World Scientific Research Journal Volume 7 Issue 5, 2021
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

Semantic Segmentation of Street Scene Based on Multi-Scale and Attention Mechanism

Tingting Xu, Guping Zheng

School of Computer and Control Engineering, North China Electric Power University, Baoding, 071003, China

Abstract

Street scene images contain various objects of different scales. The segmentation model with single scale and feature extraction and fusion can not get good segmentation and prediction results. Therefore, a semantic segmentation model based on multi-scale feature fusion and attention mechanism is proposed. Firstly, the asymmetric structure of atrous spatial pyramid pooling (ASPP) is used to optimize the extraction of different levels and scales of street scene image. Secondly, the attention mechanism is introduced into the feature maps of different scales, so that the network can focus on the salient features of each level. Finally, all the feature images are adjusted to the same size for fusion, and the key feature information of each scale object in the street scene is fully extracted to segment it effectively. The experimental results on the dataset Cityscapes show that the semantic segmentation network model based on multi-scale and attention mechanism can further improve the segmentation accuracy and optimize the segmentation results.

Keywords

Multi-scale feature fusion; Atrous convolution; Attention mechanism; Semantic segmentation.

1. INTRODUCTION

Semantic segmentation [1] is an important direction in the field of computer vision. Its main task is to mark the categories of pixels in the image, that is, to classify each pixel in the image. It is widely used in the fields of automatic driving, virtual reality and medical image analysis [2-3]. In recent years, with the rapid development of deep learning technology, the semantic segmentation model [4] based on deep learning is proposed by many researchers to solve the shortcomings of improving traditional segmentation algorithm. Fully convolutional network [5] (FCN) is the first network that formally applies convolutional neural network to image segmentation. As a basic model, it is used in image semantic segmentation task, which breaks through the limitation of input image size and realizes end-to-end segmentation. Although FCN based method has achieved good results in semantic segmentation, there are still some problems: after a series of convolution pooling operations, the resolution of the feature image will gradually decrease, resulting in the lack of spatial location information of some pixels; in the process of segmentation, the image context information is not fully considered, the local features, the rich global features and the salient feature information of the target can not be effectively used, resulting in low segmentation accuracy and inaccurate results [6]. Researchers have proposed a series of different improvement methods based on FCN. In order to increase the receptive field and extract the effective feature information of large objects, DeepLabv2 [7] combines the atrous convolution and spatial pyramid pooling method, and proposes the atrous spatial pyramid pooling (ASPP) , atrous convolution can realize the function of partial pooling

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

layer, replace it to complete the fusion of semantic information features while preserving the resolution of the image, and increase the receptive field to improve the segmentation accuracy with less parameters. ESPNet [8] based on the idea of convolution factor decomposition, constructed a method of resampling feature map through point-to-point convolution and atrous spatial pyramid convolution, which can learn representative multi-scale features in space, accelerate the induction of semantic information, and facilitate pixel classification. Although the above methods can improve the segmentation result, the continuity of local pixel information in the image will be interrupted. In order to reduce the resolution of feature map, a symmetric network structure based on encoder decoder is proposed. SegNet [9] and U-Net [10] use different mechanisms to transfer information from the encoder to the decoder. The innovation of SegNet is that when the encoder is merged, the pooling layer index is established. In addition, SegNet can recover the boundary information better while reducing the training parameters. Although this kind of method improves the segmentation accuracy, it also has some problems, such as large amount of calculation, slow speed and so on. Attention mechanism (AM) is introduced into neural network to optimize the network structure. Attention mechanism can allocate different attention weights to the input image to capture more critical feature information in the image, so it is also widely used in semantic segmentation network.

Segmentation network model based on deep learning has been applied in various fields. At present, the more popular application research of automatic driving has been widely concerned by researchers. Many algorithm models have been proposed to solve the problems of automatic driving. However, the street scene image contains various objects of different scales, and the same object will present different size scales due to the distance, for example, the scale of the near car in the street scene is much larger than that of the far car. In addition, we need to distinguish some confusing categories in the image and consider the target objects with different appearances. In order to effectively accomplish the task of semantic segmentation of street scene, it is necessary to use the multi-scale features of image to enhance the feature representation and discrimination ability of pixel level recognition.

Considering the above problems, this paper proposes a semantic segmentation model of street scene based on multi scale features and attention mechanism. Firstly, Resnet50 network [12] is used as the backbone network for image feature extraction. Then, the high-level features of the street scene image are extracted by using the atrous spatial pyramid pooling asymmetric convolution (ASPP-AC) module, and then the scale feature maps extracted from each layer of the backbone network are fed back and fused to obtain feature maps with rich semantic information. At the same time, the attention mechanism is introduced into the multi-scale features of different layers, so that the network can adaptively focus on the salient feature information of objects in each layer of the image, and enhance the understanding of multi-scale scene to the object level. Finally, the feature maps of different scales obtained from different levels are adjusted to the same scale for fusion processing, and the feature maps containing rich image information are generated, and the street scene image is classified and predicted at the pixel level. Compared with other segmentation models, the experimental results show that this method can effectively solve the problems of single scale and simple feature extraction and fusion of image segmentation model, and improve the accuracy of semantic segmentation of street scene.

2. SCENE SEMANTIC SEGMENTATION MODEL BASED ON MULTI-SCALE FEATURE FUSION AND ATTENTION MECHANISM

The framework of semantic segmentation model proposed in this paper is shown in Figure 1. The model mainly includes three parts: backbone feature extraction module, ASPP-AC module and AM module. Firstly, the first five convolution layers of Resnet50 network are used as the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

backbone network, which can effectively extract the basic feature information of each object in the street scene image. Then, the advanced feature map of the image is further obtained by using ASPP-AC module. At the same time, the feature map of the image is merged with the output feature map of 5 layers in the backbone network through the upper sampling, and a feature map with rich semantic information is obtained. In addition, the AM module including spatial attention (SA) module and channel attention (CA) module is used to optimize the output feature map of the first four convolution layers of the backbone network in parallel to obtain the feature map with more sufficient edge feature information. Finally, the extracted feature maps of different scales are adjusted to the same size for pixel level fusion, and a feature map with rich semantic information and spatial information is generated, which makes the network effectively segment various objects in street scene images.

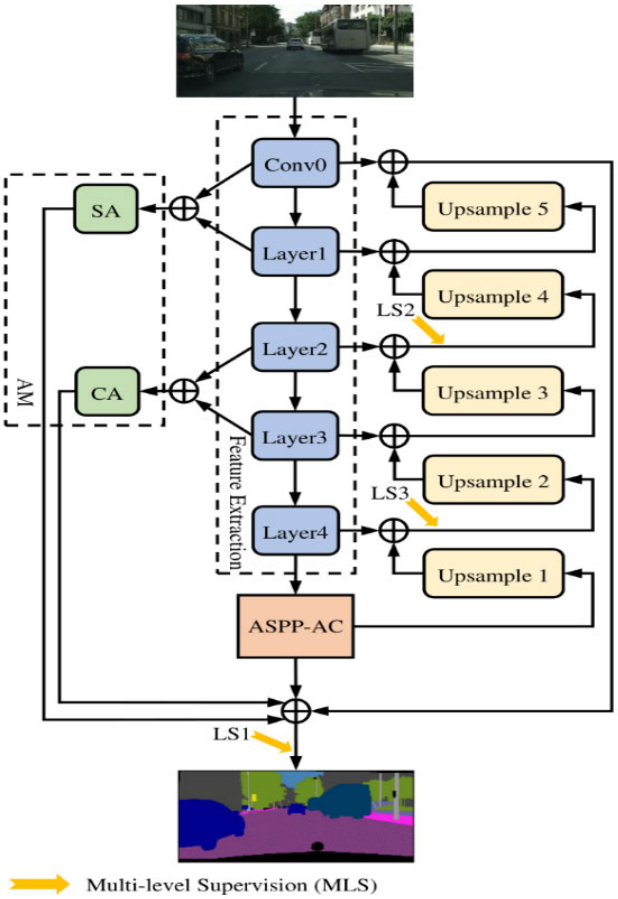


Figure 1. Semantic segmentation model framework of street scene based on multi-scale feature fusion and attention mechanism

2.1. Feature Extraction Module

Multi-scale context information [13] plays an important role in semantic segmentation, especially for objects with large scale variation. The convolutional neural network is mainly used for image classification task, which extracts multi-scale information features by stacking multiple convolution layers and pooling layers. However, the size and position of various objects in street scene images usually have great changes. In addition, although image semantic segmentation belongs to intensive classification task, it is different from image classification in structure. Simple overlay convolution network can not effectively deal with these complex changes, which will inevitably lead to low segmentation accuracy. In order to obtain rich multi-scale feature maps of street scene images, this paper uses Resnet50 network as the backbone

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

network, and uses its first five convolution structures to extract features, and obtains different scale feature maps of each layer.

The input of backbone network is the original color image with 3 channels. After the convolution and pooling operation of the first layer, the feature image FM0 with the largest resolution and the richest geometric detail information is obtained. Its size is 1/2 of the size of the original image, and the number of channels is increased to 64. Then FM0 is sent to the next layer module as the input. The residual modules of the last four layers are composed of fixed but unequal number of 1×1 convolutions and 3×3 convolutions. The specific structure can be seen in reference [12]. The input of each layer is the output of the previous layer. The characteristic diagrams obtained by each layer are FM1, FM2, FM3 and FM4. The size of each layer is reduced to 1/4, 1/8, 1/16 and 1/16 of the original image, and the channel number changes to 256,512,1024,2048, which is a multi-scale feature map. With the deepening of network structure, the extracted feature graph can be divided into low-level feature and high-level feature. Low level features are relatively large in resolution and scale, and contain more location and detail information, but their semantic information representation ability is weak and noise is more. Advanced features have stronger semantic information, but the resolution is very low, the scale is relatively small, and the perception of details is poor.

2.2. ASPP-AC Module

Image semantic segmentation is achieved by fine-tuning the classification network. However, because the classification network needs to use more pooling operations to expand its receptive field and fuse semantic information, the image resolution will be reduced and the spatial location information and dense semantic feature information will be lost. In addition, this type of segmentation model has some problems, such as single scale of feature extraction, relatively simple feature fusion processing and insufficient utilization of image feature information, which is more unfavorable for semantic segmentation of street scene image. Atrous convolution [14] is to obtain the different sensing field range by using convolution operation with different atrous coefficient d. The function diagram of atrous convolution is shown in Figure 2. Figure 2 (a) shows the atrous convolution of d = 1, that is, the standard convolution, with the receptive field range of 3×3; Figure 2 (b) shows the atrous convolution of d = 2, with the receptive field range of 5×5; Figure 2 (c) shows the atrous convolution of d = 3, with the receptive field range of 7×7. This structure avoids many pooling operations, and can effectively capture multi-scale feature information. Therefore, this paper designs a more advanced feature extraction module ASPP-AC module based on ASPP. In this module, asymmetric convolution is introduced to further optimize the feature extraction, and the feature scale is expanded to further mine the image feature information, reduce the amount of model calculation, and speed up the operation of the model. The theory of asymmetric convolution is as follows: if the rank of the two-dimensional convolution kernel is 1, then the operation can be equivalent to a series of one-dimensional convolution [15]. It is usually used to approximate the existing square convolution for model compression and acceleration. Using the additivity of two-dimensional convolution, it can be true even if the convolution kernel size is different.

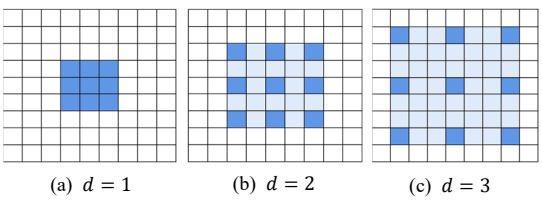


Figure 2. Function diagram of atrous convolution

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

The ASPP-AC module is shown in Figure 3. The module takes feature FM4 as input and then transfers it to five independent branches in parallel. The first branch is atrous convolution layer with convolution core of 1 and atrous coefficient of 1, and the second to fourth branches are atrous convolution layer with convolution core of 3 and atrous coefficients of 6, 12 and 18 respectively. Asymmetric convolutions of 1×3 and 3×1 are introduced into the three branches to further mine and utilize the multi-scale feature information of images and reduce the amount of computation. The fifth branch is the average pooling layer, which is to pool the features globally to get the image level features. Finally, the scale of the feature map is adjusted by 1×1 convolution to get the advanced features after feature information enhancement, which further improves the feature extraction effect of image semantic information of different scales.

The different branches of ASPP-AC module from the top to the bottom can be understood as the visual range from small to large, which helps the network to better identify objects of different scales in the image. Then, the top-down feature map is fused to get a feature map which contains all the feature information of different levels and scales. The purpose of this operation is to use high-level features to guide the fusion of low-level features, and obtain feature maps with rich semantic information, so that the pixel level classification accuracy of street scene image is higher, and the misclassification is reduced.

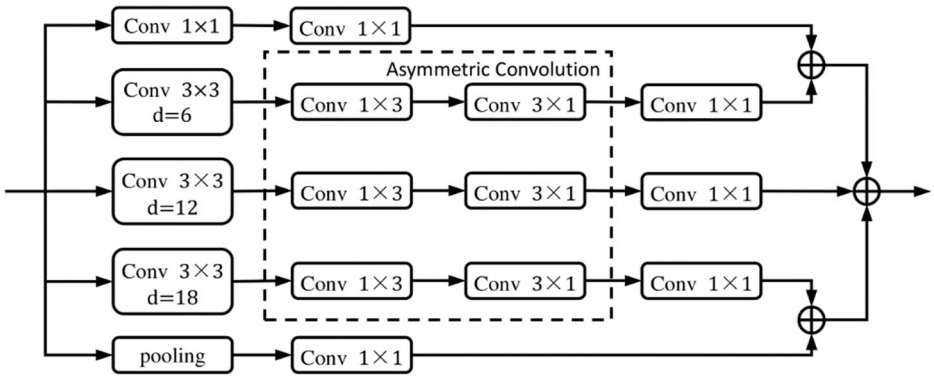


Figure 3. ASPP-AC module

2.3. AM Module

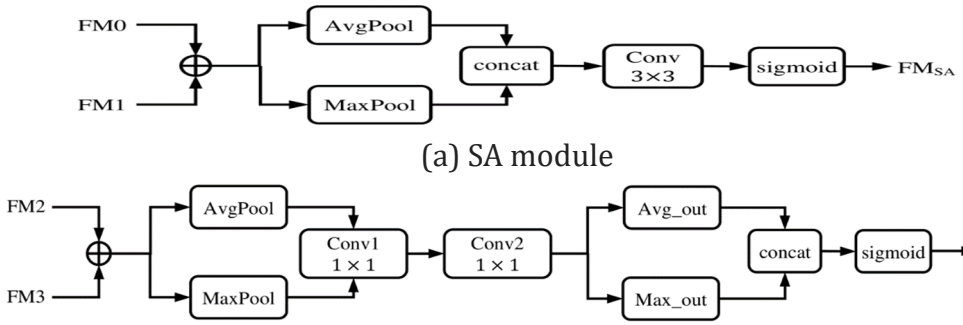
In the neural network model, attention mechanism has its unique importance, most models will use its attention mechanism characteristics to improve the performance of their own model. Attention mechanism can make feature graphs at different levels generate saliency graphs with different semantic values [16], but most existing methods integrate multi-scale features without distinction, which leads to information redundancy. More importantly, some levels of inaccurate information will lead to performance degradation or even false prediction, so it is very important to filter these feature information to obtain more valuable features. Therefore, according to the characteristics of different levels of feature information of street scene image, the AM module includes SA module and CA module. SA module can effectively focus on the feature information of image space, while CA module focuses on extracting the semantic feature information of image. The spatial and semantic feature information of image plays an important role in image segmentation.

The structure of SA module is shown in Figure 4 (a). The input of SA module is the fusion of FM0 and FM1, which is equivalent to the low-level feature of image extraction and contains rich spatial structure information. SA module uses average pooling and maximum pooling to pool the input feature map, so as to ensure feature extraction within enough receptive field and retain the background semantic information of street scene image. Then the two pooled results are fused according to the same dimension to obtain the spatial attention weight. After a 3×3

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

convolution and sigmoid normalization operation, the spatial attention feature map is obtained to obtain the feature map FMSA containing rich spatial structure information, which provides convenience for the next segmentation processing.

The structure of CA module is shown in Figure 4 (b). The input of CA module is the fusion of FM2 and FM3, which can be regarded as the advanced feature of image feature extraction. The CA module also uses average pooling and maximum pooling to process the image in parallel, and then increases its network depth and reduces its spatial dimension by two times of 1×1 convolution, and then fuses the output results of the two at the pixel level, and obtains the attention weight on the channel through the activation function sigmoid. The resulting feature map FMCA assigns a large weight to the main image information and highlights the key semantic information features of the objects in the street scene image.



(a) SA module

(b) CA module

Figure 4. AM module

2.4. Loss Function

Loss function is an indispensable part of the whole network model in deep learning. It is used to evaluate the difference between the predicted value and the real value by the back propagation of errors generated by the prediction sample and the real sample mark [17].

The loss function is also an important index to measure the generalization ability of the trained model, that is, the loss function is the reflection of the model to the degree of data fitting. The better the fitting, the smaller the value of the loss function, indicating that the more close the predicted value of the model to the real value, the better the robustness of the model. Common loss functions include: least square loss function, smoothL1 loss function used in regression, cross entropy (CE) loss function, etc.

CE loss function is a common loss function for multi classification task in convolutional neural network classification. The formula of CE loss function is:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c} log p_{i,c} \quad (1)$$

Where N represents the total number of samples; M represents the total number of categories; yi,c indicates the variable (0 or 1), if the category i is the same as the category of the sample, it is 1, otherwise it is 0; pi,c for the prediction probability of observation sample i belonging to the category c.

The semantic segmentation model of street scene image proposed in this paper also belongs to multi classification task, so CE loss function is also used to train the model. Due to the large amount of calculation and relatively slow speed of data processing in training, it takes a lot of time to achieve a good training effect, that is, to obtain a small loss. Therefore, in the process of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

model training, multi levels supervision (MLS) is used to speed up the model training speed and optimize the model performance. The loss function of MLS is as follows:

$$MLS = \alpha Loss_1 + \beta Loss_2 + \theta Loss_3 \quad (2)$$

Where Loss1, Loss2 and Loss3 are the CE loss functions, and α, β and θ are the proportional coefficients.

3. EXPERIMENT AND RESULT ANALYSIS

The running software platform of this experiment is: Windows10 64 bit operating system, python deep learning framework; hardware platform: CPU is Intel Core i7-10700, main frequency is 2.9 GHz; memory is 16GB; GPU is GeForce RTX 2060 graphics card.

3.1. Experimental Data Set and Evaluation Index

The data set Cityscapes [18] is used as the experimental training data set. Cityscapes data set, namely urban landscape data set, is the image data set of semantic understanding of urban street scene in the field of semantic segmentation. There are two sets of evaluation standards, fine and coarse. The former provides 5000 fine annotated images, while the latter provides 5000 fine annotated and 20000 rough annotated images. The size of the images in the data set is 1024×2048, mainly including images from street scenes from 50 different cities. The experiment mainly uses 5000 fine labeled images to train the model, including 2975 training set images, 500 verification set images and 1525 test set images, with a total of 19 categories.

At present, the performance evaluation methods of semantic segmentation model mainly include pixel accuracy (PA), mean intersection over union (mIou), etc., among which mIou is the standard measure of semantic segmentation [19], which is generally based on class calculation, defined as calculating the ratio of intersection and union of two sets of real value and predicted value. The calculation formula of mIou is as follows:

$$mIou = \frac{1}{k+1}\sum_{i}^{k}\frac{p_{i,i}}{\sum_{j=0}^{k}p_{i,j}+\sum_{j=0}^{k}p_{i,j}-p_{i,i}} \quad (3)$$

Where k is the number of categories of foreground objects, pi,j is the total number of pixels that originally belong to category i but are classified and predicted to category j.

3.2. Model Parameter Setting

When the model is trained through the experiment, the input image batch size is set to 2 each time, the Cityscapes dataset image and the corresponding label are cut to 1024×512 size, the iterative training is 800 times, the CE loss function is used to calculate the model training loss, and the MLS parameter coefficients α, β and θ are set to 1, 0.4 and 0.8 in turn.

After calculating the loss value of the model, we need to use the loss value to optimize the model parameters. The ultimate goal of optimizing the model is to reduce the loss as much as possible without over fitting. Because the traditional gradient descent algorithm needs to calculate all the samples every time it is updated, which results in more time-consuming, the experimental model uses the stochastic gradient descent (SGD) to update the weights and biases of the network to train the network model parameters, and the SGD uses the input samples to update the parameters every iteration. At the same time, the model is further optimized by ploy learning strategy. The formula of ploy learning strategy is as follows:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

$$LearningRate = lr \times (1 - \frac{i}{e})^p \quad (4)$$

Where lr represents the initial learning rate, i is the number of training iterations, e is the total number of training iterations, and p is the parameter momentum. During the experiment, the momentum is set to 0.9 and the initial learning rate is 0.01.

3.3. Comparative Analysis of Models

In the experiment, Resnet50 model is used as the base net (BaseNet) to extract the multi-level and multi-scale features of the image, but the simple fusion of the extracted features for segmentation and prediction will result in rough segmentation results and low accuracy. In order to get better segmentation accuracy, this paper adds am module and ASPP-AC module of feature extraction based on BaseNet. AM module can make the model pay more attention to the main object feature information and edge information in the image. ASPP-AC module can improve the model to retain the semantic feature information of the image more effectively when the feature fusion. In addition, MLS is also used to optimize the model training. The results of the comparative experiment are shown in Table 1.

Table 1. The performance comparison of each optimization module used in the proposed network structureon the Cityscapes verification set

Module	mIoU/%
BaseNet	38.1
BaseNet+AM	59.3
BaseNet+ASPP-AC	59.7
BaseNet+ASPP-AC+AM	60.6
BaseNet+ASPP-AC+AM+MLS	66.5

Table 2. The performance comparison between this method and other semantic segmentation methods on Cityscapes test set

Category	SegNet [9]	Enet [20]	ESPNet [8]	CRF-RNN [21]	FCN [5]	Ours
road	96.4	96.3	97.0	96.3	97.4	98.6
sidewalk	73.2	74.2	77.5	73.9	78.4	77.8
building	84.0	75.0	76.2	88.2	89.2	89.3
wall	28.4	32.2	35.0	47.6	34.9	40.1
fence	29.0	33.2	36.1	41.3	44.2	45.6
pole	35.7	43.4	45.0	35.2	47.4	50.8
t-light	39.8	34.1	35.6	49.5	60.1	58.9
t-sign	45.1	44.0	46.3	59.7	65.0	67.5
vegetation	87.0	88.6	90.8	90.6	91.4	91.0
terrain	63.8	61.4	63.2	66.1	69.3	67.6
sky	91.8	90.6	92.6	93.5	93.9	94.1
person	62.8	65.5	67.0	70.4	77.1	77.5
rider	42.8	38.4	40.9	34.7	51.4	55.3
car	89.3	90.6	92.3	90.1	92.6	93.7
truck	38.1	36.9	38.1	39.2	35.3	41.2
bus	43.1	50.5	52.5	57.5	48.6	54.3
train	44.1	48.1	50.1	55.4	46.5	49.1
motorcycle	35.8	38.8	41.8	43.9	51.6	52.9
bicycle	51.9	55.4	57.2	54.6	66.8	63.9
mIoU/%	57.0	58.3	60.3	62.5	65.3	66.5

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

According to Table 1, the minimum mIou after the simple base network is divided. After adding AM module and ASPP-AC module, the mIou has increased by 21.2% and 21.6% respectively. After adding two modules, the mIou reaches 60.6%, and the segmentation effect has been improved obviously. The training result of MLS model is 66.5%, and the mIou is increased by 6.5% compared with the model without MLS. The experimental data show that the application of AM and ASPP-AC modules in BaseNet benchmark network can further improve the network performance, which proves that ASPP-AC module and am module can enhance the fusion of multi- scale feature maps at different levels, obtain feature maps containing key information of images, and carry out effective segmentation. MLS can optimize the performance of the model, accelerate the training speed and improve the final segmentation accuracy of the model, which also proves the effectiveness of the proposed method.

Compared with other methods in the field of semantic segmentation: SegNet, ENet [20], ESPNet, CRF-RNN [21] and FCN, the results are shown in Table 2. The segmentation results of SegNet and ENet models are relatively low, both of which are less than 60%. SegNet has better segmentation accuracy than ENet for signal lights, but the segmentation effect of the pedestrian and vehicle is not as good as that of ENet. ESPNet model has a good segmentation effect for large areas in the image, such as streets, sky, buses and so on, but it has a relatively poor segmentation effect for small objects. CRF-RNN model uses conditional random field for back-end optimization processing, but because of the limitation of the feature expression ability of standard convolution network, although it can better recognize the target object, it often lacks consistency in the detail processing of the object boundary. The segmentation result of FCN model has reached 65.3%. The segmentation of large-scale objects in the image is better than that of small-scale objects. As can be seen from Table 2, this method achieves 66.5% in mIou index, and the segmentation effect is better than other methods. It obtains rich and significant semantic features and spatial detail features of each level from the feature extraction module, which makes the image segmentation more detailed for objects with relatively small scale, such as signal lights and signal signs, and obtains a better segmentation effect on the whole.

The comparison of segmentation results between the proposed model and other methods is shown in Figure 5. It can be seen from the experimental comparison results that SegNet semantic segmentation network mainly aims at the classification of pixels, and lacks the overall consistency between adjacent pixels, so the edge of segmentation results is not complete. There are many segmentation errors in ENet segmentation model, such as classifying sidewalks into roads and poles into buildings. The segmentation results of ESPNet, FCN and CRF-RNN models still have some object segmentation category errors. In contrast, this method not only achieves accurate segmentation of large-scale objects such as cars and small-scale objects such as signal signs in the scene image, but also smoothes the edge of segmentation, and has less misclassification of semantic categories of other objects in the image.

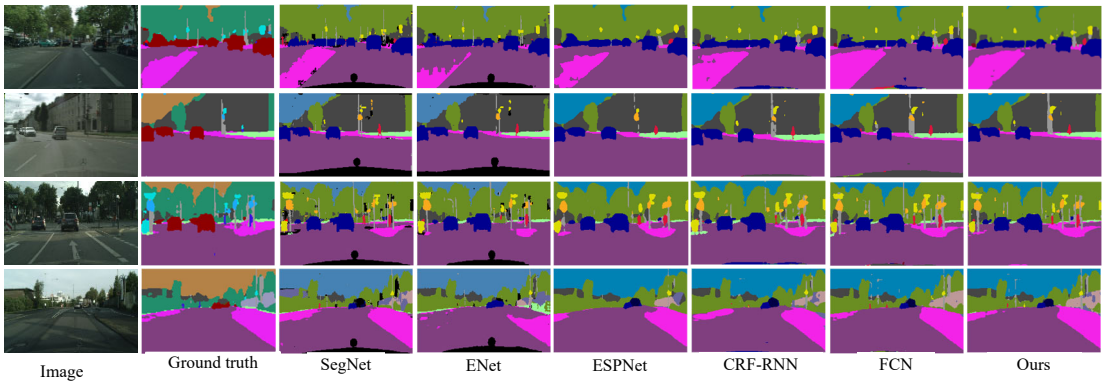


Figure 5. Comparison of segmentation effect between this method and other method

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

4. CONCLUSION

In this paper, a semantic segmentation model of street scene based on multi-scale and attention mechanism is proposed, which can effectively utilize the multi scale feature information of different levels of street scene image and optimize the feature extraction process. Firstly, Resnet50 is used as the backbone network to extract different levels of different scale feature maps of street scene image, and then ASPP-AC module with asymmetric convolution is used to get deep advanced feature map. At the same time, the feature map extracted from the backbone network is optimized by feedback fusion, which not only obtains rich image feature information, but also reduces the parameter calculation of the model to a certain extent. In addition, the AM module is used to make the network focus on the features of each level of image. Finally, all the multi scale feature maps extracted are fused at pixel level to effectively segment the street scene image. By comparing the experimental results with other segmentation network models, this method can further improve the accuracy of street scene image semantic segmentation results and improve the image segmentation effect. It provides a new idea for street scene recognition and classification technology in the field of automatic driving.

REFERENCES

[1] WU Z S,FU W P, HAN G N. Road scene understanding based on deep convolutional neural network. CEA, 2017, 53(22): 8-15.

[2] MA Shuhao, AN Jubai, YU Bo. Improved DeepLabv2 Real-time Image Semantic Segmentation Algorithm. CEA, 2020, 56(18): 157-164.

[3] Choy SK, Shu YL, Yu KW, et al. Fuzzy Model-Based Clus-tering and Its Application in Image Segmentation[J]. Pattern Recognition, 2017, 100(68): 141-157.

[4] TIAN X, WANG L, DING Q. A Survey of Image Semantic Segmentation Algorithms based on Deep Learning[J]. Jour-nalof Software , 2019, 30(2): 440-468.

[5] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39 (4): 640-651.

[6] KUANG Huiyu, WU Junjun. Survey of Image Semantic Segmentation Based on Deep Learning. CEA, 2019, 55(19): 12-21.

[7] Chen LC, Papandreou G, Kokkinos I, et al.Deeplab: Semantic image segmentation with deep convolutional nets,atrous convolution, and fully connected crfs [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4):834-848.

[8] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmenta-tion .In ECCV,2018.

[9] BADRINARAYANAN V, KENDALL A,CIPOLLA R, et al. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[10] RONNEBERGER O, BROX T, FISCHER, et al. U-net: convolutional networks for biomedical image segmenta-tion[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Interven-tion, Munich, Oct5-Oct 9,2015. Charm: Springer, 2015:234-241.

[11] DING C, WENG L G, XIA M, et al. Multi-attention Mecha-nism Network Satellite Image Segmentation Algorithm. CEA, 2021, 57(2): 223-229.

[12] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification[C]//Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Piscataway: IEEE, 2017:6450-6458.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202105_7(5).0030

[13] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing net-work[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, Jul 21-26, 2017.Honolulu: IEEE Computer Society, 2017:2881-2890.

[14] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Re-thinking atrous convolution for semantic image segmenta-tion[J]. arXiv:1706.05587,2017.

[15] Xiaohan Ding ,Yuchen Guo , Guiguang Ding, et al.ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. In CVPR,arXiv preprint arXiv:1908.03930,2019.

[16] LI H,XIONG P,AN J,et al.Pyramid attention network for semantic segmentation[J].arXiv: 1805.10180,2018.

[17] LIU W,WEN Y,YU Z,et al.Large- margin softmax loss for convolutional neural networks[C]//Proceedings of ICML,2016.

[18] Cordts M, Omran M, Ramos S ,et al .The cityscapes dataset for semantic urban scene understanding. In: IEEE conference on computer vision and pattern recognition, Las Vegas, pag-es:3213–3223,2016.

[19] ZHANG L P,ZHANG L F,DU B.Deep learning for remote sensing data: a technical tutorial on the state of the art[J].IEEE Geoscience and Remote Sensing Maga-zine, 2016, 4(2): 22-40.

[20] PASZKE A, CHAURASIA A, KIM S, et al. ENet: a deep neural network architecture for real-time semantic segmenta-tion[EB/OL].[2016-06-07].https://arxiv.org/abs/1606.02147.

[21] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional random fields as recurrent neural net-works .IEEE International Conference on Computer Vision. Santiago: IEEE, pages: 1529–1537, 2015.