ISSN 2522-3488 Vol. 10, Issue 1: 57-63, DOI: 10.25236/IJNDES.2026.100108

Design and Implementation of a Lightweight Image Semantic Segmentation Model Combining Attention Mechanism

Xia Jiayue[1,a], Long Yanbin[1,b,*]

[1]*University of Science and Technology Liaoning, Anshan, China*
[a]*2677399204@qq.com,* [b]*1034182681@qq.com*
**Corresponding author*

Abstract: *Image semantic segmentation is a key technology in computer vision, widely used in scenarios such as autonomous driving, intelligent security, and medical image analysis. However, existing semantic segmentation models often face a contradiction between large number of parameters, high computational complexity, and insufficient segmentation accuracy. To address this issue, this paper designs and implements a lightweight image semantic segmentation model incorporating an attention mechanism. This model is based on an improved DeepLabV3+ framework, using MobileNetV2 as a lightweight backbone network. A Hybrid Domain Attention Module (CBAM) is introduced in the encoding stage to enhance the representation ability of important features. Simultaneously, a Channel Hollow Spatial Pyramid Pooling Module (C-ASPP) is designed to weight each channel while extracting multi-scale contextual information. In the decoding stage, a dense neighborhood prediction module is introduced to fuse high- and low-level features, refining the segmentation boundaries. Experimental results on the PASCAL VOC 2012 and Cityscapes datasets show that the proposed model achieves an average intersection-over-union (mIoU) of 74.83% and 75.21% respectively with only 24.3 MB of parameters, representing improvements of 3.21% and 2.94% over the baseline model, achieving a good balance between segmentation accuracy and computational efficiency.*

Keywords: *Semantic Segmentation; Lightweight Network; Attention Mechanism; DeepLabV3+; MobileNetV2*

1. Introduction

Image semantic segmentation aims to assign predefined semantic category labels to each pixel in an image, and is one of the core tasks of scene understanding. With the rapid development of deep learning technology, deep neural networks, represented by fully convolutional networks (FCNs), have made significant progress in the field of semantic segmentation. However, high-precision segmentation models are usually accompanied by a large number of parameters and computational overhead, making them difficult to deploy in resource-constrained scenarios such as mobile devices and embedded systems [1].

In recent years, researchers have explored solutions from two directions: one is to design lightweight network structures, such as MobileNet and ShuffleNet, to reduce computational complexity through depthwise separable convolutions; the other is to introduce attention mechanisms to enable the network to focus on important feature regions and improve segmentation accuracy. However, how to effectively combine the two to further improve segmentation performance while maintaining model lightweightness remains a hot topic and challenge in current research [2].

This paper proposes a lightweight semantic segmentation model that combines an attention mechanism to address the above problems. The main contributions include: (1) replacing the Xception backbone in DeepLabV3+ with MobileNetV2, which significantly reduces the number of model parameters; (2) designing a channel-diffused spatial pyramid pooling module (C-ASPP), introducing channel attention on the basis of dilated convolution, and strengthening important channels in multi-scale features[3-4]; (3) introducing a hybrid domain attention module (CBAM) in the encoding stage to optimize feature representation from both channel and spatial dimensions; and (4) using a dense neighborhood prediction module to fuse high and low layer features and improve segmentation

ISSN 2522-3488 Vol. 10, Issue 1: 57-63, DOI: 10.25236/IJNDES.2026.100108

boundary details. Experimental results show that the proposed model achieves good performance on multiple public datasets[5].

2. Related Work

2.1. Overview of the Development of Semantic Segmentation Networks

The evolution of semantic segmentation networks has gone through the development process from fully convolutional networks to encoder-decoder structures, and then to multi-scale feature fusion. The FCN proposed by Long et al. first achieved end-to-end pixel-level classification. The U-Net proposed by Ronneberger et al. adopts a symmetrical encoder-decoder structure and fuses high and low layer features through skip connections, achieving good results in medical image segmentation. The DeepLab series introduces dilated convolution to expand the receptive field and proposes the Spatial Pyramid Pooling Module (ASPP) to extract multi-scale contextual information[6-7].

2.2. Lightweight Semantic Segmentation Model

To meet the real-time requirements, researchers have proposed a variety of lightweight semantic segmentation models. MobileNetV2 proposed by Sandler et al. adopts an inverse residual structure and a linear bottleneck layer, which significantly reduces the computational cost of the model. BiSeNet proposed by Yu et al. adopts a bilateral network structure to process spatial details and semantic context separately. ICNet proposed by Zhao et al. achieves fast segmentation through image concatenation. These methods have laid the foundation for lightweight semantic segmentation, but there is still room for improvement in segmentation accuracy[8].

2.3. Application of Attention Mechanism in Semantic Segmentation

Attention mechanisms enable networks to adaptively focus on important features. SENet, proposed by Hu et al., learns channel weights through a compression-activation module. CBAM, proposed by Woo et al., combines channel attention and spatial attention, achieving performance improvements in multiple visual tasks. DANet, proposed by Fu et al., uses a dual attention mechanism to capture long-range dependencies in both spatial and channel dimensions. Introducing attention mechanisms into lightweight networks is expected to achieve significant accuracy improvements with relatively small computational overhead[9].

3. Model Design

3.1. Overall Network Architecture

The overall architecture of the lightweight semantic segmentation model proposed in this paper is shown in Figure 1. The model employs an encoder-decoder structure. The encoding part includes a lightweight backbone network, MobileNetV2, and a channel-dilated spatial pyramid pooling (C-ASPP) module. The decoding part uses a dense neighborhood prediction module to fuse high- and low-level features and progressively restore resolution[10].

In the encoding stage, the input image is first processed by MobileNetV2 to extract multi-level features. Among them, low-level features (C1, resolution 1/4) contain rich spatial detail information, while high-level features (C5, resolution 1/16) have stronger semantic abstraction capabilities. The high-level features are input into the C-ASPP module, which captures multi-scale contextual information through dilated convolutions with different dilation rates and uses a channel attention mechanism to weight the features of each channel. Subsequently, the output of C-ASPP is fused with the low-level features enhanced by the CBAM module and input into the decoder. The decoder uses a dense neighborhood prediction strategy to progressively upsample and restore resolution, finally outputting the segmentation result.

ISSN 2522-3488 Vol. 10, Issue 1: 57-63, DOI: 10.25236/IJNDES.2026.100108

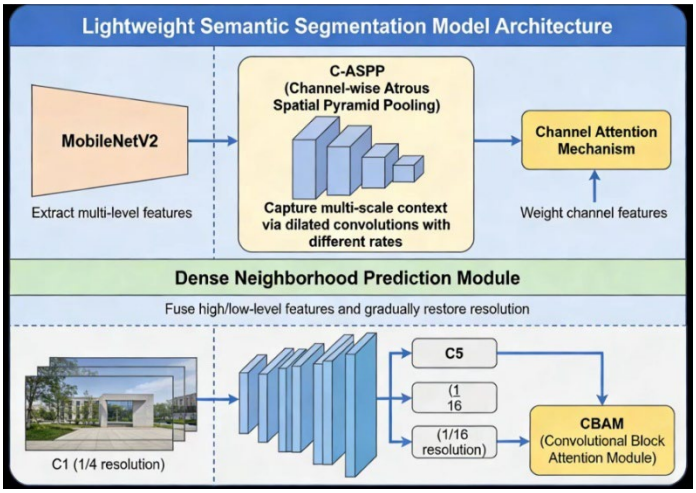

Figure 1: Overall Model Architecture Diagram

3.2. Lightweight Backbone Network Selection

This paper chooses MobileNetV2 as the feature extraction backbone mainly because of its efficient inverse residual structure and linear bottleneck layer design. The basic building block of MobileNetV2 is the inverse residual bottleneck module, which consists of three parts: first, expanding the number of channels through 1×1 convolutions; then, extracting features using 3×3 depthwise convolutions; and finally, compressing the number of channels through 1×1 convolutions. This structure significantly reduces computational cost while maintaining representational power.

Table 1 compares the performance of commonly used lightweight backbone networks. MobileNetV2 achieves a Top-1 accuracy of 72.0% on the ImageNet classification task with 3.4M parameters, achieving a good balance between computational efficiency and representational power.

Table 1: Performance Comparison of Different Lightweight Backbone Networks

Network Model	Number of parameters (M)	Computational cost (GFLOPs)	Top-1 accuracy (%)
MobileNetV1	4.2	0.57	70.6
MobileNetV2	3.4	0.30	72.0
MobileNetV3	2.5	0.22	73.3
ShuffleNetV2	2.3	0.28	71.5

3.3. Channel Hollow Spatial Pyramid Pooling

The ASPP module in the original DeepLabV3+ uses parallel dilated convolutions to extract multi-scale information, but it treats the importance of each channel feature equally, making it difficult to highlight key channels. This paper proposes the Channel Dilated Spatial Pyramid Pooling Module (C-ASPP), which introduces a channel attention mechanism on the basis of ASPP. The structure is shown in Figure 2.

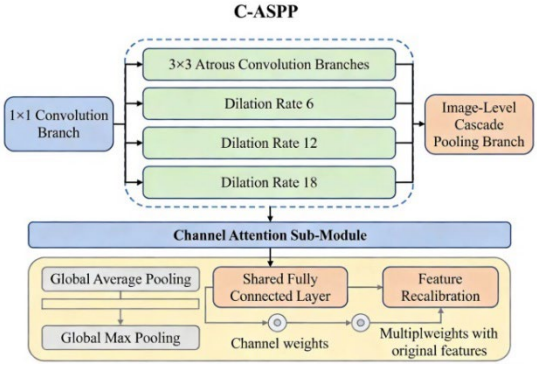

Figure 2: C-ASPP module structure diagram

ISSN 2522-3488 Vol. 10, Issue 1: 57-63, DOI: 10.25236/IJNDES.2026.100108

C-ASPP contains four parallel branches: a 1×1 convolution branch, three 3×3 dilated convolution branches (dilation rates of 6, 12, and 18, respectively), and an image cascaded pooling branch. The output features of each branch are concatenated and then input into the channel attention submodule. Channel attention aggregates spatial information through global average pooling and global max pooling, generates channel weights through a shared fully connected layer, and multiplies them with the original features to achieve feature recalibration. This design enables the model to adaptively emphasize channels containing important semantic information and suppress the responses of irrelevant channels.

3.4. Hybrid Domain Attention Module

To further enhance the feature representation capability, this paper introduces the CBAM module in the encoding stage, which acts on low-level features and high-level features respectively. CBAM contains two submodules, channel attention and spatial attention, which are connected in a serial manner.

The channel attention submodule is similar to that in Section 3.3, aggregating spatial dimension information through average pooling and max pooling, and generating channel weights through a multilayer perceptron. The spatial attention submodule performs average pooling and max pooling along the channel dimension, and generates spatial weights through a 7×7 convolution after concatenation. Through this "channel-space" sequential attention mechanism, the model can simultaneously focus on "what features are important" and "where features are important".

3.5. Dense Neighborhood Prediction Decoder

The goal of the decoder design is to effectively fuse low-level detailed information and high-level semantic information to restore fine segmentation boundaries. This paper adopts a dense neighborhood prediction strategy, refining the segmentation results through multilayer feature fusion and progressive upsampling.

Specifically, the high-level features output by C-ASPP are upsampled by 2 times and then concatenated with the low-level features enhanced by CBAM. The concatenated features are further fused through two 3×3 convolutional layers and then upsampled to the original image resolution. During training, an auxiliary supervision strategy is introduced to calculate the loss on the intermediate layer output of the decoder, accelerating network convergence and improving segmentation quality.

4. Experiments and Analysis

4.1. Experimental Setup

Datasets: Experiments were conducted using the PASCAL VOC 2012 augmented dataset and the Cityscapes dataset. PASCAL VOC 2012 contains 20 object categories and 1 background category, with a total of 10582 training images and 1449 validation images. Cityscapes focuses on urban street scene understanding, containing 19 categories, with a total of 2975 training images, 500 validation images, and 1525 test images, with an image resolution of 1024×2048.

Evaluation Metrics: Mean Intersection over Union (mIoU) and mean pixel accuracy (mPA) were used as the main evaluation metrics. Model complexity was measured by the number of parameters and computational cost (GFLOPs).

Implementation Details: The experiments were implemented using the PyTorch framework and trained on an NVIDIA RTX 3090 GPU. A stochastic gradient descent (SGD) optimizer was used with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 5e-4. The batch size was set to 16, and the training iterations were 80 epochs. A multinomial learning rate decay strategy was used, with power set to 0.9. Data augmentation included random flipping, random scaling (0.5-2.0), and random pruning.

4.2. Ablation Experiment

To verify the effectiveness of each module, ablation experiments were conducted on the PASCAL VOC 2012 dataset. The results are shown in Table 2.

ISSN 2522-3488 Vol. 10, Issue 1: 57-63, DOI: 10.25236/IJNDES.2026.100108

Table 2: Ablation Experiment Results for Each Module (mIoU/%)

Experimental Setup	Number of Parameters(MB)	mIoU(%)	ΔmIoU(%)
Baseline Model (DeepLabV3+)	42.6	71.62	$\-$
$\+ MobileNetV2$ Backbone	18.4	69.83	-1.79
$\+ C-ASPP$ Module	21.2	72.45	+2.62
$\+$ CBAM Module	22.8	73.68	+1.23
$\+$ Dense Neighborhood Prediction	24.3	74.83	+1.15

As can be seen from Table 2, after replacing the Xception backbone with MobileNetV2, the number of model parameters decreased from 42.6 MB to 18.4 MB, but the mIoU decreased by 1.79 percentage points. After introducing the C-ASPP module, the mIoU increased by 2.62 percentage points, reaching 72.45%, indicating that the channel attention mechanism can effectively enhance important features. After further adding the CBAM module, the mIoU increased by another 1.23 percentage points, indicating that the hybrid domain attention can optimize features from both spatial and channel dimensions. Finally, the introduction of the dense neighborhood prediction decoder increased the mIoU to 74.83%, verifying the effectiveness of multi-scale feature fusion.

4.3. Comparison with Mainstream Models

The model in this paper is compared with the current mainstream semantic segmentation models on the PASCAL VOC 2012 and Cityscapes datasets. The results are shown in Table 3.

Table 3: Performance Comparison of Different Models

Model	Backbone Network	Number of Parameters (MB)	VOC2012 mIoU(%)	Cityscapes mIoU(%)
DeepLabV3+	Xception	42.6	72.3	73.8
PSPNet	ResNet50	48.2	71.4	72.6
BiSeNetV2	$-	15.3	68.7	70.3$
LEDNet	$-	12.8	67.2	68.9$
Model of Reference 1	MobileNetV2	23.6	73.9	74.9
Model of Reference 9	$-	14.6	73.8	-$
Model of This Paper	MobileNetV2	24.3	74.8	75.2

As can be seen from Table 3, the model in this paper achieves mIoU of 74.83% and 75.21% on VOC2012 and Cityscapes respectively with only 24.3 MB of parameters, which is better than DeepLabV3+ and PSPNet with larger parameter numbers. Compared with BiSeNetV2 and LEDNet, which are also lightweight models, the model in this paper achieves significant accuracy improvement with slightly higher parameter numbers. Compared with the model in reference 1 , the model in this paper improves mIoU by about 0.9 percentage points with basically the same parameter number, which verifies the effectiveness of C-ASPP and dense neighborhood prediction module.

4.4. Visualization Results Analysis

Figure 3 shows a visual comparison of the segmentation results of the model in this paper on the Cityscapes validation set. As can be seen from the figure, compared with the baseline model, the model in this paper performs better in the following aspects: (1) Small objects are segmented more completely, such as traffic signs and pedestrians in the distance; (2) Object boundaries are more refined, such as vehicle outlines and road edges; (3) Large objects have better internal consistency, such as no obvious holes in the walls of buildings.

ISSN 2522-3488 Vol. 10, Issue 1: 57-63, DOI: 10.25236/IJNDES.2026.100108

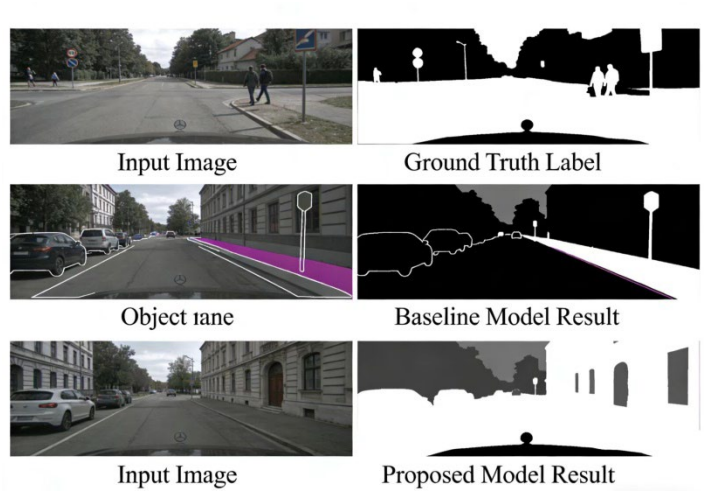

Figure 3: Visual Comparison of Segmentation Results

4.5. Discussion

The role of attention mechanism: Experimental results show that the two attention modules, C-ASPP and CBAM, bring an improvement of 2.62% and 1.23% in mIoU, respectively. It is believed that C-ASPP strengthens the important semantic information in multi-scale features through channel attention, while CBAM helps the model focus on the object region and suppress background interference through spatial attention.

Computational Efficiency Analysis: The inference speed of the proposed model on a single NVIDIA RTX 3090 is 43 frames per second, which basically meets the real-time requirements. However, further optimization is needed for deployment on mobile devices, such as model quantization and pruning.

Limitations: The model in this paper still has some segmentation errors in complex scenarios, such as confusion of similar categories (pedestrians and cyclists), and unstable segmentation under drastic changes in illumination. Future work can consider introducing time series information or multimodal data to enhance robustness.

5. Conclusions

This paper designs and implements a lightweight image semantic segmentation model that combines an attention mechanism. The model uses MobileNetV2 as a lightweight backbone, strengthens important features through channel-hole spatial pyramid pooling and a hybrid domain attention module, and optimizes the segmentation boundary using a dense neighborhood prediction strategy. Experimental results on the PASCAL VOC 2012 and Cityscapes datasets show that the proposed model achieves average intersection-union ratios of 74.83% and 75.21% respectively with only 24.3 MB of parameters, achieving a good balance between segmentation accuracy and computational efficiency.

Future work will focus on the following directions: First, exploring more efficient attention mechanisms, such as linear attention or multi-scale sliding window attention, to further reduce computational overhead; second, introducing knowledge distillation techniques to compress model size while maintaining accuracy; and third, adapting and optimizing models for specific application scenarios (such as medical imaging and remote sensing images).

References

[1] Ma Dongmei, Wang Pengyu, Guo Zhihao. A lightweight semantic segmentation based on attention mechanism[J]. Computer Engineering and Science, 2024, 46(8): 1503-1512.
[2] Yuan Manman, Lu Hao. Semantic segmentation algorithm integrating lightweight and attention mechanism[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025(1): 57-63.

ISSN 2522-3488 Vol. 10, Issue 1: 57-63, DOI: 10.25236/IJNDES.2026.100108

[3] Zhou Hanlian, Ye Qing, Liu Wenqi. A lightweight improved UNet small target aircraft cross-domain segmentation model based on domain adaptation[J]. Optoelectronic Engineering, 2026.
[4] Ma D, Wang P, Guo Z. An improved lightweight semantic segmentation network with dual attention[J]. Journal of Computer Engineering and Science, 2024, 46(9): 1620-1628.
[5] Wang Guogang, Dong Zhihao. Lightweight image semantic segmentation based on attention mechanism and dense neighborhood prediction[J]. Computer Science, 2024, 51(6A): 230300204.
[6] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.
[7] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision, 2018: 3-19.
[8] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision, 2018: 801-818.
[9] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510-4520.
[10] Yu C, Wang J, Peng C, et al. BiSeNet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European Conference on Computer Vision, 2018: 325-341.