MANet: An End-To-End Multiple Attention Network for Extracting Roads Around EHV Transmission Lines from High-Resolution Remote Sensing Images

Yaru Ren, Xiangyu Bai(✉), Yu Han, and Xiaoyu Hu

Inner Mongolia School of Computer Science, Inner Mongolia University, Hohhot 010021, China
bxy@imu.edu.cn

Abstract. Complete and accurate road network information is an important basis in the detection of EHV transmission lines, and regular updates of road distribution near transmission lines are necessary and meaningful. However, no relevant research has been found for this application area, and coupled with the fact that roads themselves are significantly challenging, extracting roads with good connectivity and integrity in remote sensing images remains a problem to be solved. Therefore, in this paper, we develop a new end-to-end road extraction network, Multiple Attention Networks (MANet). Specifically, by fusing convolutional and self-attentive approaches, we focus on global contextual features to obtain an effective feature map. In addition, the Strip Multi-scale Channel Attention (SMCA) module is specifically designed for the line features of roads, focusing on extracting row and column features, while the Edge-aware Module (EAM) is used to extract connected and complete roads, aided by edge information. Meanwhile, in order to enhance the practicality of the study, a Mengxi Transmission Line Road Dataset was constructed independently following the processing process of remote sensing images in industrial production. By conducting relevant quantitative and qualitative experiments on this dataset and the publicly available CHN6-CUG dataset, it is fully verified that the method in this paper is superior to other advanced methods and can still extract roads with strong connectivity in complex backgrounds, which has good potential and outstanding advantages in practical applications.

Keywords: Road extraction · High-resolution remote sensing images · Deep learning · Semantic segmentation

1 Introduction

As one of the major infrastructures, the road network is widely used in various industrial fields and social life. Especially in the detection and inspection of ultra-high voltage transmission lines, regular measurement of their surrounding road networks and timely updating of their detailed information are essential to ensure the safe and smooth operation of the entire power system. With the development of remote sensing technology

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
X. Yang et al. (Eds.): ADMA 2023, LNAI 14176, pp. 553–568, 2023.
https://doi.org/10.1007/978-3-031-46661-8_37

in recent years, the extraction of roads from high-resolution remote sensing images and ultra-high resolution remote sensing images has become a popular topic [1].

However, the roads themselves have complex geometric, radiometric, and topological characteristics, such as different widths, directional changes, uniform grayscale, obvious boundaries, and connectivity; they are also in the middle of complex scenes and easily obscured by obstacles such as vehicles, trees, buildings and their shadows, making the task very challenging. In addition, compared with the urban areas where road data are concentrated, the EHV transmission lines are widely distributed and span a large area, and the accessibility of the roads around them is weaker, mainly concentrated in remote areas far from the urban areas, with sparse and disorganized distribution of features and the existence of more third- and fourth-class roads as well as concrete roads and dirt roads, which are more easily integrated into the scene environment and make the extraction more difficult, with obvious differences compared with the former. Therefore, for road extraction around the ultra-high voltage transmission lines, it is important for the field to study a more advanced and suitable method to improve the model performance and further improve the accuracy and quality of road extraction.

In this paper, we propose a Multiple Attention Network (MANet) for road extraction. Even when the road distribution is very hidden and inherently tortuous, the road information can still be captured sensitively for accurate and effective extraction. The main contributions of this paper can be described as follows:

1. A new end-to-end road extraction network, MANet, is proposed, deploying a self-attention mechanism and a channel-attention mechanism, while adding target boundary information to constrain the extracted roads, effectively enhancing road connectivity and reducing disconnections;
2. A Strip Multi-scale Channel Attention (SMCA) mechanism is designed to extract features from two dimensions, row and column, respectively, for the geometric and topological features of winding and narrow roads, and perform multiscale differential fusion to improve the model's ability to perceive roads in complex scenes;
3. To the best of our knowledge, this paper is the first study to perform road extraction in the scenario around EHV transmission lines, for which a road dataset is constructed.

2 Related Work

This section reviews the relevant research methods for road extraction.

The deep learning approach emerging in recent years is data-driven and represented by the semantic segmentation task, which is widely used in the field of road extraction. It mainly relies on the color, geometric and texture features of remote sensing images for feature extraction of images to achieve almost automated road extraction. The first use of deep learning methods for road extraction from remote sensing images in the field of road extraction was by Mnih et al. They used a restricted Boltzmann machine to extract roads from remote sensing images [2]. Subsequently, Wang et al. proposed a neural dynamic tracking framework based on deep convolutional neural networks and finite state machines to extract road networks [3]. However, this method was gradually replaced because it was limited in terms of accuracy and speed, and it was prone to overfitting because the roads themselves accounted for a small proportion of the whole image. Fully

Convolutional Networks (FCN) is regarded as the pioneer of semantic segmentation, and its proposal has led to a significant improvement in semantic segmentation and realized end-to-end image segmentation [4]. Zhong et al. were the first to use FCN for research in the field of road extraction [5]. And then, a series of methods based on encoder-decoder structures were proposed one after another, such as U-Net [6], CasNet [7], D-Linknet [8], etc., which achieved multi-level feature stitching and reuse while extracting features more deeply, all achieving better results at that time. Shi et al. first introduced generative adversarial networks (GAN) into pixel-level remote sensing image classification by acting the basic segmentation network as Generator in GAN [9], and implemented the task of road area detection in Google Earth remote sensing images by GAN model. Recently, with the boom of Transformer structure [10], some researches based on this framework have also emerged, such as RoadFormer [11], HA-RoadFormer [12], etc. Compared with the above-mentioned methods based on convolutional neural networks, they have stronger ability to learn remote features and global modeling, and pay more attention to the global features of images, and all of them have also achieved considerable results.

3 Methodology

3.1 MANet Overall Framework

Figure 1 shows the overall architecture of the proposed MANet model. The whole network is designed with an encoder-decoder structure, using DeepLab V3+ [13] as the semantic segmentation model framework and improving on the original model. In the following, the proposed network structure is briefly described in terms of two components, encoder and decoder.

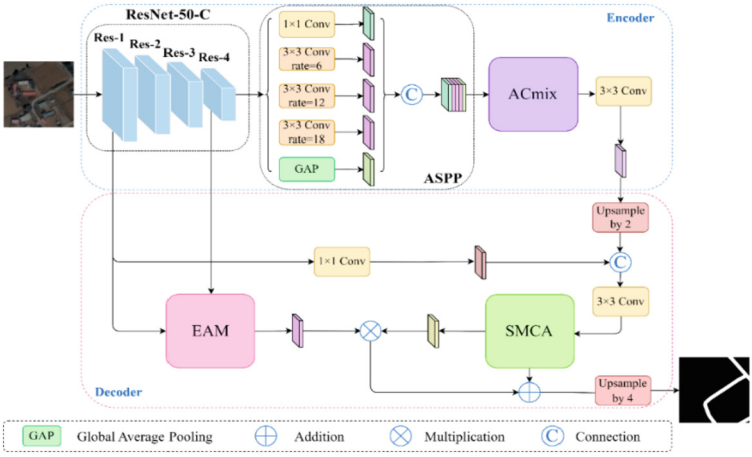

Fig. 1. Structure of our proposed MANet.

The encoder part uses ResNet-50-C [14] as the backbone network. We first use the improved residual network to extract features from four different stages, denoted as

$f_{R_i}(i = 1, 2, 3, 4)$; second, the extracted features f_{R_4} are fed into the ASPP structure to enrich the encoder module in the compiled code network by fusing multi-scale contextual information. Based on this, we further add a self-attention and convolution integration module (ACmix) [15], which effectively integrates the advantages of self-attention and convolution, capturing global contextual information effectively while paying more attention to the acquisition of local information, and at the same time, exploiting the features of the whole image as fully as possible without significantly increasing the computational effort to extract the rich information in it.

In the decoder, the different hierarchical features obtained from the encoder will be convolved and upsampled to obtain the high-resolution segmentation results. We first perform upsampling operation on the multi-scale high-level feature map extracted from the encoder to expand the feature map size by a factor of 2. After compressing the dimensionality of the shallow sub-feature map f_{R_1} in the backbone network as well, we stitch the two together to perform cross-level fusion of high bottom-level features. The feature map is then adjusted by a 3×3 convolution, after which we feed the feature map fused with rich semantic information into our proposed Strip Multi-scale Channel Attention (SMCA), which can fuse features in both row and column directions and is well suited for road feature extraction. At the same time, we consider the importance of edge information, and for the four different layers of features obtained from the initial extraction of the network, we use the shallow features f_{R_1} and the deep features f_{R_4} as the input of the Edge-aware Module (EAM) [16], and get the feature output containing the road boundary information through the detail information in the shallow layer and the semantic information in the deep layer, and then multiply it pixel by pixel with the SMCA output feature map to give the boundary information, and then pixel by pixel Then it is added together to reduce the feature segmentation map from multiple dimensions. Finally, the final road extraction result map is obtained after 4-fold upsampling operation.

3.2 ACmix-Based Encoder

When encoding feature information in the encoder, we deploy the self-attention and convolution integration module(ACmix), which mines the potential connection between convolution and self-attention from a new perspective, decomposing them into two-stage operations - the first stage divides the k $\times$ k convolution into k2 independent 1 $\times$ 1 convolution operations, in which the self-attention Query, Key, Value mapping is also composed by 1 $\times$ 1 convolution; the second stage convolution performs shift and sum operation, and self-attention computes attention weights and aggregates Value values. On this basis, it is found that they rely heavily on the 1 $\times$ 1 convolution operation in the first stage, so the first stage operation of both are fused and then the second stage calculation is performed separately, as shown in Fig. 2. The convolution and self-attention mechanisms are effectively integrated while minimizing the computational overhead, so that the local information is complementarily integrated with the global information and the model considers the road itself without forgetting the complex and rich background information in remote sensing images.

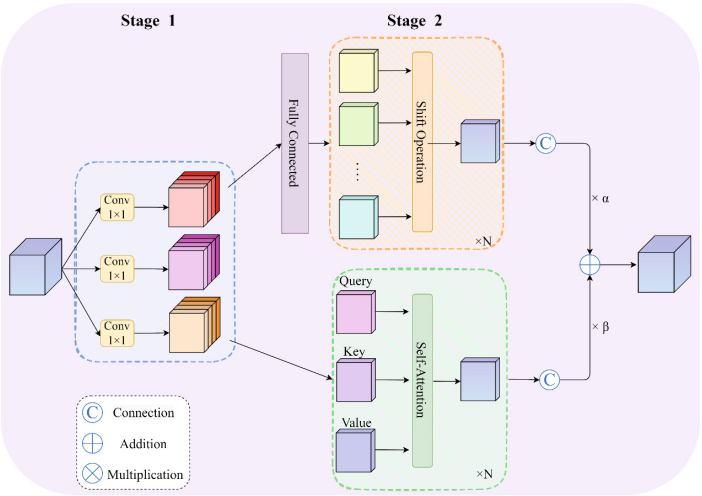

Fig. 2. Self-attention and convolution integration module.

3.3 SMCA and EAM based Decoder

Strip Multi-scale Channel Attention Module. Considering the directional extension of roads, inspired by Strip Attention Networks (SANet) [17], a Strip Multi-scale Channel Attention module is specially designed for road extraction with a three-branch structure that differentially fuses the lineal topological features that focus on roads horizontally and vertically in images from different dimensions, respectively.

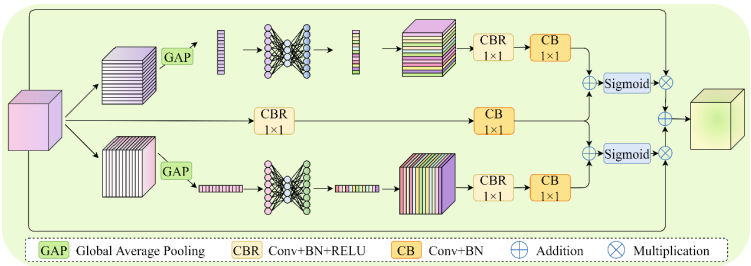

Fig. 3. Strip Multi-scale Channel Attention Module.

The module structure is shown in Fig. 3. SMCA is a three-branch structure, which first divides the feature map into three. The upper branch of the figure is the column pixel feature extraction structure, which firstly performs global average pooling operation to obtain global contextual features from the dimension of each row and performs compression processing of information; then goes through two fully connected layers to learn the attention coefficients of different channels; then the dimensionality is extended to recover the feature map size; and then two 1×1 convolution layers are passed to complete the enhancement of the global channel column direction features. On the other

hand, the second branch is directly subjected to two convolution operations to ensure the extraction of local features and avoid losing details. The feature maps of the second branch are summed pixel by pixel with those of the column branch, fused after giving more attention to the multiscale features of the column pixels, fed into the Sigmoid activation function, filtering the miscellaneous terms to a certain extent to obtain the final weights, and multiplying the original feature maps pixel by pixel, the whole process involves the upper and middle branches in the figure, so that the model captures the road distribution information in the vertical direction in space.

The whole process of column pixel feature extraction can be described as follows:

$$\begin{cases} F_1 = \frac{1}{H \cdot C} F_{CB}(F_{CBR}(F_E(F_{Linear-S}(F_{Linear-R}(F_{GAP}(f_m)))))) \\ F_2 = F_{CB}(F_{CBR}(f_m)) \\ F_C = F_S(F_1 \oplus F_2) \otimes f_m \end{cases} \tag{1}$$

where, F_1 and F_2 represent the first and second branch processing respectively, F_C is the whole column pixel branching process. First, for the first branch, the module input feature map f_m goes through the row mapping of $\frac{1}{H \cdot C}$ as well as the global average pooling F_{GAP} to obtain the $W \cdot 1$ dimensional column feature vector, followed by two fully connected layers $F_{Linear-R}$ with $F_{Linear-S}$, $F_{Linear-R}$ representing the fully connected followed by ReLU activation function, $F_{Linear-S}$ representing the post-connected Sigmiod activation function; F_E representing the extended dimensionality, recovering the size, followed by two 1×1 convolutional layers F_{CBR} with F_{CB}, F_{CBR} representing the convolutional layer, BN, ReLU, F_{CB} denotes the convolutional layer with BN layer. The second branch F_2 performs two 1×1 convolutional layer F_{CBR} with F_{CB} pairs only f_m. .. After the with operation, the two are added pixel by pixel ($\oplus$), and then after the Sigmoid function, they are multiplied pixel by pixel ($\otimes$) with the original feature map to obtain the column branch output.

Similar to the column pixel extraction branch, the row pixel extraction branch is the lower branch in the graph, except that the column mapping is done to obtain the $1 \cdot H$ row feature vector. The procedure is as follows:

$$\begin{cases} F_2 = F_{CB}(F_{CBR}(f_m)) \\ F_3 = \frac{1}{W \cdot C} F_{CB}(F_{CBR}(F_E(F_{Linear-S}(F_{Linear-R}(F_{GAP}(f_m)))))) \\ F_R = F_S(F_2 \oplus F_3) \otimes f_m \end{cases} \tag{2}$$

The following is the SMCA module process:

$$F_{SMCA} = F_C \oplus F_R \tag{3}$$

Edge-Aware Module. For the extraction of road features, we not only consider the acquisition of global information that facilitates semantic segmentation. At the same time, we also take into account the importance of edge information. Therefore, we incorporate the Edge-aware Module (EAM) to extract the edge information effectively by using shallow features and deep features as inputs. As shown in Fig. 4.

Specifically, EAM combines the feature maps f_{R_1} and f_{R_4} output from Res-1 and Res-4 in ResNet-50-C, and then changes the number of channels through a 1×1 convolutional layer, and then performs a 2-fold upsampling operation on the feature

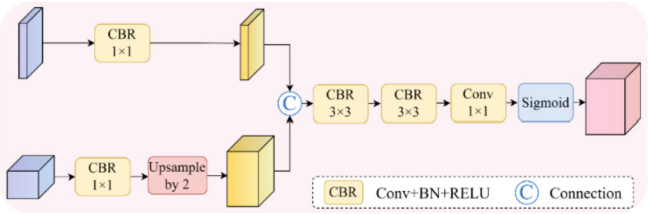

Fig. 4. Edge-aware Module.

map of the f_{R_4} branch, with $\bigcup$ denoting upsampling, so that the feature maps of the two branches are the same size and then stitched together, i.e., $\mathbb{C}$. Then after two 3×3 convolutional layers, BN with Relu, one 1×1 convolutional layer, and finally the edge extraction map is obtained via the Sigmoid function. The procedure is as follows.

$$F_{EAM} = F_S\Big(F_{Conv}\Big(F_{CBR}\Big(F_{CBR}\Big((F_{CBR}(f_{R_1}))\mathbb{C}\Big(\bigcup F_{CBR}(f_{R_4})\Big)\Big)\Big)\Big)\Big) \tag{4}$$

4 Results and Discussions

4.1 Dataset Descriptions and Training Details

Dataset Descriptions. To evaluate the effectiveness of our proposed model in detail, we conducted experiments on two sets of road extraction datasets. A detailed description of these two datasets is given below.

The first dataset is our independently constructed transmission line road dataset for the Mengxi power transmission line. The remote sensing images of this dataset are derived from the Gaofen-2 and SuperView-1 satellites, and are local images of some of the areas through which the ultra-high voltage transmission lines are located in the western region of Inner Mongolia Autonomous Region, China in 2022, covering an area size of about 3000 square kilometers, with a total of 18 scenes and a resolution between 0.5 m and 0.8 m.

The images of the public road dataset underwent finer and stricter processing and screening in the production process, with higher image clarity, and the scenes involved were mostly urban areas in developed regions, with roads mainly being highways, and roads in rural scenes were also easier to distinguish; whereas in this study, based on the perspective of industrial production applications, the data came directly from remote sensing satellites, and the image scenes were the environmental images around the channels of ultra-high voltage transmission lines. In order to make this study more practical, we obtained the original remote sensing data, followed the processing process of remote sensing images in industry, and used Arcgis, Qgis and other software to obtain usable remote sensing images through the processes of image mosaic, image color leveling, image correction, image cropping, image slicing and so on. It is undoubtedly a huge workload and inefficient to produce the required datasets for these remote sensing images using traditional manual annotation. In order to reduce the production cost and

improve the annotation efficiency, this study selects OSM data combined with professional practitioners for manual review and annotation to construct the dataset. Finally, we obtained a total of 5365 images in the Mengxi transmission line road dataset, with a size of 256 × 256, and divided them into training and testing sets according to the ratio of 8:2, which are 4292 and 1073 images, respectively. However, at present, we cannot make them public due to copyright issues. A partial image of it is shown in Fig. 5.

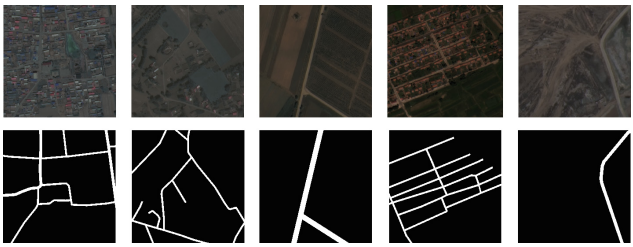

Fig. 5. Mengxi Transmission Line Road Dataset.

The second dataset is the CHN6-CUG road datase t [18], which was produced and shared by the team of Zhu from China University of Geosciences in 2021. Its remote sensing image base map is from Google Earth, and six major cities in China, namely Beijing, Shanghai, Wuhan, Shenzhen, Hong Kong, and Macau, are selected as the study area. The whole dataset is manually labeled and contains 4511 labeled images of 512 × 512 size, of which 3608 are used for model training and 903 are used for testing and result evaluation. A partial image of it is shown in Fig. 6.

Fig. 6. CHN6-CUG road dataset.

Training Details. All the experiments performed in this paper are based on the Pytorch deep learning framework and are done using the MMSegmentation [19] framework in Open-MMLab. DeepLab V3+ was used as the baseline model. The loss function used in this paper is the joint loss function of the sum of the binary cross-entropy loss function and the dice function. In this paper, Stochastic Gradient Descent (SGD) [20] is used as an optimizer for the model training process and poly learning strategy is used to update the learning rate, i.e $lr = init_{lr} * (1 - iter/maxiter)^{0.9}$. And following the relationship between learning rate and batch size in MMSegmentation, the initial learning rate is set

to 5e-3 in this paper. Momentum and weight decay are 0.9 and 0.0001, respectively. The model uses an NVIDIA GPU A100 to accelerate the training. To expand the dataset, we use data augmentation, including random cropping, coefficients from 0.5 to 2, random flipping and multi-scale augmentation. Finally, the original image is used for testing the model.

4.2 Evaluation Metrics

Different evaluation metrics reflect the performance advantages and disadvantages of the developed method from different perspectives. In this paper, six evaluation metrics are mainly used to evaluate the proposed network model, including precise, recall, road Intersection over Union, Mean Intersection over Union, F1-score and aver-age accuracy, all of which are more general evaluation metrics in semantic segmentation of remote sensing images [21].

Road extraction is considered as binary classification tasks, road areas are foreground, i.e., positive samples, and non-road areas are background, i.e., negative samples. Where, TP indicates correctly predicting positive samples as positive, FN indicates incorrectly predicting positive samples as negative, FP indicates incorrectly predicting negative samples as positive, and TN indicates correctly predicting negative samples as negative. The confusion matrix allows the calculation of the above evaluation index values.

$$\text{Precise} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{MIoU} = \frac{1}{\text{k} + 1} \sum_{\text{i}=0}^{\text{k}} \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precise} \times \text{Recall}}{\text{Precise} + \text{Recall}} \tag{9}$$

$$\text{AverageAccuracy} = \frac{1}{\text{k} + 1} \sum_{\text{i}=0}^{\text{k}} \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

4.3 Comparative Experiments

Experiments on the Mengxi Dataset. Table 1 shows that the method proposed in this paper significantly outperforms the other methods, where the bolded values represent the best results for quantitative comparison. Specifically, MANet has improved the values of

its main indexes to different degrees compared with DeepLab V3 +, which is the most outstanding performance among other methods: recall is improved by 6.37%, F1-score is improved by 4.84%, road IoU is improved by 4.42%, MIoU is improved by 2.31%, and average accuracy is improved by 3.1%. Among them, SANet has the highest accuracy rate of 75.82%, but all other values of this model are low, and the network does not perform well when all values are considered, and the method in this paper still performs best. And it is not difficult to find through the experiment that the accuracy as a single indicator does not effectively evaluate the model performance, and the value is too high even makes the performance degraded.

Table 1. Quantitative comparison between MANet and other methods on the Road Dataset of Mengxi Transmission Line.

Model	Precise	Recall	F1-score	IoU	MIoU	AA
U-Net [6]	37.42	29.55	32.95	19.73	56.5	63.32
UperNet [22]	55.87	15.42	24.17	13.74	54.18	57.35
Swin Transformer [23]	56.83	31.53	40.56	25.44	60.13	65.07
DANet [24]	58.7	34.78	43.68	27.94	61.45	66.68
ViT [25]	64.15	29.37	40.29	25.23	60.18	64.21
Deeplab V3 + [13]	66.37	39.62	49.62	33.0	64.23	69.23
SANet [17]	**75.82**	11.48	19.92	11.06	52.98	55.63
MANet (Ours)	67.42	**45.99**	**54.46**	**37.42**	**66.54**	**72.33**

Figure 7 shows the visualization comparison of the method in this paper with the above seven parties, and the listed images are derived from the Mengxi transmission line road dataset. Compared with other advanced methods, MANet obviously has a better visual effect. Taking the first row as an example, the method designed in this paper has a stronger sensitivity to the roads when the image is weakly illuminated and the road distribution is not obvious. Comparing the whole image, we can see that under the difficult situation of image extraction, the method in this paper can still extract a relatively complete road network and effectively ensure the road connectivity to a certain extent, and the edges are smoother, while other methods have a lot of fractures and burrs on the edges.

Experiments on the CHN6-CUG Dataset. To further verify the superiority of MANet, quantitative evaluation was also conducted on the publicly available CHN6-CUG road dataset in this paper, and the results are shown in Table 2.

Figure 8 shows the visualization comparison between MANet and other methods on the CHN6-CUG road dataset.

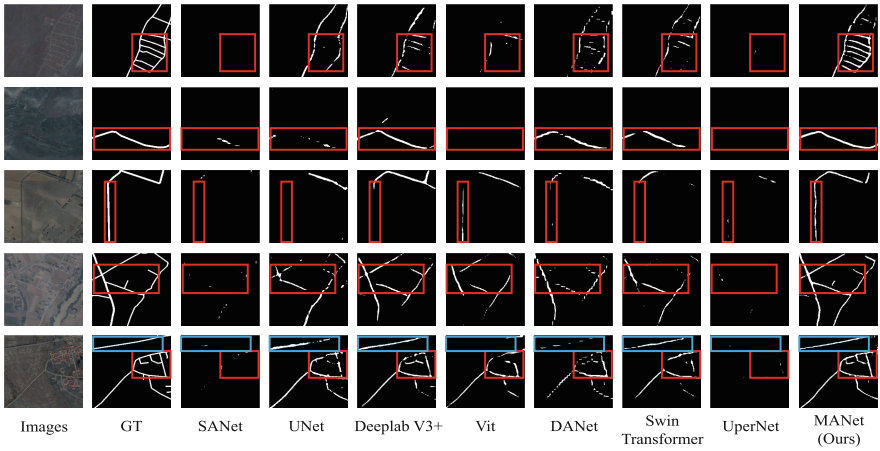

Fig. 7. Comparison of qualitative results between MANet and other advanced methods on Mengxi Transmission Line Road Dataset.

Table 2. Quantitative comparison between MANet and other methods on the CHN6-CUG Road Dataset.

Model	Precise	Recall	F1-score	IoU	MIoU	AA
DANet [24]	78.99	69.13	73.73	58.39	77.73	84.01
Deeplab V3 + [13]	**80.23**	70.36	74.97	59.97	78.58	84.65
U-Net [6]	63.51	66.66	65.01	48.2	71.96	82.71
UperNet [22]	76.35	73.46	74.88	59.85	78.45	86.04
Swin Transfomrer [23]	76.34	73.67	74.98	59.97	78.52	86.14
MANet (Ours)	77.06	**74.08**	**75.54**	**60.69**	**78.91**	**86.37**

4.4 Ablation Experiments

Experiments on the Mengxi Transmission Line Road Dataset. To validate the superior performance of the proposed network structure, a series of ablation studies were conducted in this paper to verify the effectiveness of the used and designed modules separately. Table 3 reports the quantitative evaluation of the effectiveness of each module on the Mengxi transmission line road dataset, and B denotes the baseline model Baseline, which is Deeplab V3 + for this method. Figure 9 presents the enhancement of the extracted visual effects after adding each module.

According to the model No. 1 and No. 2 in Table 3, it can be seen that the SMCA module developed in this paper has gained over the baseline model in road extraction except for the precise. From Fig. 9, it is easy to find that the module has a repair function for road breaks.

Table 4 lists several other contemporary advanced attention mechanism modules, namely SE [26], ECA [27], CBAM [28], and also includes the different effects of adding

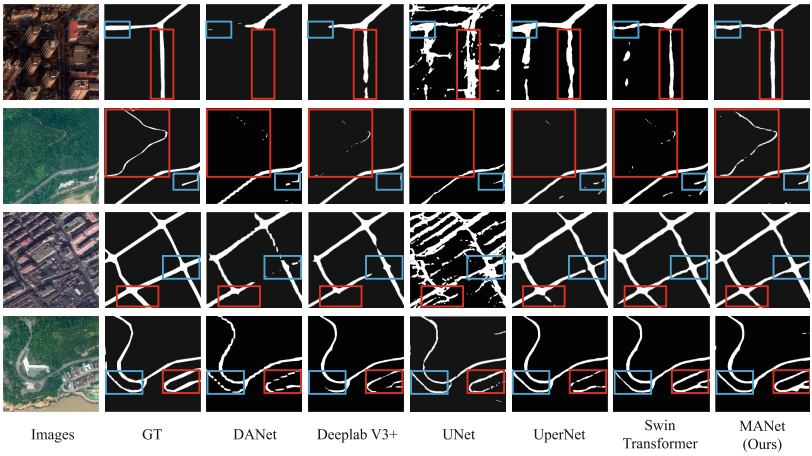

Fig. 8. Comparison of qualitative results between MANet and other advanced methods on the CHN6-CUG road dataset.

Table 3. Quantitative evaluation for key components ablation studies of MANet on the Road Dataset of Mengxi Transmission Line.

Number	Method	Precise	Recall	F1-score	IoU	MIoU	AA
No.1	B	66.37	39.62	49.62	33.0	64.23	69.23
No.2	B + SMCA	64.39	44.23	52.44	35.53	65.5	71.4
No.3	B + SMCA + EAM	65.56	**46.42**	54.35	37.32	66.46	**72.5**
No.4	MANet	**67.42**	45.99	**54.46**	**37.42**	**66.54**	72.33

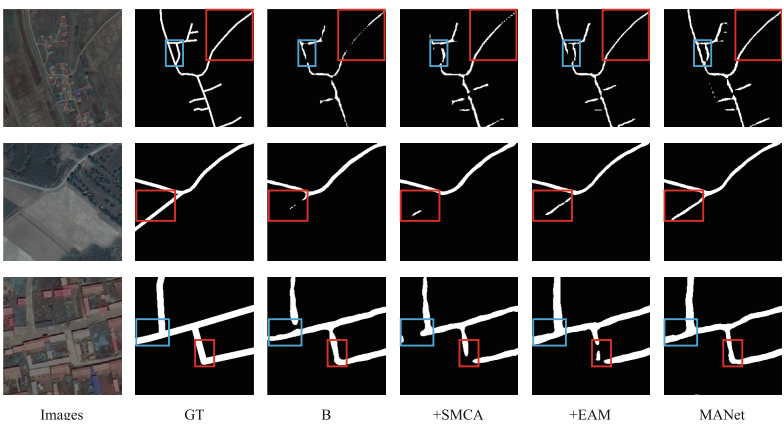

Fig. 9. Qualitative results of ablation studies for MANet on the Mengxi Transmission Line Road Dataset.

Table 4. Quantitative research results of different modules and different positions.

Location	Method	Precise	Recall	F1-score	IoU	MIoU	AA
Encoder	B	66.37	39.62	49.62	33.0	64.23	69.23
	B + SA	66.58	38.43	48.47	32.22	63.83	68.65
	B + MSCA	65.75	41.55	50.92	34.15	64.82	70.14
	B + SE	65.58	42.46	51.63	34.8	65.16	70.59
	B + CBAM	65.02	43.06	51.81	34.96	65.22	70.86
	B + ECA	65.48	42.87	51.82	34.97	65.24	70.78
	B + ACmix(Ours)	64.99	**43.19**	**51.89**	**35.04**	**65.26**	**70.92**
	B + SMCA	**68.14**	37.83	48.65	32.14	63.83	68.4
Decoder	B + SA	**66.47**	39.66	49.68	33.05	64.26	69.25
	B + CBAM	66.05	41.16	50.71	33.97	64.73	69.96
	B + ECA	65.77	42.94	51.95	35.09	65.31	70.82
	B + SMCA(Ours)	64.39	**44.23**	**52.44**	**35.53**	**65.5**	**71.4**

two modules, SA and MSCA, to different positions of the baseline network encoder and decoder, respectively, as approximated in this paper. By adding different modules at the encoder locations separately, the SMCA in this paper achieves the highest accuracy rate by 1.77% compared to the baseline, but other values decrease instead of increasing, which is not satisfactory. After analysis, it is concluded that since the features received at the encoder position itself are deep-level features and contain more global information, the feature fusion module in SMCA does not play a role and other geometric information useful for its extraction is lost when focusing on the line features instead, so SMCA is not suitable at the encoder position. By adding each module at the decoder position, our SMCA achieves the highest gain except for accuracy rate, which is better than SA, ECA & CBAM.

Experiments on the CHN6-CUG Dataset. Through ablation experiments on the CHN6-CUG Dataset, it is demonstrated that each module of our approach remains valid with the overall structure of the model on different datasets. As shown in Table 5.

Table 5. Quantitative evaluation for key components ablation studies of MANet on the CHN6-CUG road dataset.

Number	Method	Precise	Recall	F1-score	IoU	MIoU	AA
No.1	B	**80.23**	70.36	74.97	59.97	78.58	84.65
No.2	B + SMCA	79.81	70.73	75.0	60.0	78.59	84.82
No.3	B + SMCA + EAM	77.09	73.61	75.31	60.04	78.76	86.14
No.4	MANet	77.06	**74.08**	**75.54**	**60.69**	**78.91**	**86.37**

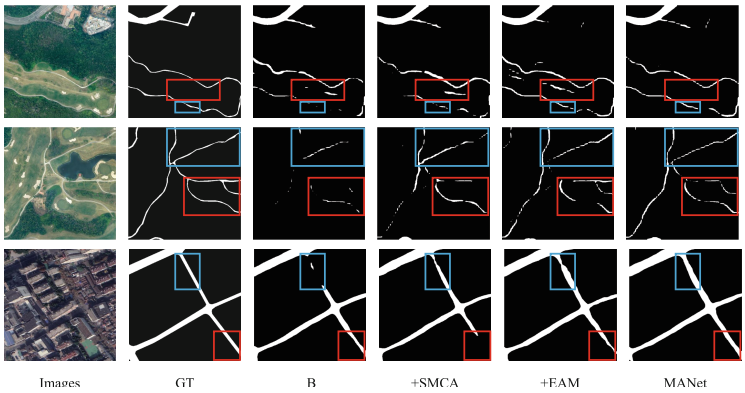

Fig. 10. Qualitative results of ablation studies for MANet on the CHN6-CUG road dataset.

The analysis of the visualization results in the Fig. 10 shows that as the model extracts the road features in depth, it filters out some targets that are extracted incorrectly and corrects the mis-checking phenomenon, making the overall extraction of the model better. At the same time, the line geometric features of road connectivity are used to connect the scattered pixels step by step to identify a more complete road network information.

5 Conclusion

In this paper, an end-to-end extraction method, MANet, is proposed for roads near EHV transmission lines. The model is evaluated in depth on the self-produced road dataset of the Mengxi transmission line and the CHN6-CUG Dataset. The experimental results and comparative analysis show that the algorithm is competitive and advantageous in the road network extraction task. For MANet, the next work can be complemented by incorporating the centerline extraction subtask to improve the network extraction performance. At the same time, the model we designed is focused on solving specific tasks and can be improved in the future to enhance its robustness and expand its applicability.

Acknowledgments. The authors gratefully acknowledge the financial supports by the National Natural Science Foundation of China under Grant No. 62077032, as well as the Inner Mongolia Science and Technology Plan Project under project No. 2021GG0159.

References

1. Hoeser, T., Kuenzer, C.J.R.S.: Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part i: Evolution and Recent Trends. **12**, 1667 (2020)
2. Mnih, V.: Machine Learning for Aerial Image Labeling. University of Toronto (Canada) (2013)
3. Wang, J., Song, J., Chen, M., Yang, Z.J.I.J.o.R.S.: Road Network Extraction: A Neural-Dynamic Framework Based on Deep Learning and a Finite State Machine **36**, 3144-3169 (2015)

4. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation (2017)
5. Zhong, Z., Li, J., Cui, W., Han, J.: Fully convolutional networks for building and road extraction: preliminary results. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (2016)
6. Ronneberger, O., Fischer, P., Brox, T.: 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, pp. 234–241. Springer (2015)
7. Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., Pan, C.J.I.T.o.G., Sensing, R.: Automatic road Detection and Centerline Extraction via Cascaded End-to-end Convolutional Neural Network **55**, 3322–3337 (2017)
8. Zhou, L., Zhang, C., Wu, M.: D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 182–186 (2018)
9. Shi, Q., Liu, X., Li, X.J.I.a.: Road Detection from Remote Sensing Images by Generative Adversarial Networks **6**, 25486–25494 (2017)
10. Vaswani, A., et al.: Attention is All You Need **30** (2017)
11. Jiang, X., et al.: Geoinformation: RoadFormer: Pyramidal Deformable Vision Transformers for Road Network Extraction with Remote Sensing Images **113**, 102987 (2022)
12. Zhang, Z., Miao, C., Liu, C., Tian, Q., Zhou, Y.J.M.: HA-RoadFormer: Hybrid Attention Transformer with Multi-Branch for Large-Scale High-Resolution Dense Road Segmentation **10**, 1915 (2022)
13. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818 (2018)
14. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 558–567 (2019)
15. Pan, X., et al.: On the integration of self-attention and convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 815–825 (2022)
16. Sun, Y., Chen, G., Zhou, T., Zhang, Y., Liu, N.J.a.p.a.: Context-Aware Cross-Level Fusion Network for Camouflaged Object Detection (2021)
17. Huan, H., Sheng, Y., Zhang, Y., Liu, Y.J.R.S.: Strip Attention Networks for Road Extraction **14**, 4516 (2022)
18. Zhu, Q., et al.: A Global Context-Aware and Batch-Independent Network for Road Extraction from VHR Satellite Imagery **175**, 353–365 (2021)
19. MMSegmentation contributors. MMSegmentation: Openmmlab Semantic Segmentation Toolbox and Benchmark (2020). https://github.com/openmmlab/mmsegmentation. Accessed 11 Aug 2020
20. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22–27, 2010 Keynote, Invited and Contributed Papers, pp. 177–186. Springer (2010)
21. Abdollahi, A., Pradhan, B., Shukla, N., Chakraborty, S., Alamri, A.J.R.S.: Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-of-the-Art Review **12**, 1444 (2020)
22. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 418–434 (2018)

23. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
24. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
25. Dosovitskiy, A., et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2020)
26. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
27. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)
28. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)