World Electric Vehicle Journal

MDPI

Article

An Algorithmic Study of Transformer-Based Road Scene Segmentation in Autonomous Driving

Hao Cui and Juyang Lei *

School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; 2577497893@outlook.com

* Correspondence: leijuyang@sues.edu.cn

Abstract: Applications such as autonomous driving require high-precision semantic image segmentation technology to identify and understand the content of each pixel in the images. Compared with traditional deep convolutional neural networks, the Transformer model is based on pure attention mechanisms, without convolutional layers or recurrent neural network layers. In this paper, we propose a new network structure called SwinLab, which is an improvement upon the Swin Transformer. Experimental results demonstrate that the improved SwinLab model achieves a segmentation accuracy comparable to that of deep convolutional neural network models in applications such as autonomous driving, with an MIoU of 77.61. Additionally, comparative experiments on the CityScapes dataset further validate the effectiveness and generalization of this structure. In conclusion, by refining the Swin Transformer, this paper simplifies the model structure, improves the training and inference speed, and maintains high accuracy, providing a more reliable semantic image segmentation solution for applications such as autonomous driving.

Keywords: autonomous driving; semantic segmentation; Transformer; attention mechanism

check for updates

Citation: Cui, H.; Lei, J. An Algorithmic Study of Transformer-Based Road Scene Segmentation in Autonomous Driving. *World Electr. Veh. J.* **2024**, *15*, 516. https://doi.org/10.3390/wevj15110516

Academic Editor: Joeri Van Mierlo

Received: 23 June 2024
Revised: 29 August 2024
Accepted: 19 September 2024
Published: 8 November 2024

1. Introduction

Image semantic segmentation plays a crucial role in various applications, including in autonomous driving. Traditional methods of image segmentation and classification have been extensively studied to generate specific performance parameters for different domains [1]. With the rise in deep learning techniques, particularly in the field of computer vision, deep learning approaches have been applied to semantic segmentation tasks, including scene understanding [2]. Deep learning methods have been reviewed for semantic segmentation in various application areas, highlighting common loss functions and error metrics [3]. In the context of autonomous driving, the application of semantic image segmentation has been explored for enhancing the capabilities of autonomous vehicles [4]. Studies have shown that the quality of ground-truth annotations can impact the performance of semantic image segmentation, with coarse annotations potentially simplifying the dataset preparation and model fine-tuning without sacrificing accuracy [5]. It is essential to benchmark the robustness of semantic segmentation models, especially for practical applications like autonomous driving, by considering a wide range of image corruptions [6,7]. Furthermore, techniques for detecting false positive and false negative samples in semantic segmentation have been reviewed, emphasizing the importance of self-monitoring machine learning algorithms based on uncertainty quantification [8]. Recent advancements in semantic segmentation include the development of novel networks, such as the Bilateral Awareness Network, which aims to capture long-range relationships and fine-grained details in very-fine-resolution urban scene images [9]. Additionally, improvements in semantic segmentation algorithms for low-light autonomous driving scenarios have been proposed, expanding the application range of autonomous vehicles by introducing semantic information and strong feature extraction capabilities [10]. Overall, the application of image semantic segmentation in autonomous driving continues to evolve with the integration of deep learning techniques, benchmarking studies for model robustness, and advancements in network architectures to address specific challenges in different driving scenarios.

In recent years, common models based on deep convolutional neural networks (DCNNs), such as FCN [11], Deeplab v3 [12], and SegNet [13], have achieved good results in traditional semantic analysis tasks but still struggle to achieve ideal performance in the segmentation of urban road scenes. The research on road scene segmentation algorithms has been a significant area of interest in the field of computer vision and machine learning. Various studies have been conducted to develop efficient and accurate methods for segmenting road scenes based on different approaches. Wang et al. [14] introduced an improved road detection algorithm that provides a pixel-level confidence map. The approach was inspired by previous work on road feature extraction and plane extraction from v-disparity map segmentation. Banik et al. [15] focused on recognizing Bangla road signs with a high percentage of accuracy, highlighting the importance of road sign recognition in intelligent vehicle systems. Peng et al. [16] proposed a scene image segmentation algorithm based on the snake model to enhance vision tracking quality for bionic robots. Jing et al. [17] discussed the significance of vision-based perception for mobile robots in understanding their surrounding environment, emphasizing the importance of visual environment-perception algorithms. Qin et al. [18] aimed to establish a speed decision model based on visual road information to promote self-explaining roads and to optimize road designs for safer driving. Wang et al. [10] presented the segmentation of road scenes based on an improved SFNet-N model, highlighting the importance of image segmentation algorithms in automated driving applications. Firkat et al. [19] proposed a novel road detection approach using a hierarchical Transformer model for autonomous driving, achieving significant improvements in performance compared with baseline methods. Duan [20] introduced a point-cloud semantic segmentation model based on geometric segmentation and graph neural networks to enhance the semantic segmentation performance in computer vision applications. Alkendi et al. [21] developed a neuromorphic vision-based motion segmentation algorithm using a Graph Transformer neural network for interpreting scene dynamics in robotic navigation systems. Given the excellent performance of Transformers in the natural-language processing (NLP) domain, there has been increasing interest in applying them to the computer vision (CV) domain, leading to considerable progress. Following the introduction of Vision Transformers (ViTs) [22], many works have explored applying Transformers to various CV tasks. The Swin Transformer [23] is the first highly favored pure Transformer structure suitable for downstream tasks. However, it has drawbacks, such as excessive parameter size, high memory consumption, and long training times. This is because downstream tasks like semantic segmentation require high segmentation accuracy, resulting in a large number of training parameters and increased training costs.

To address these issues, this paper improves the network structure based on the Swin Transformer, significantly accelerating the training speed and better locating the segmentation boundaries. Additionally, for the problem of insufficient feature information learning, traditional approaches have been adopted that involve extracting feature maps at different scales using convolutional layers or pooling layers with various parameters and then fusing these feature maps within the network. However, the multi-scale input of the image pyramid necessitates preserving a large number of gradients during computation, imposing high hardware requirements. In this study, the network is trained at multiple scales and fused at multiple scales during the testing phase, reducing parameters and memory usage. This approach improves network performance by better localizing segmentation boundaries due to the introduction of multi-scale information.

In summary, Transformer models show significant potential and advantages in the field of autonomous driving, particularly in semantic segmentation. By integrating various methods and improving network structures, more efficient and accurate segmentation can be achieved in complex urban road scenes.

The main structure of the paper is as follows: Section 2 describes the preparation and processing of the dataset; Section 3 proposes a new network structure, SwinLab; and Section 4 describes the experimental comparisons and analyses performed.

2. Preparation and Processing of the Datasets

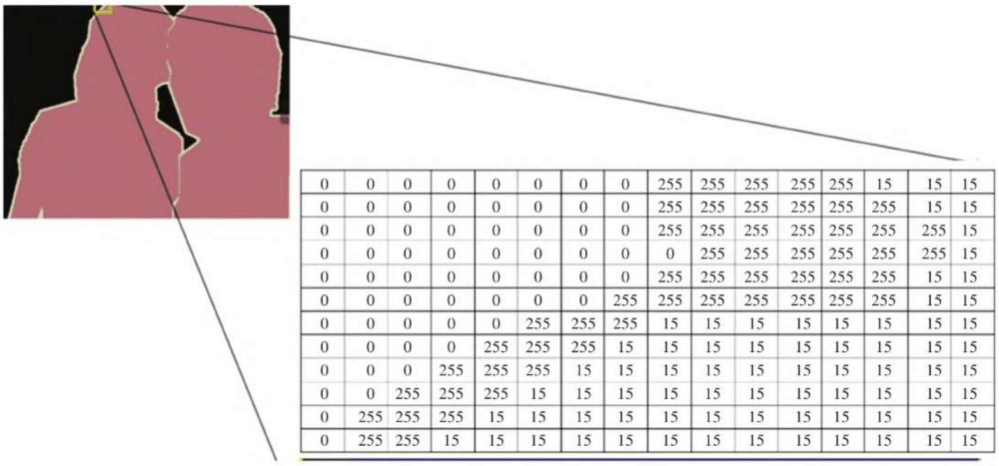
2.1. Datasets

This paper uses the Pascal VOC2012 augmented dataset for research and the Cityscape dataset for further validation.

The Pascal VOC challenge is a world-class computer vision competition. The Pascal VOC challenge can be broadly divided into categories such as the following: image classification, object detection, object segmentation, behavior recognition, and more. The Pascal VOC dataset mainly contains 20 target categories and 1 background category.

For the semantic image segmentation, Pascal VOC2012 has a total of 1464 training images, 1449 validation images, and 1456 test images. However, for semantic segmentation, especially based on Transformer backbone networks, having a large amount of data is very necessary. Therefore, this paper uses the augmented Pascal VOC dataset [24], which has a total of 10,582 training images.

Additionally, when reading the corresponding annotated images (.png) in semantic segmentation using the PIL Image.open() function, the default is the P mode (palette mode), which is a single-channel image. The pixel value at the background is 0, and at the target edges, the pixel value is 255. The target area is filled according to the category index information, as shown in Figure 1. For the person category, the target index is 15, so the pixel value in the target area is set at 15 (refer to Table 1 for specific palette information).



0	0	0	0	0	0	0	0	255	255	255	255	255	15	15	15
0	0	0	0	0	0	0	0	255	255	255	255	255	255	15	15
0	0	0	0	0	0	0	0	255	255	255	255	255	255	255	15
0	0	0	0	0	0	0	0	0	255	255	255	255	255	255	15
0	0	0	0	0	0	0	0	255	255	255	255	255	255	15	15
0	0	0	0	0	0	0	255	255	255	255	255	255	255	15	15
0	0	0	0	0	255	255	255	15	15	15	15	15	15	15	15
0	0	0	0	255	255	255	15	15	15	15	15	15	15	15	15
0	0	0	255	255	255	15	15	15	15	15	15	15	15	15	15
0	0	255	255	255	15	15	15	15	15	15	15	15	15	15	15
0	255	255	255	15	15	15	15	15	15	15	15	15	15	15	15
0	255	255	15	15	15	15	15	15	15	15	15	15	15	15	15

Figure 1. Label map in P mode.

Table 1. Index values for different categories.

category name	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
category index	0	1	2	3	4	5	6	7	8	9	10
category name	diningtable	dog	horse	motorbike	person	pottedplant		sheep	sofa	train	tvmonitor
category index	11	12	13	14	15	16		17	18	19	20

The Cityscape dataset was released in 2016 and is one of the authoritative and popular semantic segmentation datasets in the field of autonomous driving. It contains high-resolution images of a variety of road scenes, with a total of 5000 finely annotated images. The roughly labeled images total 19,998, which ensure maximum access to sufficient data information. This article uses rough labeling and a detailed annotated dataset of 24,998 images, divided into 19 categories such as buildings, pedestrians, sky, etc. The 24,998-image annotated dataset was accessed by image-set augmentation techniques.

2.2. Data Preprocessing

Based on the Transformer network architecture, compared with deep convolutional neural networks, the phenomenon of overfitting is more likely to occur. In addition to optimizing the model structure in the network, having a large amount of data can also reduce overfitting. Hence, preprocessing the images should be considered. The data augmentation operation in this article is carried out on OpenCV and includes image processing rotation within -10° to 10°, random cropping of crop-size from 0.5 to 2 times, random horizontal flipping and blurring of images, etc.

3. Model Architecture

The algorithm of this article consists of two paths, namely the encoder extraction path and the decoder extraction path. The encoder block is based on the Swin Transformer and has been improved to not only speed up the training process but also to alleviate overfitting. The Prediction Head in the decoder block is based on the ASPP+ module, considering the optimization of the module structure through skip connections and shortcut branches to better address the target multi-scale issue. Specifically, it is based on the improved Swin Transformer model, SwinLab, as the backbone network, then optimizing the ASP module and constructing the module ASP+, enabling ASP+ to understand contextual information on multiple scales. The overall model constructs feature maps of different sizes in three stages, and on the basis of the Swin Transformer, the Patch Partition and Linear E modules are removed, and are added together by Patch Merging to form one module, while the following two stages downsample the layers. The overall network model structure is shown in Figure 2. Suppose an input feature image of size $H \times W \times 3$ is input into the model of this paper, where H is the image length, W is the image width, and 3 denotes the image dimension. When the feature image is cut into patches of size 4×4 after the Overlap Patch Merging module, the tensor of each patch is $(H/4) \times (W/4) \times 48$. This is then projected onto the C-dimension through the linear embedding layer, after the Block Transformer searches for global context information. At this time, Stage 1 is completed, with an output feature map of $(H/4) \times (W/4) \times C$, and the Stage 2 and Stage 3 output feature maps are $(H/8) \times (W/8) \times 2C$ and $(H/16) \times (W/16) \times 4C$. After the encoding part, we then obtain the different levels and sizes of the semantic segmentation feature maps.

As the backbone network, the number of parameters and computation amount of the Swin Transformer is far more than that of the CNN network, and at the same time, the local correlation between images is closely related to the maintenance of image resolution. Therefore, in this paper, we improve the coding part of the Swin Transformer to increase the image relevance to ensure high resolution and to maximize the semantic information of different categories and sizes of images. A lightweight coding structure is adopted, which is more suitable for semantic image segmentation based on the Transformer and reduces the number of parameters.

3.1. Encoding Block

The encoding block has two structures, one using the $W-MSA$ structure, and the other using the $SW-MSA$ structure. In general, these two structures are used in pairs; first, using the $W-MSA$ structure, and then, the $SW-MSA$ structure. The specific encoder model is shown in Figures 3 and 4. Meanwhile, for the MLP module, the equation is as follows:

$$T2 = MLP(LN(W_{W-MSA}(LN(T1)) + T1) + t_1 \tag{1}$$

$$T3 = MLP(LN(W_{SW-MSA}(LN(T2)) + T2)) + t_2 \tag{2}$$

where t_1 and $T2$ denote the characteristic maps of the $W-MSA$ module and the MLP output of the next module thereof, and t_2 and $T3$ denote the characteristic maps of the $SW-MSA$ module and the MLP output of the next module thereof.

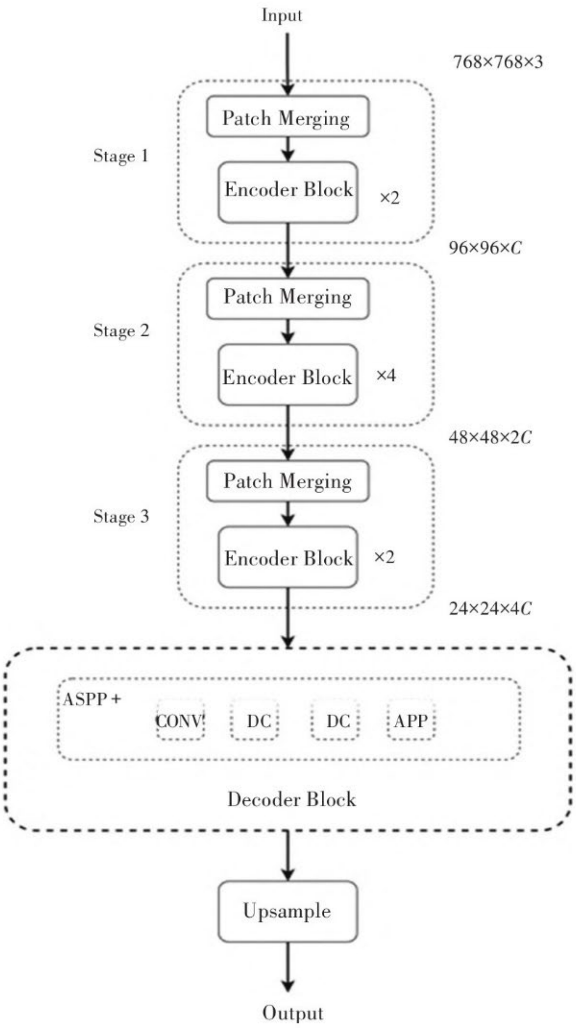


Figure 2. Overall model structure of the network.

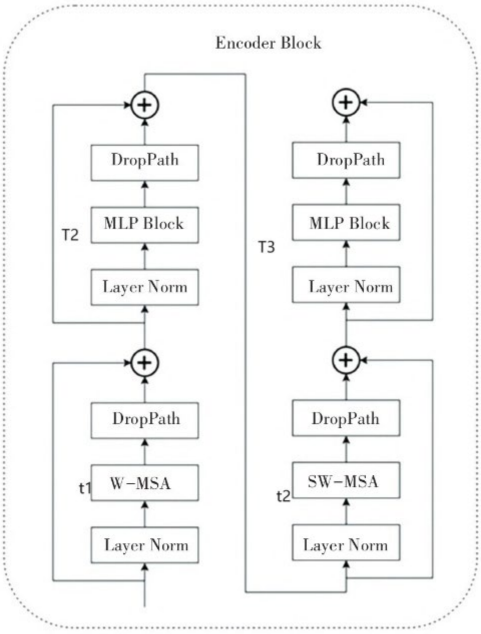


Figure 3. Diagram of the encoder structure.

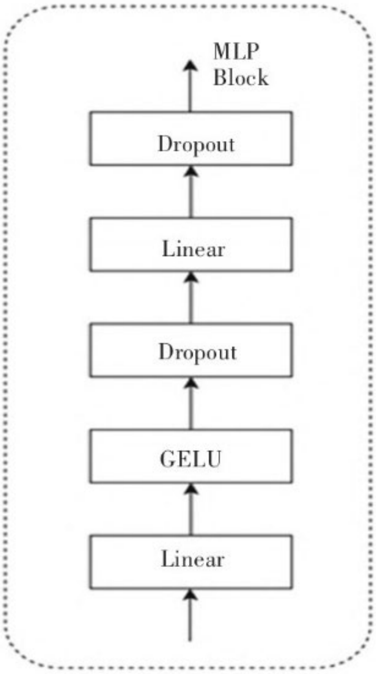


Figure 4. Diagram of the MLP structure.

3.2. Decoding Block

The decoding block includes the ASPP+ module and the Prediction Head module. ASPP+ is based on ASPP; it eliminates the dilated convolutional layer with a dilation factor of 36 and uses adaptive average pooling, totaling four parallel branches, including one 1 × 1 convolutional layer, three 3 × 3 dilated convolutional layers, and one adaptive global average pooling layer, aimed at adding global context information. Here, concat is used before concatenating the four parallel branches using the method, first using the self-attention mechanism to process the information obtained from the different branches. This is beneficial for integrating different feature information, while the dashed shortcut branch uses a 1 × 1 convolutional kernel for dimension processing. As for the Prediction Head module, it receives the ASPP+ module. After the output, a skip connection residual module [25] is added, followed by a Layer Norm layer, and then the information is merged through a 1 × 1 convolutional layer. The Prediction Head uses bilinear interpolation to restore the size of the input image [26]; the network model details are shown in Figure 5. The decoder part of the equation is shown below.

$$x_{out}^{i} = UP_{ample}\left(L_{inear}\left(x_{in}^{i}\right)\right)(i = 1,2,3,4) \tag{3}$$

$$x_{out} = C_{linear}\left(concat\left(x_{out}^{i}\right)\right)(i = 1,2,3,4) \tag{4}$$

where $x_{in}^{i}(i = 1,2,3,4)$ is the decoding part of the different layers for the input feature map after linear transformation into the same channel D_C, upsampling for the original map $(H/4, W/4)$ (the size after splicing; after the linear layer, according to the number of classes N_C of different datasets and the output segmentation results of the feature map x_{out}).

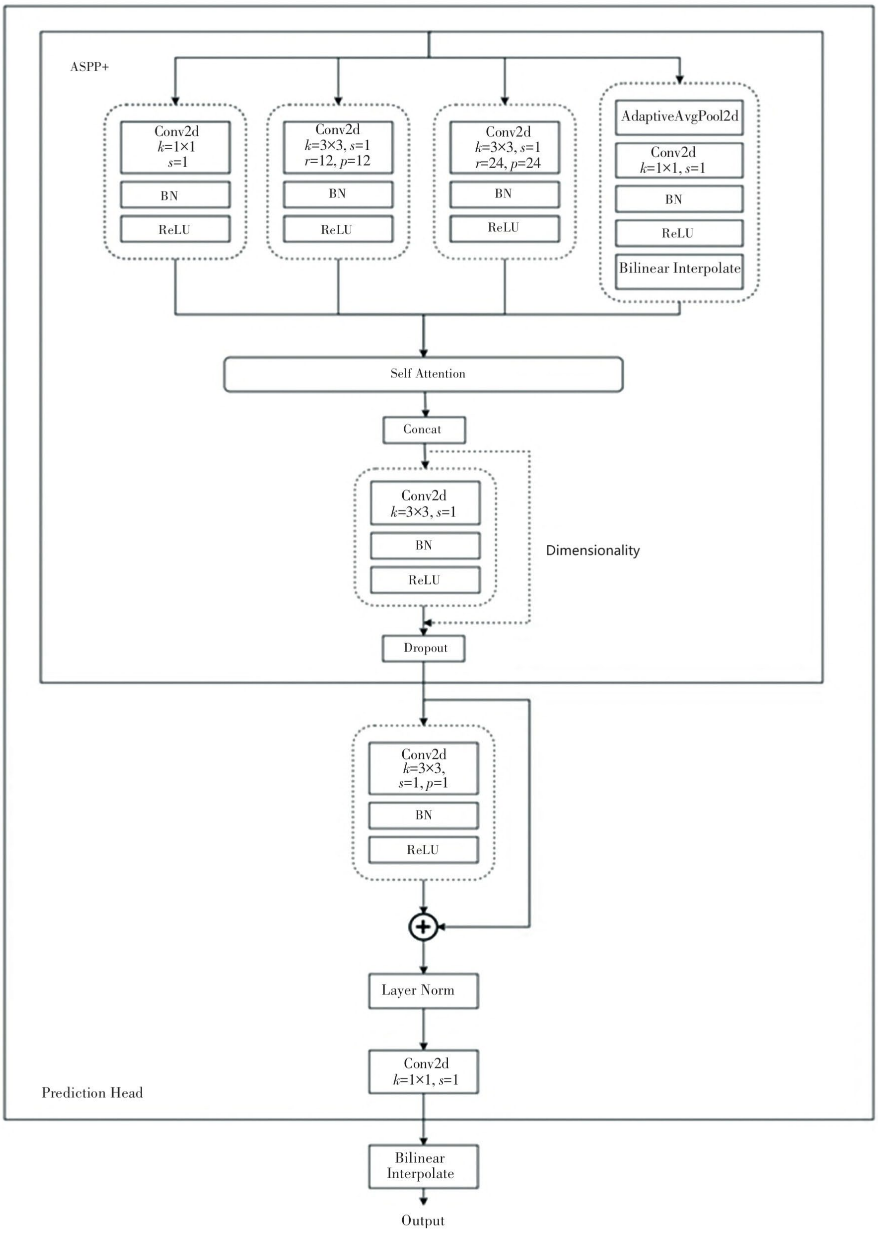


Figure 5. Diagram of the decoder structure.

4. Analysis of the Experimental Results

4.1. Evaluation Indicators

The SwinLab model evaluates the comparative metrics as follows:

(1) Runtime: The runtime includes the training time and testing time of the network model. Since the runtime depends on the hardware devices and the backend implementation, it is difficult to provide the exact runtime in some cases. However, providing information about the hardware on which the algorithm runs and the runtime is helpful for evaluating the effectiveness of the method and ensuring that the fastest execution method is tested in the same environment.

(2) Exact value: Semantic image segmentation is essentially a classification problem, i.e., the pixels are segmented according to the semantic information in the image. The binary classification confusion matrix is shown in Table 2. The rows represent the prediction results of the classification algorithm, and the columns represent the real categories of the samples. The meanings of the four forms in the table are as follows: True Positive (TP): the sample category is positive, and the prediction result is true;

False Positive (FP): the sample category is negative, but the prediction result is positive, which is a misclassification; False Negative (FN): the sample category is positive, but the prediction result is negative, which is an omission; True Negative (TN): the sample category is positive, but the prediction result is negative, which is an omission.

Table 2. Binary confusion matrix.

Prediction/Truth	True Results (Positive)	True Results (Negative)
Projected results (positive)	TP	FN
Projected results (negative)	FP	TN

Once the confusion matrix is obtained, the metrics can be classified. The formulas for Intersection over Union (*IoU*), Accuracy (*A*), Precision (*P*), and Recall (*R*) are shown in Equations (5)–(8).

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

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

The thinking of the confusion matrix is brought into the semantic segmentation multi-categorization evaluation metrics to classify the pixel points in the images. The specific segmentation metrics are as follows:

(1) Accuracy. This includes the Pixel Accuracy (*PA*), which is the ratio of correctly classified pixels in an image to the total pixels in the image; the formula is shown in Equation (9).

$$PA = \frac{\sum_{i=0}^{n} p_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} p_{ij}} \tag{9}$$

(2) Mean Precision (*MPA*). The average of the pixel accuracy of all object categories in the image; the formula is shown in Equation (10).

$$MPA = \frac{1}{n+1} \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij}} \tag{10}$$

(3) Mean Intersection Ratio (*MIoU*). The ratio of the intersection of the true values of the image segmentation results to their concatenation, averaged by class, as shown in Equation (11).

$$MIoU = \frac{1}{n+1} \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n} p_{ji} - p_{ii}} \tag{11}$$

(4) Frequency-weighted intersection and merger ratio (*FWIoU*). This is a new evaluation criterion that improves on the average intersection and merger ratio and aims to weight each pixel class according to its frequency of occurrence, as shown in Equation (12).

$$FWIoU = \frac{1}{\sum\limits_{i=0}^{n} \sum\limits_{j=0}^{n} p_{ij}} \sum_{i=0}^{n} \frac{\sum\limits_{i=0}^{n} p_{ij} p_{ii}}{\sum\limits_{j=0}^{n} p_{ij} + \sum\limits_{i=0}^{n} p_{ji} - p_{ii}} \tag{12}$$

(5) For real-time semantic segmentation architectures, the efficiency comparison of network-to-model is also essential. In addition to the above necessary evaluation metrics, comparative parameters such as milliseconds (ms), frames per second (fps) transmitted on

the screen, the number of floating-point operations performed per second (*FLOPs*), and the number of model parameters (Params) are also required to be compared. The formula is shown in Equation (13).

$$FLOPs = 2HW\left(C_{in}K^2 + 1\right)C_{out} \tag{13}$$

Following the commonly used semantic segmentation evaluation methods, this paper uses *MIoU*, *FLOPs*, and Params as evaluation metrics to analyze the experimental results on the Cityscapes and the Pascal VOC 2012 datasets.

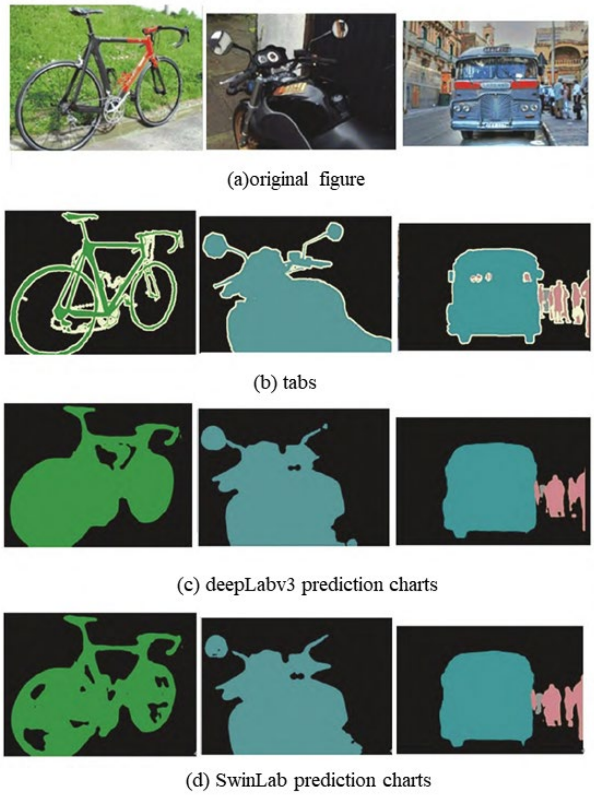
4.2. Experimental Results

In order to further validate the generalization ability of the model, this paper conducts experiments on the Cityscapes and the Pascal VOC 2012 datasets to validate the Transformer-based model. The experimental results are shown in Table 3. Compared with the DeepLabv3 model, the model in this paper can achieve better experimental accuracy. As can be seen through Table 2, the MIoC on the Cityscapes and the Pascal VOC 2012 datasets is 77.61% and 64.64%, which is higher compared with the results of the lightweight network MobileNetv2 under the model of the network DeepLabv3.

Table 3. Performance comparison of different models.

Method	Backbone	Params (M)	Flops (G)	MioU (%)	
				Cityscapes	VOC 2012
DeepLabv3	MobileNetv2	18.70	75.37	69.67	54.69
	ResNet-50	68.21	270.25	75.61	62.89
	ResNet-101	87.21	384.15	75.93	64.48
SwinLab	Transformer	29.31	30.61	77.61	64.64

The segmentation results on the Pascal VOC2012 dataset and the Cityscapes dataset are shown in Figures 6 and 7. In Figures 6 and 7, (a) to (d) show the original figure, tabs, DeepLabv3 prediction charts, and SwinLab prediction charts, respectively.

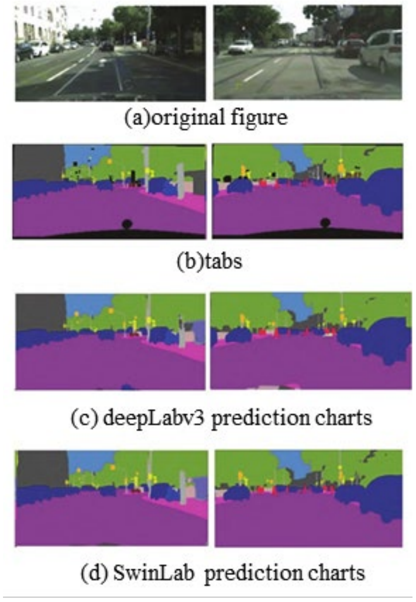

(a)original figure

(b) tabs

(c) deepLabv3 prediction charts

(d) SwinLab prediction charts

Figure 6. Pascal VOC2012 dataset.


(a)original figure

(b)tabs

(c) deepLabv3 prediction charts

(d) SwinLab prediction charts

Figure 7. Cityscapes dataset.

Compared with the feature maps cut by DeepLabv3, it can be noticed that the details of small-object segmentation are severely lost during the segmentation of the image. Furthermore, the distant view of the cut DeepLabv3 is fuzzier, and there is even is a small-object mis-segmentation phenomenon, but the model in this paper is very good at avoiding this kind of failure.

In summary, it can be seen that the model in this paper can effectively segment large-area categories, has a stronger ability to refine the segmentation of small-object categories, and even has the advantage of segmentation with detail information captured for multiple categories of cut pictures. Therefore, the Transformer-based model is suitable for downstream image-cutting tasks.

5. Conclusions and Outlook

5.1. Conclusions

For road scene recognition tasks, this paper proposes a Transformer-based SwinLab model architecture. This network architecture enhances robustness in multi-scale and multi-category segmentation by employing different sampling ratios and receptive field feature extraction to capture contextual information at multiple scales. The experimental results show that the SwinLab model, based on the Transformer, improves the effect and performance of semantic segmentation compared with traditional deep convolutional neural network (DCNN) models. Although it does not achieve state-of-the-art (SOTA) performance, it attains a mean Intersection over Union (MIoU) of 80.1 on the Pascal VOC 2012 dataset and shows promising results on the Cityscapes dataset. Additionally, the training speed, which is a focus of this paper, has been significantly improved, offering practical insights for future research. Furthermore, given the substantial influence of the experimental environment on network performance, there is considerable potential for further enhancement of the network constructed in this study.

5.2. Outlook

In this paper, we launch a research study based on the Transformer segmentation algorithm to provide a newly developed semantic segmentation for subsequent perspectives on the subsequent development of semantic segmentation, which will stimulate further research. However, there are still a lot of areas that need to be improved in terms of how to effectively weigh the segmentation accuracy and speed:

(1) Accuracy. In the Transformer-based image semantic segmentation algorithm proposed in this paper, the number of parameters is small, which is mainly due to the multi-head attention mechanism module and the lightweight decoding part. Therefore, there are still many parameters worthy of optimization to increase the accuracy of the network, such as how to further reduce the exponential amount of computation, model prediction, image preprocessing, training strategy, and so on.

(2) Dataset. The datasets used in this paper are the more commonly used datasets for semantic segmentation, namely Cityscapes and VOC 2012, which are both based on the precise labeling of the training images to complete the learning of the fixed model, although the generalization ability is limited to a certain extent, and the segmentation ability of the images with large differences in the inputs needs to be further improved. The cost of manual annotation is also high, and semi-supervised or fully supervised learning based on Transformer is also a major research direction for the future.

(3) Practical application. The model proposed in this paper achieves satisfactory results in terms of the number of parameters, but the CNN model is already very mature in practical application, while the practical application of Transformer in industrial scenarios needs to be further explored and researched, and there are still many practical problems to be solved.

Author Contributions: Conceptualization, H.C. and J.L.; Methodology, H.C.; Software, H.C.; Writing—original draft, H.C.; Writing—review & editing, H.C.; Funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sevak, J.S.; Kapadia, A.D.; Chavda, J.B.; Shah, A.; Rahevar, M. Survey on semantic image segmentation techniques. In Proceedings of the 2017 International Conference on Intelligent SUSTAINABLE systems (ICISS), Palladam, India, 7–8 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 306–313.
2. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
3. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65. [CrossRef]
4. Kaymak, Ç.; Uçar, A. A brief survey and an application of semantic image segmentation for autonomous driving. In *Handbook of Deep Learning Applications*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 161–200.
5. Taran, V.; Gordienko, Y.; Rokovyi, A.; Alienin, O.; Stirenko, S. Impact of ground truth annotation quality on performance of semantic image segmentation of traffic conditions. In *Advances in Computer Science for Engineering and Education II*; Springer International Publishing: New York, NY, USA, 2020; pp. 183–193.
6. Kamann, C.; Rother, C. Benchmarking the robustness of semantic segmentation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8828–8838.
7. Rottmann, M.; Maag, K.; Chan, R.; Hüger, F.; Schlicht, P.; Gottschalk, H. Detection of false positive and false negative samples in semantic segmentation. In Proceedings of the 2020 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, France, 9–13 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1351–1356.
8. Kamann, C.; Rother, C. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *Int. J. Comput. Vis.* **2021**, *129*, 462–483. [CrossRef]
9. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
10. Wang, H.; Chen, Y.; Cai, Y.; Chen, L.; Li, Y.; Sotelo, M.A.; Li, Z. SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 21405–21417. [CrossRef]
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.

13. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
14. Wang, B.; Frémont, V.; Rodríguez, S.A. Color-based road detection and its evaluation on the KITTI road benchmark. In Proceedings of the 2014 IEEE Intelligent Vehicles Symposium Proceedings, Dearborn, MI, USA, 8–11 June 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 31–36.
15. Banik, B.; Alam, F.I. A robust approach to the recognition of text based Bangla road sign. In Proceedings of the 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, Bangladesh, 23–24 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1–7.
16. Peng, W.; Jian, W.; Yuan, Z.; Jixiang, L.; Peng, Z. Research on Robustness Tracking of Maneuvering Target for Bionic Robot. *Int. J. Secur. Its Appl.* **2015**, *9*, 67–76. [CrossRef]
17. Jing, P.; Zheng, W.; Xu, Q. Vision-based mobile robot's environment outdoor perception. In Proceedings of the 3rd International Conference on Computer Science and Application Engineering, Sanya, China, 22–24 October 2019; pp. 1–5.
18. Qin, Y.; Chen, Y.; Lin, K. Quantifying the effects of visual road information on drivers' speed choices to promote self-explaining roads. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2437. [CrossRef] [PubMed]
19. Firkat, E.; Zhang, J.; Wu, D.; Yang, M.; Zhu, J.; Hamdulla, A. ARDformer: Agroforestry road detection for autonomous driving using hierarchical transformer. *Sensors* **2022**, *22*, 4696. [CrossRef] [PubMed]
20. Duan, S. Semantic segmentation of point cloud based on graph neural network. In Proceedings of the Third International Conference on Computer Vision and Pattern Analysis (ICCPA 2023), Hangzhou, China, 7–9 April 2023; SPIE: Bellingham, WA, USA, 2023; Volume 12754, pp. 357–361.
21. Hamandi, M.; Seneviratne, L.; Zweiri, Y. Static hovering realization for multirotor aerial vehicles with tiltable propellers. *J. Mech. Robot.* **2024**, *16*, 031004. [CrossRef]
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 568–578.
26. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.