International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 7362-7371

https://doi.org/10.70517/ijhsa463631

Research on urban scene image segmentation model based on PSPNet

Yaojie Zhang[1,*] and Yanling Li[1]
[1] Department of Computer Science, Changzhi University, Changzhi, Shanxi, 046011, China
Corresponding authors: (e-mail: zhang_yaojie@163.com).

Abstracts Deep learning-based image segmentation of urban scenes faces the problems of edge blurring and difficulty in distinguishing similar targets in practical applications. In this paper, an improved PSPNet model CBPSPNet incorporating CBAM attention mechanism is proposed to enhance the performance of urban scene segmentation by embedding a hybrid domain attention mechanism in PSPNet. The model combines the channel attention module and spatial attention module to adaptively focus on important features and suppress useless information. The experiments are validated with two datasets, Cityscapes and CamVid, using the SGD optimizer with the base learning rate set to 0.005 and power to 0.5, and 500 epochs of training. The results show that on the Cityscapes dataset, the CBPSPNet outperforms the traditional method on all evaluation metrics, with the range of evaluation metric values reaches 0.7-1, while the traditional method is only 0.6-0.9; it also exhibits faster convergence and lower loss values on the CamVid dataset. Ablation experiments demonstrate that using both average and maximum pooling together is more effective than using them individually. It is shown that the PSPNet model incorporating the CBAM attention mechanism can effectively improve the image segmentation accuracy of urban scenes.

Index Terms Urban scene image segmentation, PSPNet, CBAM attention mechanism, pyramid pooling, semantic segmentation, deep learning

I. Introduction

At present, the field of artificial intelligence is developing rapidly, and various related technologies are growing vigorously, and the technology of computer vision is widely used in various fields. In the medical field, the use of machine vision's big data analysis and high-efficiency processing functions to assist doctors in their work can not only ease the pressure of manual work but also improve a certain degree of accuracy [1], [2]. In the field of computational photography, analyzing and identifying the imaging effect can achieve accurate bokeh effect [3], [4]. And in the field of automatic driving, by analyzing the images of the scene around the car during driving and determining whether there are obstacles on the road and their categories and locations, it provides safer driving services for users [5], [6]. Among them, image segmentation, as a prerequisite, is the basis of image recognition as well as target detection [7].

With the continuous development of automatic driving technology, the status of image segmentation in this field has gradually increased, and the automatic driving technology is more widely used with urban transportation, so improving the accuracy of image segmentation in urban scenes is crucial for assisting the performance and safety of automatic driving systems [8]-[11]. However, in urban street scene images, the complex and changing scenes and the number of target types determine the difficulty of the image segmentation task [12], [13]. In the process of image acquisition, the problem of different scales of the same object and mutual occlusion between objects due to lighting variations, shooting distances and angles, all of the above problems affect the accuracy of image segmentation [14]-[17]. In addition, most of the currently known image segmentation models are more complex, the computational volume is on the large side, and the time consuming for segmenting the image is longer, so they cannot meet the actual needs of the automatic driving vehicle assistance system [18]-[20]. In order to further improve the segmentation accuracy of the network model, so that the segmentation visualization output shows more refined results in urban scenes, the integration of the unique feature learning capability of deep learning technology provides a new solution for road image segmentation [21]-[24].

The rapid development of deep learning technology has driven significant breakthroughs in the field of computer vision, in which semantic segmentation, as an important branch of computer vision, plays a key role in the fields of automatic driving, intelligent monitoring, and medical image analysis. Urban scene image segmentation, as an important application direction of semantic segmentation, faces complex challenges, including light changes, target

occlusion, edge blurring, and difficulty in distinguishing similar targets. Traditional image segmentation techniques based on rules and statistical methods can no longer meet the accuracy requirements of modern urban scene analysis, and deep learning-based methods bring new solutions to this field.

The emergence of full convolutional network (FCN) marks the entry of semantic segmentation into the era of deep learning, and the subsequent emergence of networks such as U-Net, SegNet, and Deeplab series further promotes the development of this field. However, these networks have inherent flaws in processing urban scenes: multiple convolution and pooling operations lead to a decrease in spatial resolution and a gradual loss of detail information, especially in the processing of target boundaries and fine structures.PSPNet mitigates this problem to a certain extent by introducing the pyramid pooling module, which is capable of capturing contextual information at different scales and achieves excellent performance on multiple datasets. However, PSPNet still faces challenges in complex urban environments: it is not precise enough for segmenting targets with blurred edges, is prone to confusion when dealing with different categories of similar appearances, and has limited effectiveness in segmenting small targets. These problems constrain the effectiveness of PSPNet in real urban scene analysis.

Attention mechanism, as a method to simulate the human visual system, can allow the model to adaptively focus on important information, and shows great potential in computer vision tasks.CBAM (Convolutional Block Attention Module), as a lightweight attention mechanism, by combining the channel attention and the spatial attention, it can be used in the computational low computational complexity. Aiming at the shortcomings of PSPNet in complex urban scenarios, this paper proposes CBPSPNet, an improved PSPNet model that incorporates the CBAM attention mechanism. This model maintains the original advantages of PSPNet, and by embedding the hybrid-domain attention mechanism, the network is able to better focus on the important features and inhibit the interference of ineffective information, so as to enhance the segmentation accuracy. In this study, the effectiveness of the proposed method will be verified through experiments on two mainstream urban scene datasets, Cityscapes and CamVid, and the contribution of each component will be deeply analyzed through ablation experiments. Through loss function analysis, comparison of multiple evaluation metrics and detailed analysis of experimental results, the superiority of CBPSPNet in urban scene image segmentation tasks is demonstrated to provide references and support for technology development in related fields.

II. Urban scene image segmentation model design

With more and more networks such as FCN applied to semantic segmentation, deep learning based semantic segmentation networks have been developed rapidly. However, some networks (e.g., U-Net, Deeplab series, etc.) undergo multiple operations such as convolutional pooling, which leads to the gradual loss of image details, thus affecting the accuracy of the results. Among these network models, the PSPNet network model uses ResNet50 as the feature extraction module. And the pyramid pooling module is used to collect information from different regions and different scales, which can alleviate the problem of loss of contextual information in different sub-regions to a certain extent. Therefore, the spatial accuracy is improved. However, in more complex scenes, the effect between target objects is sometimes not so ideal, and the segmentation of fuzzy edge contours and the distinction of similar parts of different target objects can easily cause the problem of information loss or target object segmentation error. In order to solve these problems and improve the segmentation effect in complex urban scenes, this paper designs an improved PSPNet model CBPSPNet (PSPNet), i.e., a deep residual network by incorporating the attention mechanism of CBAM (Convolutional Attention Module) into PSPNet.

II. A.CBAM Attention Mechanism

Attention mechanism is a machine learning data processing method widely used in tasks such as target tracking, semantic segmentation and natural language processing [25]. Attention is the human behavior of giving priority attention to useful information and ignoring secondary, low-value information as shown in actions such as listening, watching and reading. Computers can be set up with sufficient memory capacity to process information equally. However, in the field of computer vision, in order for computers to simulate human attention and thus better and faster processing of practical problems, it is necessary for computers to learn to pay attention to what people care about and ignore things or scenarios that people do not care about.The CBAM attention mechanism is used for this purpose due to its high degree of lightness, versatility, and good effect.The CBAM is used in the CNN network to performs spatial operations and channel sizing, and it contains two main modules: spatial attention module and channel attention module. The specific description is as follows:

II. A. 1) Channel Attention Module

The channel attention module is developed by using the feature relationships between channels to focus the attention on the key elements of a given input, perform global maximum pooling or average pooling on the feature

maps on different channels to obtain the maximum pooled channel attention vector and the average channel attention vector, both of which have a vector size of $C\times1\times1$. These two vectors are then fed into the MLP that shares the weights of only one hidden layer, where the the shape of the weight vector of the hidden layer is $C/r\times1\times1$. Two processed channel attention vectors are obtained, and finally these two vectors are subjected to element-by-element aggregation operation and Sigmoid activation function processing, and the elements are multiplied with the original feature map to obtain a new feature map. The schematic diagram of the channel attention module is shown in Fig. 1 below. The operational equations (1) and (2) are as follows:

$$M_c(F)=\sigma(MLP(AvgPool(f))+MLP(MaxPool(f))) \tag{1}$$

$$M_c(F)=\sigma(W_1(W_0(F_{avg}^c)+W_1(W_0(F_{\max^c}))) \tag{2}$$

where, M_c denotes the 1D channel attention map, σ is the Sigmoid activation function, MLP denotes the multilayer perceptron, W_0 and W_1 denote the weight values of the two layers of the MLP, F_{avg}^c and $F_{\max}^c$ denote the average pooled features and the maximum pooled features on the channel axes, respectively, and r denotes the proportion of the intermediate channel reduction.

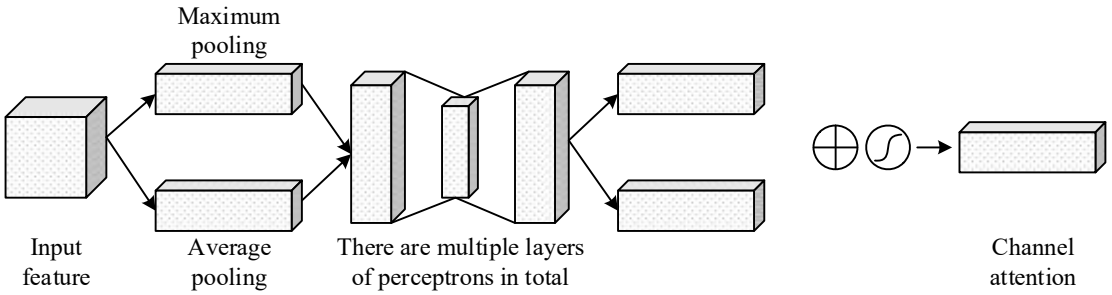


Figure 1: Channel Attention Module

II. A. 2) Spatial attention module

The spatial attention module utilizes attribute relationships between spaces to create spatial attention maps. The spatial module pays attention to the position of key information in the input, performs maximum pooling and average pooling operations on the input feature map, compresses it at the channel level to obtain two 2D feature maps, splices them into a feature map with the number of channels of 2 according to the channel dimensions, and then performs the convolution operation of a single convolutional layer, so that the processed features are consistent with the inputs in the spatial dimension. The spatial feature maps are obtained using Eq. (3) and Eq. (4), and Fig. 2 shows the schematic diagram of the spatial attention module with the following operational equations:

$$M_s(F)=\sigma(f^{7\times7}([AvgPool(F);MaxPool(F)])) \tag{3}$$

$$M_s(F)=\sigma(f^{7\times7}([F_{avg}^s;F_{max}^s])) \tag{4}$$

where, M_s denotes the 2D spatial attention map, σ is the Sigmoid activation function, $f^{7\times7}$ refers to the convolution kernel of size 7 × 7, and F_{avg}^s and $F_{\max}^s$ denote the average pooled and maximum pooled features on the spatial axes, respectively.

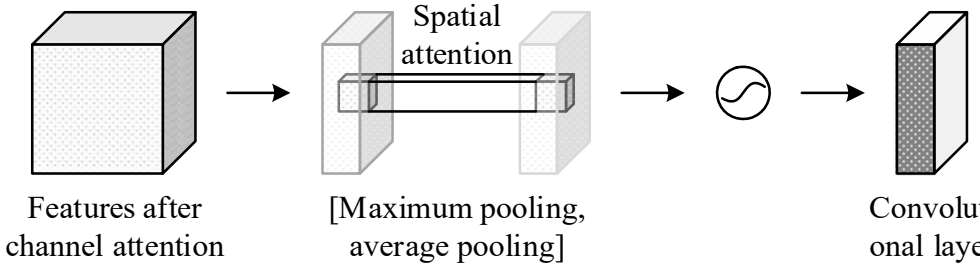


Figure 2: Spatial Attention Module

The CBAM attention mechanism not only tells us where to focus, but also improves the performance of features by focusing on important features in the spatial and channel axes and suppressing unnecessary features [26]. Applying the two modules sequentially can facilitate the individual branches to learn what and where they need to

pay attention to in each of the channel and spatial axes, which in turn helps the flow of information through the modules by learning to emphasize or suppress information. In the channel attention module, average pooling and maximum pooling are used simultaneously, which greatly improves the model's representational capability, compared to using them individually; on the contrary, in the spatial attention module, the average pooling is applied first, followed by the maximum pooling operation along the channel axis, and they are connected to generate efficient feature descriptions. After this, a convolutional layer is applied to generate a spatial attention map that encodes emphasized or suppressed locations.

II. B.CBPSPNet Network Architecture

The CBPSPNet model first acquires a large number of semantic features of the input image using a residual network based on the hybrid domain attention mechanism, and then uses pyramid pooling in the CBPSPNet model to manipulate multiple local features of different sizes, and then reduces the dimensionality of the pooling layer with a 1*1 convolution kernel, and then up-sampling it to the same dimensional size as that of the feature map, and uses global pooling to take a feature channel with weighting computation. After the feature weighted fusion operation, the final output results.

II. B. 1) Pooling module

Previous pooling strategies often ignore part of the image feature information when downsampling is performed at a fixed size. Therefore, the pooling operation process uses the pyramid pooling module, which serves to obtain a variety of contextual feature information of different sizes. Among them, the first layer, also known as global average pooling, aims to obtain the global context information of the target image. Then the second, third, and fourth layers are divided into 2*2, 3*3, and 6*6 sub-regions, respectively, and based on these sub-regions, the global average pooling operation is then performed to obtain local features. For each pooling layer is performed using 1*1 convolution, and then the pooled data of multiple sizes are upsampled to the size of the input image.

II. B. 2) Feature Fusion Module

After up-sampling after the pooling layer in turn from the bottom to the top layer placed on the feature extraction map, and then global pooling of all the feature channels on the feature points to calculate the average of the pooled data and the corresponding channel matrix to expand the number of multiplication, the design of each feature channel weights in turn, calculated all the feature channels have valuable information, and then use 1 * 1 convolution kernel on all the features to expand the sum of the cumulative, and computed sequentially for pixel points, the formula is shown below:

$$y_k = \sum_{i=1}^{N} w_k x_i \tag{5}$$

where, x_i represents the eigenvalue of channel i in the original feature map. w : the weight coefficient. N : the number of channels; y_k : the eigenvalue represented by the k th channel.

II. B. 3) Overall model structure

The CBPSPNet network model uses ResNet50 as the base feature extraction network, contains five convolutional blocks with different structures but the same convolutional operation, and the CBPSPNet network model layers are linked with jump structures. This is intended to address the case of degraded learning performance and gradient explosion in deep networks. In order to suppress the influence of useless features on the model, the network model CBPSPNet is embedded with a hybrid domain attention mechanism on the outside of each convolutional block of ResNet50 while maintaining the original structure of ResNet50 as much as possible. Compared with using the channel attention mechanism and spatial attention mechanism alone, the hybrid domain attention mechanism module combines the spatial attention module and the channel attention module internally, which enables the CBPSPNet network to acquire more useful feature information during image processing, thus making the segmentation effect better. The CBPSPNet network model embedded with the hybrid domain attention mechanism has the main processing flow for the input image:

(1) The channel attention module performs separate pooling operations (including global pooling as well as maximum pooling) on the input feature maps.

(2) The final pooling result is fed into the multilayer perceptron again for cascade summation operation.

(3) Finally, the activation function is used to generate the channel weight coefficients.

(4) Finally, multiply the weight coefficients with the results of channel weighting obtained from the original feature map, and output the feature map.

The formula for calculating the channel weight coefficients is as follows:

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (6)$$

where, M_C represents the channel weight coefficients, MLP is a multilayer perceptron.

II. C.Loss Function and SGD Optimizer

II. C. 1) Loss function

The loss function is a non-negative real-valued function that serves to accurately quantify the difference between the model's predictions and the actual values. The smaller the value of the loss function, the more accurate the model prediction and the more robust the model. During model training, the loss function calculates the difference between the predicted value and the true value for each batch of data and guides the model to update its parameters through a back-propagation algorithm, with the aim of minimizing the loss and improving the prediction accuracy. This process constitutes the core aspect of the machine learning model to learn the features of the dataset.

Commonly used loss functions when performing semantic segmentation tasks include cross entropy loss, soft cross entropy loss and dice loss. The cross-entropy loss function is one of the widely used loss functions in image semantic segmentation tasks. This function calculates the corresponding loss value by comparing the predicted category probability of each pixel in the image with the correct label. Given that the function calculates the average of the losses of all pixels in the image, the model is able to learn unbiasedly for each pixel in the image. The specific mathematical form of the cross-entropy loss function is as public below:

$$L = -\sum_{c=1}^{N} y_c \log(p_c) \quad (7)$$

In the expression, N represents the total number of categories, y_c is a vector of uniquely hot coded labels, and if a category is the same as the actual category of the sample, the corresponding component of y_c is 1, otherwise it is 0. p_c represents the probability that a pixel belongs to a C category in the prediction result. When dealing with an image semantic segmentation task, if there are two categories of objects to be segmented in an image and one of them has a larger percentage of pixels while the other has a smaller percentage, then when using the cross-entropy loss function, the category with the larger percentage will dominate the loss calculation, which may weaken the segmentation performance for the category with the smaller percentage.

II. C. 2) SGD Optimizer

Commonly used optimizers are Stochastic Gradient Descent (SGD), AdaGrad, Adam, and AdamW. Choosing the right optimizer is crucial for the tuning strategy of the network parameters, and the proper optimizer can accelerate the network training process and improve the robustness of the network. Considering that the topic of this research is urban scene segmentation, the SGD optimizer is used in this paper for this purpose. The optimizer is responsible for managing and adjusting the responsibility of trainable parameters in the model in order to make the predicted output of the model closer to the actual labels. During the parameter updating process, the gradient descent method is usually used. This method utilizes the gradient to guide the direction of parameter update. The direction of the gradient is the direction in which the function grows the fastest, so gradient descent varies along the opposite direction of the gradient, resulting in a rapid decrease in the function value, allowing the model to reach the optimization goal more quickly. The two core elements of the optimizer are the learning rate and momentum. The learning rate controls the step size of the parameter update during gradient descent, too large may lead to violent fluctuations in the parameters, making it difficult for the model to converge; too small makes the convergence speed slow and increases the number of iterations. Stochastic gradient descent is a commonly used optimization algorithm in machine learning and deep learning, which is widely used in model training. The core idea of stochastic gradient descent is that in each iteration, it only uses the gradient information of a single sample to update the model parameters, which is one of the most basic first-order optimization methods. Stochastic gradient descent calculates the gradients of the samples in the training set by randomly selecting them and updates the parameters accordingly. Its advantages are computational simplicity, only one sample's gradient needs to be processed at a time, small computational volume, the ability to update parameters according to new samples in real time, and its applicability to online learning scenarios.

III. Validation Analysis of Urban Scene Image Segmentation Model

III. A. Data sets

In order to verify the effectiveness of the PSPNet-based image segmentation model for urban scenes, the dataset Cityscapes and the experimental dataset CamVid are used as the main data sources for the experimental analysis of this research. The detailed description of the dataset is shown below:

III. A. 1) Dataset Cityscapes

Cityscapes is a large-scale urban street view dataset that can be applied to a variety of studies such as semantic segmentation, instance segmentation, panoramic segmentation and 3D vehicle detection. The dataset is collected from the viewpoints of moving vehicles in 50 European cities, spanning the spring, summer, and fall seasons, and is mainly collected on a day-to-day basis under good weather conditions. There is a dedicated official website to manage the dataset, which can be downloaded according to different tasks. The website also has a test set result submission server and a leaderboard to evaluate the researchers' algorithms. The dataset is widely used in semantic segmentation research. For the semantic segmentation task, the whole dataset contains 2500 images. 5000 finely labeled images are divided in the ratio of 6:1:3. A total of 33 categories are included, and 11 categories (buildings, cyclists, cars, fences, sidewalks, poles, pedestrians, roads, traffic signs, sky, trees) are selected, and the pixels of the images are all 1024×2048. In order to ensure the fairness of the evaluation, and to prevent the misuse of the test set to train the model, the labels of the test set are not officially disclosed, and the researchers are required to submit the segmentation results of the test set to the official server for judging. In order to unify the submission format and facilitate the visualization, the official government also provides the processing script of Cityscapes dataset, which specifies the attributes of id, trainId and color for each category, the id attribute can not be changed, and the trainId attribute can be changed according to the needs of the research.

III. A. 2) Dataset CamVid

The dataset CamVid was produced by the University of Cambridge labs and is primarily used for scene understanding tasks in autonomous driving. The dataset is captured from video and contains both daytime and nighttime road scenes. The dataset contains 701 finely labeled images, all with a resolution of 960 × 720.CamVid contains 32 categories, with the proportion of each category varying widely, and only 11 of them tend to be used in image segmentation of urban scenes, which are buildings, cyclists, cars, fences, sidewalks, utility poles, pedestrians, roads, traffic signs, sky, and trees. The images in this dataset were originally divided as a whole by the researcher. Generally the training set contains 367 images, the validation set contains 101 images and the remaining 233 images are used as the test set. There are also studies that do not take this division and use the 421/112/167 division. All images in this dataset are labeled, and the category and color mapping approach is different from Cityscapes.

III. B. Evaluation indicators

III. B. 1) IoU calculation methods

When evaluating the results of image segmentation of urban scenes, the prediction results are usually categorized into four types: true case (TP), false positive case (FP), true negative case (TN), and false negative case (FN) IoU (Intersection and Union Ratio) calculates the ratio of the intersection of the true value X and the predicted value Y to the concatenation of the true and the predicted values and is written as a formula in the following form:

$$IoU = \frac{X \bigcap Y}{X \bigcup Y} = \frac{TP}{TP + FP + FN} \tag{8}$$

For semantic segmentation with multiple categories, its average intersection and merge ratio mIoU is the sum of the ratio of intersection and merge between the predicted results of all categories and the true labels, and then the average of all categories. Assuming a total of N categories, the formula is as follows:

$$mIoU = \frac{1}{N}\sum_{k=1}^{N}\frac{TP_k}{TP_k + FP_k + FN_k} \tag{9}$$

III. B. 2) OA calculation methodology

OA (Overall Accuracy) represents the global accuracy, which is calculated without considering categories and counts the categorization of all samples. It is calculated as in the formula below:

$$OA = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k} \tag{10}$$

III. B. 3) F1 scores

In the classification evaluation index, the accuracy rate indicates the proportion of samples correctly categorized in a particular category to all samples predicted to be in that category, and the completeness rate indicates the proportion of samples correctly categorized in a particular category to the actual samples of that category in the test set. Their calculations are shown in equations (11) and (12), respectively:

$$precision = \frac{1}{N}\sum_{k=1}^{N}\frac{TP_k}{TP_k + FP_k} \tag{11}$$

$$recall = \frac{1}{N}\sum_{k=1}^{N}\frac{TP_k}{TP_k + FN_k} \tag{12}$$

The F1-score (F1-score) is calculated based on the accuracy and completeness of the predicted samples, i.e., it balances the importance of the accuracy and completeness. It is calculated as in equation (13):

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{13}$$

III. C. Experimental environment and parameters

III. C. 1) Experimental environment

All the ablation comparison experiments analyzed in this chapter are based on the Pytorch deep learning framework, which is based on the extension of torch, but the biggest difference between the two is that PyTorch adopts Python as the compilation language. And the biggest feature of PyTorch is that it has powerful GPU acceleration and automatic derivation system, which is very suitable for the development of deep neural networks. Considering that the semantic segmentation task of street view image has a large amount of computation, GPU is also used to realize efficient acceleration operation. The hardware platform for this experiment is as follows: the model of GPU is NVIDIA GeForce 4080Ti, the model of CPU is Intel Core I7-10700K CPU@4.6GHz, 512G SSD and 16G RAM. The software environment for running the experiment: computer operating system Windows 10, Torch vision version 0.9.4, Pytorch version 1.8.4, using CUDA10.1, CUDNN10.1 as the acceleration libraries for GPU.

III. C. 2) Experimental parameterization

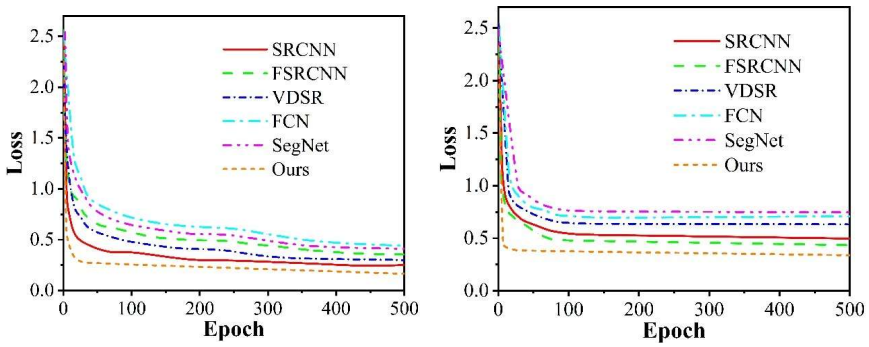
Both the module ablation experiments and the model comparison experiments included in this chapter use the Poly strategy to update the learning rate during the training process and the SGD optimizer to optimize the parameters during the training process. The specific update process during the training process can be expressed as follows:

$$cur\_lr = base\_lr \times \left(1 - \frac{cur\_epoch}{epochs}\right)^{power} \tag{14}$$

In Equation (14), cur_lr represents the current real-time learning rate during this iterative loop, $base\_lr$ represents the set base learning rate, cur_epoch represents the current number of iterative loops, $power$ represents the weights updated for the learning rate, and epochs represents the total number of iterations. In the experimental training, $base\_lr$ is set to 0.005, power is set to 0.5, and epochs is set to 500. Meanwhile, the resolution size of the input image is set to 1024×1024, and $batch\_size$ is 8.

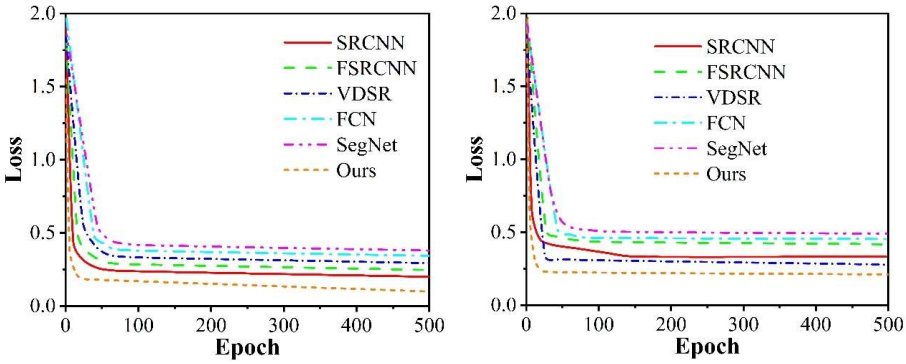
III. D. Comparative analysis of loss functions

Five traditional urban scene image segmentation models (SRCNN, FSRCNN, VDSR, FCN, SegNet) are used as control models to explore the training loss values of different urban scene image segmentation models from two aspects: the dataset Cityscapes and the dataset CamVid. The loss values of different city scene image segmentation models on the dataset Cityscapes are shown in Fig. 3, and the loss values of different city scene image segmentation models on the dataset CamVid are shown in Fig. 4, where (a)~(b) are the training set and the test set, respectively. Based on the loss value curves in Fig. 3(a)~(b), it can be seen that on the dataset Cityscapes, the loss value curves of the five traditional urban scene image segmentation models (SRCNN, FSRCNN, VDSR, FCN, and SegNet) fluctuate more obviously and converge slowly, whereas the fluctuations of the loss value curves of this paper's urban scene image segmentation model are The loss value curve fluctuation is relatively smooth, the convergence speed is rapid, in addition, the loss value is low, which can well meet the research goal of urban scene image segmentation. On the dataset CamVid, the loss value data performance in Fig. 4(a)~(b) shows that the gap between this paper's model and the five traditional urban scene image segmentation models (SRCNN, FSRCNN, VDSR, FCN, and SegNet) is relatively small, but it can be clearly seen that the loss value of the five traditional urban scene image segmentation models converges slowly during the training process. The loss value curve of the urban scene image segmentation model based on PSPNet is smoother and decreases more rapidly, and at this time, the segmentation accuracy reaches the highest and the segmentation performance is the best. In the process of increasing the number of pictures, the urban scene image segmentation model in this paper has a significantly lower loss value and faster descent speed than the five traditional urban scene image segmentation models, which makes the effect of image segmentation further improved.



(a)Training set (b)Test set

Figure 3: Comparison of model loss values on the dataset Cityscapes

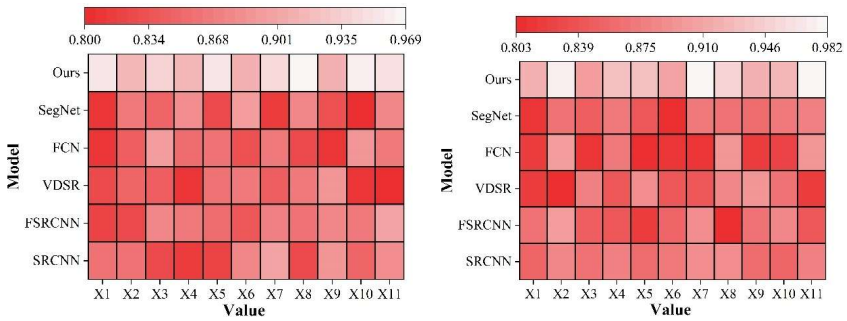


(a)Training set (b)Test set

Figure 4: Comparison of model loss values on the dataset Cityscapes

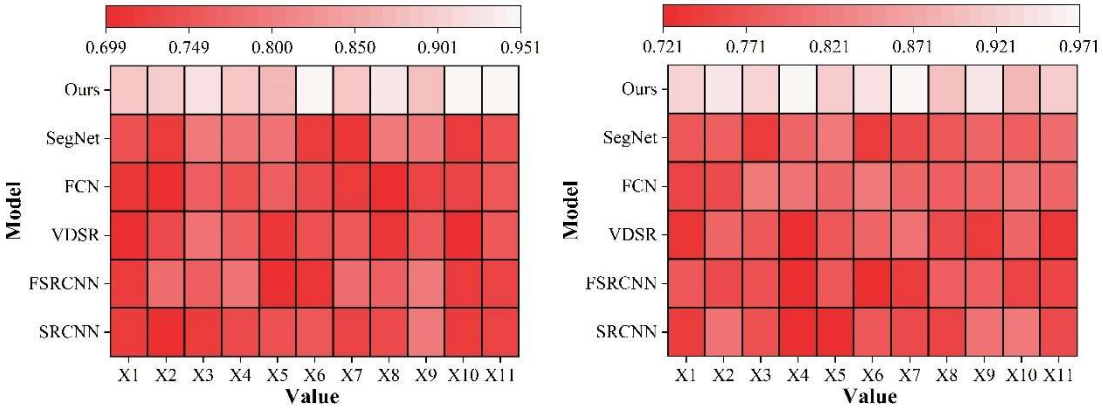
III. E. Comparative analysis of urban scene image segmentation models

In order to verify the feasibility of the urban scene image segmentation model based on PSPNet, the performance of different urban scene image segmentation models was compared and analyzed under the action of three evaluation indicators (IOU: intersection and union ratio, OA: overall accuracy rate, F1). The comparison results of IOU of different models are shown in Figure 5, the comparison results of OA of different models are shown in Figure 6, and the comparison results of F1 of different models are shown in Figure 7. In the figure, X1 to X11 represent buildings, cyclists, cars, fences, sidewalks, utility poles, pedestrians, roads, traffic signs, skies, and trees in sequence. Among them, (a) to (b) represent the dataset Cityscapes and the dataset CamVid respectively. It can be known from Figures 5 to 7 comprehensively that the evaluation index data of the five traditional urban scene image segmentation models are distributed within the range of 0.6 to 0.9, while the numerical range of the evaluation index of the urban scene image segmentation model based on PSPNet is 0.7 to 1. Through the comparison of the magnitudes of the evaluation index values, It can be concluded that the urban scene image segmentation model based on PSPNet has a higher priority. At the same time, it also confirms the real guiding value of the research scheme in this paper and provides a useful reference for the high-quality development of computer vision technology.



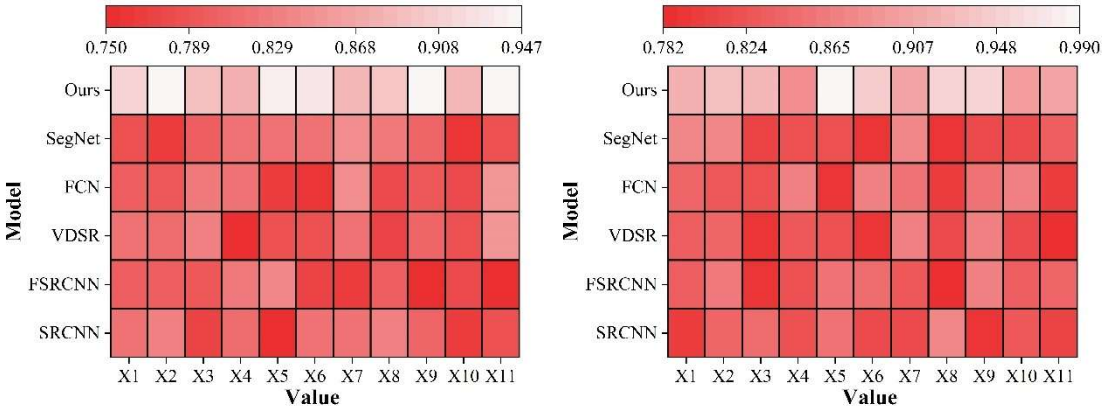
(a)Dataset Cityscapes (b)Dataset CamVid

Figure 5: The IOU comparison results of different models



(a)Dataset Cityscapes (b)Dataset CamVid

Figure 6: The OA comparison results of different models



(a)Dataset Cityscapes (b)Dataset CamVid

Figure 7: The F1 comparison results of different models

IV. Conclusion

Through extensive experiments on Cityscapes and CamVid datasets, the CBPSPNet model proposed in this paper demonstrates significant performance advantages. In loss-value comparison, CBPSPNet shows faster convergence speed and more stable training process on both datasets, and its loss-value curve fluctuation is relatively smooth, while the traditional method has obvious loss-value curve fluctuation and slower convergence speed. In terms of performance metrics evaluation, CBPSPNet achieves excellent results in the three key metrics of IoU, OA, and F1, with each metric reaching the range of 0.7-1 on the Cityscapes dataset, which significantly outperforms the performance range of 0.6-0.9 of the traditional method. On the CamVid dataset, CBPSPNet also demonstrates superior segmentation accuracy, especially in the segmentation of major urban elements such as buildings, cars, and roads.

The experimental results show that the introduction of the CBAM attention mechanism indeed improves the segmentation performance of PSPNet in complex urban scenes. The channel attention module can adaptively adjust the weights of different feature channels so that the model pays more attention to useful feature information; the spatial attention module helps the model to precisely locate the key regions and improve the boundary segmentation accuracy. The synergy of the two modules makes CBPSPNet perform well in dealing with difficult problems such as edge blurring and similar target differentiation. In addition, the combination of the pyramid pooling module and the attention mechanism further enhances the multi-scale feature extraction capability of the model, so that it can maintain good segmentation results when facing targets of different sizes. The CBPSPNet model proposed in this paper provides an effective solution for urban scene image segmentation, which has important theoretical value and practical application prospects.

References

[1] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., ... & Socher, R. (2021). Deep learning-enabled medical computer vision. NPJ digital medicine, 4(1), 5.

[2] Elyan, E., Vuttipittayamongkol, P., Johnston, P., Martin, K., McPherson, K., Moreno-García, C. F., ... & Sarker, M. M. K. (2022). Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward. Artificial Intelligence Surgery, 2(1), 24-45.

[3] VidalMata, R. G., Banerjee, S., RichardWebster, B., Albright, M., Davalos, P., McCloskey, S., ... & Scheirer, W. J. (2020). Bridging the gap between computational photography and visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(12), 4272-4290.

[4] Müller, M., Casser, V., Lahoud, J., Smith, N., & Ghanem, B. (2018). Sim4cv: A photo-realistic simulator for computer vision applications. International Journal of Computer Vision, 126, 902-919.

[5] Kanchana, B., Peiris, R., Perera, D., Jayasinghe, D., & Kasthurirathna, D. (2021, December). Computer vision for autonomous driving. In 2021 3rd international conference on advancements in computing (ICAC) (pp. 175-180). IEEE.

[6] Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. International Journal of Computer Vision, 130(10), 2425-2452.

[7] Ghosh, S., Das, N., Das, I., & Maulik, U. (2019). Understanding deep learning techniques for image segmentation. ACM computing surveys (CSUR), 52(4), 1-35.

[8] Wang, L., Li, R., Wang, D., Duan, C., Wang, T., & Meng, X. (2021). Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sensing, 13(16), 3065.

[9] Sheng, H., Cong, R., Yang, D., Chen, R., Wang, S., & Cui, Z. (2022). UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes. IEEE Transactions on Circuits and Systems for Video Technology, 32(11), 7880-7893.

[10] Dong, G., Yan, Y., Shen, C., & Wang, H. (2020). Real-time high-performance semantic image segmentation of urban street scenes. IEEE Transactions on Intelligent Transportation Systems, 22(6), 3258-3274.

[11] Sun, Y., Zuo, W., Yun, P., Wang, H., & Liu, M. (2020). FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. IEEE Transactions on Automation Science and Engineering, 18(3), 1000-1011.

[12] Zhang, Y., David, P., Foroosh, H., & Gong, B. (2019). A curriculum domain adaptation approach to the semantic segmentation of urban scenes. IEEE transactions on pattern analysis and machine intelligence, 42(8), 1823-1841.

[13] Chen, Y., Li, W., & Van Gool, L. (2018). Road: Reality oriented adaptation for semantic segmentation of urban scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7892-7901).

[14] Papadeas, I., Tsochatzidis, L., Amanatiadis, A., & Pratikakis, I. (2021). Real-time semantic image segmentation with deep learning for autonomous driving: A survey. Applied Sciences, 11(19), 8802.

[15] Muhammad, K., Hussain, T., Ullah, H., Del Ser, J., Rezaei, M., Kumar, N., ... & de Albuquerque, V. H. C. (2022). Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Transactions on Intelligent Transportation Systems, 23(12), 22694-22715.

[16] Ivanovs, M., Ozols, K., Dobrajs, A., & Kadikis, R. (2022). Improving semantic segmentation of urban scenes for self-driving cars with synthetic images. Sensors, 22(6), 2252.

[17] Kaymak, Ç., & Uçar, A. (2019). A brief survey and an application of semantic image segmentation for autonomous driving. Handbook of deep learning applications, 161-200.

[18] Wang, H., Chen, Y., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes. IEEE Transactions on Intelligent Transportation Systems, 23(11), 21405-21417.

[19] Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., ... & Dietmayer, K. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1341-1360.

[20] Song, A. (2023). Deep Learning-Based Semantic Segmentation of Urban Areas Using Heterogeneous Unmanned Aerial Vehicle Datasets. Aerospace, 10(10), 880.

[21] Zhou, J., Hao, M., Zhang, D., Zou, P., & Zhang, W. (2019). Fusion PSPnet image segmentation based method for multi-focus image fusion. IEEE Photonics Journal, 11(6), 1-12.

[22] Jiang, J., & Liu, H. (2025, April). Semantic segmentation algorithm for remote sensing images based on PSPNet. In Third International Conference on Environmental Remote Sensing and Geographic Information Technology (ERSGIT 2024) (Vol. 13565, pp. 91-96). SPIE.

[23] Sun, Y., & Zheng, W. (2023). HRNet-and PSPNet-based multiband semantic segmentation of remote sensing images. Neural Computing and Applications, 35(12), 8667-8675.

[24] Yang, Z., Yu, H., Feng, M., Sun, W., Lin, X., Sun, M., ... & Mian, A. (2020). Small object augmentation of urban scenes for real-time semantic segmentation. IEEE Transactions on Image Processing, 29, 5175-519

[25] Zhao Dongmei,Ji Guoqing & Zeng Shuiguang. (2023). Network security situation assessment based on dual attention mechanism and HHO-ResNeXt. Connection Science,35(1),

[26] Li Runyi,Wang Sen,Wang Zizhou & Zhang Lei. (2021). Breast cancer X-ray image staging: based on efficient net with multi-scale fusion and cbam attention. Journal of Physics: Conference Series,2082(1),