Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 9763027, 1 page
https://doi.org/10.1155/2023/9763027

Retraction

Retracted: A Method of Image Semantic Segmentation Based on PSPNet

Mathematical Problems in Engineering

Received 10 October 2023; Accepted 10 October 2023; Published 11 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] C. Yang and H. Guo, "A Method of Image Semantic Segmentation Based on PSPNet," *Mathematical Problems in Engineering*, vol. 2022, Article ID 8958154, 9 pages, 2022.

Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8958154, 9 pages
https://doi.org/10.1155/2022/8958154

Research Article

A Method of Image Semantic Segmentation Based on PSPNet

Chengzhi Yang and Hongjun Guo

Laboratory of Intelligent Information Processing, Suzhou University, Suzhou 234000, Anhui, China

Correspondence should be addressed to Chengzhi Yang; szxyycz@ahszu.edu.cn

Received 1 July 2022; Accepted 20 July 2022; Published 9 August 2022

Academic Editor: Zaoli Yang

Image semantic segmentation is a visual scene understanding task. The goal is to predict the category label of each pixel in the input image, so as to achieve object segmentation at the pixel level. Semantic segmentation is widely used in automatic driving, robotics, medical image analysis, video surveillance, and other fields. Therefore, improving the effect and accuracy of image semantic segmentation has important theoretical research significance and practical application value. This paper mainly introduces the pyramid scene parsing network PSPNet based on pyramid pooling and proposes a parameter optimization method based on PSPNet model using GPU distributed computing method. Finally, it is compared with other models in the field of semantic segmentation. The experimental results show that the accuracy of the improved PSPNet model in this paper has been significantly improved on Pascal VOC 2012 + 2017 data set.

1. Introduction

Artificial intelligence-oriented modern computer vision technology is widely used in image classification, face recognition, object detection, video analysis, robots, and automobile driving. At present, image semantics segmentation is one of the most popular technologies in the field of computer vision, because the above tasks require intelligent image segmentation to fully understand the content of the image [1]. The core goal of image semantics segmentation task is to segment and classify all the pixels in the image according to the advanced semantics information determined by the model. Semantic segmentation, as the cornerstone of image understanding, plays an important role in VR, navigation, auto-driving related applications, various embedded devices, and unmanned aerial vehicle applications. For example, to achieve indoor auto-navigation, the most important thing is to segment the image actually received in the robot navigation system, to determine the location of various obstacles, and to achieve true intelligence [2, 3]. In the field of auto-driving, through the vehicle camera sensor and the image detected by lidar, when inputting semantics segmentation network, it can automatically identify different targets and intelligently avoid pedestrians and vehicles, so as to drive safely; doctors often need to use the segmentation results of various organ images to give accurate judgment of human tissue lesions. In the field of cosmetic shopping, we can locate the eyes, ears, nose, and throat by accurately segmenting the face with the model and realize functions, such as automatic makeup test and accessories and even help people to automatically test their clothes and free their hands by recognizing the trunk of the human body. It is also important to use remote sensing image segmentation for marine surveying and strategic and tactical investigation in the military field [4, 5].

With the in-depth study of artificial intelligence, convolutional neural network (CNN) has also brought new ideas to the study of image semantic segmentation. Convolutional neural network can imitate the mechanism of human brain and learn abstract features from a large number of data, and then analyze and understand the information of images [6, 7]. In recent years, the scene segmentation algorithm based on convolutional neural network has also made a great breakthrough, realizing the end-to-end learning of the scene segmentation task. Image segmentation algorithm based on convolution neural network is generally a "convolution deconvolution" structure, including full convolution module and deconvolution module [8]. The

convolution module mainly makes the image pass through a series of convolution layers and pooling layers and obtains the deep semantic information of the image and obtains the feature map of the image; the deconvolution module is used to solve the problem that the resolution of the feature image is smaller than that of the original image [9]. After multiple deconvolution operations, the feature image is continuously enlarged to the same size as the original image, and the scene segmentation image with the same resolution of the original image is output. Most of the "convolution deconvolution" algorithms are improved on the basis of full convolution neural network (FCN), which is the first full convolution neural network to realize semantic segmentation [10, 11]. While the full convolution neural network has made a breakthrough, it also has some limitations. The up sampling is too rough, resulting in rough image segmentation results. Therefore, in the field of semantic segmentation, there are U-Net networks that fuse feature maps from the size of the original image to the size of 1/16 of the original image, RefineNet networks that use shallow resolution to obtain high-level semantic features, multiscale ML-CRNN by fusing feature maps of different levels, PSPNet that uses parallel structures to realize multiscale branches through spatial pyramid pooling of different coefficients, and DeformableNet based on adaptive learning. At the same time, the modules of pooling operation, expanding convolution, and pyramid structure also play an important role in the field of image semantic segmentation [12, 13].

In this paper, image semantics segmentation is studied and analyzed. The main research contents are as follows:

(1) The common deep neural network-based semantics segmentation network is described, and the application of attention mechanism in image semantics segmentation is introduced.

(2) The description of the spatial pyramid pooling method is focused, and the framework structure and training process of the pyramid scene analysis network PSPNet are detailed, and the different network models are compared experimentally.

(3) The deep neural network model based on image semantics segmentation is compared through experiments, and the parameters are optimized. The semantics segmentation model with better performance is obtained through comparison.

2. Relevant Research Work

In recent years, in the field of image segmentation, deep convolution neural network has become the mainstream method. It can automatically construct features. But the traditional segmentation technology is still in the leading position by a specific step. Therefore, many researchers choose to integrate the traditional segmentation technology into the deep learning model and combine the traditional methods and deep learning methods to solve the segmentation problem.

Since semantic segmentation is performed at the pixel level, it may lead to over segmentation of a small number of pixels. Therefore, Jungeun et al. [14] proposed a method to improve the accuracy of semantic segmentation network to solve the problem of over segmentation. By defining outliers based on confidence and semantic correlation, pixels are pruned from the segmentation results, so as to improve the accuracy of semantic segmentation. Semantic segmentation based on deep convolution neural network needs a lot of computation and annotation to train data, and heterogeneous image semantic segmentation methods need to classify each pixel. Therefore, Sheu et al. [15] designed a fast heterogeneous image semantic segmentation architecture based on multi hybrid self-coders and decoders to solve this problem and used a discrete autonomous feature extraction framework of RGB images and thermal images with a single convolution layer. Compared with the existing methods, this structure has fewer layers, lower parameters, and faster reasoning speed and has the characteristics of intersection (IOU). Ahn and Kwak [16] proposed the AffinityNet deep neural network, which propagates the local response to the adjacent areas belonging to the same semantic entity to predict the semantic affinity between a pair of adjacent image coordinates and then randomly walk according to the affinity predicted by AffinityNet to achieve semantic propagation. The whole framework only relies on image level class labels without any additional data or comments, which solves the problem of insufficient segmentation labels. Hu and Zhao [17] suggested fusing parallax information in street scene understanding task and taking the structure of parallax coding as the supplementary information of RGB image and designed four methods of summation, multiplication, concatenation, and channel concatenation to introduce them into the semantic segmentation framework. The experimental results verify the effectiveness of parallax information in street scene semantic segmentation task. Jaimes et al. [18] proposed a completely unsupervised semantic segmentation method to solve the problem that deep semantic segmentation networks (DSSNs) are not suitable for the field of label scarcity. They can find an appropriate number of semantic labels without annotation data sets. Once the semantic labels are identified, they can be used to assign semantics to new input images. This method does not need to input any parameters and can reduce the overall position estimation error in UAV positioning management.

Xu et al. [19] optimized the image semantic segmentation model in electronic devices and used cross entropy loss to determine the network structure of the first image semantic segmentation model and the second image semantic segmentation model. This method involves obtaining labeled images and unlabeled images, which improves the accuracy of image semantic segmentation model. Tian et al. [20] proposed a method to obtain the sample image of the enhanced image and fuse it. Through the fused image, the semantic segmentation image is obtained by using the semantic segmentation model. Based on the semantic segmentation image and the semantic segmentation model, a loss function is established to determine the error signal based on the semantic segmentation loss. Yan et al. [21] trained a semantic segmentation model using multiple first fog images and predefined image semantic segmentation

models and applied it to the secondary and tertiary processing of fog images. This method can select attenuation coefficient according to depth to reduce visibility and increase fog density, which improves the efficiency of semantic segmentation of fog images captured by cameras in autonomous vehicle. Yang et al. [22] proposed a semantic segmentation method involving obtaining a first semantic segmentation image corresponding to the first image data. This method can improve the image data of the target environment and improve the semantic segmentation and recognition accuracy of each object. Tao et al. [23] proposed a semantic segmentation model training method for semantic segmentation of images taken at night. The method involves obtaining the first group of labeled images taken in the sun to train the semantic segmentation model, applying the semantic segmentation model to the second group of unlabeled images taken at dusk and then labeling them. The semantic segmentation model is trained by using the first group of labeled images and the second group of labeled images. This method can automatically determine the semantic labels of objects in the image. Kollias [24] proposed a semantic segmentation network training method. This method involves acquiring the real image to collect the simulated image corresponding to the real image, adjusting the parameters of the primary semantic segmentation network according to the difference information between the real image and the simulated image, and establishing the target semantic segmentation network. This method can accurately segment the image and improve the prediction accuracy of the semantic segmentation network. Hong-Gu et al. [25] proposed an image segmentation method based on deep neural network (DNN), which can extract semantic objects with well aligned edges by using image processing technology and DNN and has a good effect on DNN based segmentation. Rao et al. [26] proposed a semantic and difference bidirectional fusion network SDBF net for 3D semantic detection of satellite images, which is composed of three main modules: semantic segmentation module (SSM), stereo matching module (SMM), and fusion module (FM). This method can effectively detect targets in satellite images and generate high-quality segmentation images and more accurately match left and right satellite images to obtain more accurate disparity maps. Its performance is significantly better than the most advanced semantic stereo methods. Liu et al. [27] proposed a generation antagonism network (FISS GAN) for fog image semantic segmentation to solve the problem of difficulty in texture extraction and expression of fog image. The network consists of edge network and semantic segmentation network. Edge GAN is used to generate edge information from fog images and provide auxiliary information for semantic segmentation GAN. Semantic segmentation GAN is used to extract and express the texture of fog image and generate semantic segmentation image. Experiments show that FISS GAN achieves the most advanced performance. Huang et al. [28] designed an image semantic segmentation network framework for joint target detection for complex indoor environment. Using the parallel operation of adding semantic segmentation branches to the target detection network, the multivision task of combining target classification, detection, and semantic segmentation is creatively realized. By designing a new loss function, the idea of transfer learning is used to adjust the training. Finally, the feasibility and effectiveness of the method are verified on the self-built indoor scene data set, which has good robustness.

3. DNN-Based Semantic Segmentation Model

In common semantic segmentation methods, most classifiers can only calculate for a single category. When there are too many categories, it will not only cause a lot of redundancy, but also affect the effect of the model. Since the deep learning based neural networks such as FCN and DeepLabv1 were proposed in 2014, significant progress has been made in the field of image semantic segmentation. It can not only directly predict multiple categories of targets, but also greatly improve the final results. Common semantic segmentation methods include DeepLab series methods, attention mechanism-based methods, and image pyramid based methods. The following mainly describes the above models and compares the effects of different models.

3.1. FCN. The main difference between full convolutional networks (FCN) and convolutional neural networks (CNN) is that FCN replaces the full connection layer in CNN with convolution for operation. Because FCN is operated by full convolution, there is no requirement for the number of neurons in the input layer. The convolution layer with local connection can accept input images of different sizes. At the same time, it does not need that the size of the training image is the same as that of the test image.

In CNN network, pooling operation will reduce the resolution of feature map, which is very effective in tasks related to image classification, because the ultimate goal of these tasks is to find the existence of a specific class, and the location of this class is irrelevant. Therefore, after each convolution block of FCN, a pooling layer will be introduced, so that more prominent and effective features can be extracted in subsequent operations. In FCN-8s, the features of different roughness will be considered at the same time. It is very necessary to make full use of the information of different resolutions generated by the encoder at different stages in semantic segmentation, because it can be used to refine the segmentation effect.

Although FCN has made great progress in the field of semantic segmentation, it still has some defects. For example, after deconvolution by FCN, the detailed information of the image will be lost, and the global consideration of the image will be lacking, and the position information of all pixels will not be fully utilized, resulting in a large difference in the utilization between local features and global features. In addition, the speed and accuracy of FCN cannot meet some real-time segmentation requirements. Therefore, researchers combined with different knowledge and methods in other fields, continuously improved FCN on this basis, and proposed a series of more effective semantic segmentation methods.

3.2. DeepLab Series Methods. To solve the problem that FCN does not consider the global information and lacks spatial consistency, which leads to the segmentation result is not fine enough, DeepLab v1 introduces the concept of hollow convolution, so that it can expand the receptive field and reduce the loss of detail information when pooling is not applicable. At the same time, another innovation of DeepLab v1 is to optimize the final segmentation effect by using fully connected conditional random field (FCCRF).

Subsequently, Chen et al. proposed DeepLab v2 based on previous research on v1. On the original basis, VGG-16 is abandoned and the RESNET model is used, and the problem of different feature targets with different scales is solved by using Atlas spatial pyramid pooling (ASPP). DeepLab v3, which was born later, removed the CRF part and introduced the multi grid strategy by referring to the hybrid divided convolution (HDC). By using different expansion rates for the continuous hole convolution, the continuous segmentation effect can be generated. In this way, the receptive field can also be improved, so that the same or even better effect can be achieved only through multiple convolutions. On the other hand, compared with the ASPP structure in v2 version, the structure in v3 version has been modified with reference to the idea of ParseNet, the input characteristic map has been pooled for global average, and its corresponding normalization layer has been added after each convolution.

In view of the fact that the effect of DeepLab v3 shows that the boundary of its prediction results is not clear, the team proposed DeepLab v3+, which uses the encoder decoder architecture of FPN and other networks for reference, selects the v3 version of the network as the encoder, and adds the decoder to restore the details of the boundary. The overall network structure is shown in Figure 1. In addition, DeepLab v3+ selects XCeption as the backbone and makes improvements on it, changing it into Aligned XCeption to improve the overall segmentation speed and accuracy.

3.3. Methods Based on Attention Mechanism. Attention mechanism means that at a certain moment, the machine only focuses on the recognition of some things and ignores others. Attention mechanism is of great significance in the field of deep learning, because models often need to deal with a large amount of data. When a small part of the data plays a role, attention mechanism can be used. This feature is also feasible in the field of semantic segmentation, such as PSANet. The structure design of convolution kernel in CNN will cause the image information to be trapped in a small area, which makes the model perform poorly in the face of complex scenes. So PSANet connects the locations of feature mapping through attention mask and designs a path for information to spread in both directions, so that the information of each location can act on other locations to assist in the overall prediction. The PSANet network architecture is shown in Figure 2.

In addition, the dual attention network (DANet) can enable the whole network to capture more image context information, so as to improve the overall segmentation effect. The dual attention refers to the position attention module and the channel attention module respectively. Similarly, OCNet also uses self-attention mechanism for research, but pays more attention to selecting a good strategy to better aggregate picture context information, so as to improve the accuracy of overall prediction, that is, an object context pooling (OCP) method.

4. PSPNet Image Semantic Segmentation Algorithm

4.1. Spatial Pyramid Pooling. Spatial pyramid pooling (SPP) is an efficient algorithm proposed by He in 2014. The two greatest advantages of SPP are as follows: (1) using SPP module, you can input any size of pictures without any operations such as cropping and scaling; (2) combined with multiscale information, the accuracy is effectively improved.

In CNN, for the convolution layer and pooling layer, the input image of any size can be convoluted and pooled, but the full connection layer requires that the size of the input image must be consistent. In order to implement the normal training mode, the input image was usually cropped and scaled previously. The SPP module can solve the problem of transition from convolution layer to full connection layer. By using convolution check of different sizes to process a picture, it can be transformed into a multi-scale problem, so as to learn different local details, which is helpful to improve the overall accuracy.

SPP structure usually uses SPP module as the connection between convolution layer characteristic diagram and full connection layer. The input of SPP module is the convoluted feature map, and the output is a fixed size (21 features) neuron. SPP uses three different scales of $1 * 1$, $2 * 2$, and $4 * 4$ to divide the same feature map into 1 block, 4 blocks, and 16 blocks and then calculates the maximum value (or average value) of each block from the 21 blocks that have been divided, so as to obtain a fixed size output neural element.

4.2. PSPNet Image Semantic Segmentation Algorithm

4.2.1. PSPNet Frame Structure. The pyramid scene parsing network PSPNet is a multiscale network. It applies the pyramid pooling module to the field of semantic segmentation, so that it can better learn the global context information of the scene and effectively improve the segmentation accuracy. PSPNet has achieved good results in the current ranking lists of semantic segmentation. PSPNet introduces richer context information into semantic segmentation and obtains a background prior. Compared with FCN, its semantic segmentation error rate is significantly reduced.

In multilayer convolutional neural network, the size of receptive field indirectly determines the degree of using image context information. Although ResNet effectively expands the receptive field through hole convolution and feature map addition, with the deepening of the level and the increase of the network depth, the actual receptive field is still smaller than the theoretical receptive field. The spatial

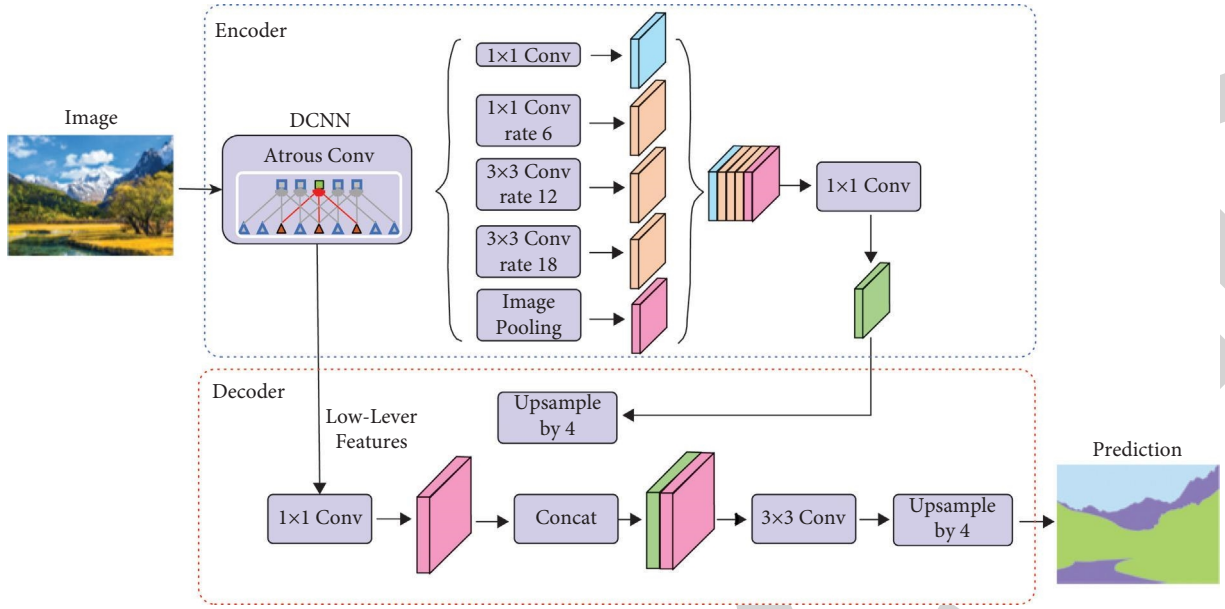

Figure 1: Encoder decoder network structure of DeepLab v3.

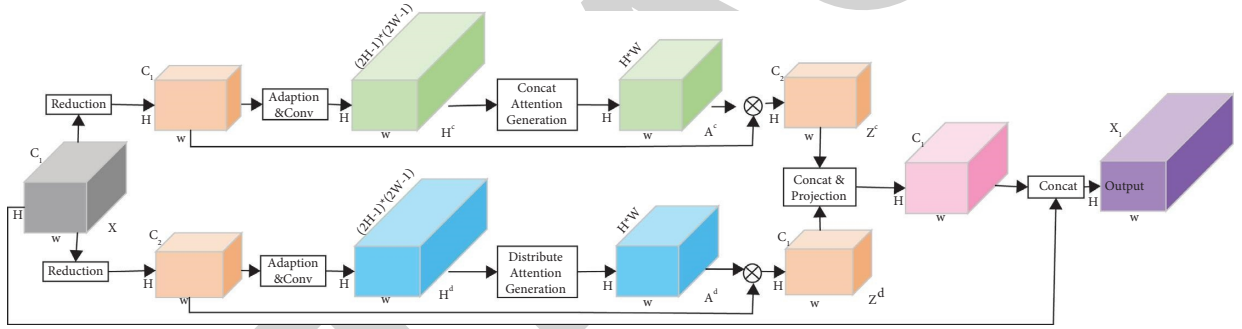

Figure 2: PSANet structure.

pyramid pooling module effectively alleviates this problem by using pooling at different scales and can expand the actual receptive field in the network. PSPNet network effectively makes use of this advantage. Its network structure is shown in Figure 3.

It can be seen from Figure 3 that, first, the feature map is obtained by extracting the input image features through CNN steps. Usually, CNN network based on ResNet structure is used. Secondly, the pyramid adaptive average pooling module is used to capture the features of different subregions at different partition scales, and the subregion features are upsampled to the same size as the global features before pooling, and the CONCAT operation is performed with the global features before pooling, so that the current feature map contains both global and local features, enriching the feature map information. Finally, the final prediction results are obtained by specific convolution and upsampling operations. In Figure 3, the part outlined by the blue dotted line is the core part of the PSPNet structure, that is, the pyramid pooling module. In the pyramid pooling module, four scales of 1×1 (red), 2×2 (yellow), 3×3 (blue), and 6×6 (green) are used for the obtained global feature map to adaptively average pool, which is used as a priori information, and further convolution, batch normalization, and Relu operations are performed in turn to learn model parameters, reduce the dimension of the feature map, and capture the feature information of local sub areas. The learned subregion features with different scales are sampled and fused with the original feature map.

4.2.2. PSPNet Training Process. In the PSPNet network model, the average of the sum of the output errors of all pixels on the sample image is taken as the training error, and the weight parameters of the network are updated according to the method of minimizing the training error.

In the process of updating the network weight parameters by back propagation, the stochastic gradient descent algorithm (SGD) is selected to update the weight through the linear combination of the negative gradient $\nabla L(w_t)$ and the last weight update value. The mathematical formula is as follows:

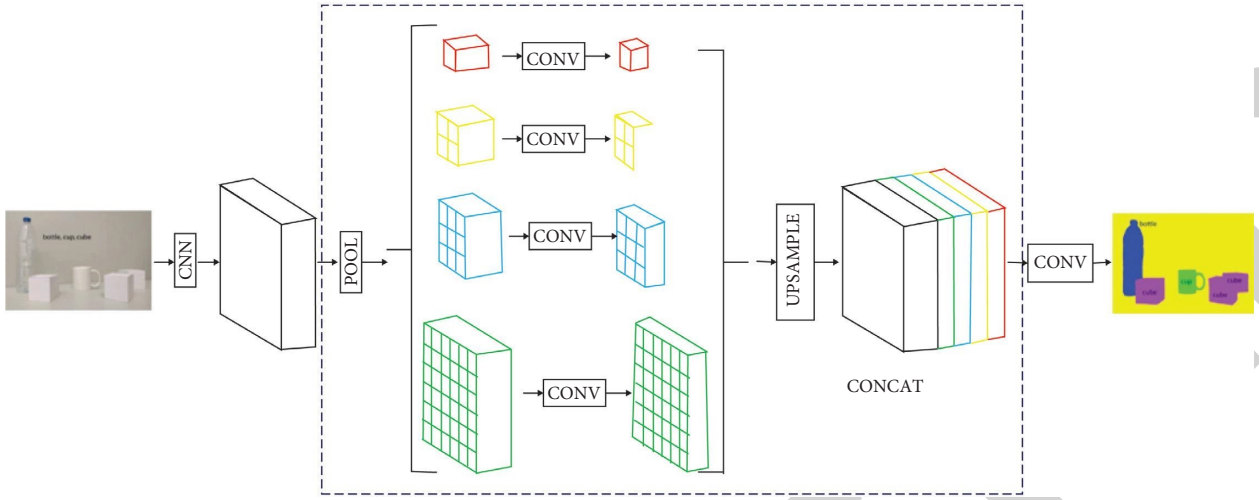

Figure 3: PSPNet network structure.

$$V_{t+1} = \mu V_t - \alpha \nabla L(W_t),$$
$$W_{t+1} = W_t + V_{t+1}. \quad (1)$$

In the formula, w_t is the weight matrix at the t iteration, v_t is the weight update value at the t iteration, and μ is the basic learning rate of the negative gradient, which is used to weight the influence of the previous gradient direction on the current gradient descent direction.

At the same time, in order to accelerate the convergence of the model, the PSPNet network model adjusts the basic learning rate. The mathematical formula of the learning rate (LR) is as follows:

$$LR = \text{base}_{lr} \cdot \left(1 - \frac{i}{i_{\max}}\right)^{\text{power}}, \quad (2)$$

where base_{lr} is the basic learning rate, i is the current number of iterations, $i_{\max}$ is the maximum number of iterations, and power is the learning rate parameter.

PSPNet network is improved on the basis of the original residual network. It is proposed to generate the initial result through another loss function and learn the residual through the final loss function. Therefore, the learning optimization problem of deep neural network is divided into two parts, and each part becomes easier to be solved and optimized. In practice, PSPNet network adds an auxiliary loss function, which can obtain better optimization results.

In addition to the Softmax cross entropy loss calculation for the output of the last layer of the network, the auxiliary Softmax cross entropy is applied to calculate the loss after the fourth stage (RES4b22 residual block). The two losses act on the network in front of them respectively. The main loss function bears the main loss, and the auxiliary loss function is used to optimize the learning process. Finally, a more reasonable total loss is obtained by increasing the weight to balance the auxiliary loss function. Two losses with different weights are propagated simultaneously to jointly optimize the parameters. The loss function of the whole network is

$$\text{Loss}(x) = \alpha L_1(x) + \beta L_2(x), \quad (3)$$

where α, β are the weight of the loss function, respectively, and are set as 1.0 and 0.4 in this paper, respectively. $L_1(x)$ is the main loss function and $L_2(x)$ is the auxiliary loss function.

5. Experimental Process and Result Analysis

5.1. Data Set Used in the Experiment. Pascal VOC is an international computer vision challenge. With the passage of time, the category and number of its data sets are increasing. Many excellent computer vision models are trained based on this data set. Pascal VOC 2012 dataset and Pascal VOC 2007 dataset are widely used. Pascal VOC 2012 + 2007 two in one data set was used in this experiment. The number of categories and labels of both are consistent, which not only makes up for the problem of less data in a single data set, but also improves the problem that voc2012 does not have a corresponding training set.

5.2. Evaluation Indicators. Pixel accuracy (PA), mean accuracy (MA), and mean intersection over union (mIou) are the most commonly used evaluation indicators in the field of semantic segmentation. MIou represents the coincidence degree between the segmentation result and its true value, which is the most representative and most frequently used evaluation index in the field of semantic segmentation.

(i) PA is the ratio between the number of correctly divided pixels and the total number of pixels. The specific calculation formula is as follows:

$$\text{PA} = \frac{\left(\sum_{i=1}^{N} X_{ii}\right)}{\left(\sum_{i=1}^{N} T_i\right)}. \quad (4)$$

(ii) MA represents the average pixel accuracy of all categories of objects, and its specific calculation formula is as follows:

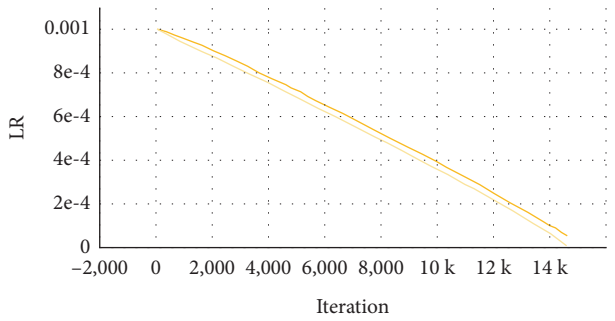

Figure 4: Learning rate curve.

$$\mathrm{MA} = \left(\frac{\sum_{i=1}^{N} X_{ii}/T_i}{N} \right). \tag{5}$$

(iii) MIou indicates the coincidence degree between the segmentation result and the true value of the original image. The specific calculation formula is as follows:

$$\mathrm{mIou} = \frac{\left(\sum_{i}^{N} = X_{ii}/T_i + \sum_{j=1}^{N} \left(X_{ji} - X_{ii} \right) \right)}{N}, \tag{6}$$

where N represents the number of categories of image pixels; and T_i represents the total number of pixels of class i; and X_{ii} represents the total number of pixels with actual type i and prediction type i; and X_{ii} represents the total number of pixels with actual type i and prediction type j.

5.3. Experimental Parameter Setting. First, set the initial parameters of model training: base size = 520, crop size = 480, workers = 4, batch size = 4, epochs = 50, learning rate = 0.0001, and weight decay = $1e-4$.

According to the training results, mIou is only 0.6 under this parameter. The reason for this phenomenon may be that the initial learning rate is too small, the model training is slow, and it is not easy to converge; at the same time, the epoch is also small, which makes the model unable to be trained to the optimal. Subsequently, the parameters were fine tuned for this phenomenon, and it was found that the accuracy was greatly improved.

Set the adjusted parameters of model training: base size = 520, crop size = 480, workers = 4, batch size = 4, epochs = 80, learning rate = 0.001, weight decay = $5e-4$.

5.4. Experimental Results and Analysis. Before the experiment, the training data and test data are normalized, and the Tensorboard visualization tool is used to observe the changes of various parameters in the network with the training process. Figures 4–9 show the learning rate (LR) curve, loss curve, point accuracy curve, mIou curve and their final values of the model during training.

It can be seen Figure 9 from the above figures that PSPNet and PSANet models perform well on this dataset,

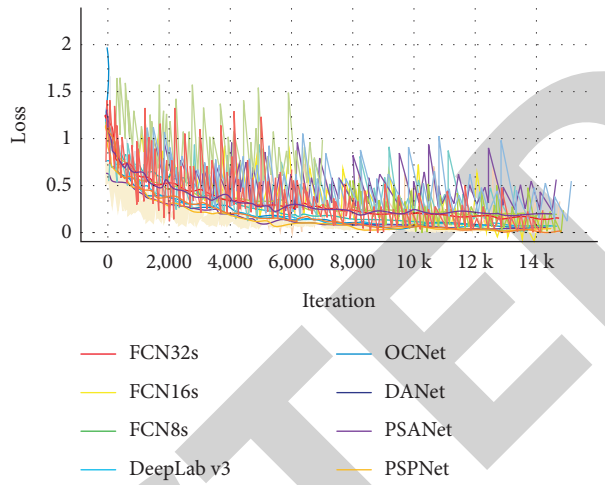

Figure 5: Loss curve and final value.

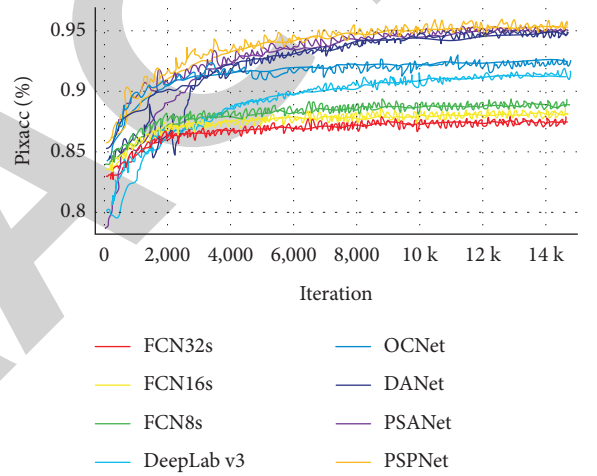

Figure 6: Point accuracy curve and final value (training set).

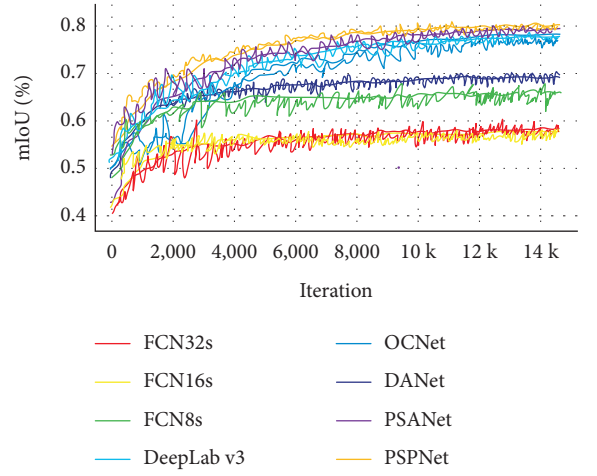

Figure 7: MIou curve and final value (training set).

while the overall accuracy of FCN8s model is significantly improved compared with FCN16s and FCN32s.

It can be found that although the accuracy of DeepLab v3 on the training set has reached 91.4% and mIou has reached

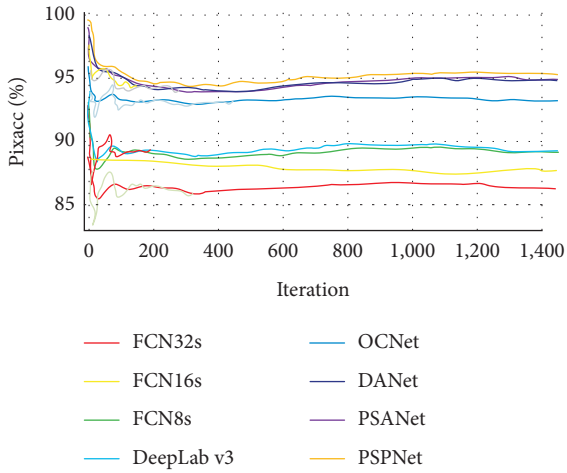

Figure 8: Point accuracy curve and final value (test set).

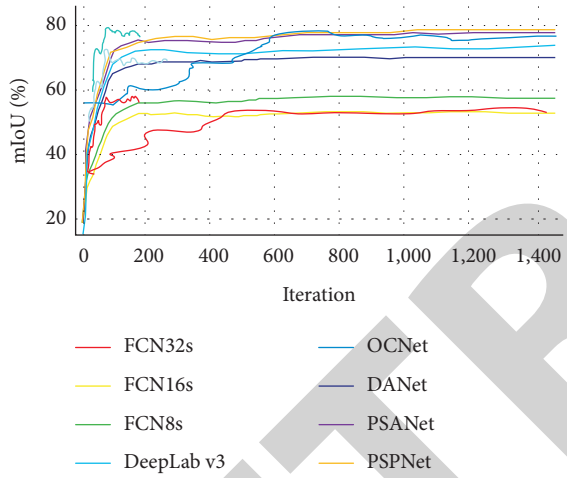

Figure 9: MIou curve and final value (test set).

Table 1: Experimental comparison results.

Model name	MIoU (%)	PA (%)
FCN32s	58.9	87.1
FCN16s	58.3	87.7
FCN8s	66.7	88.5
DeepLab v3	78.0	91.4
PSANet	79.4	94.8
OCNet	78.2	92.3
DANet	70.0	94.6
PSPNet	80.6	95.2

78%, the effect on the test set is not outstanding. The PSPNet after parameter adjustment exceeds other models. Summarize the data obtained from the experiment, as shown in Table 1.

Through comparison, it is found that in the FCN framework, when combining the output of lower layers, the output effect becomes more refined with the decrease of upsampling rate. When the backbone of FCN-8s changed from VGG 16 to ResNet 50, its overall accuracy has also been greatly improved, which proves the effectiveness of ResNet from the side. On the other hand, among the three methods based on attention mechanism, we can find that the segmentation effect of PSANet is better than that of DANet and OCNet. On the whole, PSPNet has the most outstanding effect, with mIou reaching 80.6%, PSANet followed by based on the principle of image pyramid, with mIou reaching 79.4%.

6. Conclusion and Future Work

This paper proposes a multi-layer feature fusion semantic segmentation method PSPNet based on pyramid pooling, which can effectively reduce the parameters of the model. After the image is extracted through the backbone feature extraction network, the effective global context information is obtained through pyramid pooling, and the shallow features of the corresponding size are continuously fused in the decoding process to enrich the information of the feature map. In the process of feature extraction, the channel attention mechanism and spatial attention mechanism are combined to allocate weights to different parts of the feature map, enhance the expression of features, and improve the global perception of features, so as to achieve the purpose of improving the segmentation effect. Experiments show that the proposed PSPNet model can effectively segment images and has good performance in public data sets.

The rise of deep learning promotes the rapid development of computer vision. Although the deep learning semantic segmentation algorithm has solved many segmentation problems, there are still some defects. In the future, the accuracy and speed of the current model can be further improved through optimization and improvement. In addition, in this paper, there is less research on objects with more and more subtle occlusion. In this case, the segmentation accuracy may be affected, and the difference of illumination brightness, multitarget overlap, and so on may cause false recognition. Therefore, the model generalization ability should be further studied.

Data Availability

The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Zhou and H. Zhang, "3SP-Net: semantic segmentation network with stereo image pairs for urban scene parsing," in *Proceedings of the 15th Pacific Rim International Conference on Artificial Intelligence (PRICAI)/15th Pacific Rim Knowledge Acquisition Workshop (PKAW)*, pp. 503–517, China, 2018.

[2] S. Aslan, G. Ciocca, and R. Schettini, "Semantic segmentation of food images for automatic dietary monitoring," in

Proceeding of the 26th IEEE Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, May 2018.

[3] A. Y. Noori, "A survey of RGB-D image semantic segmentation by deep learning," in *Proceeding of the 7th International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, March 2021.

[4] A. Kundu, X. Yin, A. Fathi, D. Brewington, and D. Ross, "Virtual multi-view fusion for 3D semantic segmentation arXiv," pp. 518–535, 2020.

[5] H. Defa and S. Hailiang, "Fusion of infrared and visible images based on nonsubsampled shearlet transform and block compressive sensing sampling," *Ukrainian Journal of Physical Optics*, vol. 18, no. 3, pp. 156–167, 2017.

[6] S. Kazdorf and Z. Pershina, "Algorithm of semantic segmentation of three-dimensional scenes," *Cloud of Science*, vol. 6, no. 3, pp. 451–461, 2019.

[7] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM Computing Surveys*, vol. 52, no. 4, pp. 1–35, 2020.

[8] X. Xu, G. Li, G. Xie, J. Ren, and X. Xie, "Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions," *Complexity*, pp. 1–12, Article ID 9180391, 2019.

[9] A. Kundu, X. Yin, A. Fathi et al., "Virtual multi-view Fusion for 3D semantic segmentation," in *Proceedings of the (computer vision-ECCV 2020. 16th European conference Lecture Notes in Computer Science)*, pp. 518–535, Glasgow USA, August 2020.

[10] S. Kim, L. T. Nguyen, K. Shim, J. Kim, and B. Shim, "Pseudo-label-free weakly supervised semantic segmentation using image masking," vol. 10, , pp. 19401–19411, IEEE Access, 2022.

[11] D. Guo, Y. Pei, K. Zheng, H. Yu, Y. Lu, and S. Wang, "Degraded image semantic segmentation with dense-gram networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 782–795, 2020.

[12] L. Jing, Y. Chen, and Y. Tian, "Coarse-to-Fine semantic segmentation from image-level labels," *IEEE Transactions on Image Processing*, vol. 29, pp. 225–236, 2020.

[13] P. Tuan, "Semantic road segmentation using deep learning," in *Proceedings of the 2020 Applying New Technology in Green Buildings*, pp. 45–48, Da Nang, Vietnam, March 2021.

[14] P. Jungeun, S. Chaewon, and K. Chulyun, "PESSN: precision enhancement method for semantic segmentation network," in *Proceeding of the IEEE International Conference on Big Data and Smart Computing*, pp. 347–350, Kyoto, Japan, February 2019.

[15] M.-H. Sheu, S. M. S. Morsalin, S.-H. Wang, L.-K. Wei, S.-C. Hsia, and C.-Y. Chang, "FHI-unet: faster heterogeneous images semantic segmentation design and edge AI implementation for visible and thermal images processing," vol. 10, pp. 18596–18607, IEEE Access, 2022.

[16] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision forWeakly supervised semantic segmentation," in *Proceeding of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4990, Salt Lake City, UT, 2018.

[17] H. Hu and X. Zhao, "Semantic segmentation of street scenes using disparity information," *10th International Conference on Image and Graphics (ICIG)*, vol. 11901, pp. 169–181, 2019.

[18] B. R. A. Jaimes, J. P. K. Ferreira, and C. L. Castro, "Unsupervised semantic segmentation of aerial images with application to UAV localization," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[19] X. Y. Xu, G. Q. Li, G. Xie, J. Ren, and X. Xie, "Weakly supervised deep semantic segmentation using CNN and ELM with semantic candidate regions," *Complexity*, pp. 1–12, Article ID 9180391, 2019.

[20] L. Tian, X. R. Zhong, and M. Chen, "Semantic segmentation of remote sensing image based on GAN and FCN network model," *Scientific Programming*, vol. 2021, pp. 1–11, Article ID 9491376, 2021.

[21] B. Yan, X. J. Niu, B. Bare, and W. Tan, "Semantic segmentation guided pixel fusion for image retargeting," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 676–687, 2020.

[22] D. W. Yang, Y. Du, H. Yao, and L. Bao, "Image semantic segmentation with hierarchical feature fusion based on deep neural network," *Connection Science*, vol. 34, no. 1, pp. 1772–1784, 2022.

[23] H. Tao and W. H. Li, "Image semantic segmentation based on convolutional neural network and conditional random field," in *Proceeding of the 10th International Conference on Advanced Computational Intelligence (ICACI)*, pp. 568–572, Xiamen, China, March 2018.

[24] S. Kollias, "Image segmentation and classification using semantic analysis," in *Proceeding of the 10th European Congress of Stereology and Image Analysis*, pp. 285–290, Milan, Italy, 2009.

[25] J. Hong-Gu, J. Hyeon-Woo, Y. Byung-Hyun, and C. Kang-Sun, "Image segmentation algorithm for semantic segmentation with sharp boundaries using image processing and deep neural network," *IEEE International Conference on Consumer Electronics - Asia*, p. 4, 2020.

[26] Z. Rao, M. He, Z. Zhu, Y. Dai, and R. He, "SDBF-net: semantic and disparity bidirectional fusion network for 3D semantic detection on incidental satellite images," in *Proceeding of the Annual Summit and Conference of the Asia-Pacific-Signal-and-Information-Processing-Association (APSIPA ASC)*, pp. 438–444, Lanzhou, November 2019.

[27] K. Liu, Z. Ye, H. Guo, D. Cao, L. Chen, and F.-Y. Wang, "FISS GAN: a generative adversarial network for foggy image semantic segmentation," *IEEE-CAA Journal of Automatica Sinica*, vol. 8, no. 8, pp. 1428–1439, 2021.

[28] L. Huang, M. He, C. Tan, D. Jiang, G. Li, and H. Yu, "Jointly network image processing: multi-task image semantic segmentation of indoor scene based on CNN," *IET Image Processing*, vol. 14, no. 15, pp. 3689–3697, 2020.