Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29562

LAE-DeepLabV3+: A Lightweight Network for Road Scene Segmentation

Jun Yang

Computer and Information Engineering, Xiamen University of Technology, Xiamen, China
2521101008@stu.xmut.edu.cn

Abstract. For autonomous driving technology, semantic segmentation is an important technology. It is a dense prediction task, and its performance is largely influenced by the external environment. Therefore, existing models always find it hard to achieve a balance between high-precision segmentation and real-time processing on unstructured roads. This paper presents an LAE-DeepLabV3+ model to solve these problems. It tries to improve the model's accuracy while also improving how quickly it can operate. First, change the MobileNetV2 network to replace the original Xception network. Greatly decreases the model's parameter number but still maintain the ability to get multi-scale features. Second, replacing the standard convolutions used in the ASPP module with depthwise separable convolutions can reduce the computational load and maintain the same receptive field size, which is able to mitigate the grid issue. To solve the problem that DeepLabV3+ is not good at recognizing object details and their edges, the model introduces a Shuffle Attention after the ASPP module. It makes use of this mechanism for the model to pay attention to important parts in complicated road scenes to stay lightweight. Finally, an ECA module is used as pre-processing before model output so that the model can give more weight on semantic feature channels at little cost, thus making the overall segmentation more accurate. The method reached mAP of 78.6% and FPS of 30.4 with training and evaluation on the CamVid dataset. The experimental results indicate that the proposed method successfully achieved an improvement in balancing segmenting ability and real time performance.

Keywords: Autonomous driving, Semantic segmentation, Deeplabv3+, Attention mechanism.

1. Introduction

Traffic accidents have become a major concern around the world to this day, which is hindering the economic growth of nations and endangering public safety [1]. Even though governments are carrying out many targeted measures, accident rates do not show much obvious improvement. This situation is indicative of the persistent challenges currently being experienced in the realms of contemporary traffic management and automotive driving technologies. However, autonomous driving has demonstrated immense potential in addressing these challenges. As a key driver of future transportation transformation, autonomous driving has been prioritized by research institutions and industries worldwide. By integrating advanced cameras, sensors, artificial intelligence algorithms,

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29562

and big data analytics, autonomous driving technology enables vehicles to monitor road conditions in real time and execute rapid responses when necessary, thereby mitigating or avoiding collision risks. Furthermore, the entire process, from perception and decision-making to execution, requires minimal human intervention, eliminating hazards caused by dangerous driving.

With the rapid advancement of artificial intelligence, autonomous driving has become a core technology within the new generation of intelligent transportation systems. A fundamental capability of autonomous driving systems is the real-time perception and precise understanding of complex road environments. Road semantic segmentation is a foundational task within the perception layer. The image is processed by assigning semantic categories (e.g. roads, vehicles, pedestrians) to each pixel, thereby constructing a detailed environmental map for the vehicle. This ensures safe operation in dynamic scenarios and enables the planning of subsequent driving routes.

To tackle this core task, academia has proposed numerous deep learning-based solutions [2]. However, obtaining high-precision and low-latency semantic segmentation in complex and dynamic real-world traffic road environments is a problem.

Therefore, this paper will propose an optimized model named LAE-DeepLabV3+, which is based on the DeepLabV3+ architecture. To improve the computing's efficiency, the model replaces the Xception backbone in the original encoder, which has heavy computation, with MobileNetV2 based on depthwise separable convolution. This kind of approach, thereby having this result of cutting the parameters and making no change to the capability in the way that extracts features. To solve the problem of the high computational cost of the ASPP core structure, this paper replaces all standard convolution modules in the ASPP core structure with depthwise separable convolution modules, thereby increasing the efficiency of multi-scale context information extraction. Next for feature enhancement, the Shuffle Attention mechanism (SA) is introduced. Mechanism is improving the information exchange between different feature spaces as well as channel dimension through channel grouping and channel reordering, and it enhances the model's perception of the edges and details. Finally, the ECA module is integrated into the decoder to realize the channel feature recalibration without too much computation. The main contributions of this paper are as follows:

• To reduce computational complexity, this paper adopts MobileNetV2 as the backbone network, replacing Xception in the original model. MobileNetV2's inverted residual and linear bottleneck design facilitates the maintenance of highly efficient multi-scale feature extraction capabilities whilst concomitantly effecting a substantial reduction in parameters.

• To address the high computational cost of the ASPP module, all standard convolutions are replaced with depthwise separable convolutions. This design substantially reduces the ASPP module's parameters and computations without sacrificing the receptive field. This, in turn, enables more efficient extraction of multi-scale contextual information.

• To address the limitations of DeepLabV3+ in handling object details and edges, a Shuffle Attention mechanism has been introduced after the ASPP module. This mechanism helps the model tackle the road features in complex scenes without using much extra computing power, so it can make segmentations better.

• To let the model focus more on road parts, an ECA module is introduced in the decoder. This module improves the network's segmentation accuracy without adding much to its weight.

2. Related work

Compared to traditional image segmentation algorithms [3], the general segmentation levels reached by deep learning based semantic segmentation algorithms are usually higher. They are good at

DOI: 10.54254/2755-2721/2026.TJ29562

figuring out all sorts of different things and scenes in big, changing photos and separating them out properly.

The initial research focused largely on the FCN [4] with its end-to-end pixel-level output, which changed the block-based segmentation model fundamentally with a fully convolutional network. Furthermore, U-Net [5] was very good at retaining high-resolution information because of its encoder-decoder structure and leap connections. Skip connection has become the standard model for many medical and remote sensing images. The idea of skip connection has also been often used in road scene segmentation [6]. On the other hand, the DeepLab series, such as DeepLabv3+, adopt Atrous Spatial Pyramid Pooling (ASPP) modules to better expand the field of reception so that multi-scale object segmentation is addressed [7]. Building upon this foundation, SegNet introduced an interesting decoder design [8]. By recording the maximumpooling indices when doing a pool and then using them to do non-linear up-sampling inthe decoder. It greatly improves the accurate location of boundaries.

Recent studies were built upon the former ones, so the models were created that can be used in more complex, specific cases. Low-light nighttime environment is very difficult, Wang et al. proposed the SFNet-N model [9]. In this novel approach, the integration of a light enhancement network along with a semantic segmentation network is done. This enhances the road edge's segmentation when it is in a low-light scenario. Open area detection being a challenge in the case of autonomous driving, Fan et al. proposed RoadSeg as a new data fusion semantic segmentation approach [10]. The model uses Surface Normal Estimator (SNE) module, which it can efficiently and accurately infer the surface normal information from a dense depth image. The model also assimilates the acquired information alongside the RGB features into a fusion network to detect the right open places. Facing the shortcomings of traditional detection methods in both resource consumption and precision, Wang et al. used an optimized DeepLabV3+ [11]. The model adopts MobileNetV2 as the backbone and reduces the model parameters. Waterfall-like expanding spatial pyramid pooling (WASPP) uses a waterfall-like dilated convolution strategy, using a stack of different dilated convolutions to expand the field of view. The integration of Squeeze-and-Excitation (SE) attention and Feature Pyramid Network (FPN) enhances lane line feature extraction capabilities.

3. Method

3.1. Dataset

The CamVid dataset [12], created by the University of Cambridge, primarily provides image data for autonomous driving research. It contains real-world urban road scenes recorded from vehicle front cameras, covering diverse environments and 32 categories. This dataset provides an experimental foundation for road scene semantic segmentation. Also, it gives camera pose info like on a pic, which would help for 3D rebuilding a sense of the situation better. It got 701 detailed pictures out of a recording lasting around 10 minutes, and each picture had exact words about what was in it. Image content involves multiple categories, including pedestrians, buildings, roads, and vehicles. The variety of categories makes it suitable for semantic segmentation model training and evaluation. Besides, the dataset contains diverse weather conditions, lighting conditions, and different road surfaces, which are close to reality. The dataset is divided into a training set (367 frames), a validation set (101 frames), and a test set (233 frames). By preserving temporal information from the video, CamVid serves as an ideal dataset for training and evaluating semantic segmentation models in complex environments. Figure 1 displays selected original images and their annotation results.

DOI: 10.54254/2755-2721/2026.TJ29562

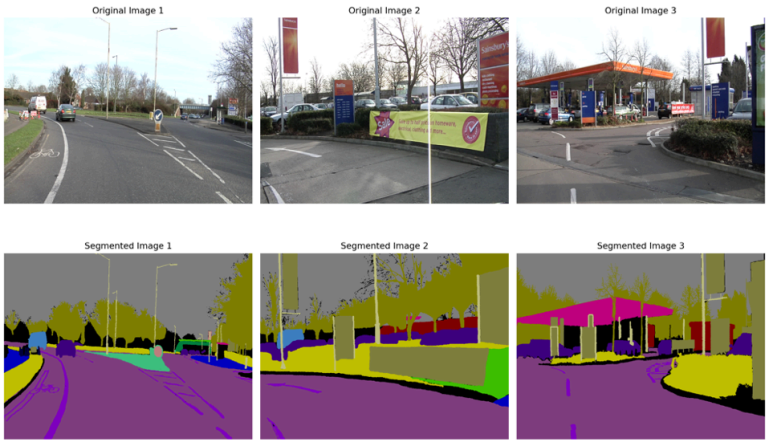


Figure 1. Selected images from the CamVid dataset

3.2. Traditional DeepLabV3+ network model

DeepLabv3+ is a typical semantic segmentation algorithm by Google [7]. Its architecture is an encoder-decoder structure. In the encoder, the features of model are that it uses Xception as the backbone network and ASPP modules. Backbone uses depthwise separable convolution, which separates normal convolutions into two steps. First, each input channel is operated independently by depthwise convolutions. Then, cross-channel feature fusion is executed by 1×1 convolutions of pointwise convolution. Design cuts down on computation, but at the expense of the ability to represent the environment. The backbone's intermediate layers output shallow semantic features, which are fed directly into the decoder. Deeper features, on the other hand, are input to the ASPP module for multi-scale fusion. ASPP, another core component, uses a multi-branch parallel structure for feature extraction. Its innovation lies in utilizing dilated convolutions with varying dilation rates. This enables the model to adaptively process objects in different sizes. The module also incorporates a global average pooling layer, which fuses features from all branches to provide the model with holistic contextual information about the image. In the decoder section, upsampling operations and feature fusion strategies integrate shallow-level semantics from the backbone network with deep-level semantics from the ASPP module, thereby generating edge and detail information in the image. The network structure is shown in Figure 2.

DOI: 10.54254/2755-2721/2026.TJ29562

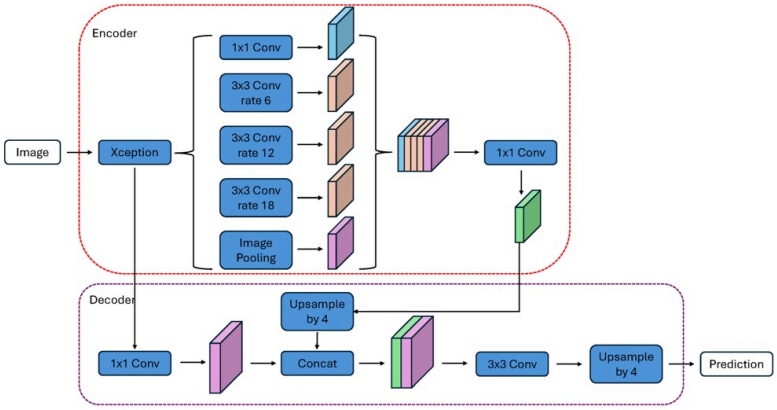


Figure 2. DeepLabV3+ network architecture

3.3. LAE-DeepLabV3+ network model

To enhance the accuracy of segmentation while reducing the model's parameters, this paper proposes a novel framework named Lightweight-Attention-Enhanced Deeplabv3+ (LAE-DeepLabV3+), as shown in Figure 3. The improvements are primarily reflected in four aspects. Firstly, the MobileNetV2 network is employed in the encoder section to decrease the number of model parameters [13]. This backbone network improves the feature extraction capability of each channel output through inverted residual modules, maintaining accuracy while reducing computational load. Secondly, the ASPP module replaces standard convolutions with depthwise separable convolutions. This mitigates the grid effect while reducing computational load. Furthermore, to make the model better represent the details but with less computation, the SA mechanism is introduced after the feature fusion in ASPP [14]. Lastly, in the decoder part, an ECA module is introduced after concat to get more accurate edges and details in the feature map after combining the channels [15].

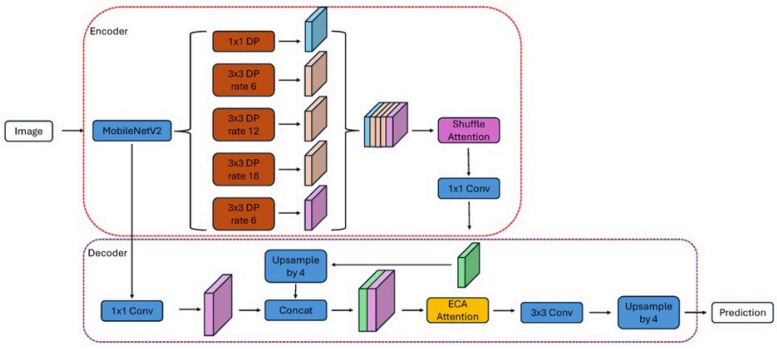


Figure 3. LAE-DeepLabV3+ network architecture

3.3.1. Feature extraction

In the input image stage, LAE-DeepLabV3+ uses lightweight MobileNetV2 in the model for feature extraction. In MobileNetV2, the intermediate layers directly transmit shallow semantic information to the decoder and use multiple convolution layers to perform deep feature extraction until entering the ASPP multi-scale fusion module. The model is much faster at inference than Xception, yet still maintains the segmentation accuracy. This is because of the basic technologies used in

DOI: 10.54254/2755-2721/2026.TJ29562

MobileNetV2: depthwise separable convolution, linear bottleneck layers, and inverted residual structure.

The depthwise separable convolution, illustrated in Figure 4, decomposes convolutions into two independent parts. Depthwise convolution first eliminates efficiently cross-channel operations and reducing parameters. Pointwise convolution then combines the outputs of the depthwise convolutions along the depth dimension to form the final fused feature map.

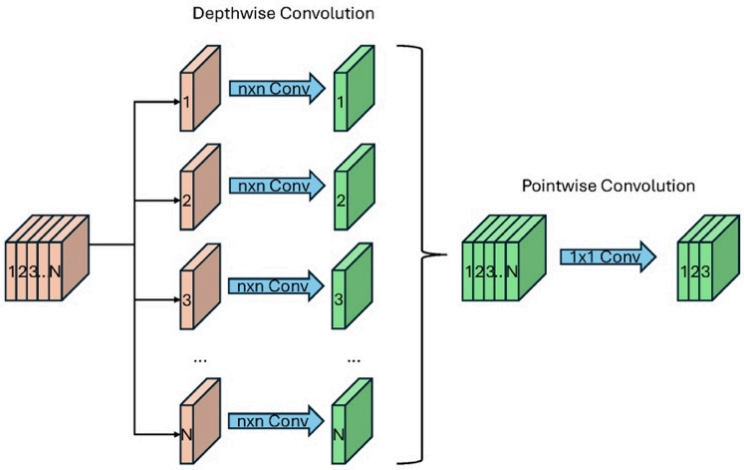


Figure 4. Depthwise separable convolution computation process

The linear bottleneck layer in MobileNetV2 employs channel compression to achieve dimensionality reduction of information. Additionally, linear activations are used throughout the bottleneck layer to reduce computational complexity. Additionally, the inverted residual structure first expands the input features to a higher dimension, then extracts features through deep convolutions, and finally compresses them back to a low dimension to enhance semantic feature information. Figure 5 shows the inverted residual structure.

DOI: 10.54254/2755-2721/2026.TJ29562

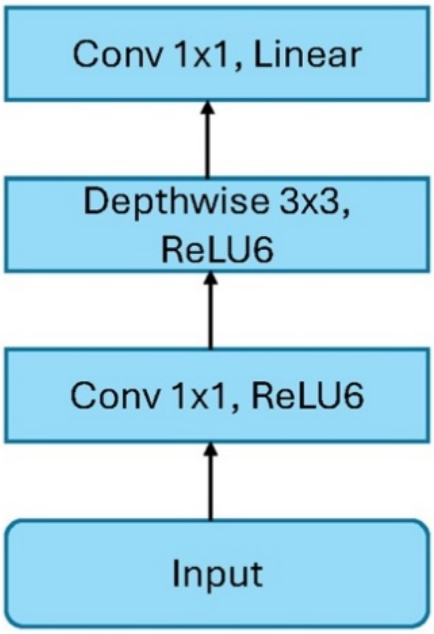


Figure 5. Inverted residual structure

3.3.2. Multi-scale feature fusion

During the multi-scale feature fusion stage, LAE-DeepLabV3+ introduces depthwise separable convolutions in the ASPP module to replace standard convolutions, called DP-ASPP. Firstly, separable convolutions extract deep features at different dilation rates, followed by processing through parallel convolutional layers. The first layer of the ASPP module employs 1×1 convolutions to reduce the number of feature map channels to 256, further decreasing computational complexity. Additionally, depthwise separable convolutions utilize varying dilation rates for feature extraction at each layer. This approach effectively preserves global detail information, thereby mitigating the impact of grid effects on fine segmentation. Finally, the processed feature maps are aggregated through a global average pooling layer to form semantically richer feature maps. The DP-ASPP module is shown in Figure 6.

DOI: 10.54254/2755-2721/2026.TJ29562

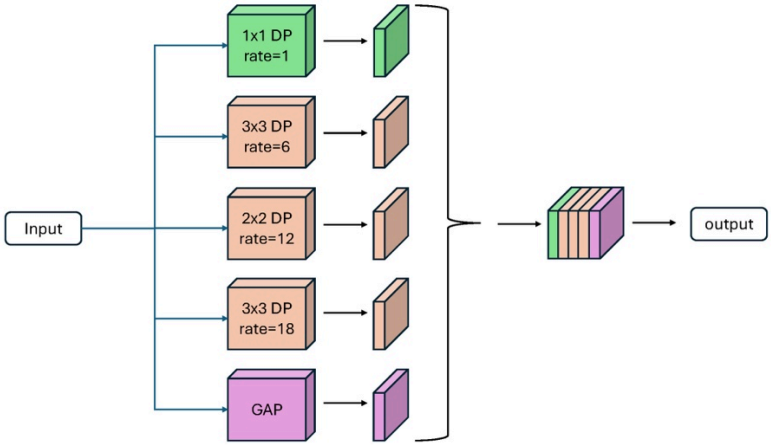


Figure 6. DP-ASPP module

After the DP-ASPP module generates semantic information, the LAE-DeepLabV3+ model incorporates a lightweight attention mechanism called SA following the DP-ASPP module. The structure of SA is illustrated in Figure 7.

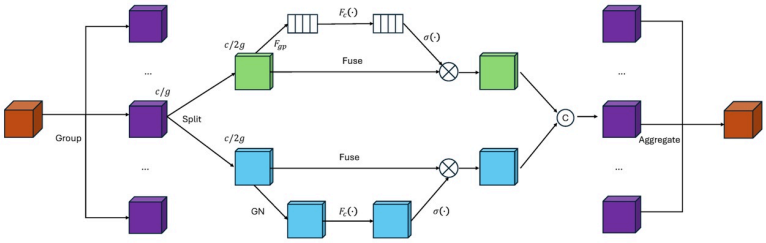


Figure 7. Shuffle Attention module

Firstly, the feature map X is divided into several groups, following the output of ASPP. This approach reduces computational complexity in subsequent network operations while enhancing the network's expressive power. Next, the SA module further decomposes each grouped sub-feature into two attention branches. In channel attention, feature X_{k_1} first undergoes global average pooling followed by the linear transformation. Subsequently, attention weights are generated by an activation function. Finally, attention weights are multiplied with X_{k_1} to obtain the enhanced feature representation $X_{k_1}^{'}$. In spatial attention, feature X_{k_2} undergoes group normalization first, followed by operations that are analogous to those employed in channel attention. These two processes are defined by Equations (1) and (2):

$$X_{k_1}^{'} = \sigma\left(F_c\left(s\right)\right) \bullet X_{k_1} = \sigma\left(W_1 \bullet s + b_1\right) \bullet X_{k_1} \tag{1}$$

where F_c is the linear function, $\sigma\left(*\right)$ represents the linear activation function, s denotes the feature after global average pooling, and the remaining two parameters W_1 and b_1 are used for scaling and offsetting the s feature.

DOI: 10.54254/2755-2721/2026.TJ29562

$$X_{k_2}^{'} = \sigma\left(W_2 \bullet GN\left(X_{k_2}\right) + b_2\right) \bullet X_{k_2} \tag{2}$$

where $X_{k_2}^{'}$ represents the spatial attention feature, $GN(*)$ represents the group norm normalization function, $\sigma(*)$ represents the linear activation function, and the remaining two parameters, W_2 and b_2 are used for linear transformation.

Following the completion of dual processing of both channel attention and spatial attention by each feature group, the two branch processing paths are concatenated to form a new feature representation. Finally, after all feature groups are fused, channel grouping is performed to rearrange the channels. This process employs a deterministic permutation function to forcibly blend feature groups with different semantics, thereby generating enhanced deep semantic information.

3.3.3. Decoder feature fusion

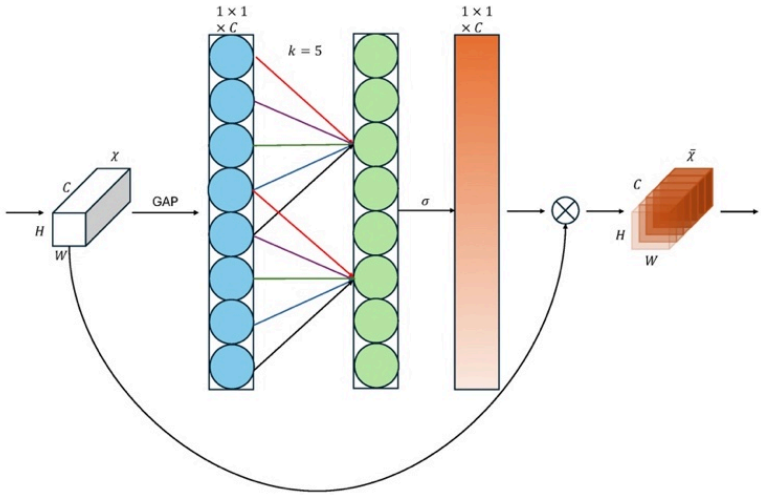


Figure 8. ECA module

The final semantic segmentation result map is fused from shallow and deep semantic information in the decoder. The LAE-DeepLabV3+ model uses the ECA attention module in the decoder to improve the ability to recognize edge details and small objects. The structure of the ECA attention mechanism can be seen in figure 8. For a feature map $X \in R^{C \times W \times H}$, where C, W and H denote the number of channels, the width and height of the feature map, respectively. The ECA module first performs global average pooling along the spatial dimension to obtain a one-dimensional vector $Z \in R^{C}$. Subsequently, it bypasses dimensionality reduction steps and directly processes the features after global average pooling. Finally, the original resolution is restored through upsampling to output a high-precision segmentation map. The relevant mathematical formula is as follows (3-4):

$$Z_c = \frac{1}{H \bullet W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{cij} \tag{3}$$

DOI: 10.54254/2755-2721/2026.TJ29562

where $\sigma(*)$ represents an activation function applied in the convolution operation, such as the Sigmoid function. It ensures that the generated channel weight values fall within the interval [0,1]. The multiplication of the generated weight vector by the original features achieves feature selection and enhancement. X_c denotes the output feature map.

$$\widehat{X_c} = \sigma\left(Conv_k\left(z\right)\right) \bullet X_c \tag{4}$$

4. Environmental results

4.1. Environment setting

The experiment platform uses the Ubuntu 20.06 system, equipped with an Nvidia T4 GPU, and has 16G computer memory. Neural networks created with Python 3.8 and the Pytorch 1.8 deep learning frameworks.

Due to the resources in this region, the training process is split into two stages: frozen and unfrozen, which uses a transfer learning method. During the frozen phase, only parameters in the tail of the model are trained. At this point, batch size and the number of training epochs are set to 16 and 50, respectively. When it moves into the unfrozen phase of training, changes have also been applied to the weights and parameters of the entire network architecture. In this phase, batch size and epochs are set to 8 and 100. From the beginning of the entire training process, the learning rate is fixed at 0.007, and the weight decay is set to 0.0002. Training loss function uses a combined dice and Jaccard loss with weights of 0.7 and 0.3, respectively, to improve the model's resistance against imbalance.

The evaluation metrics that are used in this paper are mIoU, mPA, and FPS. These figures are common benchmarks well-recognized in the field of computer vision. mIoU shows how well the model performs in terms of segmentation of different categories, so it can evaluate the overall segmentation performance of the model. mPA can be used to evaluate the accuracy of the model in segmenting each category, thus helping to find out that the model does not do a good job with some categories. FPS stands for the model's efficiency of real-time inference. It makes sure the model meets the real-time processing requirement in practice. It becomes extremely important when applied to self-driving, where the real-time requirement is very high. The definitions are given in the formulas (5-7).

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{5}$$

$$mPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}} \tag{6}$$

$$FPS = \frac{fn}{et} \tag{7}$$

DOI: 10.54254/2755-2721/2026.TJ29562

4.2. Ablation experiments

4.2.1. Impact of backbone

Table 1. Experimental results the of backbone network

Method	mIoU(%)	mPA(%)	FPS
Xception + DeepLabV3+	69.6	79.2	16.2
MobileNetV2 + DeepLabV3+	67.9	78.4	33.9

To study the impact of the backbone network on the model's real-time segmentation capability and segmentation accuracy, ablation experiments were conducted based on the CamVid dataset. The results of the experiment are shown in Table 1.

As shown in Table 1, replacing the backbone network with MobileNetV2 resulted in a more than twofold increase in the FPS metric, reaching 33.9. This indicates that the model employing MobileNetV2 as a backbone network presents excellent real-time performance. However, the reduction in computational parameters within the network inevitably leads to a decline in its feature extraction capability. For the CamVid dataset, mIoU and mPA down 1.7% and 0.8% respectively. However, FPS is the best reflection of the model's real-time performance. To replace the backbone network leads to a trivial decline in the segmentation accuracy, but it brings about a substantial enhancement in real-time segmentation results. The experiment shows that it is feasible to adopt the MobileNetV2 network in this model, which can maintain the accuracy of the segmentation results and improve the real-time of the model.

4.2.2. Impact of DP-ASPP

Table 2. Experimental results of DP-ASPP

Method	mIoU(%)	mPA(%)	FPS
Original ASPP	69.6	78.4	18.2
DP-ASPP	68.5	78.2	28.4

This ablation study compares the impact of adopting depthwise separable convolutions (DP) and ASPP modules on model segmentation accuracy and real-time performance. The results of this analysis are presented in Table 2.

As illustrated in Table 2, replacing standard convolutions in the original ASPP module with separable dilated convolutions resulted in a 10.2 FPS improvement, despite slight decreases in mIoU and mPA metrics on the CamVid dataset compared to the original network. Compared to the significant increase in FPS, the decrease in mIoU and mPA is negligible. Therefore, it can be concluded that the incorporation of depthwise separable convolutions into the ASPP module enables the model to significantly enhance its real-time performance while maintaining segmentation accuracy.

DOI: 10.54254/2755-2721/2026.TJ29562

4.2.3. Impact of SA mechanisms

Table 3. Experimental results of SA

Method	mIoU(%)	mPA(%)	FPS
DeepLabV3+	69.6	78.4	18.2
SA + DeepLabV3+	73.1	82.2	17.5

This ablation study compared the performance of the model before and after introducing the SA module. The results of the experiment are shown in Table 3.

As shown in Table 3, adding the SA module after the ASPP module did not result in a substantial decline in the model's FPS metric. On the contrary, mIoU and mPA metrics demonstrated notable improvements, with mIoU increasing by 3.5% and mPA by 3.8%. This finding suggests that the SA mechanism enhances segmentation accuracy without imposing excessive computational overhead on the model.

4.2.4. Impact of ECA mechanisms

Table 4. Experimental results of ECA

Method	mIoU(%)	mPA(%)	FPS
DeepLabV3+	69.6	78.4	18.2
ECA + DeepLabV3+	73.4	82.3	17.6

The ablation study compares the impact of implementing the ECA module and removing this attention in terms of the model’s segmentation performance and feasibility. The results of the experiment are presented in Table 4.

Table 4 shows that introducing the ECA module after the decoder’s feature concatenation layer reduces the model’s FPS metric slightly but improves the mIoU and mPA metrics by 3.8% and 3.9%, respectively. This concludes that the ECA mechanism improves the accuracy of segmentation at the cost of a little extra time.

4.3. Comparative experiments on segmentation network models

Table 5. Experimental results of semantic segmentation models

Method	mIoU(%)	mPA(%)	FPS
SegNet	67.1	77.2	16.1
Unet	66.2	76.4	12.3
DeepLabV3+	69.4	78.4	18.2
The Proposed Model	69.7	78.6	30.4

This paper conducts comparative experiments between other semantic segmentation algorithms and the LAE-DeepLabV3+ model to validate its effectiveness in segmentation accuracy and real-time performance. These experiments were performed on the CamVid dataset, with selected

DOI: 10.54254/2755-2721/2026.TJ29562

comparison algorithms including SegNet [8], U-Net [4], and the original DeepLabV3+ [7]. The comparative results are shown in Table 5.

From experimental results, it can be observed that the mIoU for the proposed model is 0.3% higher and mPA is 0.2% higher than the best performing DeepLabV3+, where the improvements were not expected in the initial design. Regarding real-time performance, the proposed model continues to demonstrate superiority over all other models, attaining a 14.3-point increase in the FPS metric compared to DeepLabV3+, which had previously achieved the highest FPS metric. In summary, the proposed improvements enhance the network's real-time processing capabilities while effectively boosting segmentation accuracy.

4.4. Segmentation performance of the LAE-DeepLabV3+

In Figure 9, the original image is displayed on the left, the segmentation map generated by the LAE-DeepLabV3+ model is shown in the middle, and the segmentation map generated by the DeepLabV3+ model is presented on the right. Apparently, the DeepLabV3+ model exhibits certain limitations during image segmentation, particularly at the boundaries between objects of different categories, where edge confusion occurs. In contrast, the proposed model demonstrates significantly enhanced edge segmentation performance. Moreover, even in low-brightness images such as those presented in Figure 9, the LAE-DeepLabV3+ model achieves clear segmentation of distinct objects.



Figure 9. Visualization of different models

5. Conclusion

Compared to the original DeepLabV3+ model, the LAE-DeepLabV3+ model presents superior real-time performance on the CamVid dataset, despite only marginal improvements in mIoU and mPA metrics. Furthermore, as evidenced by specific detection maps, the proposed model demonstrates a notable proficiency in managing the delineation between disparate objects, particularly in the context of road segmentation. The model's high-precision segmentation of road edge positions makes this model as a promising candidate for autonomous driving applications.

However, there are some limitations in this study. First is using MobileNetV2 as a backbone network, which significantly improved the processing speed but had no significant improvement in the segmentation accuracy. This suggests the presence of significant opportunities for enhancement. To address this issue, future research could explore improvements in the convolution approach by introducing more advanced convolution strategies or combining the advantages of multiple network

DOI: 10.54254/2755-2721/2026.TJ29562

architectures to further enhance accuracy. Secondly, regarding images with high weather extremity, the model performs poorly, and the segmentation results are also unstable, thus unable to be limited to certain situations. A possible option would be to use more varied datasets or apply data augmenting procedures to improve the model's stability under different environmental circumstances. And these might help make the model even better at understanding tricky roads and doing a good job of cutting words into pieces on them.

References

[1] Babić D, Babić D, Fiolic M, et al. (2022). Road markings and signs in road safety. Encyclopedia. 2(4), 1738-1752.
[2] Mo Y, Wu Y, Yang X, et al. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. Neurocomputing. 493, 626-646.
[3] Long J, Shelhamer E, Darrell T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 3431-3440.
[4] 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. Cham: Springer international publishing, 234-241.
[5] Abderrahim N Y Q, Abderrahim S, Rida A. (2020). Road segmentation using u-net architecture. 2020 IEEE international conference of moroccan geomatics (Morgeo). IEEE, 1-4.
[6] Chen L C, Zhu Y, Papandreou G, et al. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV). 801-818.
[7] Badrinarayanan V, Kendall A, Cipolla R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE transactions on pattern analysis and machine intelligence. 39(12): 2481-2495.
[8] Yu C, Wang J, Peng C, et al. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. Proceedings of the European conference on computer vision (ECCV). 325-341.
[9] Fan R, Wang H, Cai P, et al. (2020). Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. European Conference on Computer Vision. Cham: Springer International Publishing. 340-356.
[10] Wang J, Sun K, Cheng T, et al. (2020). Deep high-resolution representation learning for visual recognition [J]. IEEE transactions on pattern analysis and machine intelligence. 43(10), 3349-3364.
[11] Wang H, Gao B, Liu S, et al. (2024). Semantic Segmentation of Road Landscape Based on Improved Deeplabv3+. 2024 5th International Conference on Artificial Intelligence and Computer Engineering (ICAICE). IEEE, 158-162.
[12] Brostow G J, Fauqueur J, Cipolla R. (2009). Semantic object classes in video: A high-definition ground truth database. Pattern recognition letters. 30(2): 88-97.
[13] Sandler M, Howard A, Zhu M, et al. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition. 4510-4520.
[14] Zhang Q L, Yang Y B. (2021). Sa-net: Shuffle attention for deep convolutional neural networks. ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2235-2239.
[15] Wang Q, Wu B, Zhu P, et al. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11534-11542.