IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 8 November 2023, accepted 15 December 2023, date of publication 19 December 2023,
date of current version 27 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3344797

RESEARCH ARTICLE

Remote Sensing Image Road Segmentation Method Integrating CNN-Transformer and UNet

RUI WANG[1], MINGXIANG CAI[1], ZIXUAN XIA[2], AND ZHICUI ZHOU[3]

[1]China Transport Telecommunications and Information Center, Beijing 100011, China
[2]Heilongjiang University of Technology, Harbin, Heilongjiang 150022, China
[3]No. 1 Middle School, Jixi, Heilongjiang 150022, China

Corresponding author: Mingxiang Cai (jamtsai@whu.edu.cn)

ABSTRACT Real-time and accurate road information is crucial for updating electronic navigation maps. To address the problem of low precision and poor robustness in current semantic segmentation methods for road extraction from remote sensing imagery, we proposed a UNet road semantic segmentation model based on attention mechanism improvement. First, we introduce a CNN-Transformer hybrid structure to the encoder to enhance the feature extraction capabilities of global and local details. Second, the traditional upsampling module in the decoder is replaced with a dual upsampling module to improve feature extraction capabilities and segmentation accuracy. Furthermore, the hard-swish activation function is used instead of ReLU activation function to smooth the curve, which helps to improve the generalization and non-linear feature extraction abilities and avoid gradient vanishing. Finally, a comprehensive loss function combining cross entropy and dice is used to strengthen the segmentation result constraints and further improve segmentation accuracy. Experimental validation is performed on the Ottawa Road Dataset and the Massachusetts Road Dataset. Experimental results show that compared with U-Net, PSPNet, DeepLab V3 and TransUNet networks, this algorithm is the best in terms of MIoU, MPA and F1 score. Among them, on the Ottawa road data set, the MPA of this algorithm reached 95.48%. On the Massachusetts road data set, MPA is 92.56%. This method shows good performance in road extraction.

INDEX TERMS Road segmentation, deep learning, CNN-transformer, attention, UNet.

I. INTRODUCTION

Real-time and accurate road information is crucial for updating navigation electronic maps, and road extraction is an important issue in the field of computer vision [1]. In practical applications, roads are an important basis for transportation, and accurate road extraction can improve traffic safety and efficiency in fields such as autonomous driving and intelligent transportation [2]. With the continuous development of computer vision and artificial intelligence, road extraction algorithms based on deep learning have become a very popular research direction [3]. Compared with traditional algorithms, deep learning-based algorithms have stronger self-learning and feature extraction capabilities, better adaptability, and higher accuracy [4]. These algorithms can extract road information from images, help autonomous vehicles better understand the environment in which they operate, provide correct driving decisions, and improve the safety and efficiency of autonomous vehicles [5]. In addition, road extraction is also crucial in urban planning. By extracting road information in the city, urban planners can better understand information such as traffic flow and road layout, so as to better carry out urban planning and improve traffic conditions [6]. Therefore, road extraction can improve the safety and efficiency of self-driving vehicles and help urban planners better understand urban road layout and traffic flow, thereby better planning the city and improving traffic conditions.

Traditional road extraction algorithms mainly rely on geometric and physical principles in image processing, and usually require a large amount of manual intervention and parameter optimization [7], [8]. In recent years, deep

The associate editor coordinating the review of this manuscript and approving it for publication was Akansha Singh.

learning algorithms have achieved significant success in the field of computer vision, especially in speech and image recognition [9], [10]. Deep learning algorithms have powerful self-learning and feature extraction capabilities, and can deeply mine deep features in data. They are also used to solve road extraction problems [11], [12]. How to achieve high-reliability road extraction based on deep learning algorithms, in-depth research on the better adaptability of deep learning models to complex terrain images, and achieve high accuracy and robustness in practice have become hot issues. In recent years, commonly used semantic segmentation methods for road extraction include traditional methods and image segmentation based on deep learning [13], [14].

Traditional semantic segmentation methods use the texture, color, geometric shape and spatial structure information of the image to segment objects, dividing pixels with the same semantics into a region, and there is no intersection between each region [2], [15]. Traditional semantic segmentation can be divided into threshold-based, clustering-based, edge-based and region-based segmentation types [3], [16]. However, traditional methods have high computational complexity and long processing time, and cannot effectively handle high-resolution images with noise, multiple objects, and complex backgrounds. Therefore, the accuracy of the segmentation results is low, resulting in the limited scope of application of traditional methods [17], [18]. With the emergence of deep neural networks, image segmentation technology based on deep learning has entered a stage of rapid development beyond traditional methods [19], [20].

Currently, existing research mainly relies on designing road extraction algorithms based on convolutional neural networks (CNNs) [21], [22]. The principle of CNN is to use a large amount of road extraction training data to train weights. In the process of scanning the entire image, convolution operations are performed and further feature information is extracted from the image. Compared with traditional methods, this method has good self-learning and feature extraction capabilities, can automatically mine certain features in the data, and provide effective solutions for scenarios in complex environments. However, the CNN network has shortcomings such as inputting fixed-size images. In recent years, Long et al. [23] proposed fully convolutional networks (FCN) to introduce encoders and decoders into the field of image segmentation. Ronneberger et al. [24] proposed an improved version of FCN called UNet. This version connects the feature maps of the encoder and decoder to form a ladder structure, allowing each decoder to learn features lost in the encoder, thereby improving the segmentation capabilities of remote sensing images. Zhao et al. [25] proposed PSPNet, which uses the pyramid pooling module to significantly improve the ability to extract global information features. Chen et al. [26] proposed DeepLabV3, which uses dilated convolution to extract features, and uses ASPP module to expand the receptive field and enhance the feature extraction ability. The above networks have achieved good application results in remote sensing image segmentation tasks. Therefore, many scholars in the field of remote sensing have improved the above network and designed a network structure suitable for remote sensing image road segmentation tasks. Among them, Kong et al. [27] proposed an improved UNet network for extracting road information from remote sensing images. This method adds a stripe pool module to the down-sampling part of the coding layer to focus on local information. A hybrid pool module is added to the convolution of the coding layer to enhance its ability to obtain network context. The algorithm was verified by using the GF-2 remote sensing image data set. The results show that the algorithm can effectively extract the road. Li et al. [28] proposed an improved UNet network, which combines core attention and global attention to extract roads from remote sensing images. Experiments were conducted on the Massachusetts road dataset and the DeepGlobe CVPR 2018 road dataset. The results show that the method can effectively extract the road area occluded by the canopy and improve the connectivity of the road network. Han et al. [29] added dilated convolution to extract road information on the basis of DeepLab V3 network. Experiments show that this method has higher extraction accuracy than other methods. Liu et al. [30] proposed a DeepLab V3 + road extraction method with attention module. This method obtains more spatial context information through spatial attention and enhances the extraction of road information. The effectiveness of the method is verified by using the Cityscapes dataset. The results show that the ability of this method to extract road information is significantly better than that of the original network.

Although these methods have their own advantages in road extraction, they still have some shortcomings. There is still the problem of incomplete context information extraction during the road extraction process. And with the improvement of remote sensing image quality, problems such as lack of road details and discontinuous extraction are fully exposed under complex background information. Since its establishment in 2017, Transformer has quickly dominated the field of natural language processing (NLP) and has become an absolute leader in a short period of time. Vaswani et al. [31] proposed Transformer 's self-attention (SA) mechanism, which focuses on extracting the feature information of the object of interest in the feature map, reducing unnecessary information and improving the efficiency of feature extraction. Chen et al. [32] proposed TransUNet (a combination of Transformers and UNet), which integrates Transformer into UNet network to obtain global image connection, realize multi-scale prediction and supervision of feature maps, and combines the advantages of Transformer and UNet. The above scholars have verified that the integration of Transformer into the UNet network has a good effect, but when it is applied to the road semantic segmentation scene, it is easily affected by seasonal changes, lighting conditions and environmental impacts. Accurate segmentation in complex background is challenging. In the coding process, there is a lack of local information exchange, and the ability to extract features such as image edges and geometric shapes is relatively weak. In order to solve the above problems, we propose an improved road semantic segmentation network based on UNet. The main contributions of this work are as follows:

(1) We introduce a combined loss function of cross entropy (CE) and Dice to strengthen the constraints of the model on the segmentation results and further improve the segmentation accuracy.

(2) In the encoder, a CNN-Transformer hybrid structure is added to enhance the feature extraction ability of global information and local detail information.

(3) In the decoder, a double upsampling module is introduced to improve the feature extraction ability and segmentation accuracy. We use the hard-swish activation function to enhance the generalization and nonlinear feature extraction capabilities and prevent the gradient from disappearing.

II. METHODS

This section includes the network architecture and module details of the UNet road extraction semantic segmentation model that has been improved based on the attention mechanism in this paper. It also includes the CNN-Transformer hybrid structure, dual up-sample module, hard-swish activation function, and cross entropy + dice loss function. The improved UNet network model architecture is shown in Figure 1.

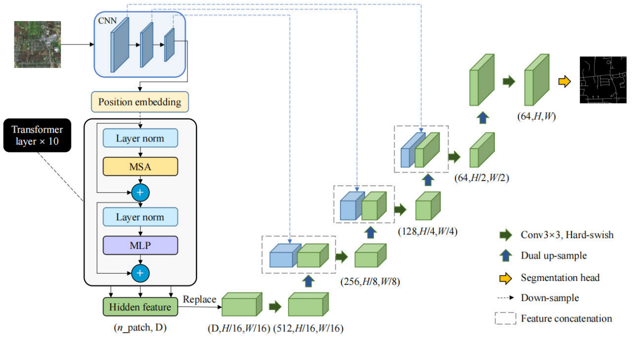

FIGURE 1. **Improved UNet architecture.**

A. CNN-TRANSFORMER HYBRID STRUCTURE

Traditional CNN has good local perception capabilities and local spatial information. Different convolution kernels can have different receptive fields. However, CNN may lose some feature information in the pooling layer and lacks correlation between local and global features. However, the Transformer network has a self-attention structure and has a strong ability to extract global information features. Therefore, we adopt a CNN Transformer hybrid structure in the encoder part to improve the local and global correlation in the CNN road information extraction process. Among them, CNN serves as a feature extractor and learns through convolution operations to obtain detailed high-resolution spatial information. Afterwards, we introduced the self-attention mechanism of the Transformer into the encoder design, which improved the limitation of convolutional operations not being able to establish long-distance models.

First, the original image is input into the CNN for feature extraction. Three layers of convolutional downsampling are performed, which reduces the size of the feature map to 1/2, 1/4, and 1/8 of the original image. Then, the downsampled images are input into the embedding layer. The images are first divided into patches, then mapped to one-dimensional vectors via linear mapping based on the given size. This outputs a vector sequence, or a two-dimensional matrix, which is then inputted to the Transformer layer for 10 iterations.

The Transformer module with global self-attention is based on the vision Transformer [33], which includes layer norm, multi-head self-attention module (MSA), and multi-layer perceptron (MLP). The operation formula of its multi-head self-attention module is shown in Eq. (1).

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{1}$$

B. DUAL UP-SAMPLE MODULE

UNet neural network belongs to the model framework of encoding and decoding. Features are learned by convolutional operations, and the feature map is smaller than the original image, so it needs to be upsampled to restore the image size in the decoder. Ordinary deconvolution causes the checkerboard effect, while the reverse max pooling and pixel shuffle upsampling methods destroy the continuity of the features. Moreover, the basic bilinear upsampling does not have learnable parameters and the effect is poor. The main function of the pixel shuffle convolution layer is to obtain a high-resolution feature map by convolving and reconstituting a low-resolution feature map with multiple channels, effectively improving the checkerboard effect. Compared to the nearest neighbor method, bilinear upsampling is smoother, faster, and requires less computation.

Tang et al. [34] used bilinear interpolation and deconvolution to construct upsampling blocks to reduce the negative effects of downsampling in the pooling stage. Fan et al. [35] proposed a dual upsampling block structure to prevent the checkerboard effect, which was used in image denoising tasks. We replaced the original upsampling scheme of UNet with a dual up-sample module, which combines the pixel shuffle convolution layer and bilinear upsampling, and added the hard-swish activation function to improve the feature extraction ability and image edge segmentation accuracy during upsampling, to prevent checkerboard effects and compensate for the loss of feature resolution caused by the Transformer. The dual up-sample module, as shown in Figure 2, is an upsampling structure composed of two channels, each containing two convolutional layers,

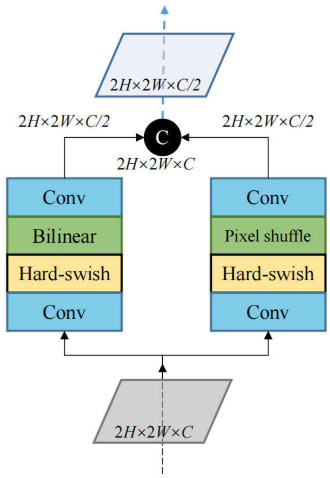

FIGURE 2. Dual up-sample module.

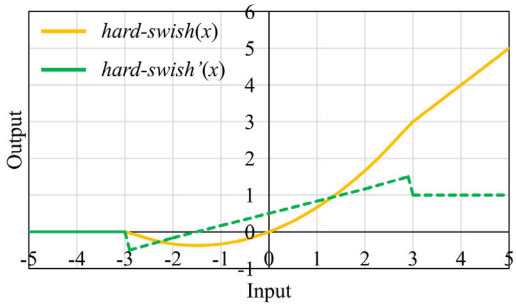

FIGURE 3. The function image of Hard-swish and its derivative.

a hard-swish layer, a pixel shuffle convolution layer, and bilinear interpolation.

C. HARD-SWISH ACTIVATION FUNCTION

In traditional neural networks, the choice of activation function has a significant impact on both training efficiency and performance of the model. The most widely used activation function is ReLU [36]. Ramachandran et al. [37] proposed the swish activation function, which performs better than ReLU in deep networks. Swish has a simple structure and is similar to ReLU, but its disadvantage is that it requires more computation.

$$hard - swish(x) = x \cdot \frac{ReLU(x+3)}{6} \tag{2}$$

We used the hard-swish activation function proposed by Howard et al. [38] in MobileNetV3, as shown in Eq. (2). Compared to swish, hard-swish has better numerical stability and computational speed, and using the hard-swish non-linear activation function did not show a significant difference in accuracy in our experiments. The function graph of hard-swish and its derivative (hard-swish'(x)) are shown in Figure 3. In practical applications, the hard-swish activation function can be reduced using a segmented method to reduce the number of memory accesses and significantly reduce waiting time, as shown in Eq. (3).

$$hard - swish(x) = \begin{cases} 0, & x \leq -3 \\ x, & x \geq 3 \\ x(x+3)/6, & other \end{cases} \tag{3}$$

D. CROSS ENTROPY AND DICE MIXED LOSS FUNCTION

During the neural network training phase, the loss function is used to calculate the difference between the iteration result and the actual value, in order to guide subsequent training in the correct direction. Improvements to the loss function mainly focus on the problem of class imbalance. We used a hybrid loss function that combines the fusion of cross-entropy and dice loss function [44]. This method combines the stability of cross-entropy loss and the ability to address class imbalance issues, without affecting the characteristics of the dice loss. The hybrid loss function we used has better stability than the dice loss and can better solve the problem of class imbalance than the cross-entropy. The formula for the cross-entropy loss function and the dice loss function are shown in Eq. (4) and (5), respectively.

$$Loss_{CE} = \frac{1}{N}\sum_{i}^{L_i} - \frac{1}{N}\sum_{i}\sum_{c=1}^{C} y_{ic}\lg(p_{ic}) \tag{4}$$

The cross-entropy loss function is used to evaluate the loss caused by pixel classification when image data is segmented. It can measure the difference between two different probability distributions under the same random variable. The higher the numerical value, the better the model's prediction result. Among them, C is the number of categories, indicating whether the category is i, if so $y_i = 1$, if not $y_i = 0$. p_i refers to the probability that sample i belongs to category C.

$$Loss_{Dice} = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{5}$$

Dice loss function is used to measure the similarity between the predicted segmentation image and the actual segmentation image. The value range is. Among them, represents the intersection of the true and the predicted value, and represents the number of elements. The comprehensive loss function of the improved UNet model is, as shown in Eq. (6). Among them,,.

$$Loss_{Total} = \lambda Loss_{CE} + \mu Loss_{Dice}. \tag{6}$$

E. EVALUATION METRICS

The commonly used evaluation indicators for semantic segmentation include pixel accuracy (PA), mean pixel accuracy (MPA), mean intersection over union (MIoU), precision (P), recall (R), F1 score, etc.

For the convenience of explaining the calculation formula for evaluation metrics, if the number of pixel points is k, p_{jj} represents the total number of pixels belonging to class i and predicted as class i (TP), p_{jj} represents the actual number of pixels belonging to class j and predicted as class j

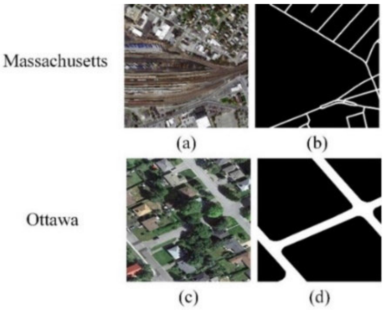

FIGURE 4. Dataset images and labels. (a) Image of Massachusetts -Dataset; (b) Corresponding label of Massachusetts -Dataset; (c) Image of Ottawa-Dataset; (d) Ottawa-Dataset corresponding label.

(TN), p_{ij} represents the total number of pixels belonging to class i but predicted as class j (FP), and p_{ji} represents the total number of pixels belonging to class j but predicted as class i.

Pixel accuracy (PA) represents the ratio of the number of correctly classified pixels in the test target to the total number of pixels in the whole test area, as shown in Eq. (7). Mean pixel accuracy (MPA) is the mean value of all PA values, as shown in Eq. (8). Here, C is the number of classes, and c is one of the classes. Mean intersection over union (MIoU) is obtained by calculating the average of intersection over union (IoU), which is the ratio of intersection to union of the true and predicted areas in the test dataset and is defined in Eq. (9). MIoU is defined in Eq. (10). Precision, recall, and F1 score are defined in Eqs. (11)~(13), respectively.

$$PA = \frac{\sum_{i=0}^{k} P_{ii} + \sum_{j=0}^{k} P_{jj}}{\sum_{i=0}^{k} P_{ii} + \sum_{j=0}^{k} P_{jj} + \sum_{i=0}^{k}\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k}\sum_{i=0}^{k} P_{ji}} \tag{7}$$

$$MPA = \frac{1}{C}\sum_{c=0}^{C} PA_c \tag{8}$$

$$IoU = \frac{\sum_{i=0}^{k} P_{ii}}{\sum_{i=0}^{k} P_{ii} + \sum_{i=0}^{k}\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k}\sum_{i=0}^{k} P_{ji}} \tag{9}$$

$$MIoU = \frac{1}{C}\sum_{c=0}^{C} IoU_c \tag{10}$$

$$P = \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij}+P_{ii}} or \sum_{j=0}^{k} \frac{P_{jj}}{\sum_{i=0}^{k} P_{ji}+P_{jj}} \tag{11}$$

$$R = \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ji}+P_{ii}} \tag{12}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{13}$$

TABLE 1. Experimental configuration.

Name	Settings
Operating system	Windows10
Experimental environment	Python 3.6.13
Deep learning framework	PyTorch 1.7.0
CUDA	CUDA 11.0
GPU	NVIDIA GeForce RTX 3080Ti (12 GB)
CPU	Intel® Core™ i7-12700
Memory	32 GB

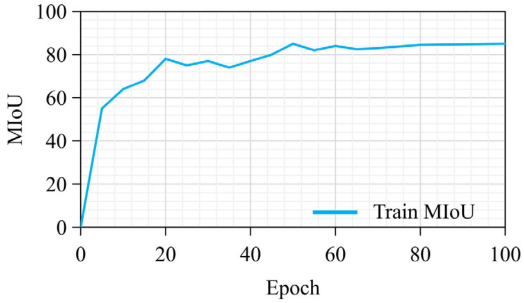

FIGURE 5. The change curve of MIoU.

III. RESULTS

A. DATASET

The road extraction images used in the experiment come from the publicly available Massachusetts road Dataset and Ottawa road Dataset. The Massachusetts roads dataset consists of 1,171 aerial images of Massachusetts, each image is 1500 × 1500 pixels in size and has a spatial resolution of 1m. The spatial resolution of the Ottawa road dataset is 0.2m, and the dataset covers various areas in cities suburbs and rural areas. We select images in dense road areas for cropping. Both data are cropped to a size of 512 × 512 pixels, and rotated 90 degrees, 180 degrees and 270 degrees for data expansion. Randomly select 30% as the validation set samples and 10% as the test set samples. The images and labels are shown in Figure 1.

B. SETUP

The experiment used the PyTorch deep learning framework, the Windows 10 operating system, an NVIDIA GeForce RTX 3080Ti GPU, and 32 GB of RAM. The detailed configuration is shown in Table 1.

C. TRAINING

We trained the Massachusetts dataset using the improved UNet network model. To measure the effectiveness of the proposed model in this paper and taking into account the performance of the experimental equipment and pre-training efficiency, the batch size for each experiment was set to 16, the initial learning rate was set to 0.01, and the epoch was set to 100. Pre-training improved the training speed of the model and effectively enhanced the network fitting,

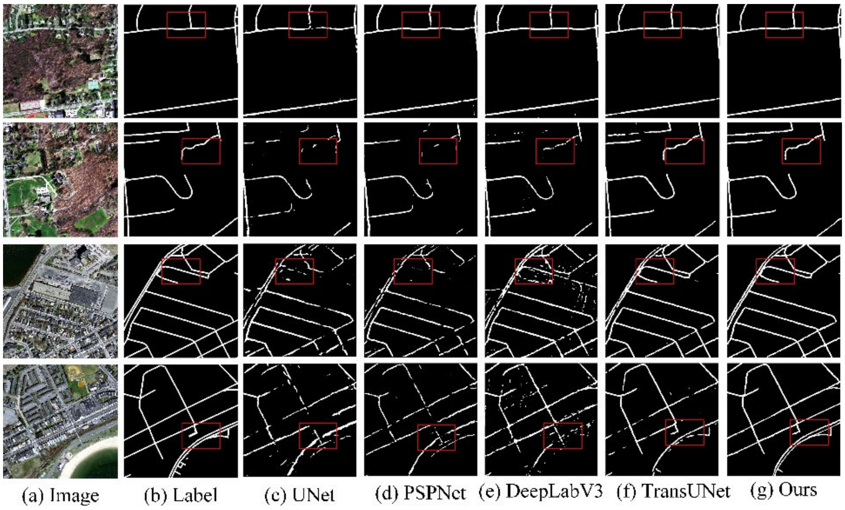

FIGURE 6. Semantic segmentation results of different models in the Massachusetts road dataset. (a) is the image; (b) is the label; (c) is the U-Net extraction result; (d) is the PSPNet extraction result; (e) is the extraction result of Deeplabv3+; (f) is the extraction result of TransUNet; (g) is the extraction result of this article.

TABLE 2. Comparison results of each model. Notes: Bold is the best, and underline is the second.

Model	MPA (%)	MIoU (%)	F1 score (%)	Time (h)
UNet	84.97	75.84	85.54	4.27
PSPNet	83.16	75.27	84.74	**3.84**
DeepLabV3	82.55	74.82	83.48	<u>4.07</u>
TransUNet	<u>90.53</u>	<u>82.36</u>	<u>89.64</u>	4.73
Ours	**92.56**	**84.97**	**91.48**	4.31

achieving the goal of improving road segmentation accuracy on a limited dataset. The MIoU of the experiment finally reached 84.97%.

The pre-trained model was trained on the Massachusetts dataset, and the training results were recorded every 5 epochs. The MIoU was calculated to measure the segmentation accuracy. Figure 4 shows the variation in MIoU over the training epochs. The model's segmentation accuracy improved as the number of training iterations increased. Above epoch 50, the MIoU stabilized above 80%, with fluctuations of less than 2 percentage points. The results indicate that the model achieved high segmentation accuracy and demonstrated good robustness.

D. COMPARISON EXPERIMENTS

In order to further evaluate the performance and effectiveness of our model, we replicated the mainstream semantic segmentation models of UNet, PSPNet, DeepLabV3 and TransUNet to verify in the Massachusetts road dataset. UNet uses an encoder-decoder framework with VGG model [39] as the feature extractor network. PSPNet and DeepLabV3 use MobileNet [40] as the feature extractor network, which can reduce the computational cost while ensuring performance. TransUNet and improved UNet models both use ResNet-50 [41] as the feature extractor network. Under the consistent experimental parameter configuration, Massachusetts dataset is used for comparison with the improved UNet model and the other four models. The experimental comparison metrics include precision, recall, F1 score, MPA, MIoU and training time, as shown in Figure 5 and Table 2.

Figure 5 shows the semantic segmentation results of different models for road extraction. The road segmentation results of the UNet and PSPNet models are comparable, both able to segment the rough area of roads, but there are still some areas with segmentation inaccuracies. The DeepLabV3 model enlarges the receptive field through atrous convolutions, but this method performs poorly in road edge segmentation, and too much detail is lost during bilinear upsampling, resulting in unsatisfactory segmentation results. Compared with the previous models, the TransUNet model has better segmentation results and higher accuracy, but there are still problems with some image edge segmentations. We address these problems through the introduction of CNN-Transformer structures and dual up-sample modules,

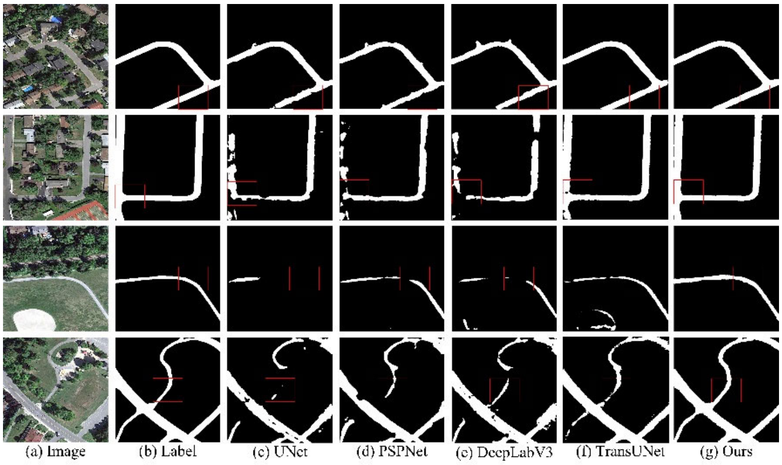

FIGURE 7. Semantic segmentation results of different models in the Ottawa road dataset. (a) is the image; (b) is the label; (c) is the U-Net extraction result; (d) is the PSPNet extraction result; (e) is the extraction result of Deeplabv3+; (f) is the extraction result of TransUNet; (g) is the extraction result of this article.

TABLE 3. Comparison results of each model. Notes: Bold is the best, and underline is the second.

Model	MPA (%)	MIoU (%)	F1 score (%)	Time (h)
UNet	86.28	78.85	84.02	4.68
PSPNet	89.20	80.32	84.74	**3.96**
DeepLabV3	84.34	78.32	76.29	<u>4.23</u>
TransUNet	<u>92.19</u>	<u>87.37</u>	<u>87.68</u>	4.92
Ours	**95.48**	**90.94**	**93.25**	4.82

effectively improving edge segmentation problems, resulting in better segmentation results and improved segmentation accuracy. The predicted results can be effectively used for road extraction.

From the perspective of evaluation metrics in Table 2, the mean pixel accuracy of the five models is 84.97%, 83.16%, 82.55%, 90.53%, and 92.56%, respectively. The overall performance of DeepLabV3 is poor, as it has low accuracy and suboptimal segmentation effect on this dataset. Meanwhile, PSPNet shows better network segmentation effect than DeepLabV3, and has the shortest training time. UNet yields a better overall performance, with its encoder-decoder architecture having been applied successfully in road extraction. Compared to UNet, TransUNet with the CNN-Transformer module improves the mean pixel accuracy by 5.56%, and the MIoU by 6.52%. The improved UNet model performs best, with an MPA of 92.56% and an MIoU of 84.97%, which represent improvements of about 2.03% and 2.61% over TransUNet, respectively. These results indicate an increase in the proportion of correctly segmented pixels and improved segmentation accuracy. In terms of training time, the PSPNet model is the fastest, followed by the DeepLabV3 and UNet models, with training times of 3.84h, 4.07h, and 4.27h respectively. The slower training time of the TransUNet model is due to the addition of an attention mechanism, which increases the number of parameters and calculations. In this study, to balance the effect of the attention mechanism, the hard-swish activation function was used to reduce memory accesses, and a cross-entropy + dice mixed loss function was introduced to reduce training time, accelerate model convergence, and ultimately achieve a training time of 4.31h.

In order to verify the superiority and generalization ability of this algorithm in road extraction, the public data set Ottawa Roads road data set was used for experimental verification. UNet, PSPNet, DeepLabV3 and TransUNet were trained with the algorithm and model of this article respectively. The experimental results of UNet, PSPNet, DeepLabV3, TransUNet and this algorithm are shown in Figure 7 and Table 3.

As shown in Figure 7, the road is clear and regular, but it is also blocked by trees. UNet road extraction results are incomplete and are greatly affected by tree occlusion. The extraction results of PSPNet and DeepLabV3 are worse than UNet, there are still discontinuities in the extraction results, and the detail processing is poor. The extraction results of TransUNet are better than U-Net and PSPNet in terms of continuity. However, it is also affected by occlusions, resulting

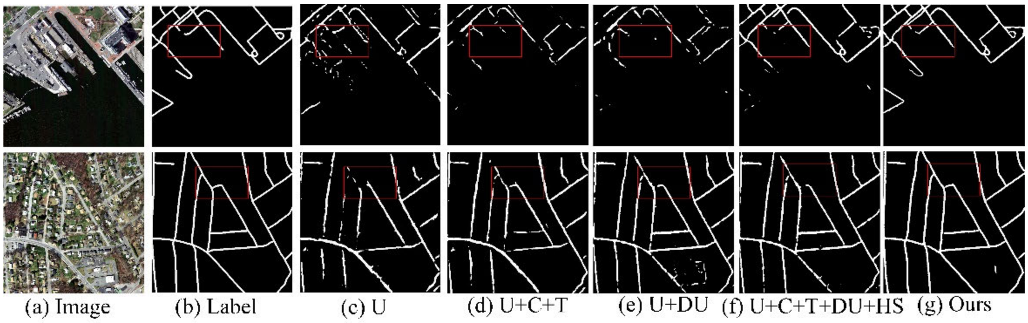

FIGURE 8. Ablation experimental image and road extraction result map. (a) is the image; (b) is the label; (c) is the U-Net extraction result; (d) adding a dual up-sample (DU) module to Scheme 1; (e) adding a dual up-sample (DU) module and a hard-swish (HS) activation function to Schemes 2; (f) changing a cross-entropy + dice mixed loss function to Scheme 4 for model training, as shown in Figure 8 and Table 4.; (g) is the extraction result of this article.

TABLE 4. Ablation study of improved UNet. Notes: Bold is the best, and underline is the second. U = UNet, CNN = C, Transformer = T, DU = dual up-sample, HS = hard-swish, and CE = cross entropy.

U	C	T	DU	HS	Loss	MPA	MIoU
✓	×	×	×	×	CE	84.97	75.84
✓	✓	✓	×	×	CE	88.46	79.52
✓	×	×	✓	×	CE	86.87	77.35
✓	✓	✓	✓	✓	CE	<u>91.12</u>	<u>82.59</u>
✓	✓	✓	✓	✓	CE + Dice	**92.56**	**84.97**

in inaccurate extraction. The method proposed in this paper can effectively extract roads. Edge details are handled well, and extraction is complete and continuous. We solve these problems by introducing a CNN-Transformer structure and a dual upsampling module, which effectively improves the problem of insufficient edge details, thereby obtaining better segmentation results and improving segmentation accuracy.

As can be seen from Table 3, the average pixel accuracy of the five models has improved in the Ottawa road data set, reaching 86.28%, 89.20%, 84.34%, 92.19% and 95.48% respectively. DeepLabV3 road extraction results are still the worst, with lower overall performance. The Miou index and F1 score of UNet and PSPNet are better than DeepLabV3. Compared with UNet, TransUNet with CNN Transformer module improves the average pixel accuracy by 5.91% and MIoU by 3.66%. This method performed the best, with MPA of 95.48% and MIoU of 90.94%, which were approximately 3.29% and 3.57% higher than TransUNet respectively. These results show that our method is also superior in road extraction in higher-resolution remote sensing images.

E. ABLATION STUDY

We selected the UNet as the baseline. To verify the feasibility and effectiveness of the improved UNet model designed in this paper, and under the same experimental conditions, the following five ablation experiment schemes were adopted:

(1) basic UNet network;

(2) adding a CNN-Transformer hybrid structure to Scheme 1;

(3) adding a dual up-sample (DU) module to Scheme 1;

(4) adding a dual up-sample (DU) module and a hard-swish (HS) activation function to Schemes 2;

(5) changing a cross-entropy + dice mixed loss function to Scheme 4 for model training, as shown in Figure 8 and Table 4.

As shown in Figure 8, the leakage extraction result of Scheme 1 is not accurate. The road extraction results are discontinuous and are greatly affected by other features. Compared with Scheme 1, Scheme 2 adds a channel CNN-Transformer hybrid structure, which has significantly improved the problem of discontinuous extraction results, but such problems still exist. In Scheme 3, based on Scheme 1, traditional upsampling is replaced by double upsampling. Compared with the extraction results of Scheme 1, the extraction results are better. Option 4 is based on Option 2 and simultaneously references up-sample (DU) module and a hard-swish (HS). It can be seen that the extraction results are significantly improved, and the extraction results are more consistent. Case 5 is the method of this article. From the results, the extraction discontinuity has been effectively solved, and the method proposed in this article has achieved good extraction results. Based on the above, the algorithm module in this study is effective and the method of this study is more suitable.

Among the results of the five schemes mentioned above, comparing Scheme 1 and Scheme 2, it can be seen that adding a CNN Transformer hybrid structure significantly improved the model's precision and MIoU by 3.49% and 3.68%, respectively. Comparing Scheme 1 and Scheme 3, it can be seen that adding the dual up-sample module has slightly improved the MPA and MIoU of the model, with increases of 1.9% and 1.51%, respectively. Verified the effect of CNN-Transformer hybrid structure and dual up-sample module on improving model performance. It can be seen from the comparison between Scheme 1 and Scheme 4 that by adding the CNN-Transformer hybrid structure, dual up-sample module and hard-swish activation function at the same time, MPA and MIoU reach 91.12% and 82.59%, with good results. It can be seen from Scheme 4 and Scheme 5 that changing cross-entropy + dice mixed loss function on the basis of Scheme 4 will further improve the model performance. From this, it can be seen that the network training strategy and model improvement plan in this study are effective and feasible.

IV. CONCLUSION

Real-time and accurate road information is the prerequisite for updating navigation electronic maps, and it is of great significance to obtain road information quickly and accurately. In order to solve the problems of discontinuous and inaccurate road extraction by the traditional UNet network, we improved the UNet network to extract road information from remote sensing images. We use UNet based on the CNN-Transformer model architecture to enhance the feature extraction capabilities of global information and local detail information. At the same time, we introduce a double upsampling module to improve feature extraction capabilities and segmentation accuracy, and introduce a hard-swish activation function to enhance generalization and nonlinear feature extraction capabilities and prevent gradient disappearance. And use the loss function of cross entropy (CE) and Dice to strengthen the model's constraints on segmentation results, further improving segmentation accuracy. We perform model validation on the Massachusetts Road Dataset and the Ottawa Road Dataset. In comparative experiments, we compared this algorithm with UNet, PSPNet, DeepLabV3 and TransUNet. The results verify that this paper has good training stability, robustness and generalization ability in road extraction semantic segmentation. In addition, the effectiveness of the improved part of this article was verified in the ablation experiment. Nonetheless, the improved UNet-based semantic segmentation algorithm is not easily applicable to mobile or embedded devices. It has the characteristics of high computational complexity, long training time and large parameter space. Therefore, future research work aims to investigate efficient and lightweight methods for image semantic segmentation that can be conveniently used on mobile devices.

REFERENCES

[1] P. Liu, Z. Xu, and X. Zhao, "Road tests of self-driving vehicles: Affective and cognitive pathways in acceptance formation," *Transp. Res. A, Policy Pract.*, vol. 124, pp. 354–369, Jun. 2019.

[2] Z. Chen, W. Fan, B. Zhong, J. Li, J. Du, and C. Wang, "Corse-to-fine road extraction based on local Dirichlet mixture models and multiscale-high-order deep learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4283–4293, Oct. 2020.

[3] M. Zeybek, "Extraction of road lane markings from mobile LiDAR data," *Transp. Res. Rec.*, vol. 2675, no. 5, pp. 30–47, 2021.

[4] S. A. Kianejad Tejenaki, H. Ebadi, and A. Mohammadzadeh, "A new hierarchical method for automatic road centerline extraction in urban areas using LiDAR data," *Adv. Space Res.*, vol. 64, no. 9, pp. 1792–1806, Nov. 2019.

[5] J. Liao, L. Cao, X. Luo, X. Sun, C. Duan, J. Li, and F. Yuan, "Road garbage segmentation with deep supervision and high fusion network for cleaning vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11190–11204, Aug. 2022.

[6] G. Shen, X. Han, K. Chin, and X. Kong, "An attention-based digraph convolution network enabled framework for congestion recognition in three-dimensional road networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14413–14426, Sep. 2022.

[7] Y. Zhang, J. Wang, X. Wang, and J. M. Dolan, "Road-segmentation-based curb detection method for self-driving via a 3D-LiDAR sensor," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 12, pp. 3981–3991, Dec. 2018.

[8] C. Wen, A. F. Habib, J. Li, C. K. Toth, C. Wang, and H. Fan, "Special issue on 3D sensing in intelligent transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 1947–1949, Apr. 2021.

[9] H. Fouchal, S. Boudra, S. Ercan, and I. Yahiaoui, "Pseudonym limitation for privacy in cooperative transport systems," *IEEE Netw.*, vol. 34, no. 3, pp. 73–77, May 2020.

[10] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[11] H. Feng, W. Li, Z. Luo, Y. Chen, S. N. Fatholahi, M. Cheng, C. Wang, J. M. Junior, and J. Li, "GCN-based pavement crack detection using mobile LiDAR point clouds," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11052–11061, Aug. 2022.

[12] L. Deng, X.-Y. Liu, H. Zheng, X. Feng, and Y. Chen, "Graph spectral regularized tensor completion for traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10996–11010, Aug. 2022.

[13] H. Yang, C. Liu, M. Zhu, X. Ban, and Y. Wang, "How fast you will drive? Predicting speed of customized paths by deep neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2045–2055, Mar. 2022.

[14] S. Guo, C. Chen, J. Wang, Y. Ding, Y. Liu, K. Xu, Z. Yu, and D. Zhang, "A force-directed approach to seeking route recommendation in ride-on-demand service using multi-source urban data," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 1909–1926, Jun. 2022.

[15] J. E. Espinosa, S. A. Velastín, and J. W. Branch, "Detection of motorcycles in urban traffic using video analysis: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6115–6130, Oct. 2021.

[16] J.-L. Yin, B.-H. Chen, and K. R. Lai, "Driver danger-level monitoring system using multi-sourced big driving data," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 12, pp. 5271–5282, Dec. 2020.

[17] Y. Jung, S.-W. Seo, and S.-W. Kim, "Curb detection and tracking in low-resolution 3D point clouds based on optimization framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3893–3908, Sep. 2020.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[19] B. Bakker, B. Zablocki, A. Baker, V. Riethmeister, B. Marx, G. Iyer, A. Anund, and C. Ahlström, "A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4791–4800, May 2022.

[20] M. Maboudi, J. Amini, M. Hahn, and M. Saati, "Object-based road extraction from satellite images using ant colony optimization," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 179–198, Jan. 2017.

[21] J. Leng, Y. Liu, D. Du, T. Zhang, and P. Quan, "Robust obstacle detection and recognition for driver assistance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1560–1571, Apr. 2020.

[22] M. Yang, Y. Yuan, and G. Liu, "SDUNet: Road extraction via spatial enhanced and densely connected UNet," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108549.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 015, pp. 234–241.

[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[26] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[27] X. Kong, C. Wang, S. Zhang, J. Li, and Y. Sui, "Application of improved U-Net network in road extraction from RS image," *Remote Sens. Inf.*, vol. 37, no. 2, pp. 97–104, 2022.

[28] J. Li, Y. Liu, Y. Zhang, and Y. Zhang, "Cascaded attention DenseUNet (CADUNet) for road extraction from very-high-resolution images," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 5, p. 329, May 2021.

[29] L. Han, "Road extraction of high resolution RS imagery based on DeepLab V3," *Remote Sens. Inf.*, vol. 36, no. 1, pp. 22–28, 2021.

[30] R. Liu and D. He, "Semantic segmentation based on Deeplabv3+ and attention mechanism," in *Proc. IEEE 4th Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC)*, vol. 4, Jun. 2021, pp. 255–259.

[31] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Long Beach, CA, USA: Curran Associates, 2017, pp. 6000–6010.

[33] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[34] Z. Tang, W. Jiang, Z. Zhang, M. Zhao, L. Zhang, and M. Wang, "DenseNet with up-sampling block for recognizing texts in images," *Neural Comput. Appl.*, vol. 32, no. 11, pp. 7553–7561, Jun. 2020.

[35] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin transformer UNet for image denoising," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 2333–2337.

[36] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Jun. 2011, pp. 315–323.

[37] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.

[38] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[39] V. Mnih, *Machine Learning for Aerial Image Labeling*. Univ. Toronto, 2013.

[40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

RUI WANG graduated from Wuhan University. He is currently with the China Transport Telecommunications and Information Center. His research interests include the application of traffic data in the insurance industry, spatial big data management, and the study of image recognition algorithms based on machine learning.

MINGXIANG CAI graduated from Wuhan University. His research interests include spatio-temporal information mining and the application of fundamental models in traffic big data.

ZIXUAN XIA is currently pursuing the degree in surveying and mapping engineering with the Heilongjiang University of Technology. His current research interests include flight calibration and the validation of forestry applications of high-definition aerial systems and the study of vegetation cover differentiation based on the image dichotomous model.

ZHICUI ZHOU received the master's degree from Shaanxi Normal University. Her research interests include ecology, artificial intelligence, and image recognition algorithms based on machine learning.