Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.023217

check for updates

ARTICLE

A Road Segmentation Model Based on Mixture of the Convolutional Neural Network and the Transformer Network

Fenglei Xu[#], Haokai Zhao[#], Fuyuan Hu[*], Mingfei Shen and Yifei Wu

Suzhou University of Science and Technology, Suzhou, 215009, China

*Corresponding Author: Fuyuan Hu. Email: fuyuanhu@mail.usts.edu.cn

[#]These authors contributed equally to this work

Received: 15 April 2022 Accepted: 07 June 2022

ABSTRACT

Convolutional neural networks (CNN) based on U-shaped structures and skip connections play a pivotal role in various image segmentation tasks. Recently, Transformer starts to lead new trends in the image segmentation task. Transformer layer can construct the relationship between all pixels, and the two parties can complement each other well. On the basis of these characteristics, we try to combine Transformer pipeline and convolutional neural network pipeline to gain the advantages of both. The image is put into the U-shaped encoder-decoder architecture based on empirical combination of self-attention and convolution, in which skip connections are utilized for local-global semantic feature learning. At the same time, the image is also put into the convolutional neural network architecture. The final segmentation result will be formed by Mix block which combines both. The mixture model of the convolutional neural network and the Transformer network for road segmentation (MCTNet) can achieve effective segmentation results on KITTI dataset and Unstructured Road Scene (URS) dataset built by ourselves. Codes, self-built datasets and trainable models will be available on https://github.com/xflxfl1992/MCTNet.

KEYWORDS

Image segmentation; transformer; mix block; U-shaped structures

1 Introduction

Accurate and robust road image segmentation can play a cornerstone role in computer-assisted driving and visual navigation. Due to the deep learning revolution, the segmentation accuracy has achieved impressive results.

Motivated by the successes of CNN-based classifiers [1–3] and the help of various optimization algorithms [4–6], the task of semantic segmentation overcame many difficulties. Past researchers used image patches to eliminate all the redundant computation [7]. The most famous fully convolutional networks [8] extend image-level classification to pixel-level classification. Dilated convolutions were introduced in [9] to perform multi-scale information fusion. The typical U-shaped network, U-Net [10], obtained great success in a variety of medical imaging applications. The aforementioned techniques proved the excellent learning ability of CNN.

Currently, although the CNN-based methods lead the trend in the field of image segmentation, they have more improvement space. Meanwhile, Transformer is showing revolutionary performance improvements in the CV field. In [11], vision transformer (ViT) is proposed to perform the image recognition task. Taking image patches as the input and using self-attention mechanism, ViT can even achieve better performance compared with the CNN-based methods. LeVit [12] designed a patch descriptor to improve calculation efficiency and ensured the accuracy. CaiT [13] optimized the Transformer architecture, which significantly improved the accuracy of the deep Transformer. These methods show that CV and NLP are expected to be unified under the Transformer structure, and the modeling and learning experience of the two fields can be deeply shared, thereby accelerating the progress of their respective fields.

Motivated by the Transformer's success, we take an approach based on U-shaped encoder-decoder and design a network architecture that combining the convolutional structure and Transformer structure (MCTNet). This method performs road detection pixel-level segmentation tasks well, and the reasons are as follows:

- We propose a fusion structure, which combine the result of CNN structure and Transformer structure. This method can gather the advantage of CNN's ability to establish the relationship between neighboring pixels and Transformer's ability to establish the relationship between all pixels.
- We find the respective characteristics of Transformer and CNN, which Transformer focuses on the main area of the road image and CNN focus on the details of edge area. Therefore, We design a post-processing adding with prior knowledge of the road scene to deal with the results, which make the accuracy of road area a certain degree of improvement.
- We built a harder task on structured and unstructured road area detection to test our method. The dateset we built contains 2000 road scene including gravel pavement, soil-covered pavement, water-covered pavement and highways. The experimental evaluation on KITTI and URS datasets proves our model validity.

2 Related Work

Road Detection In recent years, more and more researches on autonomous driving at home and abroad have accumulated a certain research foundation. OFA-Net [14] used a strategy called "1-N Alternation" to train the model, which can make a fusion of features from detection and segmentation data. RoadNet-RT [15] speeded up the inference time by optimizing Depthwise separable convolution and non-uniformed kernel size convolution. ALO-AVG-MM [16] extracted multiples side-outputs and used filtering to improve network performance. Volpi et al. [17] proposed a new evaluation framework for online study about segmentation.

Transformer In the field of NLP, transformer-based methods have achieved the most advanced performance in various tasks. Vision transformer (ViT) [11], which achieved an impressive speed-accuracy trade-off in image recognition tasks was motivated by the success of Transformer. Deit [18] introduced several training strategies to make ViT perform better. On the basis of these methods, Swin Transformer [19] restricted self-attention calculations to non-overlapping partial windows, while allowing cross-window connections, which brought greater efficiency and flexibility. Compared to previous work, Swin Transformer is able to significantly reduce the computation. Besides, Swin-Unet [20] based on U-net and Swin Transformer stood out from medical image segmentation tasks.

Self-Attention to Mix CNN Given the ability to leverage long-term dependence, transformers are expected to help atypical convolutional neural networks overcome their inherent shortcomings of spatial induction bias. However, most of the recently proposed Transformer-based segmentation methods only use Transformer as an auxiliary module to help encode the global context as a convolutional representation, and have not studied how to best combine self-attention (the core of Transformer) with convolution. NnFormer [21] has an interleaved architecture based on a combination of self-attention and convolution experience. Reference [22] was an interleaved architecture based on the experience of self-attention and convolution. It achieved SOTA performance on ImageNet-1k classification without bells and whistles.

For now, the study in the field of autonomous driving has achieved considerable results. The development of transformer-based methods can provide a research foundation for autonomous driving road detection. Therefore, this paper focuses on road detection tasks based on U-net, Transformer and CNN.

3 Methodology

3.1 Architecture Overview

The overall architecture is presented on Fig. 1. Two main pipelines are trained together. The details of two pipelines are presented on Fig. 2. We follow the base of [10]. The CNN pipeline is divided into encoding block and fusion block. The backbone of CNN pipeline encoder is resnet101. The image is down-sampled by convolution, the features are mapped to different scales, and then the image is up-sampled and restored by deconvolution. The extracted context features are fused with multiscale features from encoder via fusion block to complement spatial information.

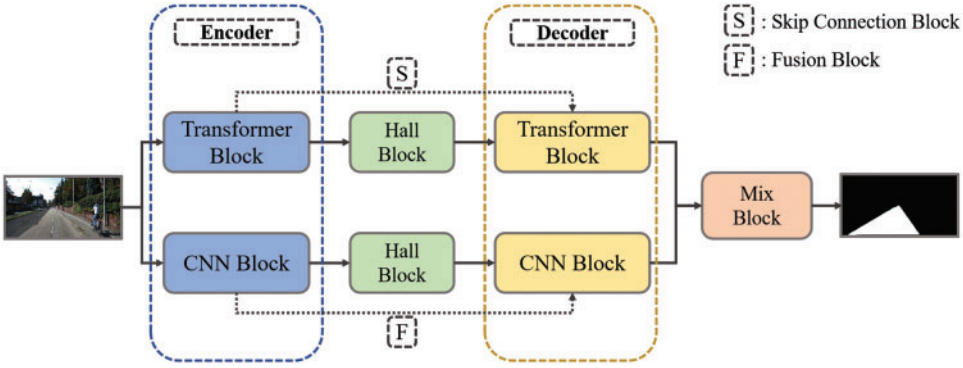


Figure 1: Network structure. The pipeline at the top is transformer-based U-net, and the pipeline at the bottom is CNN-based U-net

Our second part is based on Swin-Unet [20], which consists of encoder, hall block, decoder and skip connections. The backbone is inspired by Swin Transformer block. For the encoder, the images are split into non-overlapping patches to transform the inputs into sequence embeddings. The decoder is composed of Transformer blocks and Patch Expanding layer. The skip connections have the same function with fusion block. In contrast to patch merging layers, a patch expanding layer is designed to perform up-sampling. Then a linear projection layer is applied on these up-sampled features to output the pixel-level segmentation predictions. The output results of the two pipelines will be sent to Mix Block for further processing to obtain higher accuracy. Both the Hall Model of two pipelines are to

learn the deep feature representation. The feature dimension and resolution are kept unchanged. The difference is that one uses convolution and the other uses Swin Transformer block.

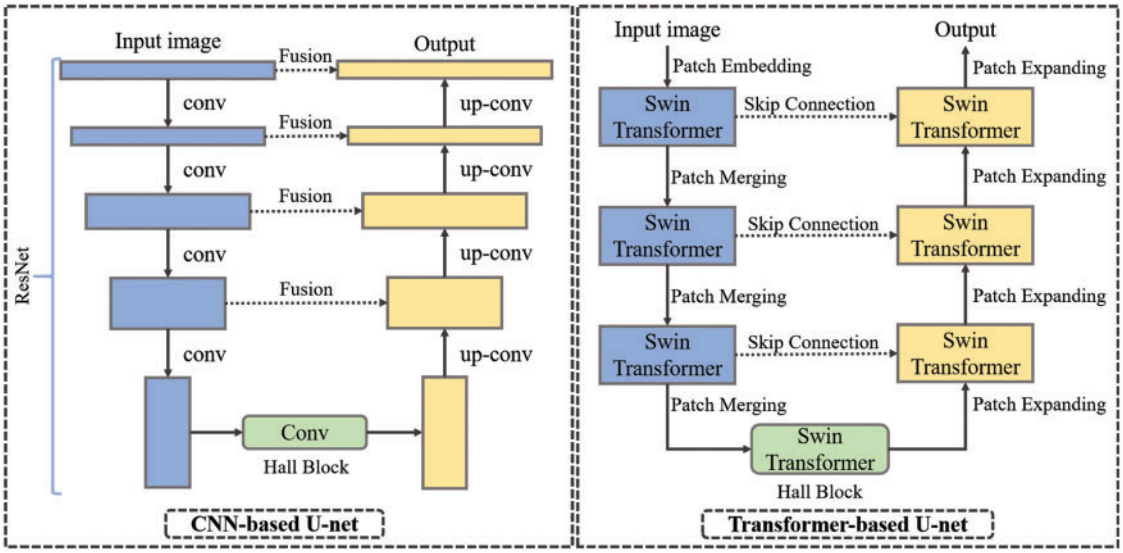


Figure 2: Details of network structure. In the section, blue rectangles indicate down-sampling process, yellow rectangles are up-sampling process, and green rectangles are hall block for deep feature representation

3.2 Embedding Block

The embedding block is the common patch processing method for Transformer structures. In this block, patches will be encoded as spatial information. Besides, Position embedding is able to preserve the location information of the image patches. ViT also encodes the position embedding at the input, and it is optional for Swin Transformer, because a relative position encoding is made by Swin Transformer when calculating attention. As a result, the embedding block only contains convolutional layer.

3.3 Swin Transformer Block

Swin transformer block is constructed based on shifted windows. In Fig. 3, two consecutive Swin transformer blocks are presented. With the shifted window partitioning approach, consecutive Swin transformer blocks are computed as:

$$\hat{z}^{l} = W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \tag{1}$$

$$z^{l} = MLP\left(LN\left(\hat{z}^{l}\right)\right) + \hat{z}^{l} \tag{2}$$

$$\hat{z}^{l+1} = SW - MSA\left(LN\left(z^{l}\right)\right) + z^{l} \tag{3}$$

$$z^{l+1} = MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{4}$$

where $\hat{z}^{l}$ and z^{l} represent the outputs of the (S)W-MSA module and the MLP module of the l^{th} block. Similar to the previous works, self-attention is computed as follows:

$$\text{Attention}\left(Q, K, V\right) = \text{SoftMax}\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V \tag{5}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ denote the query, key and value matrices. M^2 and d represent the number of patches in a window and the dimension of the query or key, respectively. And, the values in B are taken from the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$.

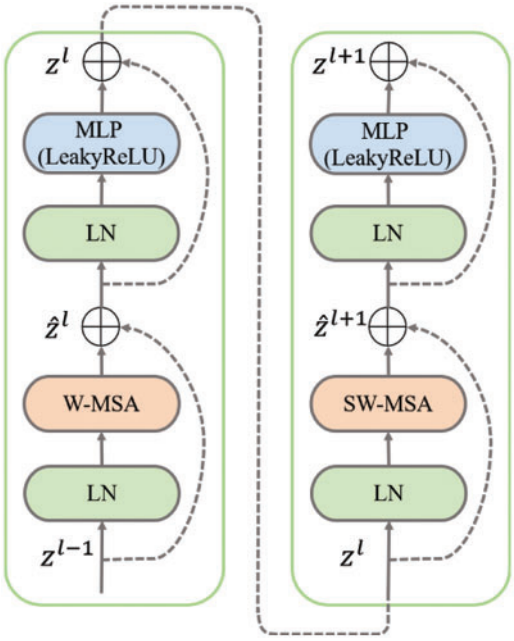


Figure 3: Transformer structure

3.4 Skip-Connection Block/Fusion Block

The fusion of deep and shallow information in FCN [8] is through the addition of corresponding pixels, we follow this method and apply it as the fusion block of CNN pipeline. U-net structure is through splicing. In the addition method, the dimension of the feature map has not changed, but each dimension contains more features. For ordinary classification tasks, which do not need to be restored from the feature map to the original resolution, this is an efficient selection; splicing retains more dimensional/location information, which allows the subsequent layers to freely choose between shallow and deep features, which benefits semantic segmentation tasks. Thus, the Skip-connection block of Transformer pipeline is depend on splicing.

3.5 Mix Block

As shown in Fig. 4, the Transformer pipeline focus on the main area of road detection. The results from CNN pipeline have a remarkable road edge detection effect, but it has quite a few false detection of non-road areas. Thus, we design a mix method below:

$$\omega_F I_F(x,y) = \left(\omega_f I_T(x,y) + \left(1-\omega_f\right) I_C(x,y)\right) \times \left(1 + \overbrace{2\left(\ln 2 - \ln\left(1 + \left|I_T(x,y) - I_C(x,y)\right|\right)\right)}^{\text{consistency}}\right) \tag{6}$$

$|\,.\,|$ means to calculate the absolute value, ω_f and ω_F the weight coefficient with a value between 0 and 1. $I_C(x,y)$ and $I_T(x,y)$ respectively represent the road confidence map based on CNN pipeline and Transformer pipeline.

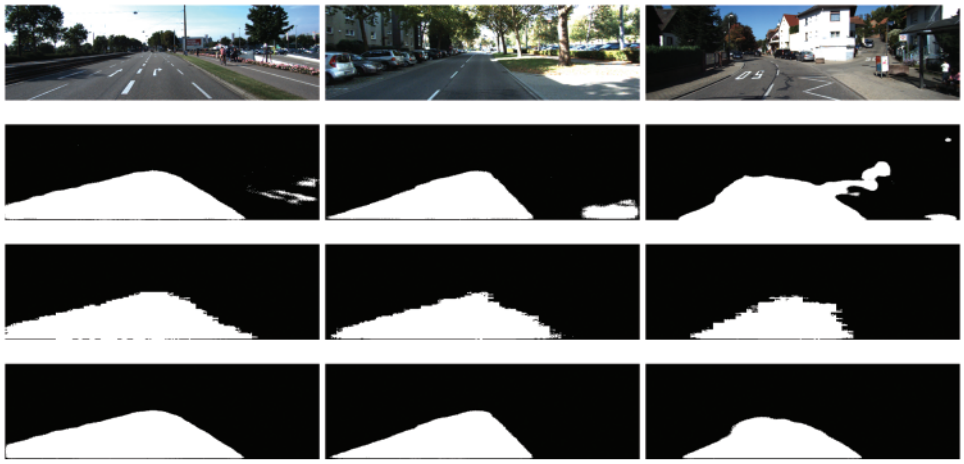


Figure 4: The results of different pipelines. The first row is the input; the second row is the result of CNN pipeline; the third row is the result of Transformer pipeline and the forth row is the result of mix block

The purpose of the above formula is to fuse two road confidence maps and enhance the pixels with similar probabilities. The first half of the formula is the general probability fusion formula, the second half adds consistency as marked in the formula (6). After the consistency item is strengthened, at any point (x, y), if the two confidence graphs have similar values, the consistency item is approximately equal to 1.386. This means that if the judgments of the two confidence graphs at point (x, y) are both roads, $I_C(x, y)$ will be enlarged. Conversely, if the two confidence maps both show that the point (x, y) is a background pixel, the road surface probability of the pixel remains 0 after the fusion because the weighted fusion item on the left is 0. In other cases, when the two fusion confidence maps obtain two completely different confidence values at (x, y), then the value of $I_T(x, y)$ will be suppressed by the consistency item after normalization. Here, a constant factor of 1 is added before the consistency term to ensure that the influence of the weighted fusion term is preserved, instead of getting a zero value when fusing two completely different confidence values. In the end, we use some prior knowledge to remove some irrelevant pixel value predictions, such as some areas at the top of the image.

4 Experiments

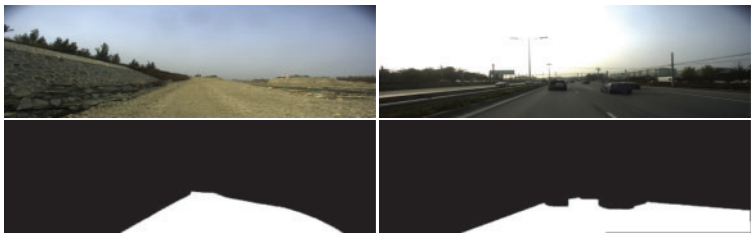
4.1 Datasets

KITTI Dataset Our model is mainly estimated on KITTI-ROAD dataset, which includes 289 training and 290 testing structured road scenes. We split the original training set for supervised data. The new training and validation set contain 230 and 59 road scenes respectively. Ablation study is conducted on new data split. The final result comparison with other existing models is based on KITTI testing set.

URS Dataset To further verify our model efficiency, we examine the results on self-built dataset (Unstructured Road Scene dataset). As shown in Fig. 5, the dataset we built contains 2000 road images of different types containing gravel pavement, soil-covered pavement, water-covered pavement and cement pavement. These images include almost all types of roads. Unlike other open datasets, such as KITTI, this dataset contains a large number of unconventional road image, which provides more abundant and diversified challenges and tests for road segmentation tasks. Its applicable scenarios are

also wider. We use 1400 images for training, 400 images for validation, and 200 images for test. The distribution of the number of image types is shown in the Table 3.

(a) cement pavement (b) water-covered pavement (c) soil-covered pavement



(d) gravel pavement (e) highway

Figure 5: Different kinds of self-built dataset. The first row is the RGB image; the second row is the ground-truth

4.2 Implementation Details

On KITTI dataset, we train two main pipelines separately. While network training, the input image size is set as 512 in both pipeline. Data augmentation is realized by rotating or cropping images.

We train MCTNet with batch size of 1 on one GPU (GeForce GTX 1080 8 GB), and use SGD optimizer with momentum = 0.9 and weight decay = 0.0005. Learning rate is set at 0.001 and decays by a factor of 10, attached every 6 epochs of 30 epochs in total. For Transformer pipeline, the stage of encoder and decoder is 4 and the number layers of them are [2,6,8,16]. In Mix block, the ω_f is set to 0.2 and the ω_F is set to 0.15 to generate evaluation result. The loss function is a common cross-entropy function used for semantic segmentation.

On self-built dataset verification, the same network setting is utilized. In Mix block, the ω_f is set to 0.15 and the ω_F is set to 0.1 to generate evaluation result.

4.3 Ablation Study

For the purpose of exploring the influence of different factors on the accuracy, we conducted ablation studies on KITTI dataset. Specifically, optimizer, input sizes, and model scales are discussed below.

Effect of optimizers, model scales and image sizes: For CNN pipeline, we explore the effect of different optimizers and weight decay (WD). The experimental results in Table 1 indicate that the SGD combined with weight decay can obtain better segmentation accuracy. For Transformer pipeline, we discuss the effect of network deepening on model performance. Similar to [19], we try different scales of the network. It can be seen from Table 1 that increase of model scale improves the performance of the model. Considering the accuracy, we adopt the large-size model to perform road image segmentation.

The testing results of the network with 224, 512 input resolutions as input are presented in Table 1. As the input size increases from 224 to 512 and the patch size remains the same as 4, the input token sequence of Transformer will become larger, and more semantic information is used to improve the ability of the Transformer pipeline. The CNN pipeline has the same effect, so the experiments in this paper use 512 resolution scale as the input.

Table 1: Ablation study in different pipelines (%)

Pipelines	Image size	Optimizer	Scales	MaxF	AP	PRE	REC	FPR	FNR
Transformer	224	×	Tiny	86.41	75.23	85.43	87.91	2.99	12.09
	224	×	Middle	87.08	77.10	87.64	87.13	2.48	12.87
	224	×	Large	85.96	73.53	86.17	86.68	2.85	13.32
	512	×	Tiny	86.56	76.32	86.71	86.97	2.67	13.03
	512	×	Middle	86.67	75.05	87.71	86.15	2.41	13.85
	512	×	Large	87.04	77.47	88.05	86.61	2.35	13.39
CNN	224	Adam	×	86.43	75.39	83.43	89.84	3.89	10.16
	224	Adam-WD	×	86.13	75.60	83.65	89.01	3.77	10.99
	224	SGD-WD	×	90.62	80.35	89.17	92.29	2.52	7.71
	512	Adam	×	93.17	84.84	91.51	94.95	1.91	5.05
	512	Adam-WD	×	93.77	86.17	92.98	94.66	1.55	5.34
	512	SGD-WD	×	94.52	86.81	93.69	95.43	1.41	4.57

Effect of fusion degree: The experimental results in Table 2 indicate that the results of Transformer pipeline do improve detection accuracy, although most of the effects are contributed by the CNN pipeline. The fusion degree balances the contributions of both and indicates each advantages.

Table 2: Ablation study of fusion degree (ω_F & ω_f). MIOU is used to evaluate the effect of mix block

ω_F	0.05	0.05	0.05	0.15	0.15	0.15	0.25	0.25	0.25
ω_f	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
MIOU (%)	88.38	89.63	87.33	89.38	90.03	88.43	87.58	88.03	86.33

Table 3: The distribution of the number of road image types

Road image types	Quantity ratio	Number
Cement pavement	0.50%	10
Water-covered pavement	2.50%	50
Soil-covered pavement	51.90%	1038
Gravel pavement	43.20%	864
Highway	1.90%	38

4.4 Main Results

Tables 4 and 6 show the performance of our method (MCTNet) on KITTI test set, compared to other listed methods. According to Table 4, the model proposed in this chapter has obvious advantages in road segmentation compared with other models on KITTI dataset. Table 5 shows the result of mixing CNN pipeline and Transformer pipeline. It can be noted that MCTNet method is more effective because of combination of CNN and Transformer.

Table 4: Main indicators results of BEV segmentation on KITTI test set (%)

Benchmark	MaxF	AP	PRE	REC	FPR	FNR
UM_ROAD	90.44	83.88	89.14	91.77	5.10	8.23
UMM_ROAD	94.40	88.26	91.85	97.10	9.47	2.90
UU_ROAD	89.64	81.10	86.75	92.72	4.61	7.28
URBAN_ROAD	92.07	84.87	89.81	94.45	5.91	5.55

Table 5: Best indicators results on KITTI train set (%)

Results	MaxF	AP	PRE	REC	FPR	FNR
CNN	94.52	86.81	93.69	95.43	1.41	4.57
Transformer	87.04	77.47	88.05	86.61	2.35	13.39
Mix result	94.59	88.45	95.48	93.79	0.96	6.21

Table 6: Segmentation accuracy of different methods on KITTI dataset (%)

Method	MaxF	AP	PRE	REC	FPR	FNR
ALO-AVG-MM [16]	91.15	83.82	89.07	93.33	5.22	6.67
RoadNet-RT [15]	91.99	92.54	92.75	91.24	3.25	8.76
OFA net [14]	92.08	82.73	87.87	96.72	6.08	3.28
MCTNet (ours)	92.07	84.87	89.81	94.45	5.91	5.55

Table 7 shows the performance of our method (MCTNet) on self-built dataset. As shown in Fig. 6, the performance prove the ability of the network. Obviously, the blurred boundaries, the brightness interference of the field light, and the absence of artificial dividing lines make this task even more challenging. Table 8 shows the performance of our method and some classical segmentation models on URS dataset. Fig. 7 also shows that our method outperforms the others, and our method basically detects the main passable areas of the road. In the future, we will continue to experiment to test the effectiveness of our method and the challenge of this dataset.

Table 7: Segmentation accuracy on URS dataset. ω_F and ω_f represent the degree of mixing

Dataset	Pipelines	Image size	ω_F/ω_f	MaxF	MIOU
URS	CNN	512×512	-	78.34%	73.03%
	Transformer		-	76.23%	69.38%
	Mix result		0.15/0.1	79.68%	74.33%

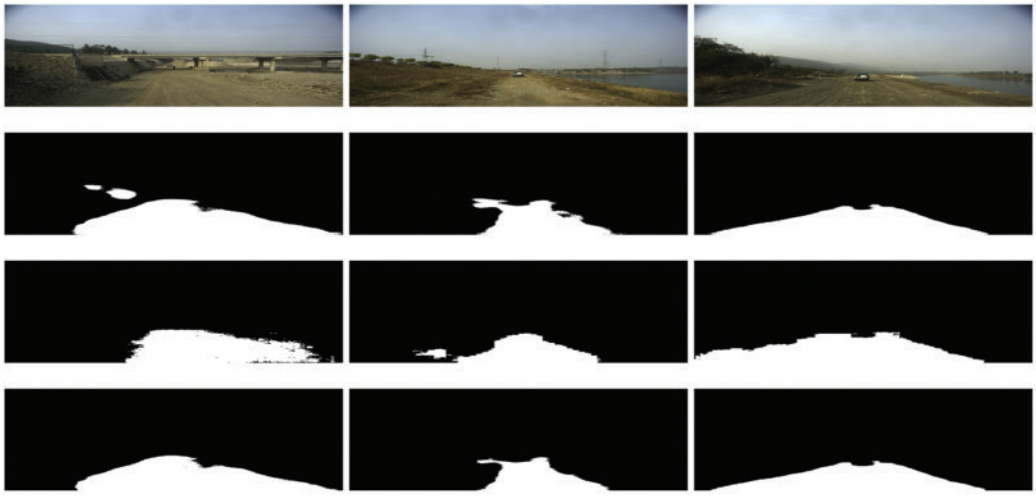


Figure 6: The results of different pipelines on URS dataset. The first row is the input; the second row is the result of CNN pipeline; the third row is the result of Transformer pipeline and the forth row is the result of mix block

Table 8: Segmentation accuracy of different methods on URS dataset

Dataset	Methods	Image size	Batch	Epoch	MIOU
URS	FCN [8]	512 × 512	8	80	71.04%
	PSPNet [23]		8	80	72.48%
	MCTNet (ours)		1	30	74.33%

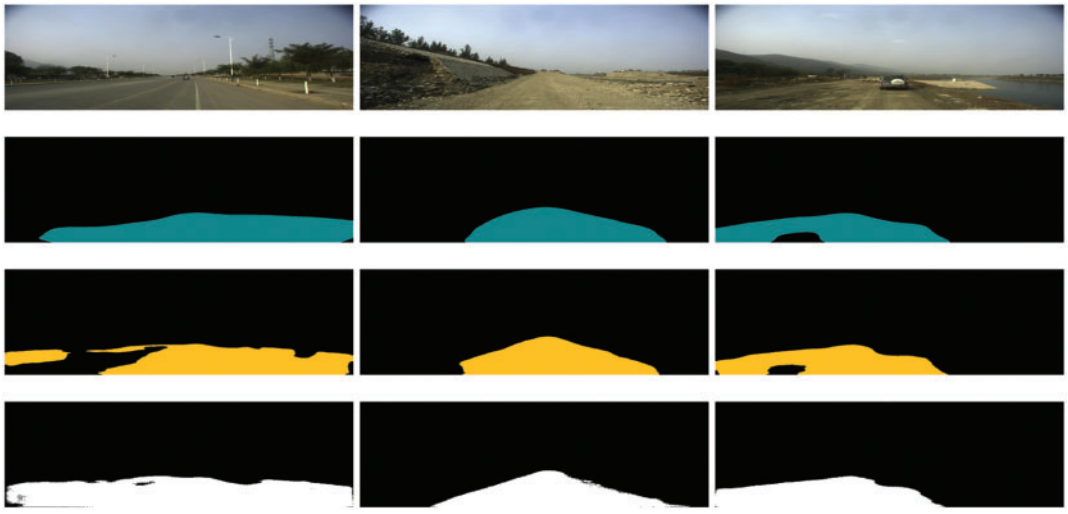


Figure 7: The results of different method on URS dataset. The first row is the input; the second row is the test result of FCN method on self-built dataset; the third row is the test result of PSPNet method on self-built dataset and the forth row is the result of our method (MCTNet)

4.5 Summary

Through the above experiments, we can naturally find that the convolution module of CNN and Swin Transformer Block have different representation effects. It is a better idea to combine them in a single network, and some researchers are already experimenting with it, such as [21]. We have tried to make the convolutional layer and Swin Transformer Block appear in the network at the same time, but we have not obtained excellent results for the time being. We will continue to explore their potentiality for road segmentation tasks.

5 Conclusion

We present a mix method of CNN and Transformer for road segmentation and achieve high performance on KITTI and self-built datasets. The method make full use of the CNN's performance and Transformer's advantages in the road segmentation task. To meet the demand of autonomous driving, a post-processing adding with prior knowledge of the road scene is also vital. Besides, the Unstructured Road Scene dataset we built shows us a unique road segmentation task scene, which brings diversity and applicability. With regard to future research, we will investigate more methods of combining CNN and Transformer for better performance.

Funding Statement: This work is supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX21_1427) and General Program of Natural Science Research in Jiangsu Universities (21KJB520019).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12),* 2481–2495. DOI 10.1109/TPAMI.34.
2. Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528. Santiago.
3. Papandreou, G., Chen, L., Murphy, K., Yuille, A. (2015). Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. arXiv preprint arXiv:1502.02734.
4. Wu, Z., Jiang, B., Karimi, H. R. (2020). A logarithmic descent direction algorithm for the quadratic knapsack problem. *Applied Mathematics and Computation, 369,* 124854. DOI 10.1016/j.amc.2019.124854.
5. Wu, Z., Karimi, H. R., Dang, C. (2019). An approximation algorithm for graph partitioning via deterministic annealing neural network. *Neural Networks, 117,* 191–200. DOI 10.1016/j.neunet.2019.05.010.
6. Wu, Z., Karimi, H. R., Dang, C. (2019). A deterministic annealing neural network algorithm for the minimum concave cost transportation problem. *IEEE Transactions on Neural Networks and Learning Systems, 31(10),* 4354–4366. DOI 10.1109/TNNLS.5962385.
7. Li, H., Zhao, R., Wang, X. (2014). Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification. arXiv preprint arXiv:1412.4526.
8. Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440. Boston.
9. Yu, F., Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

10. Ronneberger, O., Fischer, P., Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Munich, Springer.
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. et al. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
12. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A. et al. (2021). Levit: A vision transformer in convnet's clothing for faster inference. arXiv preprint arXiv:2104.01136.
13. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H. (2021). Going deeper with image transformers. arXiv preprint arXiv:2103.17239.
14. Zhang, S., Zhang, Z., Sun, L., Qin, W. (2020). One for all: A mutual enhancement method for object detection and semantic segmentation. *Applied Sciences, 10(1),* 13. DOI 10.3390/app10010013.
15. Bai, L., Lyu, Y., Huang, X. (2020). Roadnet-RT: High throughput CNN architecture and SOC design for real-time road segmentation. *IEEE Transactions on Circuits and Systems I: Regular Papers, 68(2),* 704–714. DOI 10.1109/TCSI.8919.
16. Reis, F. A., Almeida, R., Kijak, E., Malinowski, S., Guimarães, S. J. F. et al. (2019). Combining convolutional side-outputs for road image segmentation. *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. Budapest.
17. Volpi, R., de Jorge, P., Larlus, D., Csurka, G. (2022). On the road to online adaptation for semantic image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19184–19195. New Orleans.
18. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. et al. (2021). Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning*, pp. 10347–10357. Vienna, PMLR.
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y. et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030.
20. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X. et al. (2021). Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537.
21. Zhou, H. Y., Guo, J., Zhang, Y., Yu, L., Wang, L. et al. (2021). Nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201.
22. Zhao, Y., Wang, G., Tang, C., Luo, C., Zeng, W. et al. (2021). A battle of network structures: An empirical study of cnn, transformer, and mlp. arXiv preprint arXiv:2108.13002.
23. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890. Honolulu.