A Driving Area Detection Algorithm Based on Improved Swin Transformer

Shuang Liu[1]*, Ying Li[2], Huankun Sheng[3]
College of Computer Science and Technology, Jilin University, Changchun, 130012, Jilin, China[1, 2, 3]
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education,
Jilin University, Changchun, 130012, Jilin, China[1, 2, 3]

***Abstract*—Drivable area or free space detection is an essential part of the perception system of an autonomous vehicle. It helps intelligent vehicles understand road conditions and determine safe driving areas. Most of the driving area detection algorithms are based on semantic segmentation that classifies each pixel into its category, and recent advances in convolutional neural networks (CNNs) have significantly facilitated semantic segmentation in driving scenarios. Though promising results have been obtained, the existing CNN-based drivable area detection methods usually process one local neighborhood at a time. The locality of convolutional operation fails to capture long-range dependencies. To solve this problem, we propose an improved Swin Transformer based on shift window, named Multi-Swin. First, an improved patch merging strategy is proposed to enhance feature interactions between adjacent patches. Second, a decoder with upsampling layer is designed to restore the resolution of the feature map. Last, a multi-scale fusion module is utilized to improve the representation ability of global semantic and geometric information. Our method is evaluated and tested on the publicly available Cityscapes dataset. The experimental results show that our method achieves 91.92% IoU in road segmentation detection, surpassing state-of-the-art methods.**

***Keywords*—CNNS; driving area detection; multiscale fusion; semantic segmentation; Swin Transformer**

I. INTRODUCTION

With the rapid development of computer technology, autonomous driving has entered into real life. Driving area detection aims to accurately determine the current accessible area of vehicles in complex road environments using relevant technologies, which is a critical research area within the field of autonomous driving. Given the crucial role of the drivable area detection algorithm in ensuring the safety and efficiency of vehicle driving on the road, there is an urgent need to improve the accuracy of road detection.

The existing driving area detection methods can be divided into traditional methods and learning-based methods. Traditional methods use the pavement features of 2D images to segment roads. For example, Shi et al. [1] use the road color characteristics and vanishing points to detect the road boundary. Some researchers use edge detection operators to extract the edge boundary of the road and segment the road surface [2], [3]. Though traditional methods can detect driving areas in real time, they are not suitable for complex situations where the road surface features are not obvious.

Learning-based methods typically rely on semantic segmentation to achieve their goals. Semantic segmentation is a pixel-level technology that acts on each pixel of an image to predict its category. This prediction preserves the edge and semantic information of the original image, which is beneficial for enabling autonomous vehicles to understand the scene. As an exemplary approach, ERFNet [4] has demonstrated remarkable performance in road segmentation by incorporating residual layers and decomposition convolutions. Additionally, SNE-RoadSeg [5], data fusion CNN architecture, leverages RGB images and inferred surface normal information to accurately detect driving areas. Despite the success of existing learning-based techniques, the convolutional feature extraction is often criticized for its inability to capture long-range dependencies, which can impede the semantic segmentation performance.

Compared to convolution-based feature extraction methods, Transformer [6] can learn the relationship between global pixels, rather than just their local neighborhood. Additionally, the number of operations required to calculate the correlation between two positions is independent of the distance. For instance, the Swin Transformer [7] has achieved impressive results in image classification, object detection, and semantic segmentation thanks to its window attention and layered design. However, Transformer has not yet been applied to driving area detection. It should also be noted that the current fusion strategy reduces the information interaction between adjacent patches during the down sampling process.

This paper aims to address the limitation of convolution in capturing long-range dependency information. To achieve this, a pure attention model is proposed to replace the convolution operation with a gradually decreasing spatial resolution. To be specific, the input image is first divided into patches of the same size and the corresponding position encoding for each pixel is generated using a linear embedding layer. An encoder composed entirely of Swin Transformer is used to process the patches and a new patch fusion strategy is proposed to improve the information interaction between adjacent patches in the same window. A multi-scale fusion module is then employed to enhance the expression ability of the global semantic and geometric information of the feature map obtained from the encoder. Finally, a decoder with an up-sampling operation is designed to restore the resolution of the feature map and complete pixel-level segmentation prediction. The road segmentation experiment is conducted on

*Corresponding Author.

publicly available Cityscape dataset [8], and the experimental results prove the effectiveness of the proposed method.

II. Related Works

A. Semantic Segmentation

Convolution neural network (CNN) [9] is a kind of feedforward neural network with convolution computation and depth structure, which was originally designed for image classification tasks. In 2015, Long et al. [10] first applied convolution operations to semantic segmentation tasks in FCN. They use 1x1 convolution to replace the full connection layer in the convolutional network. And the feature map is upsampled to achieve end-to-end network segmentation. U-Net [11] adopts a fully symmetrical encoder-decoder structure on the basis of FCN, and deepens the decoder by stacking convolutional layers. It effectively improves the performance with only a small amount of training data. SegNet [12] transfers the maximum pooling index to the decoder, which improves the segmentation resolution and shows better performance than FCN. Furthermore, the emergence of the residual layer [13] can avoid the degradation of the deep network and achieve very high accuracy with network that stack a large number of layers [14]. DANet [15] uses the Xception network as the backbone, and adds a full connection module based on the attention mechanism at the end to retain the more receptive field. SENet [16] automatically obtains the importance of each channel by explicitly modeling the interdependence between feature channels and divides the attention mechanism into two very key operations, Squeeze and Excitation. While decreasing the number of parameters and computational requirements, the accuracy of the algorithm is improved.

B. Multi-Scale Fusion

Recently, several approaches have been presented to tackle the limited receptive field problem in FCNs and their variations. DeepLab [17] applies atrous spatial pyramid pooling (ASPP) in the spatial dimension and leverages conditional random fields (CRFs) to refine the output results. FPN [18] asserts that small targets require the use of larger-scale feature maps due to inadequate resolution information provided by smaller ones, but downsampling losses in deeper images lead to excessive information loss, potentially disregarding small target details. Zhao et al. [19] propose utilizing dilated convolutions to augment the ResNet architecture. Their PPM module facilitates multi-scale feature fusion by acquiring diverse background information across regions. DeepLabV3+ [20] builds on an enhanced Xception [21] backbone and incorporates the decoder module from DeepLabV3 [22], further integrating low-level and high-level features to improve segmentation boundary accuracy. Qin et al. [23] introduce an autofocus convolutional layer, an attentive variant of ASPP, to enhance multi-scale feature extraction capabilities. This layer dynamically learns the weights of different branches via an attention mechanism, adapting the receptive field size for effective multi-scale feature extraction. Gu et al. [24] utilize dual parallel encoders to extract information at varied scales, subsequently merging them using a decoder. With a UNet backbone, each encoder processes images of dissimilar resolutions to acquire feature maps at differing scales.

C. Vision Transformer

ViT [25] uses Transformer for vision tasks for the first time. The 2D image is divided into patches of the same size and expanded into 1D sequences by pixels. The position coding of each pixel is obtained through the linear embedding layer and then input into the encoder. It shows the great potential of Transformer in the field of vision. However, ViT must first be pretrained on a large-scale dataset. Different from ViT, DEiT [26] uses an appropriate training method and distillation technique to solve this problem. DEiT can learn inductive biases based on CNN thanks to the distillation principle, which enhances its capacity to interpret image-type data. SETR [27] achieves excellent semantic segmentation performance with three optional decoding algorithms and a Transformers-based encoder.

The global attention used in Transformer requires a lot of computing resources. Swin Transformer adopts sliding window and layered architecture to solve this problem. The sliding window restricts the attention calculation to one window, introduces the locality of CNN convolution operation and reduces the amount of calculation. It achieves the impressive results on multiple tasks in the visual field. SegFormer [28] combines Transformer encoder and MLP decoder. The position encoding will result in performance degradation because the testing and training resolution are different. To address this issue, SegFormer utilizes a 3x3 deepwise convolutional layer to transmit positional information. The proposed MLP decoder is utilized to combine local and global attention by aggregating the multi-scale features of the encoder output.

D. Driving Area Detection

The existing driving area detection methods can be divided into traditional methods and deep learning-based methods. The information about the pavement features in the 2D image is extracted and segmented by the conventional driving area detection technique. For example, Shi et al. [1] identified road borders using vanishing points and road color attributes. Gao et al. [29] proposed a real-time vision technique based on the color cue training model of continuous frames to identify the driving area in the presence of shadows, lane markers, or unstable lighting. Yao et al. [30] identify drivable area with Support Vector Machine (SVM) and achieve 82.51% F1-score on KITTI dataset [31]. Deep learning driving area detection makes use of semantic segmentation as a key tool. A multitask CNN network was introduced by Pizzati et al. [32] to determine the available space in each lane. The network can operate in real-time thanks to ROS-based calculation. Qiao et al. [33] built the architecture using the characteristic pyramid network and the spatial pyramid pool module based on the ResNet network. It was able to achieve 84.58% IoU on the BDD100K datasets. Choi et al. [34] proposes a network using accumulated decoder features, called ADFNet, which operates using only decoder information, with no skip connections between encoder and decoder. Han et al. [35] proposed a new partitioned network, EdgeNet. It includes a class aware edge loss module and a channel attention mechanism. More than

70% of IoU was obtained on the Cityscapes dataset. In order to overcome the issues of limited anti-noise ability and inadequate segmentation of small-scale objects, Dong et al. [36] proposed an approach using a generative adversarial network (GAN [37]) in conjunction with an ERFNet model.

While these approaches have yielded good experimental results in the drivable area detection domain, they do not address the problem of poor long-range information reliance due to convolutional kernel restrictions.

III. METHOD

A. Architecture Review

The architecture of the driving area algorithm of the improved Swin Transformer proposed in this paper is shown in Fig. 1, which is composed of an encoder, a decoder, and a multi-scale fusion module. The network processing flow is as follows: First, the RGB input images are separated into identically sized, non-overlapping patches. Second, linear embedding layer generates patch embedding. Then, the encoder takes these embeddings as input to generate feature maps. Next, a multi-scale fusion module is introduced between the encoder and decoder to improve the representation ability of the feature map. After that, the decoder restores the original image resolution. Finally, pixel-level segmentation prediction is produced via a 1x1 convolutional Layer. Below, we'll go into more depth about each module.

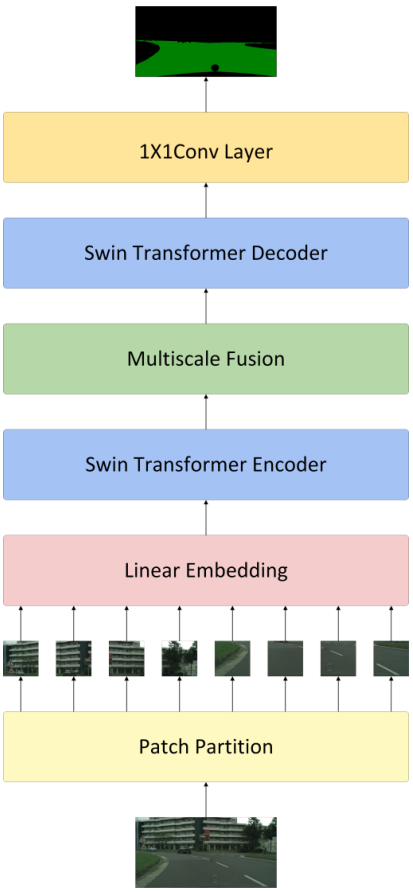

Fig. 1. Network structure.

B. Preprocessing

The input of multi-head self-attention (MSA) is 1D sequence, but there is a mismatch between 2D image and 1D sequence. The input image needs to be sequentialzed. Expanding the image pixel values into a 1D sequence is a direct way. However, the computing complexity increase sharply if the input is a high-resolution image. To solve this problem, we divide the input images into size 4X4, non-overlapping patches, which is similar to prior works [7], [21]. By further mapping each vectorized patch into a C dimensional embedding space with a linear embedding layer, we obtain a 1D sequence of patch embeddings for an input image.

C. Encoder

As shown in Fig. 2, the encoder consists of Swin Transformer blocks and patch merging layers. The supplied image is split into 4X4 patches. The Swin Transformer blocks perform feature representation learning on the input images, and generate a feature map. The patch merging layers down sample the received feature map to expand the receptive field. The layer processing of our proposed encoder is shown in Table I.

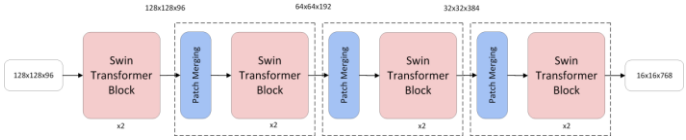

Fig. 2. Detailed display of encoder.

TABLE I. LAYER DISPOSAL OF OUR PROPOSED ENCODER

Layer	Type	Out-F	Out-Res
1	Patch Partition	48	$\frac{H}{4}$ x $\frac{W}{4}$
2	Linear Embedding	C	$\frac{H}{4}$ x $\frac{W}{4}$
3-4	Swin Transformer Block	C	$\frac{H}{4}$ x $\frac{W}{4}$
5	New Patch Merging	2C	$\frac{H}{8}$ x $\frac{W}{8}$
6-7	Swin Transformer Block	2C	$\frac{H}{8}$ x $\frac{W}{8}$
8	New Patch Merging	4C	$\frac{H}{16}$ x $\frac{W}{16}$
9-10	Swin Transformer Block	4C	$\frac{H}{16}$ x $\frac{W}{16}$
11	New Patch Merging	8C	$\frac{H}{32}$ x $\frac{W}{32}$
12	Swin Transformer Block	8C	$\frac{H}{32}$ x $\frac{W}{32}$

1) Swin Transformer block: Fig. 3 illustrates the structure of Swin Transformer block. It is consisted of LayerNorm layer (LN), multi-head self-attention (MSA), residual connection, and MLP layer with nonlinear GELU. As indicated in Fig. 3, Swin Transformer block is computed as follows:

$$\hat{y}^{l} = W - MSA(LN(y^{(l-1)})) + y^{(l-1)} \quad (1)$$

$$y^{l} = MLP(LN(\hat{y})) + \hat{y}^{l} \quad (2)$$

$$\hat{y}^{(l+1)} = SW - MSA(LN(y^{l})) + y^{l} \quad (3)$$

$$y^{(l+1)} = MLP(LN(\hat{y}^{(l+1)})) + \hat{y}^{(l+1)} \quad (4)$$

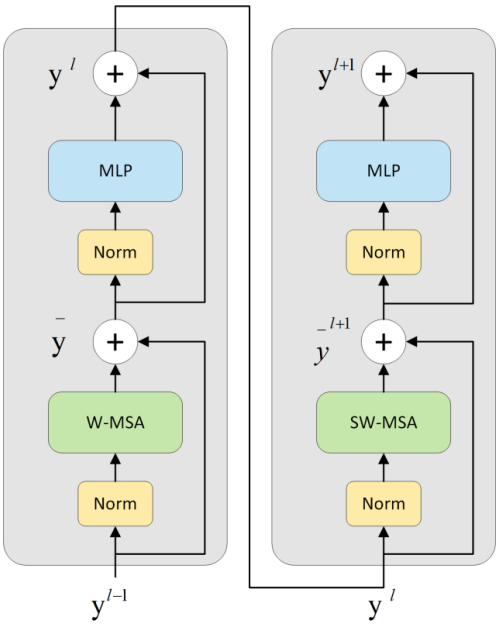

Fig. 3. Illustration of Swin Transformer block inside encoder and decoder.

where, ŷ is the output after (S) W-MSA, and y is the output after MLP. Instead of global attention, the window attention mechanism is used to reduce computational complexity. Compared to the quadratic complexity of global attention, the computational complexity of the small window grows linearly. The window-based multi-head self-attention (W-MSA) and shift window-based (SW-MSA) are utilized to improve cross-window connection. Self-attention formula is as follows:

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

where, Q, K, and V represent query, key, and value matrix, d represents dimension, and B represents offset. Attention is shown in Fig. 4.

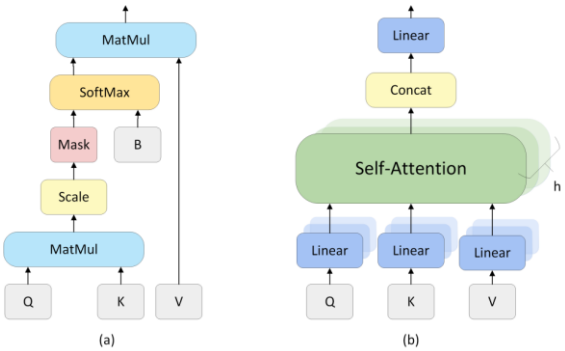

Fig. 4. (a) Self-Attention. (b) Multi-Head self-attention.

2) New patch merging: The patch merging process is shown in Fig. 5. This module aims to down sample the feature map received from the Swin Transformer blocks. It reduces calculation, and realizes hierarchical design. After the merging layer, the resolution of feature map becomes half of the original. First, the pixel values are taken at intervals in the row and column directions of the feature map to form four new tensors. As indicated in Fig. 5, two adjacent pixels on the new feature map are not adjacent in the original feature map, which reduces information interaction during fusion. To improve the interaction between adjacent pixel points, we add a pooling layer in the fusion stage. A new tensor with a channel dimension of 5C is created by concatenating the output of the pooling layer and the feature maps generated from the down sampling. The resolution of the feature map is finally changed using the fully connected layer. The problem of lack of information interaction caused by capturing pixels at intervals is relieved since the feature map produced by the pooling layer has the global features of the input.

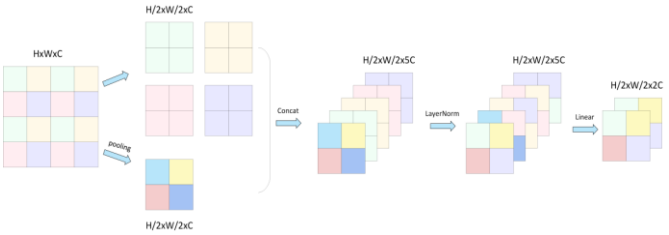

Fig. 5. The process of patch merging.

D. Multi-scale Fusion

The objects in the image are range in size, and each object has a unique set of features. Shallow features can be used to differentiate simple objects, while deep features can be used to separate complex targets. Combining data from various levels is better suited for complicated tasks since the shallow network prioritizes details while the high-level network prioritizes semantic information. In order to improve the capacity to convey global semantic and geometric information, we design a multi-scale fusion module. It combines the output of each group of Swin Transformer Blocks between the encoder and decoder. The multi-scale fusion module is shown in Fig. 6. Given four feature maps produced by Swin Transformer blocks at different stages, the down sampling operation is first performed on the three feature maps with high-resolution. Secondly, 1x1 conv layer is used to map the feature maps of different dimensions to the same dimension. Finally, four groups of feature maps are concatenated to form a new feature map. The obtained feature map has stronger representation ability because it fuses feature information from different levels.

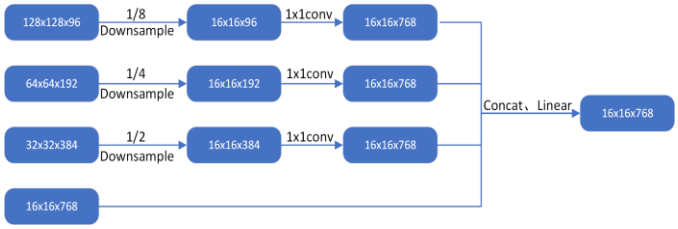

Fig. 6. Details of multi-scale fusion.

E. Decoder

Similar to the encoder, the decoder is composed of Swin Transformer blocks and patch extension layers. Fig. 7 depicts details of the decoder. Among them, the Swin Transformer block is consistent with the encoder, and the patch expansion layer upsamples the feature maps. It has been suggested by SETR that restoring the resolution to its original size in one-step might be interfered by noise. Instead of one-step upscaling, we consider a progressive upsampling technique. Each time a patch expansion layer is applied, the input feature map is increased to 4x resolution. Then feature map resolution is restore to its original size using a 2X upsampling layer at the end of the decoder. To output pixel-level segmentation prediction, a 1x1 convolutional layer is employed. Table II shows the layer processing of our proposed decoder.

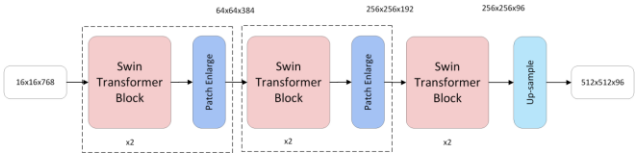

Fig. 7. Detailed display of decoder.

TABLE II. LAYER DISPOSAL OF OUR PROPOSED DECODER

Layer	Type	Out-F	Out-Res
1-2	Swin Transformer Block	4C	$\frac{H}{32} \times \frac{W}{32}$
3	Patch Enlarge	4C	$\frac{H}{8} \times \frac{W}{8}$
4-5	Swin Transformer Block	2C	$\frac{H}{8} \times \frac{W}{8}$
6	Patch Enlarge	2C	$\frac{H}{2} \times \frac{W}{2}$
7-8	Swin Transformer Block	C	$\frac{H}{2} \times \frac{W}{2}$
9	Up-sample	C	H x W

IV. EXPERIMENT

A. Dataset and Experimental Setup

The dataset chosen for this study is Cityscapes. The primary goal of Cityscapes dataset is to provide an image segmentation dataset in an unmanned driving environment, so that researchers can evaluate the performance of algorithms to understand the semantic information of the urban environment. Cityscapes provides 5000 fine annotation images and 20,000 rough annotation images, with a total of 33 categories of annotation items, including 50 street scenes of different cities in various scenarios, backgrounds, and seasons. There are 19 commonly employed categories. Drivable area detection aims to identify the driving area on the road, so we only use the datasets that contain annotations about the road. As a result, there are only two categories in this experiment: drivable area and background. Fig. 8 displays cityscape datasets. There are fine-labeled and coarse-labeled images in the Cityscapes dataset. Although the segmentation accuracy of coarse labeled images is not as good as that of fine labeled images, they still contribute to model training. Therefore, we train the ADE20K pretrained model released by Swin Transformer on roughly labeled Cityscapes images. And use it as a pre-trained model to train fine-label dataset.

Python 3.8 and Pytorch 1.12.1 are used to implement the model. The window size based on shift window attention is set to be 7, the patch size is set to 4, and the input image size is set to 512x512. We trained our model on a NVIDIA RTX2080Ti GPU. Our backpropagation model is optimized using the AdamW optimizer with a momentum of 0.9 during training. The batch size is 8 and the learning rate is 1e-4.

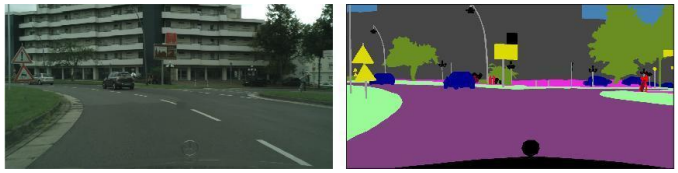

Fig. 8. Cityscapes datasets.

B. Evaluation Metrics

The commonly used Intersection over Union (IoU) [38] index, pixel level accuracy, and precision are used to evaluate the experimental results. We use the pixel level for the three test indicators. IoU is the ratio of the intersection sum of the predicted result and the true value:

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN} \tag{6}$$

Pixel level accuracy is the ratio of correctly classified pixels to the total number of pixels in the image:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

The precision rate is the probability that all predicted positives are actually positives:

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} \tag{8}$$

where, TP, TN, FP and FN represent the pixel level true positive, true negative, false positive and false negative indicators,respectively. Positive refers to the labeled part (driving area), while negative refers to the part of the non-object label (which can be directly understood as the background).

C. Performance Evaluation

In this section, we qualitatively compare our proposed model with state-of-the-art semantic segmentation models. Each model was trained for about 200 epochs until convergence of the loss function. Evaluation was performed on the Cityscapes test set, consisting of 1525 images, at a resolution of 512x512. Experimental results for the Cityscapes dataset are shown in Table III. Our method demonstrates superior performance in category IoU when compared to HRNet [39] and U-Net, and excels in pixel-level accuracy and precision over other models. Specifically, it outperforms HRNet by 0.2%, 0.9%, and 0.63% in these metrics, and has a 0.15%, 0.88%, and 0.44% advantage over U-Net. Despite having a minor disadvantage in the IoU metric against DeepLabV3+, our method ranks first in the other two metrics with respective leads of 0.26% and 0.16%. These results support our claim that our method improves classification accuracy.

The semantic segmentation outcomes on the Cityscapes dataset are depicted in Fig. 9, where (a) is Original driving scene images, (b) is Ground truth annotations, (c) is Road segmentation results of our model. (d) Road segmentation results of DeepLabV3+, (e) is Road segmentation results of HRNet and (f) is Road segmentation results of UNet. The segmentation results reveal that our approach can accurately demarcate the road and surrounding objects within the driving region. In contrast to other techniques, our method provides better predictions for the edges of the drivable area and background. This superiority can be attributed to the fact that the attention mechanism captures long-range semantic information, achieving better performance than convolutional networks in edge detection. Hence, our method excels in learning edge pixels, resulting in higher pixel prediction accuracy and overall performance than other networks.

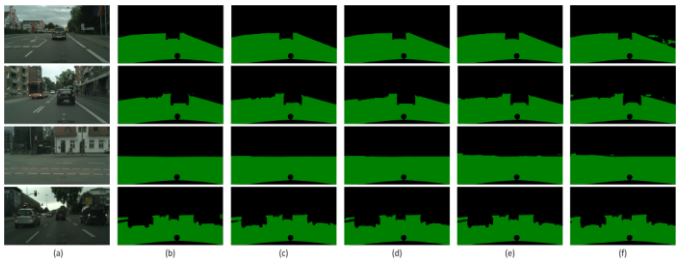

Fig. 9. Examples of road segmentation results on cityscapes dataset.

TABLE III. EVALUATION RESULTS ON THE CITYSCAPES TEST SET FOR ROAD SEGMENTATION

Models	IoU (%)	PA (%)	Precision (%)
Multi-Swin	91.92	**96.79**	**96.19**
HRNet	91.72	95.89	95.56
DeepLabV3+	**92.25**	96.53	96.03
U-Net	91.77	95.91	95.75

D. Ablation Study

In this section, we conduct ablation experiments on our proposed driving area detection algorithm to verify the effectiveness of different modules. From the perspective of patch fusion strategy, multi-scale fusion module and decoder, we conduct comparative experiments.

1) New patch merging: To substantiate the efficacy of the suggested patch fusion strategy, we replaced the patch merging layers with those from the original Swin Transformer, leaving the remaining networkarchitecture unmodified. The experimental outcomes are outlined in Table IV. From the data presented in Table IV, one can observe that our method surpasses the original Swin Transformer patch merging technique in all three examined metrics, resulting in improvements of 0.31%, 0.12%, and0.27%, respectively. This modification mitigates the insufficiency of information interaction during the fusion procedure to some degree, ultimately improving road segmentation accuracy.

TABLE IV. EXPERIMENTAL RESULTS OF FUSION STRATEGY

Models	IoU (%)	PA (%)	Precision (%)
Multi-Swin	**91.92**	**96.79**	**96.19**
Swin	91.61	96.67	95.92

2) Multi-scale fusion module: Our multi-scale fusion module's effectiveness is proven throughseveral experiments, including: a) substituting the proposed module with alternative multi-scale fusion techniques like ASPP and PPM, and b) removing the fusion module entirely. ASPP relies on multiple parallel dilated convolutional layers operating at varying dilation rates to extract features at different scales, which are then processed independently and merged into the final result. By constructing convolutional kernels with varying receptive fields through different dilation rates, ASPP captures object information across scales. On the other hand, PPM is designed to gather background information from multiple regions, addressing the lack of effective strategies to exploit global context in feature fusion. The experimental results are displayed in Table V. As illustrated in the table, our method yields the most favorable outcomes across all three metrics. Relative to ASPP, our fusion module shows significant improvement across all three metrics with gains of 0.51%, 1.28%, and 0.86%, respectively. When comparing against PPM, our fusion module exhibits performance advantages of 0.87%, 1.53%, and 1.21% for the same three metrics. Therefore, our proposed multi-scale fusion module aligns better with our network's design and enhances the accuracy of drive area detection.

TABLE V. EXPERIMENTAL RESULTS OF MULTI-SCALE FUSION MODULE

Models	IoU (%)	PA (%)	Precision (%)
Multi-Swin	**91.92**	**96.79**	**96.19**
Detachment	90.85	94.89	94.63
ASPP	91.41	95.51	95.33
PPM	91.05	95.26	94.98

3) Network structure: Swin Transformer Blocks serve as the primary components of both the encoder and decoder. To assess the effectiveness of our decoder, we substituted it with a Multi-Layer Perceptron (MLP), which includes an input layer, output layer, and several hidden layers. The MLP decoder utilizes GELU as a nonlinear activation function and restores the resolution of the input feature map to its initial dimensions. Results presented in Table VI reveal that our method outperforms MLP decoders across all three tested metrics, yielding boosts of 1.89%, 1.46%, and 2.05% for IoU, PA, and Precision, respectively. This validates the efficacy of our proposed decoder.

TABLE VII. EXPERIMENTAL RESULTS OF DECODER STRUCTURE

Models	IoU (%)	PA (%)	Precision (%)
Multi-Swin	**91.92**	**96.79**	**96.19**
MLP	90.03	95.33	94.14

E. Discussion

This paper presents the outcomes of four distinct experiments: a comparison test using cutting-edge techniques, as well as three separate sets of experiments employing the suggested drivable region recognition algorithm for purposes of elimination. These experiments show that our proposed algorithm achieves exceptional results in the realm of detecting drivable areas, with measurements such as Precision and PA reaching high levels of 96.19% and 96.79%, respectively, placing them at the forefront of comparable efforts. Additionally, the value of IoU was determined to be 91.92%. The experiments carried out for the purpose of eliminating variables confirmed the effectiveness of the various components put forth in this paper. Specifically, the novel patch fusion strategy served to enhance the interplay between neighboring points of interest, the multi-scale fusion module successfully combined more contextually relevant semantic information, thus increasing the expressiveness of feature maps, and lastly, the decoder employed in this work effectively restored the resolution of the feature map layer by layer, thereby mitigating any potential interference caused by noise and better suiting the overall architecture of the network described in this paper. Ultimately, these findings suggest that the methodology introduced in this study improves upon the accuracy of detecting drivable regions and could play a valuable role in furthering the application of deep learning within the domain of autonomous driving.

V. CONCLUSION

In this paper, an enhanced Swin Transformer based semantic segmentation algorithm is proposed. The proposed method is based on encoder-decoder framework. Different from other semantic segmentation networks, we use Swin Transformer as the main body of encoder and decoder. It is applied to the field of drivable area detection for the first time. Meanwhile, the patch merging strategy is improved to enhance the feature interaction between adjacent patches. We design a decoder with an upsampling layer to recover the resolution of the feature maps. Finally, a multi-scale fusion module between the encoder and decoder is used to optimize the expressiveness. We use the publicly accessible Cityscapes dataset for training and testing, and compare our algorithm with state-of-the-art semantic segmentation networks to demonstrate the feasibility and usability of the proposed method. Experimental results indicate that the enhanced Swin Transformer-based method outperforms other well-known algorithms in terms of IoU metrics, achieving higher levels of pixel level accuracy and precision.

Even while the Multi-Swin method has produced ground-breaking results in terms of the accuracy of drivable area recognition, it still has several issues that need to be fixed. In particular, compared to conventional convolutional networks, the use of attention mechanisms places a greater demand on processing power when handling high-resolution pictures. Consequently, future research priorities will be on efficiently lowering the computational complexity and resource usage of the method without compromising or improving detection accuracy. Moreover, this paper's suggested modifications mostly focus on improving the accuracy of the model's predictions, leaving unexplored the possibility of improving the model's real-time responsiveness. The only static images in the training dataset at this time are two-dimensional ones, which is different from the dynamic visual information found in real-world application situations. Thus, future research must immediately focus on improving the algorithm's real-time performance optimization. Concurrently, in order to make sure that the model can better respond to the real-time decision-making requirements in real-world autonomous driving scenarios, video sequences or continuous dynamic picture data must be added for training.

REFERENCES

[1] Jinjin Shi, Jinxiang Wang, and Fangfa Fu. Fast and robust vanishing point detection for unstructured road following. *IEEE Transactions on Intelligent Transportation Systems*, 17(4):970–979, 2015.

[2] ASM Shihavuddin, Kabir Ahmed, Md Shirajum Munir, and Khandakar Rashed Ahmed. Road boundary detection by a remote vehicle using radon transform for path map generation of an unknown area. *International Journal of Computer Science and Network Security*, 8(8):64–69, 2008.

[3] Madoka Otuka, Kenichi Kamino, and Tameharu Hasegawa. Detection of the road area at the ordinary road. In *MVA*, pages 518–521, 2005.

[4] Eduardo Romera, Jose M Alvarez, Luis M Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2017.

[5] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *European Conference on Computer Vision*, pages 340–356. Springer, 2020.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[12] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[18] Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[20] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[21] Franc¸ois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[22] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[23] Yao Qin, Konstantinos Kamnitsas, Siddharth Ancha, Jay Nanavati, Garrison Cottrell, Antonio Criminisi, and Aditya Nori. Autofocus layer for semantic segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*, pages 603–611. Springer, 2018.

[24] Feng Gu, Nikolay Burlutskiy, Mats Andersson, and Lena Kajland Wil´en. Multi-resolution networks for semantic segmentation in whole slide images. In *Computational Pathology and Ophthalmic Medical Image Analysis: First International Workshop, COMPAY 2018, and 5th International Workshop, OMIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings* 5, pages 11–18. Springer, 2018.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[26] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herv´e J´egou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[27] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[28] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[29] Yuan Gao, Yixu Song, and Zehong Yang. A real-time drivable road detection algorithm in urban traffic environment. In *International Conference on Computer Vision and Graphics*, pages 387–396. Springer, 2012.

[30] Jian Yao, Srikumar Ramalingam, Yuichi Taguchi, Yohei Miki, and Raquel Urtasun. Estimating drivable collision-free space from monocular video. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 420–427. IEEE, 2015.

[31] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[32] Fabio Pizzati and Fernando Garc´ıa. Enhanced free space detection in multiple lanes based on single cnn with scene identification. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2536–2541. IEEE, 2019.

[33] Donghao Qiao and Farhana Zulkernine. Drivable area detection using deep learning models for autonomous driving. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5233–5238. IEEE, 2021.

[34] Hyunguk Choi, Hoyeon Ahn, Joonmo Kim, and Moongu Jeon. Adfnet: accumulated decoder features for real-time semantic segmentation. *IET Computer Vision*, 14(8):555–563, 2020.

[35] Hsiang-Yu Han, Yu-Chi Chen, Pei-Yung Hsiao, and Li-Chen Fu. Using channel-wise attention for deep cnn based real-time semantic segmentation with class-aware edge information. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):1041–1051, 2020.

[36] Chaoxian Dong. Image semantic segmentation method based on gan network and erfnet model. *The Journal of Engineering*, 2021(4):189–200, 2021.

[37] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[38] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018.

[39] K Sun, Y Zhao, B Jiang, T Cheng, B Xiao, D Liu, Y Mu, X Wang, W Liu, and J Wang. High-resolution representations for labeling pixels and regions. arxiv 2019. *arXiv preprint arXiv:1904.04514*, 2019.