Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 1420946, 9 pages
https://doi.org/10.1155/2022/1420946

WILEY | Hindawi

Research Article

Image Semantic Space Segmentation Based on Cascaded Feature Fusion and Asymmetric Convolution Module

Xiaojuan Li and Xingmin Ma

System Department Two, North China Institute of Computing Technology (NCI), Beijing 100083, China

Correspondence should be addressed to Xingmin Ma; maxingmin1983@163.com

Received 1 March 2022; Revised 14 March 2022; Accepted 23 March 2022; Published 20 April 2022

Academic Editor: Kalidoss Rajakani

With the rapid development of deep convolutional neural networks, the results of image semantic segmentation are remarkable, and the segmentation effect is greatly improved. The pooling layer of the convolutional neural network will reduce the resolution of the feature map, which makes the convolutional neural network lose a lot of spatial information while extracting semantic features. How to integrate semantic features with semantic information and spatial information will become an important factor to improve the performance of semantic segmentation. Firstly, this paper improves the global attention upsampling module and uses the improved global attention upsampling module to form a multiscale global attention up-mining module in a new connection way. The upsampling module of multiscale attention establishes the relationship between high-level features and lower-level features at a longer distance. Compared with PANet, the method proposed in this paper deepens the close relationship between semantic information and spatial information. Experiments show that the segmentation effect of the feature fusion method based on cascade is better than that of the feature fusion method based on weight. The segmentation effect of the two fusion methods is improved by 8.3% and 5.7% compared with the PANet on the PASCAL VOC 2012 dataset and by 4.5% and 3.6% on the Cityscapes dataset, respectively. The research results of this paper make the high-level semantic information and shallow feature information cooperate to improve the segmentation effect.

1. Introduction

Since the development of the semantic segmentation model based on deep learning, how to make the segmentation model achieve higher accuracy has been paid more attention. Image semantic segmentation is a semantic level segmentation, that is, each image is composed of numerous pixels, and semantic segmentation is to accurately determine which category each pixel belongs to, to classify it into different categories [1]. Before the development of deep learning, the early image semantic segmentation methods mainly include threshold segmentation, region segmentation, and edge detection. However, because these segmentation methods are manually designed to extract features, they are time-consuming and labor-intensive and can only segment the category of the object, and even some methods cannot distinguish the category of the segmentation [2]. In the machine learning era, texton forest and random forest are mainly used as classifiers in semantic segmentation. Later, it was found that the convolutional neural network (CNN) had a good effect on image classification and segmentation, so on this basis, the improved convolutional neural network FCN (fully convolutional networks) became a hot research topic [3].

After multiple convolutions and pooling operations, the low-resolution feature map loses a lot of spatial information but gets some rich semantic information. On the contrary, the low-resolution feature map can retain more spatial information and less semantic information which is conducive to classification after only a few convolutions and pooling operations. The UNet adds a long skip connection in each layer of encoder and decoder to realize information complementation, but the UNet realizes information complementation between decoder and encoder in the same layer and the same resolution feature map. UNet + + uses nested and

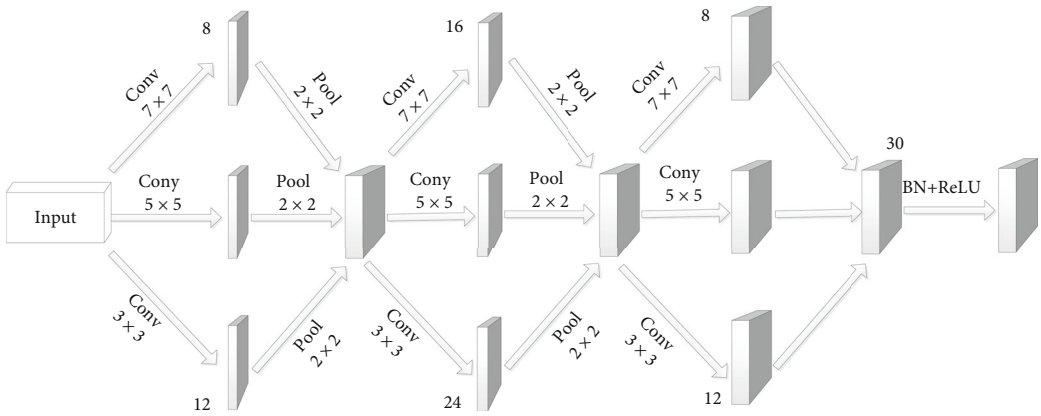


Figure 1: Structure diagram of asymmetric convolution module.

dense skip connections to capture fine-grained information from multiple different high-resolution feature maps at the encoder to recover the lost spatial location feature information at the decoder [4].

In 2014, Long et al. proposed FCN (fully convolutional networks) [5]. Before FCN was proposed, the working principle of the semantic segmentation model constructed by the convolutional neural network was mainly to generate feature maps of different depths through the convolution layer and then to map them into corresponding feature vectors by using the fully connected layer. To limit the size of input data, the model constructed by the FCN network removes the fully connected layer in some feature networks such as VGG16 and GoogLeNet so that the network structure only has a convolution layer. Thus, many feature maps with different depths and sizes can be generated. However, the semantic segmentation model requires the input data size to be consistent with the output data size to recover the size. FCN also uses deconvolution to upsample the deepest feature map to facilitate segmentation [6].

Because of the convolution operation, the resolution of the image is reduced, which is not conducive to dense prediction. To solve this problem, this paper proposes a multiscale global attention upsampling block (MGAU). The MGAU is formed by connecting GAU modules in a new connection mode. This method constructs a new semantic segmentation model with MGAU module and asymmetric convolution block (ACB). An indirect guidance is proposed, that is, the higher level features weight the lower level features at a longer distance through the channel weights. PANet does not consider whether indirect guidance can improve semantic information and spatial information. Experiments show that the use of spatial attention can greatly improve the effect of semantic segmentation.

The main innovations of this paper are the following:

(1) An image semantic segmentation method based on asymmetric convolution and multiscale global attention upsampling is proposed

(2) Use asymmetric convolution to improve the expression ability of the backbone network, and use high-level semantic features to directly and indirectly guide low-level feature maps at the decoder to strengthen the relationship between semantic information and spatial information

(3) Two fusion methods are used in this paper: one is to give different weights to different feature maps, and the other is the cascade feature fusion method

2. Related Work

2.1. Asymmetric Convolution Module. In 2019, Ding et al. proposed an asymmetric convolution module in ACNet. ACB uses one-dimensional asymmetric convolution to enhance the expressive power of square convolution to replace the square convolution layer. At the same time, the ACB module can also enhance the robustness of the model [7]. In 2020, Li et al. [8] improved the ACB module in MACU-Net as an independent module so that the ACB can be combined with any structure without adjustment. The structure of ACB is simple and does not require a lot of additional computation. Because the ACB module has stronger expressive power than the square convolution kernel and has the idea of multiscale feature extraction, it can extract more context features [9]. Therefore, the asymmetric convolution block is used in this method. And the effectiveness of the ACB module has been verified in many fields, such as image denoising, high-resolution remote sensing image segmentation, and medical image segmentation. The asymmetric convolution module structure is shown in Figure 1.

In Figure 1, the ACB module can extract feature information from different directions from horizontal, vertical, and rectangular convolution kernels, thus enhancing the ability to extract feature information. The ACB module is composed of three convolution kernels of different sizes and shapes, which is similar to the idea of multiscale feature extraction [10]. The sizes of the convolution kernels are $d \times d$, $d \times 1$, and $d \times d$, respectively. $d \times d$ is a square convolution kernel. $d \times 1$ is a vertical convolution kernel. $1 \times d$is a horizontal convolution kernel. Three convolution kernels with different shapes can obtain the spatial feature information in the rectangular area, horizontal direction, and vertical direction, respectively [11]. After the convolution operation, the elements of the three branches are added so that the output feature map can fuse the feature information extracted from the input

feature map with convolution kernels of different shapes and directions. Finally, the fused feature map is normalized by batch processing, and the nonlinear activation function ReLU is used to improve the expression ability of the feature map [12]. The value of d chosen in the method of this paper is 3. See formulas (1) and (2) for the ACB module in this method.

$$x_i = F_{3\times3}(x_{i-1}) + F_{1\times3}(x_{i-1}) + F_{3\times1}(x_{i-1}), \tag{1}$$

$$x_i = \mathrm{ReLU}(\mathrm{BN}(F_n)), \tag{2}$$

where x_i is the input feature diagram of ACB module and x_{i-1} is the output feature diagram of ACB module. $F_{3\times3}$, $F_{3\times1}$, and $F_{1\times3}$ represent 3×3, 3×1, and 1×3 convolution kernel, respectively. And ReLU$(\cdot)$ and BN$(\cdot)$ represent activation function ReLU and batch normalization operation, respectively .

2.2. Global Attention Upsampling Module. The structure of the global upsampling module has two inputs and one output. The two inputs are high-level semantic features $X \in R^{H\times W\times C}$ and low-level semantic features $Y \in R^{H'\times W'\times C'}$, where H, W, and C are the width, height, and channel number of the high-level semantic feature X. H', W', and C' are the width, height, and channel number of the low-level semantic feature Y. The semantic information of high semantic feature X is richer than that of low semantic feature Y, which is shown as $C > C'$; but its spatial location information is largely lost in downsampling, which is shown as the resolution of X, which is smaller than that of Y, i.e., $H < H'$ and $W < W'$ [13]. The idea of the global upsampling model is to use the high-level semantic features with rich semantic information to guide the low-level semantic features with a large amount of spatial information in the upsampling process. The implementation steps are as follows:

(1) Use channel attention to calculate the channel weight of high-level semantic feature X, whose dimension is $1\times1\times C$

(2) The number of channel weights of the high-level semantic features is reduced to be the same as the number of channels of the lower-level semantic features through a convolution operation, that is, $C \longrightarrow C'$

(3) Multiply the low semantic feature Y with the channel weight $1\times1\times C'$ to obtain a new feature $H'\times W'\times C'$

(4) The output feature out is the element-wise addition of the high-level semantic feature and the new feature map generated in the third step after N times of upsampling operation [14]. The above process is shown as

$$\mathrm{out} = \mathrm{Conv1}\times1(\mathrm{Attention}((X)) \times \mathrm{Conv3}\times3(Y) + \mathrm{Upsample}(X), \tag{3}$$

where Conv1 × 1$(\cdot)$ and Conv3 × 3$(\cdot)$ denote the convolution operation 1×1 and the convolution operation 3×3, respectively. Conv3 × 3$(\cdot)$ contains convolution BN and ReLU. Attention$(\cdot)$ represents attention operations, and the attention operations in the global attention upsampling module are implemented using global average pooling. Upsample$(\cdot)$ indicates the upsampling operation.

3. Improvement of Global Attention Upsampling Module

Although the global upsampling module in the PANet network has a good effect in image semantic segmentation, it is proved in CBAM (convolutional block attention module) that global average pooling is not the optimal choice for channel attention. Global maximum pooling can extract some unique features of objects and infer more meaningful feature information [15]. Therefore, CBAM uses both global average pooling and global maximum pooling to improve the expressiveness of the network [16]. Inspired by CBAM, this paper uses global average pooling and global maximum pooling to replace global average pooling in the global upsampling module, as shown in Figure 2.

In this paper, the ablation experiment proves that the improved global attention upsampling module performs better than the global attention upsampling module proposed in PANet, which indicates that the improved global attention upsampling module is effective.

4. Multiscale Global Attention Upsampling Module

Two different feature fusion methods are proposed in this paper. One is the feature fusion of multiple features of De2 and De3 according to different weights. Feature fusion is performed by cascading multiple features of De2 and De3. The two different fusion methods are explained below.

4.1. Feature Fusion Based on Weight. To facilitate the understanding of feature fusion with different weights in MGAU, the multifeature graph aggregation module in the MGAU module in Figure 3 was rebuilt in the way of feature fusion with different weights.

In the MGAU module, W represents the weight, such as W_{52} represents the weight of the feature map when the new feature generated by GAU5 is fused on the DE2 layer [17]. De2 is the second layer of the decoder. ACB3, ACB4, and ACB5 are the high-level semantic features, and ACB2 is the low-level feature. ACB3, ACB4, and ACB5 generate their own new feature maps F_{32}, F_{42}, and F_{52} through the GAU module and ACB2, respectively. Since different high-level features have different influence on the same low-level feature, that is, the information fusion capabilities of indirect guidance and direct guidance of the same low-level feature map are different. Different weights W_{i2} are given to the newly generated feature map $\{F_{i2}\}_{i=3,4,5}$ during feature fusion at the De2 level, and the formula is shown as

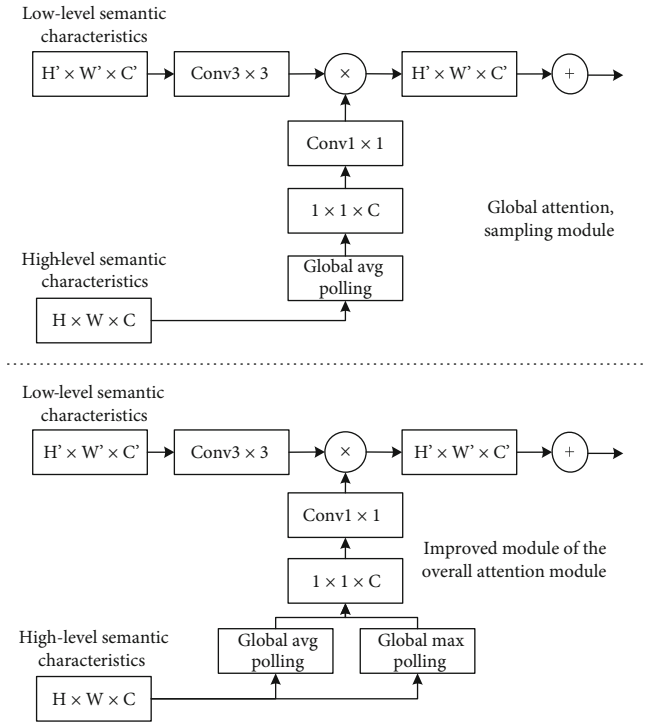


Figure 2: Global attention with its improved module structure diagram.

$$F_j = \sum_{i=j+1}^{5} W_{ij} \times \mathrm{GAU}(X_i, Y_j) = \sum_{i=j+1}^{5} W_{ij} \times F_{ij}, \quad (4)$$

where GAU(·) denotes the global attention upsampling module. X and Y represent high and low semantic features, respectively. W_{ij} represents the weight of the new feature map F_{ij} generated by the GAU module at the *j th* layer of the decoder. F_{ij} represents the new feature map generated by the GAU module. F_j represents the feature map generated by weight concatenation at the *j th* layer of the decoder [18].

4.2. Feature Fusion Based on Cascade. Figure 4 simplifies the feature extraction content in Figure 3 but refines the content in the MGAU model to facilitate the understanding of the cascading method fusion feature graph.

F_{ij} in the MGAU represents the feature map generated by GAU module for high-level semantic features of layer i and low-level features of layer j [19]. The MFF module is used to aggregate multiple new feature graphs generated on a certain layer of the decoder side. The specific structure is shown in Figure 5.

In Figure 5, the input of the MFF module of the De2 layer at the decoder side are F_{32}, F_{42}, and F_{52}, which are, respectively, the feature maps of ACB3, ACB4, and ACB5 after weighted guidance of ACB2 through channel attention. F_{42} and F_{52} are the high-level feature map ACB4, ACB5 indirectly guide the low-level feature map ACB2, and F_{32} is the high-level feature map ACB3 that directly guides the low-level feature map ACB2 [20]. The feature aggregation steps on the De2 layer are the following:

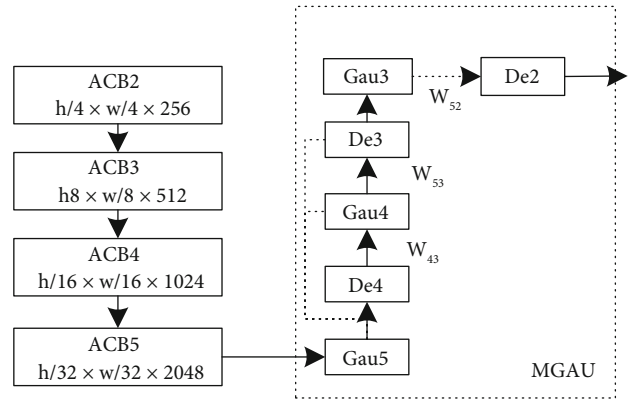


Figure 3: Structure diagram of feature fusion method based on weight.

(1) The feature map F_{32} generated by direct guidance and the feature maps F_{42} and F_{52} generated by indirect guidance are cascaded, and the channel number of the cascaded feature map is 256×4

(2) A new feature map F_2' is obtain by reducing that numb of channels 256×4 from 256 layers through a convolution kernel 1×1

(3) Pass the feature map F_2' through the spatial attention module to generate a new feature map

5. Experiment and Analysis

5.1. Experimental Dataset. In this paper, two datasets, PASCAL VOC 2012 and Cityscapes, which are commonly used

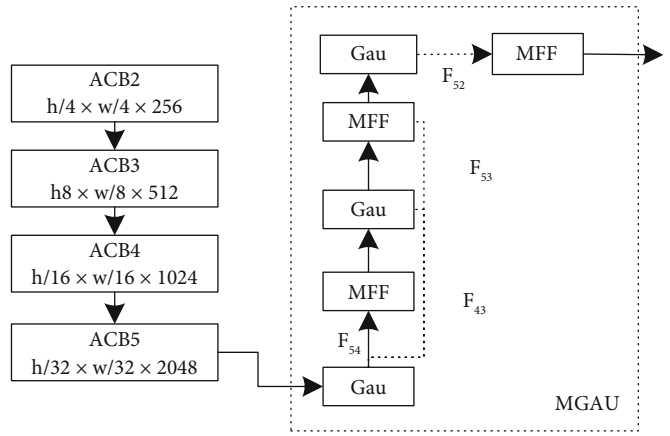


Figure 4: Structure diagram of feature fusion method based on cascade.

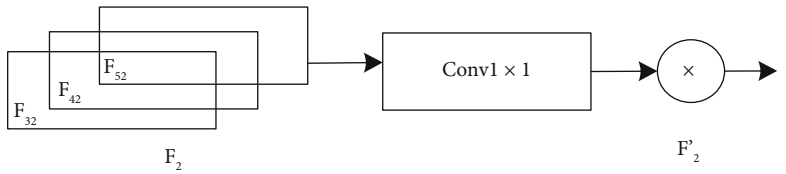


Figure 5: Multifeature fusion module of De2 layer.

Table 1: Hardware and software experimental environment settings.

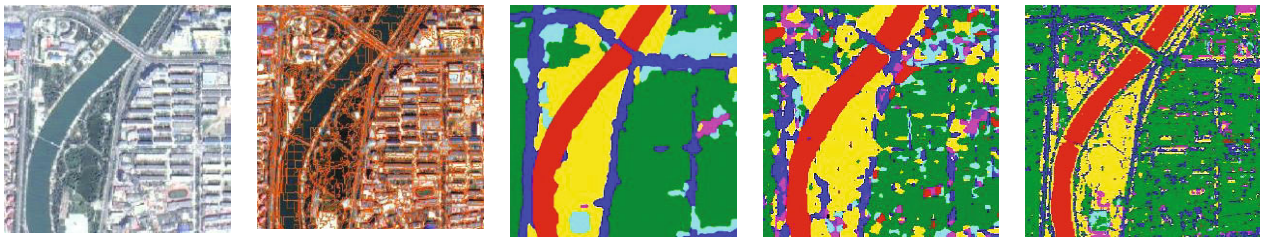
Experimental environment	Tool configuration
CPU	Ntel (R) I7-10900K 10-core/20-thread CPU @ 3.7 GHz
Running memory	32G
GPU	Quadro P2000
Video memory	5G
System type	64 bit
Operating system	Ubuntu16.04
Programming language	Python
Deep learning framework	PyTorch

Table 2: Dataset performance comparison.

Number	Method	ResNet101	Dilated ResNet101	ACB	FPA	GAU	MGAU	MIoU (%)
1	PANet		√			√		78.52
2	PANet		√		√	√		78.93
3	Our-weight		√				√	78.76
4	Our-weight		√	√			√	80.21
5	Our-weight	√		√			√	80.33
6	Our-connect		√				√	79.93
7	Our-connect		√	√			√	79.98

Table 3: Comparison of different semantic categories of IoU.

Category	FCN8s (%)	UNet (%)	PANet (%)	Our-connect (%)	Our-weight (%)
A1	76.8	66.1	87.49	88.69	87.78
A2	34.2	31.7	83.49	85.45	84.89
A3	68.9	49.2	89.58	90.82	89.88
A4	49.4	35.6	71.45	70.83	72.40
A5	60.3	39.2	70.84	75.74	73.04
A6	75.3	70.3	93.12	93.33	93.04
A7	74.7	62.4	78.03	77.55	80.19
A8	77.6	60.6	92.71	94.19	92.67
A9	21.4	16.6	48.71	47.20	47.52
A10	62.5	41.5	83.42	89.94	87.65
A11	46.8	32.5	56.10	57.69	52.93
A12	71.8	50.3	88.35	88.76	89.25
A13	63.9	48.3	82.17	84.28	86.85
A14	76.5	65.0	90.37	86.49	87.31
A15	73.9	67.8	88.39	88.76	88.81
A16	45.2	34.5	61.83	66.15	63.39
A17	72.4	54.3	85.93	88.48	89.96
A18	37.4	27.2	50.63	54.06	51.20
A19	70.9	64.6	86.55	86.48	86.47
A20	62.2	49.9	79.35	80.34	80.06



(a) Original picture (b) Real value map (c) PANet (d) Our-weight (e) Our connect

Figure 6: PASCAL VOC 2012 AUG prediction results.

in the field of image semantic segmentation, are selected for evaluation.

The PASCAL VOC 2012 dataset is a general dataset for training and testing models or methods in the fields of image classification, segmentation, and object detection. The PASCALVOC2012 has a training set, a test set, and a validation set in the field of image segmentation, including 1464 images in the training set, 1449 images in the test set, and 1456 images in the validation set [21].

Cityscapes are semantically understood pictures for city street scenes. It is a high-resolution dataset and a commonly used dataset in the field of semantic segmentation. The image resolution of the Cityscapes dataset is 1024×2048. It has 5000 manually annotated images in the semantic segmentation domain, including 2975 images in the training set, 1525 images in the test set, and 500 images in a validation set. These images have 19 semantic categories [22].

5.2. Evaluation Standard. In the field of image semantic segmentation, there are four commonly used evaluation indexes, namely, pixel accuracy (PA), mean pixel accuracy (MPA), intersection over union (IoU), and mean intersection over union (MIoU). PA represents the ratio of the number of correctly classified pixels to the number of all pixels in the image, which is the simplest evaluation method, and its formula is shown as

Table 4: Performance comparison of Cityscapes datasets.

Methods	MIoU()
FCN-8 s	62.56
UNet	58.69
PANet	72.55
Our-weight	76.36
Our-connect	74.56

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} N_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} N_{ij}}. \tag{5}$$

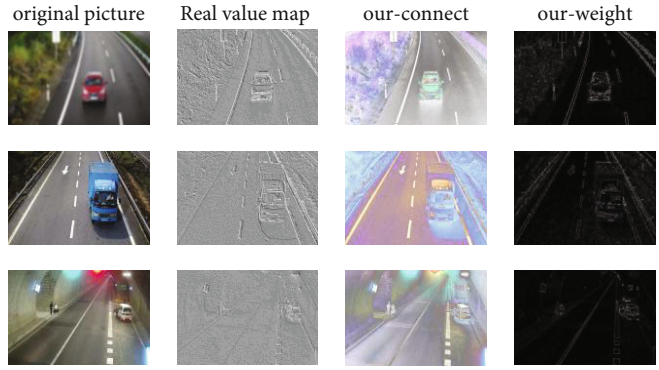


Figure 7: Prediction results of Cityscapes.

MPA is the pixel calculation accuracy of each semantic class and represents the ratio of the number of correctly predicted pixels of different semantic categories to the total number of pixels of the semantic category. The formula is shown as

$$\mathrm{MPA} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{N_{ii}}{\sum_{j=0}^{k}N_{ij}}. \tag{6}$$

IoU represents the ratio of the intersection and union of the two sets of labeled categories and predicted categories in the truth graph, and its formula is shown as

$$\mathrm{IoU} = \frac{\sum_{i=0}^{k}N_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}N_{ij} + \sum_{i=0}^{k}\sum_{j=0}^{k}N_{ji} - \sum_{i=0}^{k}N_{ii}}. \tag{7}$$

MIoU is a subdivision of IoU, which represents the IoU of each class and then averages it as

$$\mathrm{MIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{N_{ii}}{\sum_{j=0}^{k}N_{ij} + \sum_{j=0}^{k}N_{ji} - N_{ji}}. \tag{8}$$

$k+1$ in formula (8) represents the number of all semantic categories in the dataset, including k object categories and 1 background category. N represents the number of pixels in the picture. For example, N_{ij} represents N_{ij} pixel predicted to be in the *j th* class but actually belongs to the *i th* class. MIoU is a standard metric used to evaluate the segmentation effect of semantic segmentation tasks. MIoU and PA are used as the evaluation methods in this paper.

5.3. Experimental Setup. Table 1 shows the software and hardware environment settings required by the experiment.

The poly learning rate strategy was selected in the experiment, and its formula is shown as

$$lr = init\_rate \times \left(1 - \frac{iter}{\max\_iter}\right)^{power}. \tag{9}$$

lr represents the learning rate. *init_rate* represents the initial learning rate. *iter* and max_*iter* represent the current iteration number and the maximum iteration number, respectively. *power* is the momentum. In the PASCAL VOC 2012 dataset, the setting of *init_rate* was 4 e^{-3}, and *p ower* was set to 0.9. A stochastic gradient descent optimizer was used, a batch size of 8 was employed, and the weight decay was 0.0001. 100 rounds were iterated per training at the time of the ablation experiment and 150 rounds at the time of the formal experiment. Set to $1e^{-4}$ in the Cityscapes dataset *init_rate*. Use a stochastic gradient descent optimizer by employing a batch size of 8, a weight decay of 5 e^{-4}, a momentum of 0.99, and 200 iterations per experiment.

5.4. PASCAL VOC 2012. The PASCAL VOC 2012 AUG dataset was used for the experiments, and the resolution was set to 512×512. In the feature fusion method based on weight, the optimal weight ratio is W_{53}, W_{43}, W_{52}, W_{42}, $W_{32} = (0.3, 0.7, 0.1, 0.2, 0.7)$. Our-weight means that the feature fusion method based on weight is used in the MFF module. Our-connect means that feature fusion is performed in the MFF module using the cascading-based feature fusion method. Dilated ResNet101 denotes a ResNet with dilated convolution.

The method proposed in this paper uses two different ways to fuse the MFF feature map. The feature fusion method based on weight and the feature fusion method based on cascade will be compared with the PANet segmentation method, respectively. The experimental results are shown in Table 2.

In Table 2, it can be seen from groups 1, 3, and 6 that when the ResNet101 with dilated convolution is used as the backbone network, the MGAU module has a better segmentation effect than the GAU module whether the feature map is fused using the weight-based method or the cascade-based method in the MFF module. The second, fourth, and seventh experiments show that the semantic segmentation method using ACB + MGAU is better than the PANet. And while using the ResNet101 as the backbone network, the effect is better than the segmentation effect of the ResNet101 with dilated convolution.

Table 3 describes the IoU and MIoU of some current methods and the methods proposed in this paper in each category. It can be seen that the cascade and weight methods proposed in this paper are higher than other methods in 14 and 13 of the 20 semantic classes, respectively.

The segmentation effect diagram of this method is shown in Figure 6.

Our-weight and our-connect represent the feature fusion method based on weight and the feature fusion method based on cascade used in the feature fusion stage.

5.5. Cityscapes. The Cityscapes dataset is a high-resolution image of 1024×2048. Too much computing resources are needed to input the original resolution into the model for prediction. Due to the limited video memory of the experimental equipment, the input image is cropped into pixels 512×512 in the experiment. The comparison between the experimental results of this method in the validation set of Cityscapes and the experimental results of the existing semantic segmentation methods is shown in Table 4.

In the experiment, the proposed method does not use any other techniques or auxiliary datasets to improve segmentation results. Figure 7 is the segmentation effect diagram obtained by the method in this paper on the verification set.

In the fusion module, the method of weight and cascade is better than FCN-8 s and UNet. Because the input image resolution of PANet is 768×768, the effect of the method proposed in this paper is slightly worse than that of PANet.

6. Conclusion

In this paper, a semantic segmentation method based on asymmetric convolution and multiscale global attention upsampling is proposed.

(1) Asymmetric convolution is used to improve that feature expression ability, and a multiscale upsampling module is used to complete the direct and indirect influence of high-level semantic features on low-level features.

(2) In the feature fusion stage, two fusion methods are given: feature fusion based on weight and feature fusion based on the cascade. In the ablation experiment of feature fusion based on weight, the rule of different weights to achieve a good segmentation effect is found, which also proves that indirect guidance of low-level features is helpful to the segmentation effect, and is verified in the PASCAL VOC 2012 dataset.

(3) The experimental results show that the method of direct guidance and indirect guidance of high-level semantic features is better than the method of only direct guidance of low-level semantic features, and the fusion method based on cascade is better than the fusion method based on weight.

The difference between the proposed MGAU module and the GAU module is that the MGAU module not only provides direct guidance for low-level features, but also provides indirect guidance to establish a closer relationship between high-level and low-level features. It can integrate semantic and spatial information to a greater extent, thus improving the ability of semantic segmentation.

In future work, we mainly focus on the case that there are too few original images in the public dataset and personal dataset used in this paper. There is no subdivision of each large category, and the effectiveness of the segmentation effect of the model has not been verified when there are too many samples and categories in the dataset. The semantic segmentation models of the Segnet network, DeeplabV3 network, and DGCN network will be used for training and result evaluation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contributions

The authors of the manuscript "Image Semantic Segmentation Based on Convolutional Neural Network" declare the following contribution to the creation of the manuscript. Xiaojuan Li contributed the conceptualization and resource. Runze Wang contributed the methodology and writing. Xingmin Ma provided the supervision and resource.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE*, vol. 39, no. 4, pp. 640–651, 2017.

[2] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: real-time joint semantic reasoning for autonomous driving," *IEEE, In 2018 IEEE Intelligent Vehicles Symposium(IV)*, pp. , 20181013–1020, 2018.

[3] J. Liu, C. Yu, B. Yang, C. Gao, and N. Sang, "CSENet: Cascade semantic erasing network for weakly-supervised semantic segmentation," *Neurocomputing*, p. 453(1), 2021.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[5] Z. Zhou, M. M. R. Siddiquee, and N. Tajbakhsh, "UNet++: a nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, Springer Verlag, 2018.

[6] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. C. Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021.

[7] Z. Ke, J. Sun, K. Li et al., "MODNet: real-time trimap-free portrait matting via objective decomposition," pp. 6–10, 2020.

[8] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[9] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[10] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[11] L. Jiao, L. Huo, C. Hu, and P. Tang, "Refined unet: Unet-based refinement network for cloud and shadow precise segmentation," *Remote Sensing*, vol. 12, no. 12, p. 2001, 2020.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactionson pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[13] X. Hu, X. Zhang, and J. Qu, "Resource storage and management method of massive remote sensing data supported by the big data architecture," *J. Geo Inf. Sci*, vol. 18, pp. 681–689, 2016.

[14] Y. Zhou, W. Zhao, and Y. Fan, "Real-time rendering and interactive visualization of remote sensing big data," *Journal of Geo-Information Science*, vol. 18, no. 5, pp. 664–672, 2016.

[15] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensingimage caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

[16] J. Chen, C. Wang, and Y. Tong, "AtICNet: semantic segmentation with atrous spatial pyramid pooling in image cascade network," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, 2019.

[17] S. M. Kwon, X. Li, and A. D. Sarwate, "Low-rank phase retrieval with structured tensor models," pp. 54–62, 2022.

[18] Y. Zhuge, G. Yang, P. Zhang, and L. Huchuan, "Boundary-guided feature aggregation network for salient object detection," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1800–1804, 2018.

[19] W. Guan, T. Wang, J. Qi, L. Zhang, and L. Huchuan, "Edge-aware convolution neural network based salient object detection," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 114–118, 2019.

[20] U. Shalit and G. Chechik, "Efficient coordinate-descent for orthogonal matrices through givens rotations," *Computer Science*, pp. 548–556, 2013.

[21] X. Hu and D. Simpson, *Incomplete Orthogonal Factorization using Givens Rotations*, pp. 89–93, 2013.

[22] Z. Z. Bai, I. S. Duff, and J. Yin, "Numerical study on incomplete orthogonal factorization preconditioners," *Journal of Computational & Applied Mathematics*, vol. 226, no. 1, pp. 22–41, 2009.