A Deep Neural Network for Road Extraction with the Capability to Remove Foreign Objects with Similar Spectra

Haiqing He [1, 2], Yan Wei [1], Fuyang Zhou[1], Hai Zhang[1]

[1] School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China (hyhqing, weiyan08210921, fuy_zhou, haizhang2024) @163.com

[2] Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China

Keywords: Road extraction, Deep neural network, Dual channel model, Feature fusion, Similar spectra.

Abstract

Existing road extraction methods based on deep learning often struggle with distinguishing ground objects that share similar spectral information, such as roads and buildings. Consequently, this study proposes a dual encoder-decoder deep neural network to address road extraction in complex backgrounds. In the feature extraction stage, the first encoder-decoder designed for extracting road features. The second encoder-decoder utilized for extracting building features. During the feature fusion stage, road features and building features are integrated using a subtraction method. The resultant road features, constrained by building features, enhance the preservation of accurate road feature information. Within the feature fusion stage, road feature maps and building feature maps designated for fusion are input into the convolutional block attention module. This step aims to amplify the features of different channels and extract key information from diverse spatial positions. Subsequently, feature fusion is executed using the element-by-element subtraction method. The outcome is road features constrained by building features, thus preserving more precise road feature information. Experimental results demonstrate that the model successfully learns both road and building features concurrently. It effectively distinguishes between easily confused roads and buildings with similar spectral information, ultimately enhancing the accuracy of road extraction.

1. Introduction

The significance of road extraction lies in its ability to automatically identify and delineate roads from satellite images or maps. This process is essential for various applications such as urban planning, transportation management, infrastructure development, environmental monitoring, and disaster response. Accurate road extraction facilitates navigation systems, helps improve traffic flow analysis, aids in updating maps, and supports various location-based services.

Traditional road extraction methods are labor-intensive, relying on designing features based on road texture, shape, edges, and other characteristics. The accuracy of extraction is not high, and the robustness is poor. Moreover, this method is not suitable for road extraction in complex scenarios. Support vector machine (SVM) classification method (Simler, 2011) and Markov random field classification method (MRF classification method) (Li et al., 2017), formulate rules according to the spectral and spatial characteristics of roads, extract road fragments from images, and further refine them. He et al. proposed a color-based road detection algorithm by combining boundary estimation results from grayscale images with road region extraction results from color images (He et al., 2004). SIMLER et al. proposed an SVM technique using spectral and spatial features to extract roads from aerial images with a spatial resolution of 0.5m. YAGER et al. used SVM to extract roads from aerial images with a spatial resolution of 0.45m by utilizing important features such as edge length, intensity and gradient (YAGER and Sowmya, 2003). Wegner et al. proposed a high-order conditional random field (CRF) model for road network extraction (Wegner et al., 2013).

Currently, deep learning is highly favored in the field of semantic segmentation. An increasing number of studies are using deep learning to tackle various problems. Among them, the most commonly used are Convolutional Neural Network (CNN) and Fully Convolutional Networks (FCN) (Long et al., 2015). Additionally, with the advancement of deep learning, transformers have also been widely employed. U-Net architecture adopts cascaded upsampling and combines multiple loss functions for road extraction (Ronneberger et al., 2015). To minimize information loss, LinkNet directly connects the encoder to the decoder (Chaurasia and Culurciello, 2017). D-LinkNet adopts the LinkNet architecture and utilizes shortcut connections in the central part to combine atrous convolution blocks into several parallel branches (Zhou et al., 2018). Due to occlusions from buildings and shadows, discontinuities occur in roads. Therefore, in the CoANet model, a connectivity attention module (CoA) is designed to address the continuity issues in roads (Mei et al., 2021). Zhang et al. proposed a semantic segmentation neural network for road extraction that combines the advantages of residual learning and U-Net (Zhang et al., 2018). Since CNNs struggle to capture global representations, transformers are used to obtain comprehensive contextual information. Therefore, Seg-Road (Tao et al., 2023) and DPENet (Chen et al., 2023) models effectively combine local and global information using a dual-encoder structure for road extraction. SemiRoadExNet is a semi-supervised road extraction framework that employs Generative Adversarial Networks (GANs) and utilizes multiple discriminators to ensure consistency in feature distributions between labeled and unlabeled data, enhancing the generalization capability of the model (Chen et al., 2023).

However, road extraction methods based on deep learning still have limitations. These include issues with road connectivity, object occlusion, and difficulty distinguishing between objects with similar spectral characteristics (such as roads and buildings). Therefore, this paper proposes a dual-encoder-decoder structure to simultaneously learn features of roads and buildings, with building features suppressing the learning of road features. Additionally, Convolutional Block Attention Module (CBAM) is employed to enhance features, reduce semantic information loss, and more effectively utilize extracted information. Our contributions are as follows:

- We propose a model consisting of a dual-encoder-decoder architecture to simultaneously learn features of roads and buildings, which are prone to confusion.
- We employ CBAM to enhance features, extract, and leverage more shared information.
- We adopt an exclusion strategy for feature fusion, using building features to suppress the learning of road features, thereby reducing misclassification during road extraction.

2. Methods

Our proposed model primarily consists of two encoder-decoder structures and a CBAM module integrated with element-wise subtraction for feature fusion, as illustrated in Fig. 1.

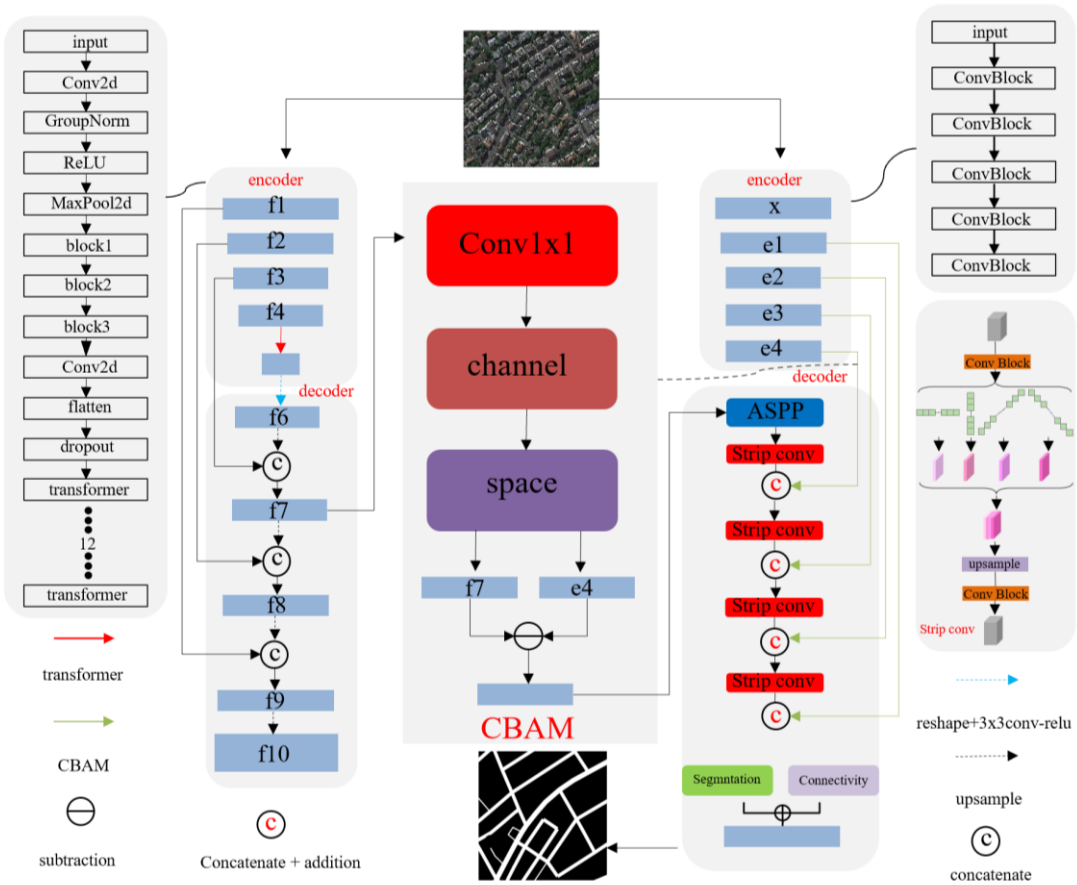

Figure 1. Architecture of the proposed network.

2.1 Dual Encoder-Decoder Network Model Structure

The dual encoder-decoder architecture is mainly based on the encoder-decoder structures of the CoANet (Mei et al., 2021) and TransUNet (Chen et al., 2021) models.

The first encoder-decoder structure is dedicated to extracting road features, with the decoder segment utilizing a pre-trained ResNet101. It comprises five modules, with the first module containing a convolutional layer, batch normalization layer, activation function, and max-pooling layer. The subsequent four modules consist of convolutional layers, normalization layers, and activation functions, with layer depths of 3, 4, 23, and 3, respectively. The last two modules employ dilated convolutions with dilation rates of 2 and 4 to extract denser features, resulting in five feature maps. The deepest feature map, along with the feature map extracted from the second encoder-decoder structure for feature fusion, is input into CBAM. This amplifies features from different channels, extracts crucial information from various spatial positions, performs feature fusion using element-wise subtraction, and then feeds the fused features into Atrous Spatial Pyramid Pooling (ASPP) to increase the receptive field. Finally, it enters a decoder containing four strip conv modules, each capturing contextual information in four different directions. Finally, the feature map is inputted into a decoder with four strip conv modules, each containing four different directional strip convolutions to capture contextual information. Connected after the decoder are the segmentation

branch and the connectivity branch. The connectivity branch integrates SE (Squeeze-and-Excitation) for attention-weighted processing of feature maps across different channels.

To distinguish between easily confused road and building features, we employ parallel encoder-decoder structures to avoid issues such as information loss due to small feature maps caused by deepening the model. The second encoder adopts ResNetV2, comprising three modules, each composed of convolutional layers, batch normalization layers, and activation functions. The encoder part produces three feature maps and one for obtaining global contextual information. The feature map used to acquire global contextual information is serialized, passed through twelve transformer layers, with positional encoding added. The resulting sequence is reshaped, and then passed through convolutional and activation functions. The resulting feature map is upsampled and fused with the feature map of the same shape from the encoder structure. The fused feature map is upsampled and then merged with the feature map of the same shape from the encoder structure. Then, the fused features are upsampled and fused once again. Finally, the fused feature map is upsampled, and segmentation is performed to obtain the predicted binary image.

2.2 Feature Fusion

The feature fusion structure diagram is shown in Fig. 2. The deepest feature map from ResNet101 and the feature map used for feature fusion from the second encoder-decoder structure are inputted into CBAM. Initially, they pass through a channel attention module, followed by a spatial attention module. The channel attention module performs max-pooling and average-pooling operations on the input feature maps, then combines them through a Multilayer Perceptron (MLP) and element-wise addition. The resulting feature map is multiplied element-wise with the input feature map to produce the input feature map for the spatial attention module. The spatial attention module conducts max-pooling and average-pooling operations on the feature maps, fuses them based on channels, applies a convolution operation, and passes through a sigmoid to obtain the final feature map. The road and building feature maps processed by CBAM are fused using element-wise subtraction to yield the final fused feature. The formula is as follows:

$$F = F_{road} - F_{build} \tag{1}$$

where F_{road} = the road feature map processed by CBAM
F_{build} = the building feature map processed by CBAM
F = the fused feature
- = the element-wise subtraction

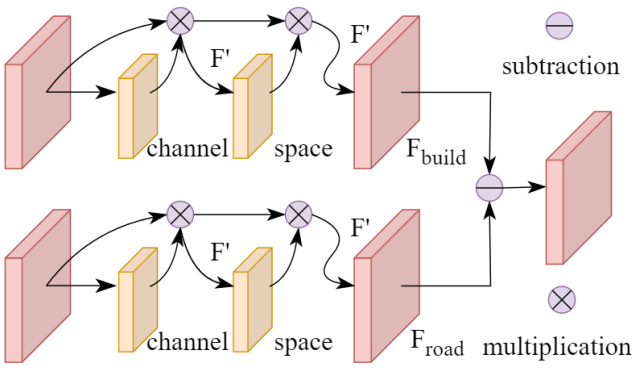

Figure 2. Feature fusion structure diagram.

2.3 Loss Function

The loss function plays a crucial role in the model training process. After each batch of data is input into the model and predicts values through forward propagation, the loss function calculates the difference between the predicted values and the ground truth. Then, through backpropagation, the parameters in the model are updated to minimize this difference, thereby allowing the model to converge and achieve the training objective. Therefore, the choice of loss function is also critical during model training.

This paper employs a combination of BCE (Binary Cross-Entropy) and Dice loss functions to address the foreground-background class imbalance issue in images. Based on the model architecture shown in Figure 1, the loss function L mainly consists of two parts: L_1 and L_2. L_1 represents the loss function of the first encoder-decoder structure, while L_2 represents the loss function of the second encoder-decoder structure. Since the first encoder-decoder part includes both linking branches and segmentation branches, the loss function in L_1 also comprises two parts: L_s and L_c. The linking branch is used to determine the connectivity between the current pixel and its surrounding eight pixels. Therefore, the loss function is formulated as follows.

$$\begin{aligned}
L_B &= -\frac{1}{H*W}\sum_{i=1}^{H*W}\left[y_i * log\left(y'_i\right) + \left(1-y_i\right) * log\left(1-y'_i\right)\right] \\
L_D &= \frac{2*\sum_{i=1}^{H*W}\left(y_i * y'_i\right)}{\sum_{i=1}^{H*W} y_i^{\,2} + \sum_{i=1}^{H*W} y'^{\,2}_i} \\
L_S &= L_B + \alpha\left(1-L_D\right) \\
L_C &= L_{d1} + \alpha' L_{d1} \\
L_{d1} &= -\frac{1}{Q_0 * H * W}\sum_{c=1}^{Q_0}\sum_{i=1}^{H*W}\left[y_c * log\left(y'_c\right) + \left(1-y_c\right) * log\left(1-y'_c\right)\right] \\
L_{d3} &= -\frac{1}{Q_0 * H * W}\sum_{c=1}^{Q_0}\sum_{i=1}^{H*W}\left[y_c * log\left(y'_c\right) + \left(1-y_c\right) * log\left(1-y'_c\right)\right] \\
L_1 &= L_S + \alpha'' L_C \\
L_2 &= L_B + \alpha\left(1-L_D\right) \\
L &= L_1 + \alpha'''\left(1-L_2\right)
\end{aligned} \tag{2}$$

where L_B = BCE loss function
L_D = Dice loss function
y_i = the ground truth
y'_i = the prediction of the segmentation branch
Q_0 = the number of surrounding pixels
y_c = the connectivity of the pixel with its surrounding pixels
y'_c = the prediction of the linking branch
α = adjust the weights of the BCE and Dice loss functions
α' = adjust the weights of the two linking branch loss functions
α'' = adjust the weights of the segmentation branch and linking branch loss functions
α''' = adjust the weights of the two encoder-decoder structure loss functions

3. Experiments

3.1 Datasets

Aerial Image Segmentation Dataset (Kaiser et al., 2017): The aerial images are divided into aerial remote sensing images from Google Maps and pixel-level buildings, roads and background labels from OpenStreetMap. Sourced from the website (https://zenodo.org/records/1154821#.XH6HtygzbIU). It covers Berlin, Chicago, Paris, Potsdam and Zurich. The image of part of the Zurich area was selected as the data set, which was cropped to 512*512 pixels, and the road and building parts in the label were extracted separately as the labels of their respective channel models. Among them, 8070 images were used as the training set, 1020 images were used as the verification set, and 1080 images were used as the test set.

3.2 Implementation Details

Training is conducted under the PyTorch deep learning framework, with training platform parameters as shown in Table 1 and model parameter settings for training as shown in Table 2. The poly learning rate decay strategy used in Table 2 is as shown in the formula.

$$lr = init\_lr * \left(\frac{1 - iter}{\max\_iter} \right)^{power} \tag{3}$$

where lr = the learning rate
init_lr = the initial learning rate
iter = the number of iterations
max_iter = the maximum number of iterations
power = the exponent used to control the rate at which the learning rate decreases as the number of iterations increases

Platform	Configuration
CPU	Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz
GPU	NVIDIA GeForce RTX 3090 24.0GB
Memory	16GB
DL Framework	Pytorch V1.12.0
Compiler	PyCharm 2023.1
Program	Python V3.8.0
Parallel computing	CUDA V11.3.1
DL Accelerator	cuDNN V8.3.2

Table 1. Training platform configuration

Training Settings	
Optimizer	SGD
LR Policy	Poly
Loss Functions	BCE/Dice
Initial learning rate	0.001
Momentum	0.9
Weight decay	5e-4
Batch size	4
Epoch Loss Functions	50

Table 2. Training settings

3.3 Evaluation Metrics

Road extraction is commonly perceived as a binary classification problem. The commonly used model performance evaluation metrics are Overall Accuracy (OA), Precision, Recall, Intersection over Union, and F1-score. We adopt the average Precision, the average Recall, the average Intersection over Union, and F1-score for both foreground and background, along with OA, to evaluate the road extraction performance of our proposed model. The formulas for OA, Precision, Recall, Intersection over Union, and F1-score are as follows.

$$\begin{aligned} OA &= \frac{(TP + TN)}{(TP + TN + FP + TN)} \\ Precision &= \frac{TP}{(TP + FP)} \\ Recall &= \frac{TP}{(TP + FN)} \\ F1\_score &= \frac{(2 * Precision * Recall)}{(Precision + Recall)} \\ IoU &= \frac{TP}{(TP + FP + FN)} \end{aligned} \tag{4}$$

where TP = the number of pixels correctly predicted as roads
FP = the number of pixels incorrectly predicted as roads
TN = the number of pixels correctly predicted as background
FN = the number of pixels missed in predicting as roads
OA = the proportion of correctly predicted pixels to the total number of pixels
Precision = the proportion of correctly predicted samples to the predicted samples
Recall = the ratio of correctly predicted positive samples to the total number of true positive samples
F1-score = the harmonic mean of precision and recall
IoU = represents the ratio of the intersection of the actual region and the predicted region to the union of the actual region and the predicted region

3.4 Comparative Analysis of other Modules

To validate the feasibility of the model proposed in this paper for road extraction, we compared it with five other state-of-the-art models (PSPNet50 (Zhao et al., 2017), TransUNet (Chen et al., 2021), DeepLabV3 (Chen et al., 2017), D-LinkNet (Zhou et al., 2018), and CoANet (Mei et al., 2021)) on the Aerial Image Segmentation Dataset. The accuracy results of the comparison are shown in Table 3.

Model	OA (%)	mPre (%)	mRecall (%)	mIoU (%)	mF1 (%)
PSPNet50	90.16	80.40	77.22	70.76	78.78
TransUNet	89.78	79.42	71.72	65.44	75.37
DeepLabV3	90.69	80.52	80.43	73.26	80.47
D-LinkNet	89.71	79.54	78.36	64.75	78.95
CoANet	90.55	79.84	88.61	74.32	84.00
Proposed	91.26	81.98	89.21	75.34	85.44

Table 3. Compare with five other state-of-the-art models

The comparison results of model accuracy in Table 3 indicate that our proposed model outperforms others on all five metrics: OA, mPre, mRecall, mIoU, and mF1. Among the six networks evaluated, CoANet exhibited the second-best performance in identifying roads, with superior performance in other metrics. DeepLabV3 outperformed others in five metrics. However, TransUNet, PSPNet50, and D-LinkNet show average performance across OA, mPre, mRecall, mIoU, and mF1 metrics. TransUNet has the lowest mPre, mRecall, and mF1, while D-LinkNet has the lowest OA and mIoU.

To further validate the feasibility of CBAM in road extraction, ablation experiments were conducted, and the accuracy results are shown in Table 4.

Model	OA (%)	mPre (%)	mRecall (%)	mIoU (%)	mF1 (%)
A	90.30	81.39	88.54	74.02	84.81
Proposed	91.26	81.98	89.21	75.34	85.44

Table 4. Comparison of CBAM ablation experiments

In Table 4, Model A only inputs two feature maps used for feature fusion into CBAM, whereas our proposed model inputs four feature maps obtained from the first encoder and the feature map obtained from the second encoder-decoder used for model fusion into CBAM. The accuracy comparison results in Table 4 demonstrate that our proposed method achieves the highest scores on metrics OA, mPre, mRecall, mIoU, and mF1. The OA of proposed is 0.96% higher than that of Model A, mPre is 0.59% higher, mRecall is 0.67% higher, mIoU is 1.32% higher, and mF1 is 0.63% higher.

To more intuitively assess the feasibility of the integrated feature module, the heatmap of the feature map was visualized, with the results displayed in Fig. 3.

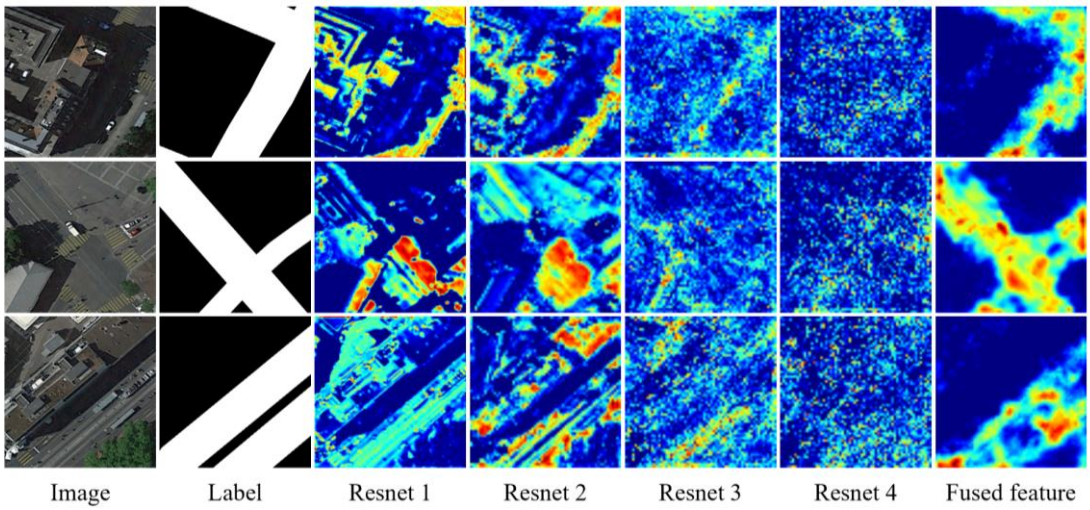

Figure 3. Heat maps of each layer of model encoder and fusion features.

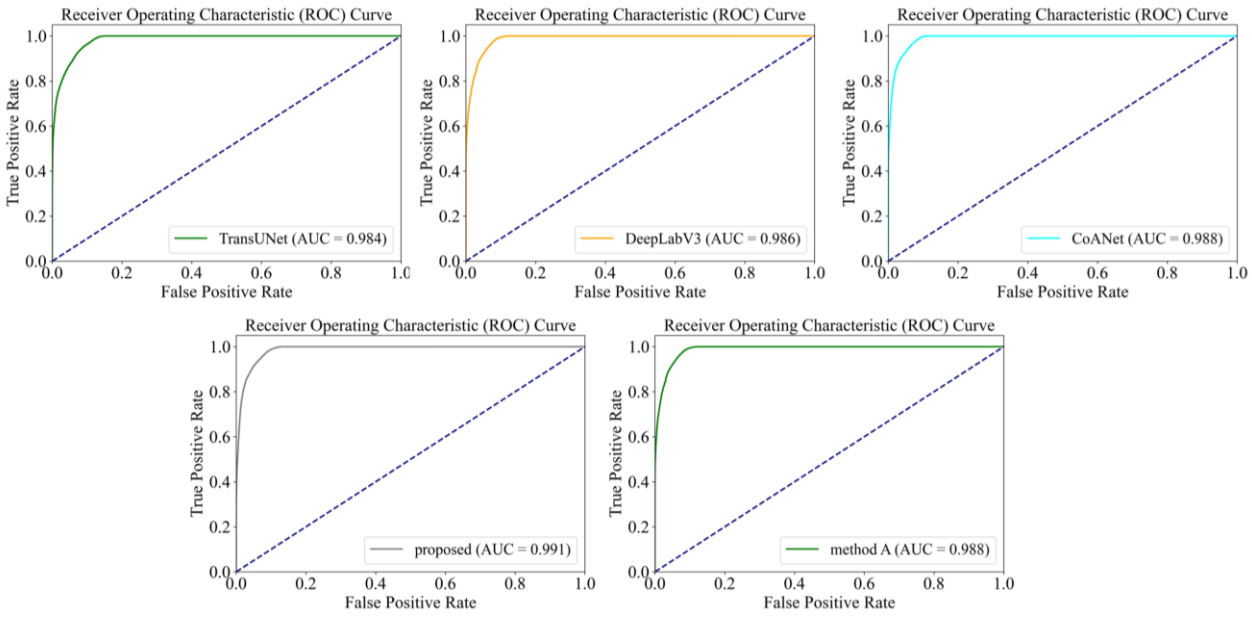

Figure 4. ROC curves and AUC values of 5 models.

The heat map in Fig. 3 demonstrates that after applying the feature fusion module, it becomes possible to distinguish between the features of roads and buildings, resulting in more accurate feature extraction. This indicates the effectiveness of the feature fusion module we utilized. However, there are still areas for improvement, particularly in extracting features at the edges and in regions obscured by trees and shadows.

In Fig. 4, the ROC curve plot has the true positive rate on the y-axis and the false positive rate on the x-axis. The closer the curve approaches the upper-left corner, the better the performance of the model. The AUC value represents the area under the ROC curve, with a larger AUC indicating better model performance. It can be observed from the graph that our proposed model performs the best.

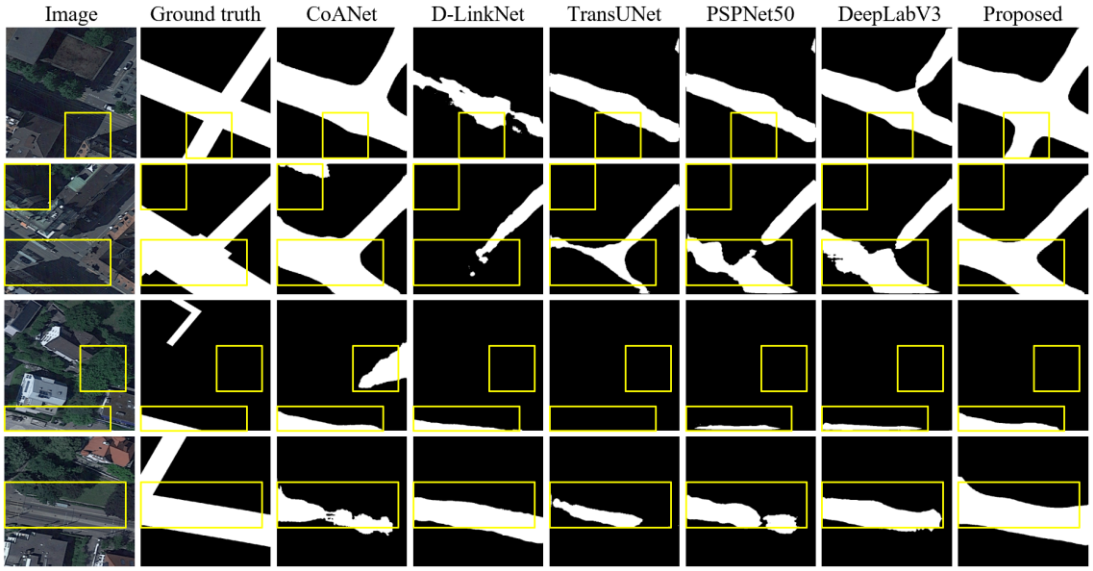

Figure 5. Comparison of model results.

The comparative results of various models in Fig. 5 demonstrate that our proposed method achieves the best extraction performance in shadow areas, with fewer instances of misextraction and strong performance in road connectivity. However, its performance in edge processing remains suboptimal.

4. Conclusion

In this paper, we propose a neural network model with a dual encoder-decoder architecture. By employing a dual-channel framework, we conduct separate learning of road and building features. We incorporate the CBAM attention mechanism with element-wise subtraction to amplify features from different channels, extracting crucial information from distinct spatial locations of roads and buildings. This facilitates feature fusion to distinguish between roads and buildings that are prone to confusion in images. The experimental results indicate that our proposed method shows improvement compared to other state-of-the-art network models in distinguishing between complex roads and buildings, with better road extraction performance.

Despite optimizing the model's performance, shortcomings still exist. Results from the Aerial Image Segmentation Dataset indicate subpar performance in extracting road edges. Additionally, the model is relatively large, resulting in longer training times. Therefore, in future research, our objective is to address road edge extraction issues and modify the model into a lightweight version to achieve more efficient, rapid, and accurate road extraction.

References

Simler, C., 2011. An improved road and building detector on VHR images. *Proc. IEEE Int. Geosci. Remote Sens. Symp*, 507-510. doi.org/10.1109/IGARSS.2011.6049176.

Li, Y., Zhang, R., Wu, Y., 2017. Road network extraction in high-resolution SAR images based CNN features. *Proc. IEEE Int. Geosci. Remote Sens. Symp*, 1664–1667. doi.org/10.1109/IGARSS.2017.8127293.

He, Y., Wang, H., Zhang, B., 2004. Color-based road detection in urban traffic scenes. *IEEE Trans. Intell. Transp. Syst*, 5(4), 309–318. doi.org/10.1109/ITSC.2003.1252047.

Yager, N., Sowmya, A., 2003. Support vector machines for road extraction from remotely sensed images. *International Conference on Computer Analysis of Images and Patterns*, Berlin, Heidelberg, Springer Berlin Heidelberg, 285-292. doi.org/10.1007/978-3-540-45179-2_36.

Wegner, J.D., Montoya-Zegarra, J.A., Schindler, K., 2013. A higher-order CRF model for road network extraction. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 1698–1705. doi.org/10.1109/CVPR.2013.222.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 3431–3440. doi.org/10.1109/TPAMI.2016.2572683.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. *Proc. Med. Image. Comput. Comput. A-ssist. Interv*, Cham,

Switzerland: Springer, 234–241. doi.org/10.1007/978-3-319-24574-4_28.

Chaurasia, A., Culurciello, E., 2017. LinkNet: Exploiting encoder represent-tations for efficient semantic segmen-tation. *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, 1–4. doi.org/10.48550/arXiv.1707.03718.

Zhou. L., Zhang. C., Ming. W., 2018. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition. Workshops (CVPRW)*, 182-186. doi.org/10.1109/CVPRW.2018.00034.

Mei, J., Li, R.J., Gao, W., Cheng, M.M., 2021. CoANet: Connectivity attention network for road extraction from satellite imagery. *IEEE Transactions on Image Processing*, 30, 8540-8552. doi.org/10.1109/TIP.2021.3117076.

Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual U-Net. *IEEE Geosci. Remote Sens. Lett*, 15(4), 749–753. doi.org/10.1109/LGRS.2018.2802944.

Tao, J., Chen, Z., Sun, Z., Guo, H., Leng, B., Yu, Z., Wang, Y., He, Z., Lei, X., Yang, J.P., 2023. Seg-Road: A Segmentation Network for Road Extraction Based on Transformer and CNN with Connectivity Structures. *Remote Sens*, 15(6). doi.org/10.3390/rs15061602.

Chen, Z.Y., Wang, C., Li, D.L., Luo, Y., Wang, J., Li, J., 2023. DPENet: Dual-path extraction network based on CNN and Transformer for accurate building and road extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 10(5). doi.org/10.1016/j.jag.2023.103510.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. doi.org/10.48550/arXiv.2102.04306.

Chen, H., Li, Z., Wu, J., Xiong, W., Du, C., 2023. SemiRoadExNet: A semisupervised network for road extraction from remote sensing imagery via adversarial learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198, 169–183. doi.org/10.1016/j. isprsjprs.2023.03.012.

Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation From Online Maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11), 6054 – 6068). doi.org/10.1109/TGRS.2017.2719738.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2881-2890. doi.org/10.1109/CVPR.2017.660.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. doi.org/10.48550/arXiv.1706.05587.