DeepLabV3+ Model Incorporating Attention Mechanisms and Deep Separable Convolution

Dongmei Ma[1, 2, a], Pengyu Wang[1, 2, b, *]

[1]School of Physics & Electronic Engineering, Northwest Normal University, Lanzhou 730070, China

[2]Gansu Province Intelligent Information Technology and Application Engineering Research Center, Gansu, China

[a]madongmei@nwnu.edu.cn, [b]787162047@qq.com

Abstract

DeepLabV3+ is a new lightweight network that can get better accuracy with a smaller model. However, DeepLabV3+ also suffers from segmentation discontinuities and obscured semantic information that cannot be easily segmented. In this paper, we improve DeepLabV3+ by using MobileNetV2 as the backbone network, replacing the null convolution in the ASPP module with deep separable null convolution, and passing a lightweight convolutional attention (CBAM) module at the output. To solve the model performance problem caused by data imbalance, we introduce the focal_loss. On the PASCAL VOC2012 dataset using mixed accuracy(fp16 and fp32). The average cross-merge ratio was 70.9%, the category average pixel accuracy was 80.92%, and the total number of parameters was 3.87×106. This achieved a good balance between accuracy and number of parameters.

Keywords

Semantic segmentation; DeepLabV3+; MobileNetV2; Attention module; Channel attention.

1. INTRODUCTION

A key component of computer vision is semantic segmentation, which seeks to identify the target class for each pixel point in an image. This method is frequently employed in a variety of difficult domains, including autonomous driving, face expression recognition, and remote sensing imaging techniques.

Researchers have presented a variety of picture semantic segmentation techniques that have significantly improved image segmentation with the advancement of technology and deep learning. One of them, the semantic image segmentation algorithm of FullyConvolution Networks (FCN)[5], conducts pixel-level classification and thereby resolves the challenge of semantic picture segmentation.

The FCN can accept input images of any size and use a deconvolutional layer to upsample the feature map of the final convolutional layer to restore it to the same size as the input image, in contrast to the classical CNN, which uses a fully-connected layer in the convolutional layer to obtain a fixed-length feature vector for classification. Chen et al. studied the DeepLabv1 model, which adds null convolution on top of VGG, in order to sense additional coordinate and position information and broaden the convolutional receptive field. To extract multi-scale features of various sizes from perceptual fields for multiple branching heterogeneous hole convolution.

Moreover, DeepLabv1[1] is improved by DeepLabv2[2] by the use of post-processing using a fully linked CRF probability map model, which results in reasonably accurate contour full connections. DeepLabv3[3] builds on DeepLabv2 by deleting the CRF module, using the cascade module, and adding batch normalization (BN) to the ASPP module. In order to overcome the issue of long-distance downscaling utilizing global averaging pools, the ASPP module offers batch Normalization (BN). The loss of significant weights across long distances is decreased by using global averaging. DeepLabv3+[4] thoroughly mimics the encoder-decoder structure and takes into account both shallow and deep semantic information to improve object edges. optimizing shallow and deep semantic information to optimize object edge details.

The attention mechanism has considerably enhanced the efficiency of DCNNs in recent years (AAM). The attention mechanism introduces a model that focuses primarily on specific portions of an image rather than the entire image in an effort to simulate human attention while reducing computational complexity of image processing and boosting performance. Convolutional Block Attention Module (CBAM), a compact attention module that uses both spatial dimension and channel dimension for attention operations, was presented in 2018. CBAM can be trained end-to-end with simple CNNs and can be smoothly integrated into CNNs because it is an end-to-end generic module. This saves parameters and processing resources while enabling plug-and-play module integration into current network architectures.

2. RELATED THEORIES

2.1. Lightweight Feature Extraction Network MobileNetv2

MobileNetv2[13], a lightweight convolutional neural network, is designed with the main goal of improving the representational power of the network by adding Linear Bottleneck and Inverted Residual on top of v1. as depicted in Figure 1.

If the size of the convolution kernel is m×h×w, the output feature map after convolution is S×I×J, and the size of the input feature map is N×H×W. The parametric quantities of the deep separable convolution and the traditional convolution process are

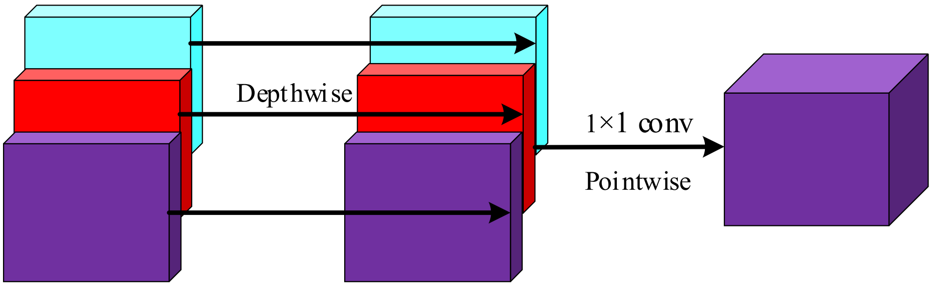

Figure 1. Schematic diagram of the depth-separable convolution process

If the size of the convolution kernel is m×h×w, and the size of the input feature map is N×H×W. The parametric quantities of the deep separable convolution and the traditional convolution process is shown in (1) and (2).

$$F_c = (h \times w + s) \times N \tag{1}$$

$$F_d = h \times w \times N \times S \tag{2}$$

As can be seen from equations (1) and (2), the number of parameters for deep separable convolution is significantly lower than that of ordinary convolution. We even replaced the hole convolution in the ASPP module with a deep separable hole convolution, which further reduces the number of parameters while also balancing performance.

2.2. CBAM

The final feature map is created by multiplying both feature map information with the prior original input feature map for adaptive feature correction. The CBAM[14] module can serially generate attentional feature map information in both channel(CAM) and spatial(SAM) dimensions. Any backbone network can use the small, lightweight CBAM module to operate better. The process of CBAM is shown in Figure 2.

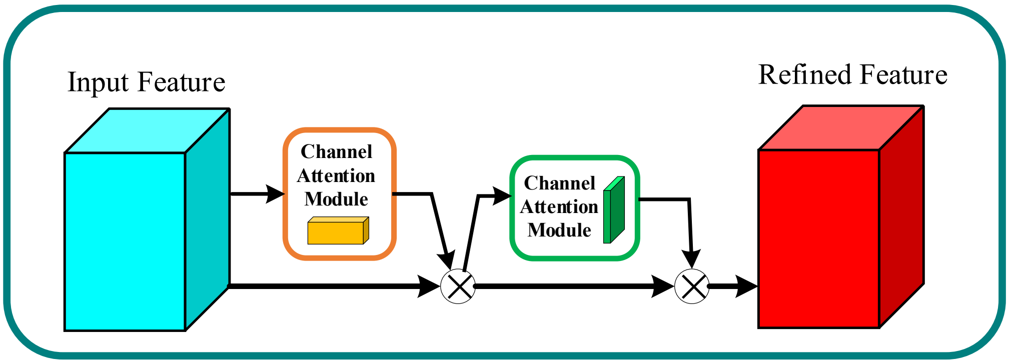

Figure 2. CBAM

The Channel Attention Module (CAM) implementation process is shown in Figure 3.

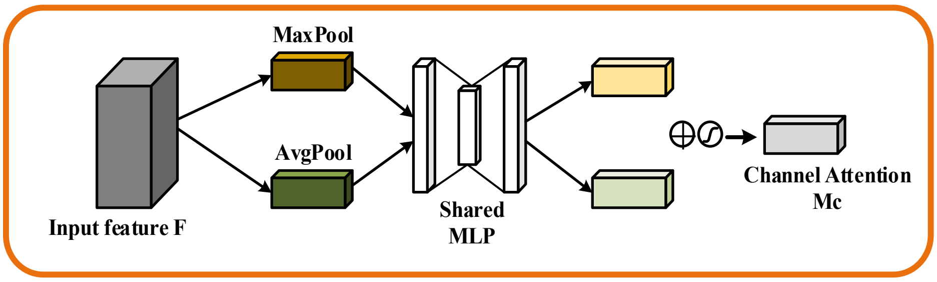

Figure 3. CAM implementation

The channel attention calculation formula is summarized in Equation 3:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_2(F^c_{avg})) + W_1(W_0(F^c_{\max}))) \end{aligned} \quad (3)$$

The process of Spatial Attention Module (SAM) implementation is shown in Figure4.

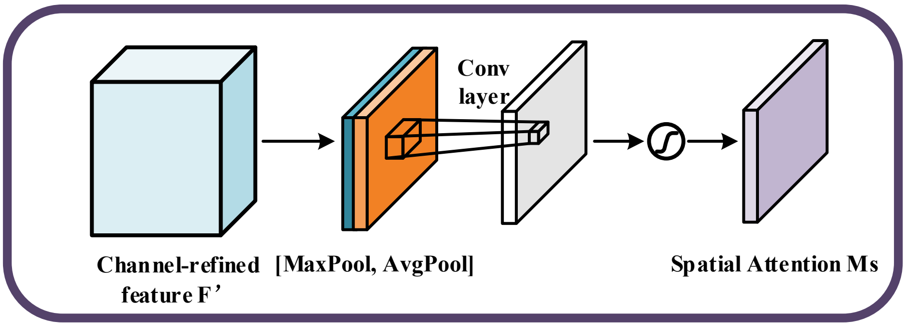

Figure 4. SAM implementation

Spatial attention is calculated by equation 4:

$$\begin{aligned} Ms(F) &= \sigma(f^{7\times7}([AvgPool(F);MaxPool(F)])) \\ &= \sigma(f^{7\times7}(F^s_{avg};F^s_{\max})) \end{aligned} \tag{4}$$

All that needs to be done is compress the spatial dimension of the feature map in order to more effectively calculate channel attention features. Prior to this, the average pooling technique was employed, which can learn the target object's degree information. CBAM investigated whether maximum pooling can also learn the object's discriminative features. As a result, both approaches are used in the channel attention module, and experiments have shown that using both approaches simultaneously is more successful than using either approach separately.

2.3. Loss function focal_loss

The effect of utilizing the cross-entropy loss function is not optimal due to the issue of inhomogeneity of the positive and negative samples, thus it is required to balance the cross-entropy first[15]. With the formula described in, focal loss modifies the balanced cross-entropy loss function to lessen the weight of easy-to-classify data and concentrate on training difficult examples (5).

$$F_p(di) = (1-di)^r \log(di) \tag{5}$$

The weights help to deal with the imbalance of the categories. Where (1-di)r is the adjustment factor and r ≥ 0 is the adjustable focus parameter. The larger the value of r, the smaller the loss of well-categorized samples will be, and the attention of the model can be focused on those samples that are difficult to classify. A large r expands the range of samples for which a small loss is obtained. At the same time, when r = 0, this expression degenerates into a Cross Entropy Loss function.

3. NETWORK STRUCTURE OF DEEPLABV3+

DeepLabV3+[4] network uses multi-branch parallel structure and multi-scale fusion to improve the segmentation effect while optimizing the spatial information, which is a network of considerable accuracy in the field of semantic segmentation network structure. The DeepLabV3+ network is structured with encoding-decoding regional semantic segmentation.

The encoded region is made up of a void space pyramid pool (ASPP) module and a backbone feature extraction network module. The backbone module is used to extract the basic depth features whereas the ASPP module improves feature extraction capabilities. The ASPP module extracts feature data at multiple scales using a multi-branch parallel structure. Whereas employing a serial network structure might only produce a single perceptual domain, a parallel network structure might produce fields of vision of various sizes. This makes it simpler to tell between items of different sizes in an image. The spatial structure of the input image is recovered by upsampling the high-dimensional feature layers produced by the decoding stage[7].The spatial information lost during the downsampling process cannot be fully recovered by the traditional upsampling method.

In deepLabV3+, deep and shallow features are combined to improve the accuracy of semantic segmentation[12]. The deep feature map has a resolution of 32 × 32 × 256 over the entire coding region, while the shallow feature map in the selected backbone network has a spatial resolution of 128 × 128 × 256 . Both have 256 channels, however, the semantic content of the deep feature maps is different. The number of channels in the shallow feature map is reduced by using 1 × 1 convolution to reduce the weights of the shallow features. Then, the final output image is obtained by up-sampling it four times and then convolving it 3×3 to restore it to its original size by a Softmax classifier. To reduce the feature weights of the shallow features, 1×1 convolution is used to reduce the number of channels in the shallow feature map. Then, the feature map is upsampled four times and 3×3 convolved to bring it back to its original size, and the final output image is generated by the Softmax classifier.

The ASPP module in DeepLabV3+ can capture contextual information efficiently. This structure cannot detect anisotropic contextual information. The novel parallel multi-branching module provided in this study has seven branches.

ASPP: In order to extract features, a 1 × 1 convolution, global average pooling (GAP) at the image level, and a null convolution with three different expansion rates are used. Where the number of channels C = 256, the size of the convolution kernel for the null separable convolution is 3 × 3, and the expansion rate r ＝｛6, 12, 18｝.

The improved DeepLabV3+ network is shown in Figure 5.

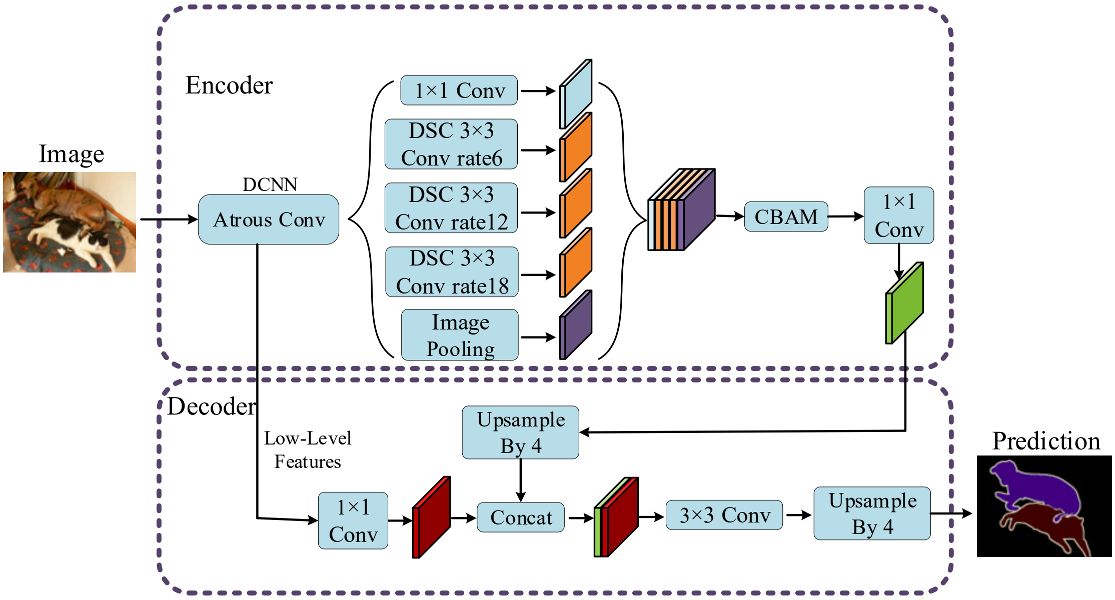

Figure 5. The improved DeepLabV3+

4. EXPERIMENTS AND RESULTS ANALYSIS

4.1. Experimental details

The performance and algorithm performance in semantic segmentation tasks can be measured by pixel accuracy, class average pixel accuracy (mPA), frequency-weighted intersection ratio and average intersection ratio. The average intersection ratio (mIoU) is again the most widely used metric. In this paper, mIoU and mPA are also used as evaluation metrics. mIoU can represent the degree of overlap between the predicted and true values of different classes, which divides the intersection and merge between the true and predicted values of a class and then sums them to take the average value, as in equation (6). mPA represents the ratio of the number of correctly classified pixels to all pixels calculated, as in equation (7).

$$mIoU = \frac{1}{k+1}\left[\sum_{c=0}^{k}\frac{X_{cc}}{\sum_{d=0}^{k}(X_{dc}-X_{cc})+\sum_{d=0}^{k}X_{cd}}\right] \tag{6}$$

$$mAP = \frac{1}{k+1}\sum_{c=0}^{k}\frac{X_{cc}}{\sum_{d=0}^{k}X_{cd}} \tag{7}$$

Where k+1 is the number of categories, X_{cc} represents the number of pixels whose predicted outcome is class c and the actual class is class c, and N represents the total number of categories of pixels in the graph. X_{dc} represents the actual category for the number of pixels that are class d and the expected outcome is class c, and X_{cd} represents the number of pixels whose actual class is class c and the predicted outcome is class d.

Crop size: Before training, the size of the image crop needs to be increased to prevent filter weights with large expansion rates from being mostly filled with zero regions. In this paper, we train and test on the PASCAL VOC2012 dataset with a cropped image size of 512 × 512[6]. No distortion is achieved by adding gray bars to the images using resize.

the chip is 12th Gen Intel(R) Core(TM) i5-12400F 2.50 GHz, the memory is 16 GB, NVIDIA GeForce RTX 3070 graphics card, 8G video memory. 100 epochs are performed in voc2012, the first 50 epochs are freezing phase, batch_size is 8, the last 50 epochs are thawing phase, sgd is used as optimizer, batch_size is 4, initial learning rate is 7e-3, learning rate The descent mode is step, momentum is 0.9, weight_decay is set to 1e-4, and the loss function focal_loss is added.

4.2. Analysis of experimental results

In order to obtain an effective semantic segmentation model, we investigated the impact of the loss function focal_loss module and the adoption of the step learning rate approach on model performance and computational complexity through experiments on the PASCALVOC2012 dataset. The results are shown in Table 1.

Table 1. Ablation experiments with loss function and learning rate descent method

Models	Learning rate decline mode	MIoU / %	mAP
BM	cos	72.31	82.52
BM	step	72.52	82.60
BM+focal_loss	cos	72.80	82.52
BM+focal_loss	step	73.48	82.85

As can be seen from Table 2, when the learning rate decrease method is modified to step on the benchmark model, the model gets a small improvement, and the mIoU improves by 0.21%, and when the focal_loss module is used on the benchmark model, the mIoU improves by 0.49%, and when the focal_loss module is used on the benchmark model and the learning rate decrease method is used to STEP, the model gets a larger improvement, and the mIoU improves by 1.17%. When the focal_loss module is used on the benchmark model and the step learning rate is decreased, the model is improved substantially and the mIoU is improved by 1.17%, and the mAP is basically maintained or increased slightly in this series of experiments.

In order to verify the effectiveness of CBAM for lightweight convolutional attention. Further thawing experiments were conducted in this paper. It was conducted with a focal_loss loss function and a learning rate descent of STEP, and the data are shown in Table 2.

Table 2. Ablation experiments on CBAM

Models	Floating point calculation volume /GFLOPs	MIoU/%	mPA
BM	53.026	73.48	82.85
BM+SEnet	53.027	73.61	82.65
BM+ECAnet	53.027	73.62	82.61
BM+CBAM	53.027	73.80	83.06

From Table 2, it can be concluded that with the focal_loss module and using the step learning rate descent approach, the addition of SEnet at the output of the ASPP module and ECAnet is less effective than the addition of the hybrid attention channel CBAM on a benchmark model with essentially the same number of parameters and floating point computations. mIoU of the model improves by 0.32% with the addition of CBAM and mPA improved by 0.21%.

Based on the above experimental results, after using the focal_loss loss function, the setp learning rate descent method and the hybrid attention mechanism CBAM, the final experiments concerning Mixed accuracy(fp16 and fp32) and depth-separable convolution(DSC) were conducted in this paper, and the results are shown in Table 3.

Table 3. Mixed accuracy(fp16 and fp32) and depth-separable convolution analysis

Models	Mixing accuracy	MIoU/%	mPA	Total params
BM	—	73.48	82.85	5.818×106
BM+ECAnet	√	73.87	83.58	5.826×106
BM+SEnet	√	73.61	82.65	5.826×106
BM+CBAM	√	73.87	82.78	5.829×106
BM+ECAnet+DSC	√	70.74	80.3	3.860×106
BM+CBAM+DSC	√	70.9	80.92	3.868×106

From Table 3 it is obtained that in using the loss function focal_loss and the learning rate descent method step the ASPP module achieves good results through a CBAM attention mechanism while replacing the ASPP null convolution with a deeply separable null convolution. The overall mIoU and mAP, although lower than though, reduce the amount of parameters in the model by almost 34%. According to the test results, the new model proposed in this paper achieves superior semantic segmentation results with less computational resource consumption.

Table 4. IoU in each category on the PASCAL VOC 2012 test set

Class	Benchmark model	New model
background	93.08	92.46
airplane	84.58	81.19
,bicycle	42.19	53.34
bird	81.81	80.69
boat	61.61	64.49
bottle	70.62	68.12
bus	93.45	90.6
car	84.7	82.79
cat	87.14	85.44
chair	33.69	30.51
cow	80.37	80.18
dining table	50.33	52.48
dog	79.77	77.13
horse	79.56	77.91
motorcycle	80.65	79.39
person	80.88	78.01
potted plant	57.45	51.33
sheep	80.43	77.66
sofa	43.66	45.23
train	84.45	80.46
tvmonitor	67.93	59.58
mIoU	72.31	70.9

A comparison of the cross-parallel ratio IoU of the benchmark model and the new model for 21 categories tested on the PASCAL VOC 2012 dataset is reported in Table 4. As can be seen from Table 4, the mIoU of the model decreases due to the inclusion of the depth-separable null convolution, so it decreases for most of the categories, but there is a substantial increase in IoU for categories like bicycle, boat, dining table, and sofa. So the model can be used for specific kinds of segmentation.

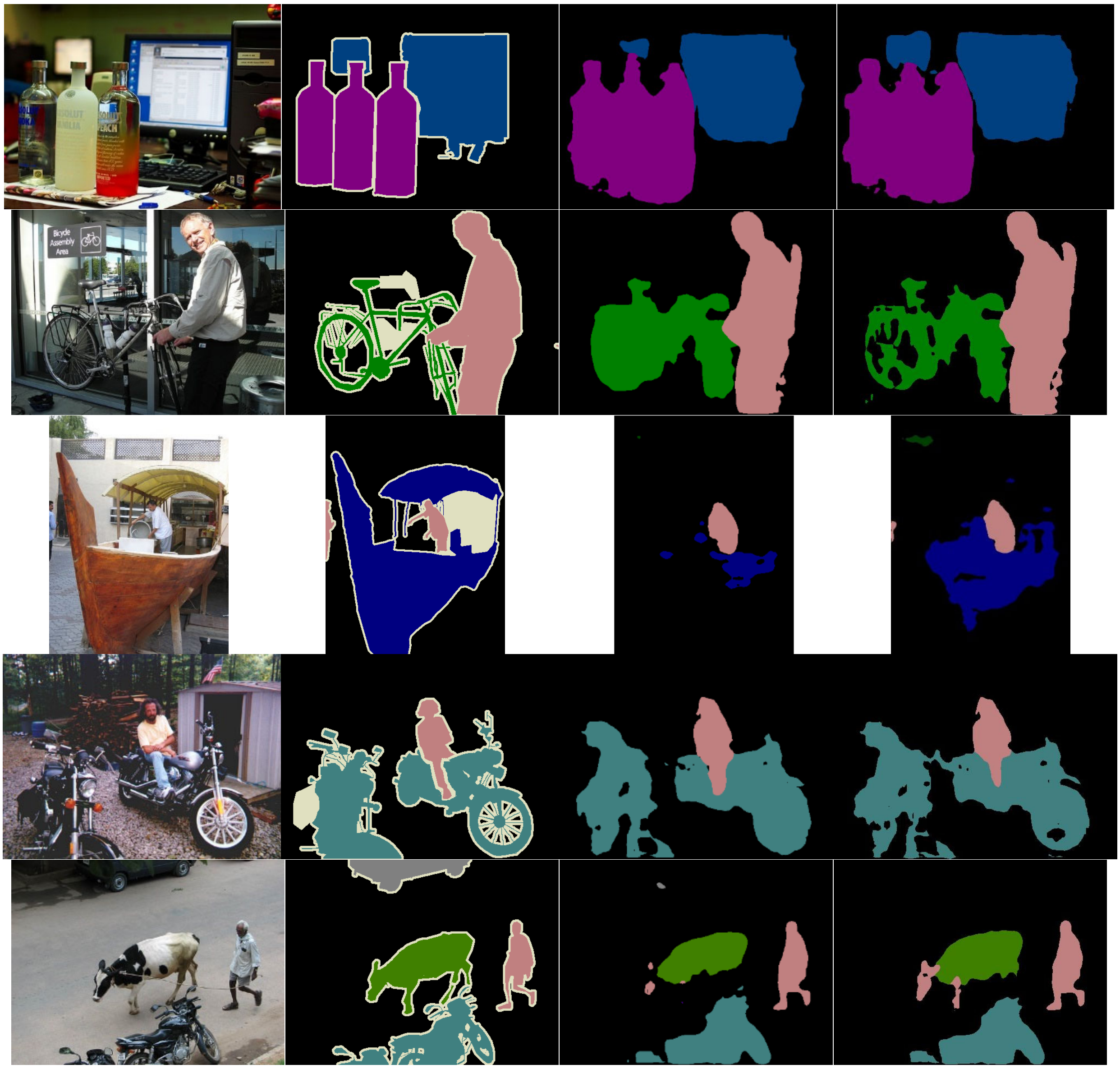

Figure 6. Partial visualization example of model

Five sets of images were arbitrarily selected from the PASCAL VOC2012 dataset for the relevant tests, and the visualized results are shown in Figure 6., it can be seen that the benchmark model has more serious loss of edge information, while the improvement is more obviously achieved in this paper's ECAnet model, such as trains, people on motorcycles, dog legs, etc.

As obtained in Figure 4, we can clearly see in Figure 6 that the segmentation of bicycles, boats, etc. is better than the baseline model. In summary s, the new model proposed d in this paper has a very low parametric number of y while also y has a high accuracy.

5. CONCLUSION

In this paper, we propose an efficient semantic segmentation method based on DeepLabV3+ improvement. The backbone feature extraction network is set as a MobileNetV2 network. The output of the ASPP module of DeepLabV3+ is passed through a CBAM, and the null convolution in the ASPP module is replaced by a depth-separable null convolution. Comparison and ablation experiments show that the optimized model has a much lower number of parameters while maintaining the accuracy of model . However, the model needs further improvement in terms of real-time performance, and this work will focus on the accuracy, complexity, and speed of inference. The next work will focus on the accuracy, complexity, and speed of inference so that the model is simultaneously high precision, lightweight, and efficient.

ACKNOWLEDGMENTS

This paper was financially supported by National Natural Science Foundation of China(No.61961037).

REFERENCES

[1] Zhu Wenbin et al. A sub-region Unet for weak defects segmentation with global information and mask-aware loss[J]. Engineering Applications of Artificial Intelligence, 2023, 122.

[2] Bobkov A. V. and Aung Kh.. Real-Time Person Identification by Video Image Based on YOLOv2 and VGG 16 Networks[J]. Automation and Remote Control, 2022, 83(10) : 1567-1575.

[3] Chen L C , Papandreou G , Schroff F , et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. 2017.

[4] Hadinata Patrick Nicholas et al. Multiclass Segmentation of Concrete Surface Damages Using U-Net and DeepLabV3+[J]. Applied Sciences, 2023, 13(4) : 2398-2398.

[5] Wang Hao et al. A novel feature attention mechanism for improving the accuracy and robustness of runoff forecasting[J]. Journal of Hydrology, 2023, 618.

[6] Ma D, Li P, Huang X,et al. Efficient semantic segmentation based on improved DeepLabV3+[J]. Computer Engineering & Science, 2022, 44(04): 737-745.

[7] Zhang J, Liu J,Xie F, et al. Semantic Segmentation of Station Selection and Line Selection in Remote Sensing Image Based on Improved Deeplabv3+ Network[J]. Control Engineering of China, 2022, 29(03): 558-563.

[8] Zhao W, Chen Y, Xiang S,et al. Research on Image Semantic Segmentation Algorithm Based on Improved DeepLabv3+[J/OL]. Journal of System Simulation: 1-12[2023-03-12].

[9] Pang B,Bang Z,Yang M, et al. Research on Defect Recognition of Power Patrol Images Based on YOLOv5[J]. Sichuan Electric Power Technology. 2022, 45(05): 48-53+94.

[10] Wang Y,Wang Y, Wu H, et al. Image Semantic Segmentation Based on Improved Deeplabv3[J]. Computer Simulation, 2022, 39(10): 148-152+158.

[11] Wang J, Wang J, Cao W, et al. Optimization of Street View Semantic Segmentation Algorithm Based on Deeplabv3+[J]. Modern Computer, 2022,28(08): 42-47.

[12] Chen H, Sun Z, Kong W. Semantic Image Segmentation Based on Fusion of Deep Neural Networks and Atrous Convolution[J]. Journal of Chinese Computer Systems, 2020, 41(01): 166-170.

[13] Kumar Ujjwal and Arora Deepak and Sharma Puneet. Face Mask Detection Using MobileNetV2 and VGG16[M]. Springer Nature Singapore, 2023 : 669-677.

[14] Li Enlin et al. A novel deep learning method for maize disease identification based on small sample-size and complex background datasets[J]. Ecological Informatics, 2023, 75.

[15] Li Tao et al. Lightweight End-to-End Neural Network Model for Automatic Heart Sound Classification[J]. Information, 2021, 12(2) : 54-54.