Research on Semantic Segmentation Algorithm Based on Lightweight DeepLabV3+ Network

Jiayu Chen
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: c921204185@163.com

Zhongsheng Wang
State and Provincial Joint Engineering Lab. of Advanced Network, Monitoring and Control
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: wzhsh1681@163.com

***Abstract*—This paper presents an improved version of the DeepLabV3+ network to address issues such as large parameter count, difficulties in mobile deployment, limited receptive field, and insufficient utilization of low-level semantic information in existing deep learning semantic segmentation networks. The main enhancement approach is as follows: we utilize the lightweight MobileNetV2 as the backbone feature extraction network, while an improved multi-scale atrous convolution module (AS-ASPP) and convolutional block attention mechanism (CBAM) are introduced. Tests conducted on the PASCAL VOC 2012 dataset demonstrate that the optimized model retains merely around one-tenth the parameters of the original network, while attaining superior segmentation precision and computational effectiveness. Specifically, it reaches a mIoU of 73.21% and a Precision of 80.56%, with the training time reduced by approximately 50% and the inference speed significantly improved.**

Keywords-Semantic Segmentation; DeepLabV3+; Lightweight Network; Attention Mechanism

I. INTRODUCTION

As a cornerstone of computer vision, semantic segmentation enables pixel-level comprehension and scene analysis, with broad applications across fields such as medical diagnosis, remote sensing, and autonomous systems. Specifically, in medical image analysis, accurate segmentation of MRI and CT scans aids physicians in lesion detection, organ boundary delineation, and treatment strategy formulation[1]. In remote sensing, semantic segmentation enables fine-grained recognition and classification of urban infrastructure, vegetation, and road networks, supporting applications such as land use assessment, environmental monitoring, and smart city development. For autonomous driving and intelligent surveillance, it allows the system to interpret complex visual scenes by distinguishing roads, vehicles, pedestrians, and obstacles, thus ensuring safety and decision-making accuracy.

As deep learning advances steadily, convolutional neural networks (CNNs) have demonstrated outstanding ability in hierarchical feature extraction, laying a solid groundwork for semantic segmentation. Among various CNN-based models, the DeepLab series has achieved notable success by introducing atrous (dilated) convolutions and atrous spatial pyramid pooling (ASPP) modules, effectively capturing multi-scale contextual information. DeepLabV3+ further enhances segmentation precision through an encoder–decoder structure, which refines boundary and small-object details. Despite its excellent performance, the traditional DeepLabV3+ architecture commonly employs heavyweight backbones such as Xception or ResNet, leading to excessive computational cost and memory consumption, thereby restricting its deployment in real-time or resource-constrained environments.

To address these limitations, this study proposes a lightweight enhanced DeepLabV3+ variant. It swaps out the original backbone for a

more efficient feature extractor while incorporating an enhanced multi-scale feature fusion module, and embeds an attention mechanism to optimize spatial and channel-level feature responses. These improvements collectively reduce the model's parameters and floating-point operations while maintaining strong segmentation accuracy, Providing a practical and efficient solution for semantic segmentation applications in real-world scenarios.

II. RELATEDWORK

A. Convolutional Neural Network

With the rapid advancement of deep learning, Convolutional Neural Networks (CNNs) have become a core technology in semantic segmentation due to their exceptional capability for hierarchical feature representation and adaptability to complex visual patterns. The pioneering work by Long et al. The Fully Convolutional Network (FCN) [2] was introduced, achieving end-to-end pixel-level prediction for the first time and thus becoming a landmark in the evolution of semantic segmentation. Building upon this foundation, Ronneberger et al. developed the U-Net model, featuring a symmetric encoder–decoder architecture derived from FCN. By introducing skip connections to bridge low-level and high-level feature maps, U-Net significantly enhanced segmentation precision, particularly in medical imaging applications where fine-grained boundary delineation is essential.Further progress was made with the SegNet model proposed by Badrinarayana et al. [3], which employs a deep convolutional encoder to extract hierarchical features and adopts transposed convolution layers in the decoder for upsampling, ensuring that the output segmentation map matches the original image dimensions [4]. Meanwhile, the DeepLab series, proposed by Chen et al. [5], introduced several crucial innovations. DeepLabV2 and DeepLabV3 integrated dilated (atrous) convolutions alongside the Atrous Spatial Pyramid Pooling (ASPP) module, enabling effective capture of multi-scale contextual information, thus improving segmentation robustness for targets of varying sizes [6]. On this foundation, DeepLabV3+ further extended the framework by adding an encoder–decoder structure, enhancing boundary refinement and the segmentation of small or intricate objects [7].

B. Feature Extraction Network

To advance the lightweight and practical deployment of semantic segmentation models, numerous studies have focused on designing efficient backbone networks that can achieve a favorable balance between accuracy and computational efficiency. Representative lightweight convolutional neural networks, such as MobileNet, ShuffleNet, and Xception, have been proposed to effectively reduce the number of parameters and floating-point operations without severely compromising model performance [8]. In parallel, a variety of optimization techniques—including parameter pruning, model quantization, and knowledge distillation—have been extensively explored to further compress model size and accelerate inference [9]. Despite these efforts, a key challenge persists in the field: how to preserve or even enhance segmentation precision while minimizing computational resource consumption, especially under real-time and embedded deployment constraints.

In these lightweight architectures, the unique inverted residual structure introduced by MobileNetV2 differs fundamentally from traditional ResNet designs. Traditional residual networks typically follow a standardized workflow of "downsampling feature extraction upsampling", while MobileNetV2 innovatively uses a reverse sequence of "channel extension depth separable convolution channel reduction". As shown in Figure 1, the input feature map is first expanded in dimension through 1 × 1 point by point convolution, significantly increasing the number of channels and providing a richer information foundation for subsequent feature extraction; Subsequently, a 3 × 3 depth separable convolution is applied, which separates spatial convolution from channel convolution to ensure effective capture of spatial features while reducing computational cost to about 1/9 of traditional convolution; Finally, another layer of 1x1 convolution is used to compress the channel dimension, and a linear activation function is specifically employed to avoid information loss caused by non-linear transformations such as

ReLU on low dimensional features. When the stride is set to 1 and the spatial size of the input and output feature maps is exactly the same, residual shortcut connections will be inserted between the input and output tensors. This key mechanism not only stabilizes the gradient propagation path of deep networks, but also effectively alleviates the common gradient vanishing or exploding problems in training, ultimately significantly improving the convergence speed of the network and the robustness of feature representation, making it more suitable for resource limited scenarios such as mobile devices.

Chen et al. first introduced the ASPP module in DeepLabV2, where multi-scale contextual information is effectively captured through dilated convolution, leading to a notable improvement in segmentation accuracy. DeepLabV3 refined this design by integrating larger, more diverse dilation rates, along with batch normalization and global average pooling layers.

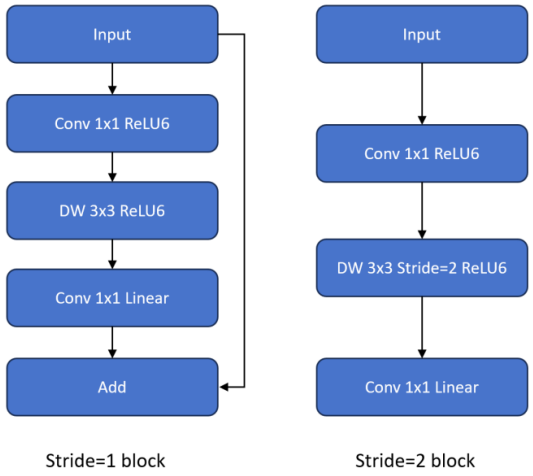

Figure 1. Inverted Residual Sructure Model

These enhancements boosted the representation capability of high-level semantic features, enabling the model to better handle objects with significant scale variations. On this foundation, DeepLabV3+ combined the ASPP module with an encoder–decoder architecture, which further improved the delineation of object boundaries and the segmentation performance of small or intricate targets.

Despite these advancements, the conventional ASPP module still faces several challenges in practical applications. The fixed configuration of dilation rates may lead to sparse sampling, resulting in incomplete capture of local texture details. Moreover, when large dilation rates are applied to high-level feature maps, redundant or irrelevant information may be introduced, thereby degrading the segmentation quality of small-scale objects and cluttered scenes. Consequently, achieving an optimal balance between global semantic understanding and local feature preservation has become a critical direction for the continued improvement of the ASPP module in lightweight and high-performance semantic segmentation networks.

III. Technical Model

This paper proposes an improved DeepLabV3+-based semantic segmentation model. It employs the lightweight MobileNetV2 as the backbone feature extractor and integrates an attention mechanism post-Atrous Spatial Pyramid Pooling (ASPP) module, balancing computational efficiency with feature representation capability. The network retains an encoder-decoder framework (see Figure 2).

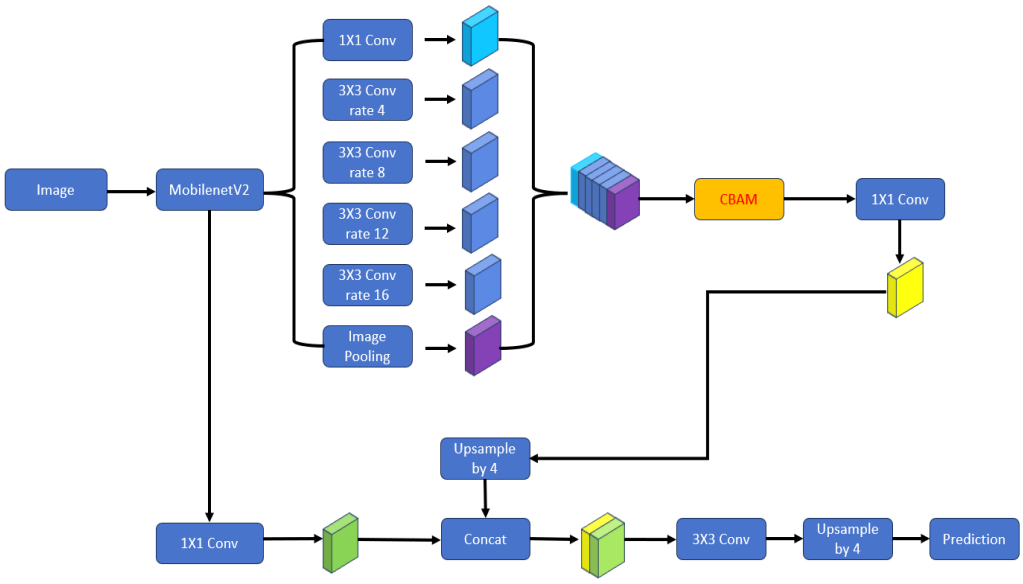

Figure 2. Improved DeepLabV3+ Model

During the encoder phase, the input image is first processed by MobileNetV2 to extract hierarchical multi-layer feature representations. MobileNetV2 leverages depthwise separable convolutions to reduce parameters and computational costs substantially, while preserving robust feature extraction capabilities—enabling the encoder to efficiently capture both low-level texture details and high-level semantic information. The generated high-level feature maps are then fed into the improved ASPP module. Unlike the original DeepLabV3+ setup, this model adjusts the dilation rates of parallel atrous convolutions to 4, 8, 12, and 16. This refined design boosts the network's adaptability to targets of different scales: smaller rates focus on fine-grained local features, while larger ones capture wider contextual information. Their balanced combination effectively reduces local detail loss from overly large receptive fields, thus improving segmentation accuracy for small and complex targets.

As part of the attention enhancement process, a Convolutional Block Attention Module (CBAM) is added post-ASPP module to refine feature representations further. This CBAM incorporates two sequential submodules: Channel Attention and Spatial Attention. The Channel Attention mechanism adaptively assigns weights to different feature channels, enabling the network to prioritize those with richer semantic content. In contrast, the Spatial Attention mechanism redistributes weights across spatial locations, emphasizing critical target regions while suppressing background noise. The joint operation of these two attention mechanisms allows the model to capture more comprehensive contextual information from both channel and spatial dimensions without introducing significant computational overhead, maintaining the model's lightweight nature.

During the decoder phase, low-level shallow features are fused with the high-level semantic features produced by the ASPP+CBAM module. Before fusion, a 1×1 convolution is applied to the low-level features for dimensionality reduction, enabling them to match the channel dimensions of the high-level features. The resulting features are then concatenated with the high-level feature

maps after a four-fold upsampling. Subsequently, multiple 3×3 convolutional layers and progressive upsampling operations are applied to gradually restore spatial resolution, ultimately producing the final pixel-level semantic segmentation map.

Through the integration of an efficient backbone, an improved ASPP module, and an attention-enhanced fusion mechanism, the proposed model achieves superior segmentation accuracy and feature discrimination under constrained computational budgets. Experimental results demonstrate that the model exhibits enhanced robustness in complex visual scenes and improved performance in identifying small-scale objects, confirming its effectiveness and practicality in real-world semantic segmentation tasks.

A. Feature Extraction Network

In semantic segmentation tasks, the backbone network is the core component of the entire architecture. It is usually a deep convolutional neural network that gradually reduces the spatial resolution of input images via multi-layer convolution and pooling operations, while enhancing feature abstraction. This hierarchical feature extraction allows the network to capture rich contextual and high-level semantic representations, which provide essential support for accurate pixel-level classification in subsequent segmentation stages. By leveraging these high-level feature maps, the segmentation model can effectively distinguish and assign semantic labels to different regions within an image, achieving precise and fine-grained scene understanding.

Table 1 presents a comparative analysis of several widely used backbone networks—such as VGG, ResNet, Xception, and MobileNet—evaluated on the ImageNet dataset. The comparison includes their respective floating-point operations (FLOPs), parameter counts, and segmentation accuracy. These metrics highlight the trade-off between computational cost and feature representation capacity. Traditional deep architectures such as VGG and ResNet generally offer high representational power but require substantial computational resources, making them less suitable for real-time or embedded applications. In contrast, lightweight models like MobileNet achieve a favorable balance between efficiency and accuracy, making them promising candidates for semantic segmentation tasks that demand both precision and computational efficiency.

TABLE I. PERFORMANCE COMPARISON OF COMMONLY USED BACKBONE NETWORDS

Model	FLOPs(G)	Parameters(M)	Accuracy%
VGG-16	15.7	138.4	71.3
ResNet18	1.8	18.6	69.8
ResNet34	3.6	21.8	71.5
ResNet50	3.8	25.6	74.9
Xception	**31.1**	22.9	**79.0**
MobileNetV1	0.56	4.2	69.0
MobileNetV2	**0.32**	**3.5**	71.3
MobileNetV3	0.31	5.4	73.3

From Table 1, it can be seen that the parameter count of MobileNetV2 network is only 3.5M, and its overall structure is lightweight and concise, but its segmentation performance is not inferior to other complex networks. In contrast, although the Xception network has a segmentation accuracy about 7.7% higher than MobileNetV2, its parameter count has increased by 19.4M, and FLOPs are about 97 times that of MobileNetV2. This indicates that although Xception has certain advantages in accuracy, its huge computational

overhead and complex structure limit its application in resource constrained scenarios. Hence, in order to boost the computational efficiency of the DeepLabV3+ model, this paper adopts MobileNetV2— which features lightweight architecture and high accuracy— as the backbone network for the revised model.

B. AS-ASPP Module

Within the original DeepLabV3+ model, the ASPP module employs dilated convolutions at rates 6, 12, and 18 for capturing multi-scale contextual details. However, these large dilation rates cause sparse sampling and loss of fine details, which reduces segmentation accuracy for small or thin objects [10]. Although a wide receptive field helps extract global semantics, it weakens sensitivity to local features and may produce redundant information in high-level maps.

To address these issues, this paper proposes an Adaptive-Scale ASPP (AS-ASPP) module, as shown in Figure 3. The module adopts four parallel dilated convolutions with rates of 4, 8, 12, and 16 to achieve balanced multi-scale feature extraction. Smaller dilation rates preserve fine texture and boundary details, while larger ones capture global context. By integrating features from different receptive fields, AS-ASPP enriches semantic representation and enhances segmentation performance for small objects and complex scenes.

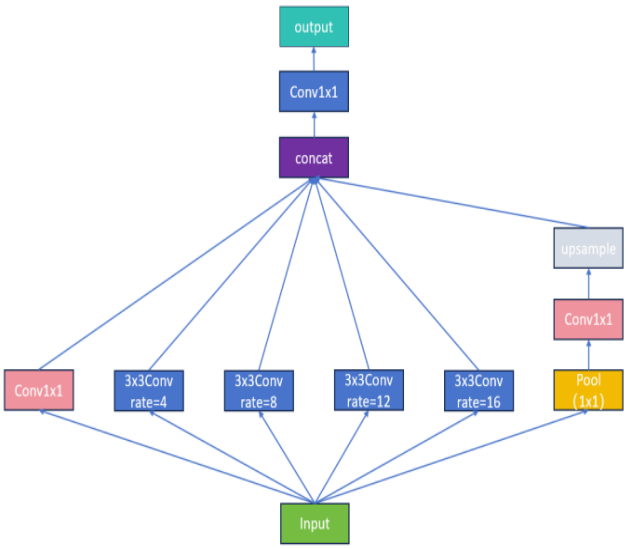

Figure 3. AS-ASPP Model

C. Convolutional Attention Mechanism module.

In semantic segmentation, although high-level features contain rich semantic information, they often exhibit uneven distribution and weak responses to fine details. To enhance feature representation, this paper introduces the Convolutional Block Attention Module (CBAM) into the improved DeepLabV3+ framework, enabling adaptive enhancement along channel and spatial dimensions to guide the network toward target-relevant features.

CBAM is a lightweight and efficient attention mechanism that strengthens feature representation through joint channel and spatial attention while maintaining low computational cost [11]. As shown in Figure 4, it consists of two submodules: a channel attention module that adaptively adjusts the importance of different feature channels, and a spatial attention module that highlights key target regions and suppresses background noise. The combination of both enables the model to extract more discriminative and context-aware features, thereby improving segmentation performance without significantly increasing model complexity.

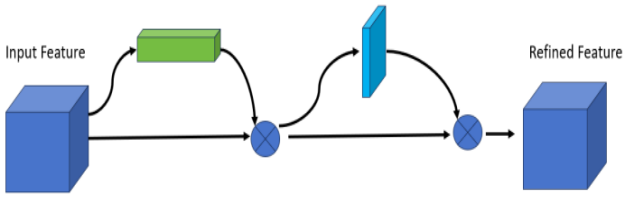

Figure 4. CBAM Model

As illustrated in Figure 5, the channel attention module initially uses Global Average Pooling (GAP) and Global Maximum Pooling (GMP) to extract channel-level statistical features. Their outputs pass through two shared fully connected layers to generate attention weights, which are then added and normalized via a Sigmoid function. These weights rescale each channel of the feature map, enhancing important features and suppressing redundant ones, thereby improving feature representation.

The spatial attention module, depicted in Figure 6, aims to refine feature responses across spatial positions.

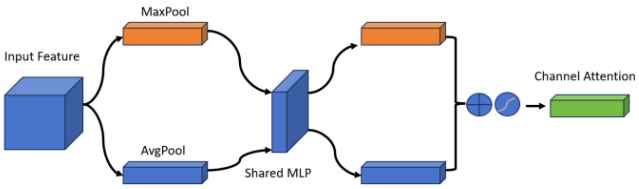

Figure 5. Channel Attention Model

Along the channel axis, max and average pooling operations are performed first, resulting in two spatial feature maps that capture complementary information. These maps are concatenated and passed through a convolution layer and Sigmoid activation to generate a spatial attention map that assigns importance scores to each location. By applying these weights to the original feature map, the module enhances critical regions while reducing the influence of background noise, achieving improved spatial feature representation.

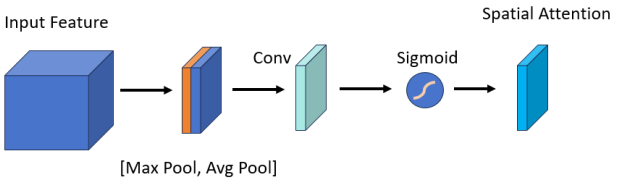

Figure 6. Spatial Attention Model

D. Loss function.

In semantic segmentation, the Cross Entropy Loss Function adjusts weights for hard-to-classify samples and prioritizes them—especially for complex images, blurry edges, or irregularly shaped targets—helping improve the model's accuracy on details and edges to boost overall segmentation accuracy and robustness. To address related issues, this paper replaces the benchmark model's Cross Entropy Loss with FocalLoss [12].

Equation (1) denotes the binary cross entropy loss function, where p is the predicted probability of a positive sample and y is the true label. For simplicity, p is used to represent the probability that a sample belongs to its true category (see Equation (2)). From Equations (1) and (2), the expression of the Cross Entropy (CE) function is derived as Equation (3). Based on Equation (3), the expressions of the FocalLoss function are obtained as Equations (4) and (5).

$$CE(p,y)=\begin{cases}-\log(p), y=1\\ -\log(1-p), otherwise\end{cases} \quad (1)$$

$$p:\begin{cases}p, y=1\\ 1-p, otherwise\end{cases} \quad (2)$$

$$CE(p,y)=CE(p_t)=-\log(p_t) \quad (3)$$

$$\alpha : \begin{cases} a, y = 1 \\ 1 - a, otherwise \end{cases} \tag{4}$$

$$FL(p_t) = -a_t(1-p_t)^{\gamma}\log(p_t) \tag{5}$$

Within the proposed framework, parameter α is employed to mitigate the class imbalance issue between positive and negative samples, where its specific value directly regulates the loss contribution of samples with a label of 1. Meanwhile, adjusting parameter γ allows for the implementation of differentiated loss assignment: samples that are easier to classify (usually associated with larger predicted values) will exhibit a substantially reduced contribution to the overall loss.

IV. Experimental Results And Analysis

A. Experimental Dataset

To confirm the efficacy of the algorithm presented in this study, semantic segmentation experiments were performed on the publicly available PASCAL VOC2012 dataset. As a widely accepted benchmark in the semantic segmentation domain, the VOC2012 dataset comprises 20 foreground classes and 1 background class, including diverse scenarios like animals, plants, and urban scenes, transportation vehicles, indoor items, and human figures. It contains 2913 training images, 1464 validation images, and 1456 test images in total, all of which are annotated at the pixel level. This comprehensive annotation enables the dataset to fully reflect the performance of semantic segmentation models across multiple categories and scenarios. In this research, both the training and validation processes adopted the standard data partitioning provided by VOC2012, and the segmentation performance of the improved model was compared with that of the original DeepLabV3+ model through a well-designed experimental setup.

B. Experimental Configuration

The operating system used in this article is Windows 11. In the experiment, the running environment was PyCharm, using Intel (R) Core (TM) i7-12700H processor and Nvidia Ge Force RTX 3060 graphics cards. The model was trained on Python 3.10, Python 2.5.1, and Cuda12.6, as shown in Table 1.

TABLE II. EXPERIMENTAL CONFIGURATION TABLE

Name	Related Configurations
Operating System	Windows11
Memory	16GB
GPU	Nvidia Ge-Force RTX 3060
CPU	Core（TM）i7-12700H
Operating Environment	PyCharm
CUDA Version	Cuda12.6

C. Evaluation

Five evaluation metrics are adopted in this study to assess the proposed optimization algorithm's performance: mean Intersection over Union (mIoU), mean Pixel Accuracy (mPA), precision, model parameter count (Params), and floating-point operations (FLOPs). Notably, mIoU evaluates the algorithm's overall performance on the dataset by determining the average ratio of intersection to union between the predicted segmentation outcomes and the ground truth labels. Its calculation formula is presented as follows:

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

Herein, $k+1$ he total number of categories including the background class. i represents the ground truth label (i.e., the actual category that the sample belongs to), and j stands for the model's predicted label (i.e., the category that the model determines the sample belongs to). P_{ij} (a commonly used notation in this scenario to clarify the correspondence of probabilities) indicates the probability that the model predicts a sample with the ground truth label i as label j

Mean Pixel Accuracy (mPA) is defined as the average ratio between correctly classified pixels and the total number of pixels over the course of training. Its mathematical formulation is provided as follows:

$$mPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{\mathrm{ij}}} \tag{7}$$

Precision denotes the ratio of samples that the model accurately classifies into positive categories relative to all samples it predicts as positive. The specific formula is:

$$P = \frac{TP}{TP+FP} \tag{8}$$

Among them, TP is a positive example for predicting the true category, and FP is a positive example for predicting the true category. The number of model parameters (Params) refers to the memory required by the model. The term "floating-point computation" describes the number of floating-point operations carried out per second, and it directly indicates the computational load required to run a system or process.

D. Experimental Results

In order to verify that the optimized MobilNetV2 backbone network used can effectively improve detection speed while reducing the number of parameters, a comparative experiment was conducted with the Xception backbone network, and the results are shown in Table 3.

TABLE III. COMPARISON OF BACKBONE NETWORK PERFORMANCE

Backbone	Precision/%	mIoU/%	Params/M	FLOPs(G)
Xception	78.31	73.87	42.0	167.00
MobileNetV2	77.46	72.31	3.5	53.02

Table 3 shows that after adopting the lightweight MobileNetV2 as the backbone, the model's segmentation accuracy experienced a slight decline, but the parameter count was greatly reduced and the processing speed increased markedly. This demonstrates the efficiency and lightweight advantage of MobileNetV2, providing a solid basis for subsequent improvements using multi-scale feature enhancement and attention mechanisms.

To further evaluate the effectiveness of the proposed modules and their combined impact, four ablation experiments were conducted, and the results are summarized in Table 4. The experimental data indicate that both the AS-ASPP and CBAM modules independently enhance the model's performance by addressing different technical limitations: the AS-ASPP module strengthens the extraction of multi-scale contextual features, enabling the model to better

capture objects of varying sizes in complex road scenes, while the CBAM module improves the network's ability to focus on key semantic regions by adaptively adjusting attention along the channel and spatial dimensions, resulting in a Precision of 79.65% and an mAP of 83.89%. When the two modules are integrated, they complement each other and achieve the best overall performance, with mIoU reaching 73.21%, Precision 80.56%, and mAP 84.11%. These results confirm that combining multi-scale feature enhancement with attention mechanisms can significantly boost the model's capability to learn and identify road features, effectively improving segmentation precision and robustness while maintaining computational efficiency, thus laying a solid foundation for subsequent real-world applications.

TABLE IV. RESULTS OF ABLATION EXPERIMENT TABLE

Group	MobileNetV2	AS-ASPP	CBAM	mIoU%	Precision/%	mAP/%
①	√			72.31	77.46	82.67
②	√	√		72.58	78.17	83.49
③	√		√	72.86	79.65	83.89
④	√	√	√	73.21	80.56	84.11

As shown in Table 5, the optimized model proposed in this study not only achieves a significant improvement in segmentation accuracy for road extraction tasks but also greatly reduces model complexity compared with the original DeepLabV3+ network. Specifically, the total number of parameters is reduced to approximately one tenth of the original model, accompanied by a substantial decrease in FLOPs, which effectively lowers computational costs. Owing to this lightweight design, the inference speed of the model is notably enhanced, allowing more images to be processed per second and reducing overall training time by nearly half. These results indicate that the optimized model achieves a favorable balance between efficiency and precision, maintaining high segmentation accuracy while significantly improving computational performance. The integration of a lightweight backbone, multi-scale feature enhancement, and attention mechanisms enables the network to extract more discriminative road features under limited computational resources, making it highly suitable for real-world scenarios that demand both accuracy and speed. Overall, the proposed method provides a practical and efficient solution for semantic segmentation in complex road scenes, offering strong potential for deployment in remote sensing applications and intelligent transportation systems.

TABLE V. MODEL PERFORMANCE COMPARISON

Model	Precision/ %	MIoU/ %	Params/ M	FLOPs(G)	Time
DeepLabV 3+	79.39	71.12	42	167.0	18h53m in
Ours	80.56	73.21	3.8	32.4	9h6min

As illustrated in Figure 7, the visual segmentation results of the model proposed in this study, compared with those of the DeepLabV3+ benchmark on the PASCAL VOC 2012 dataset, clearly demonstrate its superior ability to capture fine details and maintain structural integrity across various scene types. The PASCAL VOC 2012 dataset includes a diverse range of objects and environments, providing a rigorous and comprehensive test for evaluating segmentation robustness and generalization. From the visual

comparison, it can be observed that the proposed model performs markedly better in delineating slender targets such as aircraft tail fins, flying birds, and other elongated objects, which are often prone to fragmentation, boundary discontinuity, or misclassification in traditional models. The improved segmentation continuity is mainly attributed to the stripe pyramid pooling structure, which enhances the model's capability to capture fine-grained spatial features and preserves the geometric coherence of narrow structures. In scenes involving human–object interactions, such as people holding tools or carrying bags, the model can accurately distinguish between overlapping or adjacent targets, successfully preserving the separability of objects that are easily confused in DeepLabV3+. For more complex objects like motorcycles and bicycles that exhibit intricate structures and partial occlusions, the optimized model effectively mitigates adhesion between objects and the background, producing clearer segmentation boundaries and better visual integrity, particularly around components such as wheels and handlebars. These improvements stem from the combined effect of multi-scale feature extraction and attention-guided refinement, which enable the network to selectively emphasize salient semantic regions while suppressing irrelevant background noise. Furthermore, the boundary information enhancement mechanism strengthens edge perception, allowing the model to sharpen contour details and maintain precise object boundaries. Overall, by integrating multi-scale context modeling, directional spatial pooling, and attention-based feature optimization, the proposed model achieves higher segmentation accuracy, better structural consistency, and improved robustness compared with DeepLabV3+, providing a more reliable and visually coherent solution for semantic segmentation tasks in complex road and natural scenes.

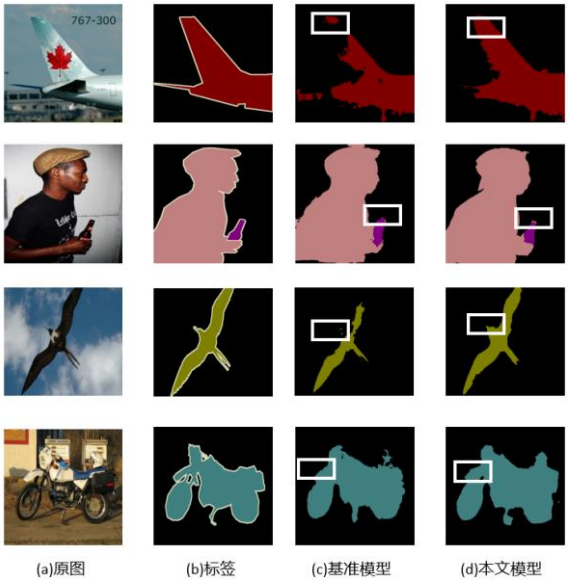

Figure 7. Visualization Comparison of Results

V. Conclusions

To address the issues of excessive parameters and high computational complexity in traditional DeepLabV3+ models, this study proposes a lightweight improved model. This model integrates three key components: the MobileNetV2 backbone network, the AS-ASPP module, and the CBAM attention mechanism. Experimental results demonstrate that the proposed model maintains high segmentation accuracy while significantly reducing both parameter count and FLOPs. Ablation experiments further validate the independent effects of each improved module as well as their synergistic enhancement. Compared with the original DeepLabV3+ model, the improved model achieves higher scores in both mIoU and Precision metrics, while also demonstrating significant advantages in training efficiency and inference speed. Overall, the lightweight model proposed in this study strikes a favorable balance between model lightweighting and segmentation performance, providing a practical and feasible

optimization solution for semantic segmentation tasks and their real-world applications.

REFERENCES

[1] Ke Yin, Chen Dan. Image Semantic Segmentation Method Based on Improved DeepLabv3+ Network [J]. Intelligent Computer and Applications, 2025, 15(04): 17-24.

[2] Ren Ziyu, You Xindong, Teng Shangzhi, et al. Image Boundary Restoration Semantic Segmentation Based on DeepLabv3+[J]. Journal of Beijing Information Science and Technology (Natural Science Edition), 2024, 39(06): 17-24.

[3] Yuan Manman, Lu Hao. Semantic Segmentation Algorithm Integrating Lightweight and Attention Mechanisms [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(01): 57-63. DOI: 10.16055/j.issn.1672-058X.2025.0001.008.

[4] Wang Na. Research on Image Semantic Segmentation Model Based on Improved DeepLabv3+[D]. Lanzhou University of Technology, 2023.

[5] Li Mingyi. Research on Image Semantic Segmentation Based on Improved DeepLabV3+ Network Structure [D]. Dalian Jiaotong University, 2022.

[6] Liu Yongfeng. Research on Image Semantic Segmentation Method for Nighttime Road Obstacles with Attention Mechanism Integration [D]. Central South University, 2022.

[7] Lu Jianbo, Peng Jungui, Huo Leigang, et al. A Lightweight DeepLabV3+ Remote Sensing Image Segmentation Method [J]. Guangxi Science, 2025, 32(02): 374-385.

[8] Zhou Huaping, Deng Bin. DeepLabv3+ Lightweight Image Segmentation Algorithm Integrating Multi-level Features [J]. Computer Engineering and Applications, 2024, 60(16): 269-275

[9] Yao Yan, Hu Likun, Guo Jun. Lightweight Semantic Segmentation Algorithm Based on Improved DeepLabv3+ Network [J]. Advances in Laser and Optoelectronics, 2022, 59(04): 200-207.

[10] Zhang Ying. Research and Implementation of Image Semantic Segmentation Algorithm Based on Improved DeepLabV3+[D]. Heilongjiang University, 2025.

[11] Xin Kai. Research and Application of Image Semantic Segmentation Model Based on DeepLabV3+ [D]. Zhejiang University of Technology, 2024.

[12] Kong Xiangming. Research and Application Based on Improved DeepLabv3+ Image Semantic Segmentation Model [D]. Wuhan University of Light Industry, 2024.