed a sequence of innovative Transformer-based methods. For example, a multi-scale deformable Transformer network (MDTNet) was proposed in [12] to capture global features and minimizes errors in road segmentation. A Bi-Directional Transformer Network with a hybrid encoder-decoder structure was proposed in [13] to simultaneously handle global and local information, enhancing road feature extraction capabilities in remote sensing images

Manuscript received April 27, 2024.

Z. Chen, Yucai Chen , D. Li and Y Chen are with the Department of Computer Science and Technology, Huaqiao University, Xiamen, JS 361021, China (e-mail: chenziyihq@hqu.edu.cn; cyc_1003@163.com; scholar.dll@hqu.edu.cn; and ywchen@hqu.edu.cn).

Lujuan Gao is with the Department of Dermatology, Zhongshan Hospital (Xiamen), Fudan University, Xiamen, Fujian, China (e-mail: gao_lujuan@fudan.edu.cn)

J. Li and L. Xu are respectively with the Department of Geography and Environmental Management and the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: junli@uwaterloo.ca and l44xu@uwaterloo.ca).

C. Wang is with the School of Information Science and Engineering, Xiamen University, Xiamen, FJ 361005, China (e-mail: cwang@xmu.edu.cn).

This work was supported by Natural Science Foundation of Fujian Province (No.2023J01135), National Natural Science Foundation of China (No.62001175), Fundamental Research Funds for the Central Universities of Huaqiao University (No.ZQN-911), the National Natural Science Foundation of China (No.42201475), and the Natural Science Foundation of Fujian Province (No.2021J05059).

and aiding in obtaining more detailed segmentation results. Yang et al. [14] revealed an innovative technique for delineating roads from remote sensing images, merging enhanced semantic representations with contextual data, which improves inference capabilities in occluded areas, making the model more robust in handling occlusion and complex scenes.

Traditional methods integrating CNNs and Transformers, either in a straightforward serial or parallel fashion, or by tweaking the attention mechanism, often grapple with efficiently balancing local nuances against global context. This imbalance results in suboptimal feature fusion, giving rise to either redundant or missing information, thereby constraining model efficacy. Furthermore, the inflexible patching approach of conventional Transformers frequently overlooks pivotal details in irregular regions, diminishing their adaptability to intricate visual landscapes. To surmount these challenges, this study unveils a novel road extraction framework termed DBARCT, wherein the principal contributions are encapsulated as follows:

1) We improve the local spatial details learning capability of Transformer models by integrating the Transformer architecture with Convolutional Neural Network(CNN) architecture into a novel dual-branch encoder-decoder architecture, such that the resulting model is efficient at learning both the local edge information and the global context information that are highly complementary for accurate road extraction.

2) We enhance the global context learning capability of Transformer models by integrating the regular patching approach in traditional Transformer models with a new irregular patching approach, such that it can better capture the global spatial information correlations that might be ignored by the regular patching approach.

3) We verify the performance of the component within our suggested model and its benefits compared to other advanced methods through comprehensive testing with the LRSNY dataset. Experimental data show that this model not only improves the accuracy, but also has good performance in edge preservation and detail description.

II. Method

This section introduces our model, DBARCT, for road delineation from remote sensing imagery using a dual-branch random coding Transformer network. We outline the core components of the dual branches and the decoder, and provide an overview of the loss functions used.

A. Network Architecture

As illustrated in Figure 1, our devised double branch architecture and random block coding Transformer for delineating roads from remote sensing images primarily comprises three critical modules: the feature processing block, encoder, and decoder. DBARCT utilizes an encoder-decoder architecture that includes two types of encoder branches: one based on a Transformer and the other on a convolutional neural network.

B. Feature Processing Block

In this study, we draw inspiration from the input-output structure of Vision Transformer to handle the transformation between images and sequences. The core input of the Transformer model is a one-dimensional array of feature embeddings, represented as $Z \in R^{L\times C}$, where L denotes the extent of a sequence and C represents the number of channels or features in a hidden layer of a neural network. Therefore, it is necessary to serialize the input image $x \in R^{H\times W\times C}$ into Z.

Image serialization simplifies the image into a one-dimensional array by flattening its structure. For an image of size $H\times W\times C$, direct flattening would result in a very long vector. Given the computational complexity of Transformer, directly processing each pixel is impractical. Hence, we adopt a more efficient approach to handle images. Specifically, we adopt an encoder structure, typical in semantic segmentation tasks, which downsampling the 2D image $x \in R^{H\times W\times C}$ into a feature representation $x_f \in R^{\frac{H}{16}\times\frac{W}{16}\times C}$. Based on this down-

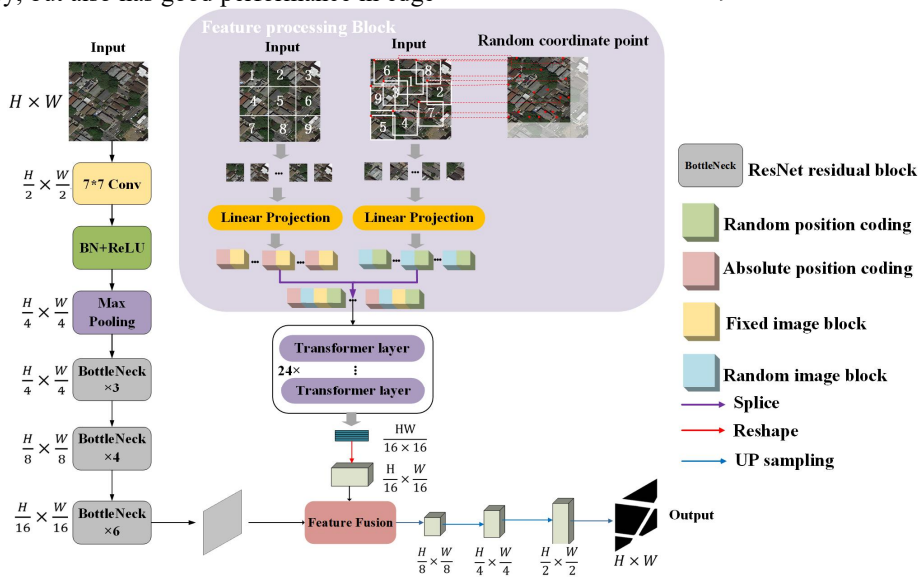

Fig. 1. Structure of DBARCT.

sampling operation, we configure the input sequence length L of the Transformer to $\frac{H}{16}\times\frac{W}{16}$ which is $\frac{H\times W}{256}$. Thus, the output sequence generated by the Transformer can be directly adjusted to fit the dimensions of the target feature map x_f.

To convert the image into a sequence required by the Transformer, we employ a non-overlapping patching approach. We partition the original image into grid blocks. Each block is flattened and transposed into a space of d dimensions by applying a linear transformation function. This process produces a linear array of pixel group representations for the image. These one-dimensional sequences are fed into convolutional layers to fully learn features. In the Transformer architecture, each position i within a patch is assigned a unique positional embedding p_i. This embedding is then combined with the respective embedding vector, resulting in the formation of the input sequence e_i to encode spatial details of the patches $E=\{e_1+p_1,e_2+p_2,\cdots,e_L+p_L\}$ .In this manner, although the Transformer itself has an unordered self-attention mechanism, spatial information is still preserved. To enhance the resilience and broad utility of the model, we introduce random patching operations. By generating a fixed number of random coordinate pairs, we extract image blocks corresponding to the patch size from the image. These randomly cropped blocks are collected into tensor patches, whose dimensions correspond to the count of grid blocks in the non-overlapping patches. After adding positional encoding to the processed patches, flattening is performed, and then integrated with the original non-overlapping patches to create the completed sequence $E_1=\{e_1+p_1,\cdots,e_{L+L}+p_{L+L}\}$. This fusion strategy combines ordered and random information, aiding the model in capturing richer context and detail information.

C. Encoder

1) Transformer-based Encoder: To fully leverage the advantage of Transformer in handling global dependencies, we designed a Transformer-based encoder to process the one-dimensional embedding sequence. This encoder endows the model greater global awareness , effectively addressing the issue of limited receptive fields in traditional CNN. It features several layers, each integrated with a multi-head self-attention mechanism and a feedforward neural network module, as depicted in Figure 1. This structure allows the model to identify distant relationships within the input sequence, demonstrating outstanding ability in learning feature representations. After the image passes through the feature processing module, we feed its results into the Transformer encoder for global encoding. By stacking multiple Transformer encoder layers, the model gradually extracts more abstract and deep feature representations, providing robust support for subsequent tasks.

2) CNN-based Encoder: Although Transformer excel in global information processing, there is still a gap compared to CNNs in extracting local features. To address this limitation, we integrate the efficient local feature extraction capability of ResNet50. ResNet50 is widely used in image processing applications because of its outstanding performance and stability. By modifying the structure of ResNet50, we stack multiple stages containing residual connections to retain more feature information. These stages serve as bottleneck layers, aiding in further refining key information from the features extracted from ResNet50. This project takes ResNet50 as the core and gives full play to its advantages in local feature extraction to improve the recognition effect of the model. The strided convolutions and max-pooling operations in ResNet50 help rapidly reduce feature dimensions while retaining important local information. Additionally, the first convolutional layer adopts a 7×7 kernel, which covers a broader receptive field compared to the conventional 3×3 kernel, thereby enhancing the model's robustness to noise.

D. Decoder

The decoder constitutes another key component of the model. To ensure the integrity and accuracy of features during transmission, we employ a progressive upsampling strategy. This technique involves alternating between convolution and upsampling procedures to progressively enlarge the feature map, thereby avoiding potential feature loss that may occur with direct upsampling from low to high resolution. During the decoding phase, we initiate by fusing features from the output of the dual-branch encoder. Subsequently, we incrementally reinstate the feature map to its original full-resolution dimensions by integrating convolutions with a series of four upsampling steps. Each upsampling operation doubles the size of the output of the previous layer to control the accumulation of noise and distortion during upsampling. After each upsampling step, we apply convolution layers to refine and optimize the feature maps, ensuring spatial consistency and feature quality. Through this approach, we can minimize the introduction of noise during the upsampling process and mitigate potential adversarial effects of large-scale upsampling. This streamlined design integrates Transformer and ResNet50 encoders with a progressive upsampling decoder, enhancing our model's ability to balance global and local feature extraction for superior performance in road extraction tasks.

E. Loss Function

The Binary Cross-Entropy (BCE) loss function is extensively employed across diverse applications, particularly effective in road extraction tasks. In this research, the BCE loss quantitatively evaluates the discrepancy between model forecasts and true labels during training. In binary classification, it computes the negative weighted sum of the output probabilities against the actual values, with weights based on the true labels. Mathematically, it is defined as follows:

$$L=\frac{1}{N}\sum_{i}^{N}\left(g_i\log p_i+(1-g_i)\log(1-p_i)\right) \quad (1)$$

where L signifies the loss , p_i indicates the probability that the i-th image predicted by the model belongs to the positive class , g_i denotes the actual label for the i-th pixel in the samples, and N denotes the aggregate count of images .

III. Experiments

A. Implementation and Metrics

1) Dateset: In this document, our approach is based on publicly available LRSNY datasets. This dataset showcases imagery of New York City's central region, captured at a detailed resolution of 0.5 meters per pixel. It encompasses a total of 1368 images, each with dimensions of 256×256 pixels, methodically segmented into three distinct groups: training, validation, and testing. The dataset includes 716 training images, 220 validation images, and 432 testing images.

2) Evaluation Metrics: We utilize five metrics to evaluate the semantic segmentation outcomes: Overall Accuracy (OA), Precision, Recall, F1 score, and Intersection over Union (IoU). OA is gauged by the ratio of pixels accurately pinpointed to the complete pixel count appraised. Precision assesses the proportion of pixels precisely designated as roads within their correct locales, while Recall indicates the fraction of true positive identifications made. The F1 score calculates the weighted equilibrium between Precision and Recall, offering a consolidated measure of the two values, and the IoU quantifies the ratio of overlap to the combined area of both the predicted road segments and the actual conditions on the ground. Following the determination of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for the pixels, these evaluation metrics can be formulated as:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{5}$$

$$F1 = \frac{2 \times recall \times precision}{precision + recall} \tag{6}$$

3) Training Settings: During the training, we utilize SGD as the optimizer, configure the batch size to 8, perform 5000 training iterations, and establish the starting learning rate at 0.1. To modify the learning rate, a cosine annealing technique was applied. gradually decreasing it during training. To combat overfitting, we utilized several data enhancement techniques throughout the training phase. These methods encompassed horizontal and vertical flips, random rotations by 90 degrees, and arbitrary shifts and rotations of the images.

B. Ablation Study

In this segment, we assess the efficacy of the DBARCT model for road extraction by conducting ablation analyses on the LRSNY dataset to evaluate the impact of specific model components on overall model efficacy. The benchmark model adopts a feature preprocessing module with random coding and a network structure with Transformer as encoder. By systematically modifying the subcomponents in the model, we can evaluate the influence of these modifications on the model's overall effectiveness. The outcomes are presented in Table I. BL (Baseline), RC (random block), C (convolution).

TABLE I
ABLATION RESULTS ON THE LRSNY ROAD DATASET

Method	OA(%)	Precision(%)	Recall(%)	IoU(%)	F1(%)
BL	95.68	89.13	82.27	74.77	85.57
BL+RC	95.67	88.74	82.65	74.80	85.59
BL+RC+C	95.83	89.69	82.67	75.50	86.04
BL+RC+C+ResNet50	**98.13**	**94.98**	**92.88**	**88.53**	**93.91**
BL+RC+C+ResNet101	98.07	94.83	92.62	88.17	93.71
BL+RC+C+ResNet152	98.09	94.74	92.84	88.29	93.78

Table I shows that by incorporating random coding and convolution processing into the baseline model, scores on each indicator were improved. When comparing different ResNet networks as decoders, it is evident that ResNet50 stands out in terms of competitiveness, with its IoU increasing to 88.53%. This further proves that DBARCT can efficiently extract road features.

C. Comparison with Existing Techniques

We benchmarked our proposed model against leading-edge segmentation techniques, including SegNet [15], ResidualU-Net [16], DANet [17], DeepLabV3 [18], DeepLabV3+[19], Bias U-Net [20], D-LinkNet [21], DDU-Net [22], DPSDA-Net [23], and SwinUnet [24]. On the LRSNY road dataset, our method outperformed previous highs, boosting OA by 0.3%, Precision by 0.56%, Recall by 0.13%, IoU by 1.63%, and F1 Score by 0.92%.The comparative assessment on the LRSNY dataset are presented in Table II. Moreover, the sections highlighted in bold in each table indicate the best scores attained on that evaluation metric.

D. Visualization Analysis

To explore the merits and limitations of our suggested strategy, this section will offer a comparative evaluation of the extraction efficacy of DBARCT versus other rival methods on the LRSNY road datasets. The extraction results of the third row in Figure 2 are presented, highlighting the challenge for other methods to accurately identify the gaps between multiple roadways. In contrast, our method successfully and comprehensively extracts intersecting road networks. Our approach can supply more precise road data for subsequent research.

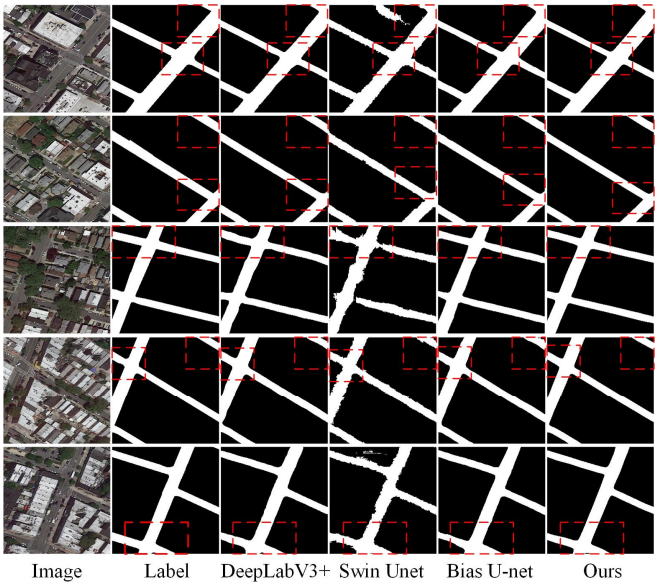

Fig. 2. The results of various approaches on samples from the LRSNY dataset.

TABLE II
COMPARATIVE ASSESSMENT OF VARIOUS APPROACHES ON THE LRSNY DATASET

Method	OA(%)	Precision(%)	Recall(%)	IoU(%)	F1(%)
SegNet	97.54	92.80	91.29	85.25	92.04
Residual U-Net	97.53	92.72	91.27	85.16	91.99
DeepLabV3	97.51	93.23	90.59	85.00	91.89
DeepLabv3+	97.80	93.09	92.75	86.78	92.92
DANet	97.66	94.42	90.32	85.74	92.32
Swin-Unet	97.22	90.96	91.19	83.61	91.07
Bias U-Net	97.83	93.42	92.57	86.90	92.99
D-LinkNet	/	93.56	89.29	84.85	90.69
DPSDA-Net	/	92.34	91.72	86.36	91.76
DDU-Net	/	91.69	90.38	84.58	90.16
Ours	**98.13**	**94.98**	**92.88**	**88.53**	**93.91**

IV. CONCLUSION

To address the key limitations of Transformer models in road extraction, this paper has presented a novel double branch architecture and random block coding Transformer approach (DBARCT). The proposed approach improved the local spatial details learning capability of traditional Transformer models by integrating the Transformer architecture with the CNN architecture. The resulting model thereby is efficient at learning both the local edge information and the global context information, both of which are critical for accurate road extraction. The proposed model further enhanced the global context learning capability of traditional Transformer models by integrating the regular patching approach with a new irregular patching approach. Extensive experiments were conducted to evaluate the proposed model in comparisons with the other cutting-edge methodologies using the LRSNY dataset, demonstrating that the proposed methodology markedly surpasses other contemporary techniques. Our future plans include integrating transformer-based local feature extraction to boost object identification accuracy, especially for roads.

REFERENCES

[1] Z. Chen, L. Deng, Y. Luo *et al.*, "Road extraction in remote sensing data: A survey," *International Journal of Applied Earth Observation and Geoinformation,* vol. 112, pp. 102833, 2022.

[2] M. Guo, H. Liu, Y. Xu *et al.*, "Building extraction based on U-Net with an attention block and multiple losses," *Remote Sensing.,* vol. 12, no. 9, pp. 1400, 2020.

[3] H. Guan, Y. Yu, D. Li *et al.*, "RoadCapsFPN: Capsule Feature Pyramid Network for Road Extraction From VHR Optical Remote Sensing Imagery," *IEEE Transactions on Intelligent Transportation Systems.,* vol. 23, pp. 11041-11051, 2022.

[4] Y. Xu, H. Chen, C. Du *et al.*, "MSACon: Mining Spatial Attention-Based Contextual Information for Road Extraction," *IEEE Transactions on Geoscience Remote Sensing.,* vol. 60, pp. 1-17, 2022.

[5] C. Li, L. Fu, Q. Zhu *et al.*, "Attention Enhanced U-Net for Building Extraction from Farmland Based on Google and WorldView-2 Remote Sensing Images," *Remote Sensing.,* vol. 13, pp. 4411, 2021.

[6] Z. Chen, C. Wang, J. Li *et al.*, "Adaboost-like End-to-End multiple lightweight U-nets for road extraction from optical remote sensing images," *Int. J. Appl. Earth Obs. Geoinformation.,* vol. 100, pp. 102341, 2021.

[7] D. Zhou, G. Wang, G. He *et al.*, "Robust Building Extraction for High Spatial Resolution Remote Sensing Images with Self-Attention Network," *Sensors.,* vol. 20, no. 24, pp. 7241, 2020.

[8] G. Zhou, W. Chen, Q. Gui *et al.*, "Split Depth-Wise Separable Graph-Convolution Network for Road Extraction in Complex Environments From High-Resolution Remote-Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing.,* vol. 60, pp. 1-15, 2022.

[9] X. Tan, Z. Xiao, Q. Wan *et al.*, "Scale Sensitive Neural Network for Road Segmentation in High-Resolution Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters,* vol. 18, no. 3, pp. 533-537, 2021.

[10] S. Wang, H. Yang, Q. Wu *et al.*, "An Improved Method for Road Extraction from High-Resolution Remote-Sensing Images that Enhances Boundary Information," *Sensors,* vol. 20, no. 7, pp. 2064, 2020.

[11] J. Chen, H. Hong, B. Song *et al.*, "MDCT: Multi-Kernel Dilated Convolution and Transformer for One-Stage Object Detection of Remote Sensing Images," *Remote. Sens.,* vol. 15, pp. 371, 2023.

[12] P. Hu, S. Chen, L. Huang *et al.*, "Road Extraction by Multiscale Deformable Transformer From Remote Sensing Images," *IEEE Geoscience Remote Sensing Letters.,* vol. 20, pp. 1-5, 2023.

[13] L. Luo, J. Wang, S. Chen *et al.*, "BDTNet: Road Extraction by Bi-Direction Transformer From Remote Sensing Images," *IEEE Geoscience Remote Sensing Letters.,* vol. 19, pp. 1-5, 2022.

[14] Z. Yang, D. Zhou, Y. Yang *et al.*, "TransRoadNet: A Novel Road Extraction Method for Remote Sensing Images via Combining High-Level Semantic Feature and Context," *IEEE Geoscience Remote Sensing Letters.,* vol. 19, pp. 1-5, 2022.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis Machine Intelligence.,* vol. 39, pp. 2481-2495, 2015.

[16] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters,* vol. 15, no. 5, pp. 749-753, 2018.

[17] J. Fu, J. Liu, H. Tian *et al.*, "Dual Attention Network for Scene Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3141-3149.

[18] L.-C. Chen, G. Papandreou, F. Schroff *et al.*, "Rethinking Atrous Convolution for Semantic Image Segmentation," *ArXiv.,* vol. abs/1706.05587, 2017.

[19] L.-C. Chen, Y. Zhu, G. Papandreou *et al.*, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. 2018, pp. 801-818.

[20] Z. Chen, C. Wang, J. Li *et al.*, "Reconstruction Bias U-Net for Road Extraction From Optical Remote Sensing Images," *IEEE Journal of Selected Topics in Applied Earth Observations Remote Sensing.,* vol. 14, pp. 2284-2294, 2021.

[21] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182-186.

[22] Y. Wang, Y. Peng, W. Li *et al.*, "DDU-Net: Dual-decoder-U-Net for road extraction using high-resolution remote sensing images," *IEEE Transactions on Geoscience Remote Sensing.,* vol. 60, pp. 1-12, 2022.

[23] L. Zhao, L. Ye, M. Zhang *et al.*, "DPSDA-Net: Dual-Path Convolutional Neural Network with Strip Dilated Attention Module for Road Extraction from High-Resolution Remote Sensing Images," *Remote. Sens.,* vol. 15, pp. 3741, 2023.

[24] H. Cao, Y. Wang, J. Chen *et al.*, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in European conference on computer vision. Springer,Oct. 2022, pp. 205-218.