Parallel Dense Merging Network with Dilated Convolutions for Semantic Segmentation of Sports Movement Scene

Dongya Huang[1,2], and Li Zhang[3,4*]
[1] Department of Physical Education, Nanjing Vocational Institute of Railway Technology, Nanjing 210094, China
[2] College of Education and Liberal Arts, Adamson University, Manila 1000, Philippines
[e-mail: nty7401@126.com]
[3] Sports College, Nanchang Institute of Science and Technology, Nanchang 330108, China
[4] Graduate School, University of Perpetual Help System DALTA, Las Pinas City, 1704, Philippines
[e-mail: nanchanggongxue@126.com]
[*]Corresponding author: Li Zhang

Received June 16, 2022; revised Septebmer 7, 2022; accepted October 23, 2022; published November 30, 2022

Abstract

In the field of scene segmentation, the precise segmentation of object boundaries in sports movement scene images is a great challenge. The geometric information and spatial information of the image are very important, but in many models, they are usually easy to be lost, which has a big influence on the performance of the model. To alleviate this problem, a parallel dense dilated convolution merging Network (termed PDDCM-Net) was proposed. The proposed PDDCMNet consists of a feature extractor, parallel dilated convolutions, and dense dilated convolutions merged with different dilation rates. We utilize different combinations of dilated convolutions that expand the receptive field of the model with fewer parameters than other advanced methods. Importantly, PDDCM-Net fuses both low-level and high-level information, in effect alleviating the problem of accurately segmenting the edge of the object and positioning the object position accurately. Experimental results validate that the proposed PDDCM-Net achieves a great improvement compared to several representative models on the COCO-Stuff data set.

Keywords: Sports movement scene, convolutional neural network, semantic segmentation.

http://doi.org/10.3837/tiis.2022.11.002 **ISSN : 1976-7277**

1. Introduction

In the field of computer vision, scene understanding can deduce the knowledge or semantics of relevant scenes from pictures or videos from four levels of detection, positioning, recognition and understanding. Semantic segmentation, as a high-level task of scene understanding, also appears to be particularly important. Its purpose is to assign corresponding labels to each pixel, and it is widely used in scenes such as automatic driving [1], medical image analysis [2], motion pose capture [3], and so on.

Traditional semantic segmentation methods are limited by artificial features, which are difficult to meet the high performance and high precision requirements of visual tasks. With the development of deep learning, many visual tasks including semantic segmentation begin to be completed by deep architecture. General semantic segmentation methods benefit from the emergence of convolutional neural networks. Representative models such as FCN (Fully Convolutional Networks) [4] firstly achieve end-to-end semantic segmentation tasks. U-net [5] is used to solve the simple segmentation problem of small samples and is widely used in the segmentation of medical images. Segnet [6] and U-net both adopt two stages of encoding and decoding, but the difference is that pooling with index is adopted. PSPnet (Pyramid scene parsing network) [7] uses spatial pyramid pooling to capture multi-level semantic features. Deeplabv3 [8] adds global average pooling to the previous one, emphasizing global features. The deepening of network layers and the change of structure can improve the accuracy of segmentation, but these complex network connections often need a lot of computing cost, which is difficult to meet the needs of the actual scene.

At the same time, real-time semantic segmentation has become another direction of attention for researchers and scholars. In recent years, real-time semantic segmentation focuses on lightweight backbone networks and multi-branch structures. Wu et al. [9] introduced spatial sparsity on the basis of FCN, removed residual elements, lost a little accuracy, but reduced computation overhead by 25 times. On the mobile end, Adam et al. [10] developed the lightweight network E-Net (efficient neural network), which did not use any bias and reduced kernel calls and memory operations. For high-resolution images, ICNet (image cascade network) [11] uses cascading image input to obtain semantic information from low resolution, medium resolution, and high-resolution images through three branching structures, so as to obtain high-quality segmentation results. In addition, Eduardo [12] et al. used downsample block to reduce the size of the input image at the beginning and save a lot of calculation. DFANet (Deep feature aggregation network) [13] not only adopts a lightweight backbone network, but also designs cross-level feature aggregation modules to ensure accuracy. In practical application scenarios, a real-time scene segmentation task is a trade-off between segmentation accuracy and operational efficiency.

For sports application scenarios, the semantic segmentation task faces two parts of the problem [14]-[15]. On the one hand, there is a lack of sports scene annotation data set. On the other hand, semantic segmentation requires high precision segmentation speed. For example, Mueller et al. [16] synthesized data sets through GAN (Generative Adversarial Nets), but the quality of data caan not be measured. Weiss et al. [17] verified the practicability of transfer learning to solve the labeling problem, and found that the experimental results were valid only when the distance between the target and source domains is close enough. Therefore, the manual selection of athletes in sports scenes can be used to establish data sets. Cioppa et al. [18] developed a general method called ARTHUS (Adaptive real-time human segmentation) to generate adaptive real-time game-specific networks without requiring any manual labeling. In order to solve the balance between performance and speed, Xie et al. [19] studied

knowledge extraction from high-performance low-speed networks to low-performance high-speed networks. In addition, motion segmentation, as a higher-level semantic task, also has reference significance for semantic segmentation of sports scenes. For 2D motion segmentation, optical flow [20] and scene flow [21] are used to achieve motion segmentation. Pia et al. [22] can make accurate segmentation for pedestrians in complex scenes by combining classical geometric knowledge with CNN's pattern recognition ability.

Therefore, in this paper, we propose a novel parallel dense dilated convolution merging network for semantic segmentation of sports movement scenes. The contributions can be summarized in the following threefold parts.

(1) We propose a lightweight architecture based on dilated convolution, and add an efficient encoder-decoder structure to extract salient features of the athletes, which is finally applied to the task of semantic segmentation of sports movement scenes.
(2) In the proposed PDDCM-Net architecture, we utilize parallel convolutions with different dilation rates to extract multi-scale information (or features) to enhance the model's ability for recognizing objects at various scales. In addition, the dense dilated convolution merging (DDCM) module is proposed to increase the receptive field of the model and complex salient features are further extracted to improve the performance of the model.
(3) We propose an end-to-end semantic segmentation model by adopting ResNet50 as a feature extractor. The low-level features and high-level features are effectively combined to realize the complementarity of geometric information (such as edges) and spatial or context information, thus alleviating the problem of blurred object edge segmentation. Compared to several representative methods, our PDDCM-Net model achieves significant segmentation results on the COCO-Stuff dataset.

2. Related Work

This encoding and decoding way was popularly utilized in many tasks of computer vision. Inspired by FCN [4], this structure is used in semantic segmentation. The encoding stage refers to the extraction of deep semantic features along with the reduction of image resolution, and the decoding stage refers to gradually restoring feature maps with low-resolution to the original image size through an up-sampling strategy, and the final output prediction result graph size is the same as the input image size. In the U-Net [5] a skip connection was added in the decoding stage, so that the geometric information of the image extracted in the encoding stage can be combined with the extracted high-level semantic features, so as to alleviate the problem of loss of fine-detailed information in the scene. The SegNet model [6] recorded the position index of the maximum value when performing max pooling in the encoding phase, and used the corresponding pooling index to assist up-sampling in the decoding phase, which not only reduced the amount of computation, but also better preserved image edge information. By using the encoder-decoder structure, a network (termed the discriminative feature network (DFN) [23]) combined average pooling operation with channel attention to enhance the expressiveness of features. In the LANet [24], an average pooling operation was used in the encoding stage to enhance the expressiveness of features, so that better output results can be obtained. Compared with LANet, DDCM [25] employed dilated convolution to extract multi-scale information in the encoding step to improve the model performance. It can be seen from the above methods that the encoder-decoder structure generally delivers satisfactory results in semantic segmentation, so this structure is also employed in our model.

In sports scene semantic segmentation, in addition to the model extracting deep semantic features, coherent context information is also very important, which requires the model to have a strong ability to discriminate objects at different scales, therefore, multi-scale techniques are usually employed to enhance the continuity of contextual or spatial information. For example, the DeepLab [26]-[27] series of network models achieve good segmentation results by adding some tricks to exploit multi-scale information to enhance the continuity of contextual information. In the Deeplabv3+ [27], the dilated convolution operation and the pyramid pooling operator were utilized to realize the extraction of multiple scale features. Similarly, a network (termed pyramid scene parsing network (PSPNet) [7]) combined the idea with different scale pooling methods in a parallel way, which not only saved the model memory consumption but also implemented multi-scale feature extraction. Compared with PSPNet, SENet [28] achieved the extraction of salient features by using global pooling and residual structure, in addition, it also uses convolution kernels of different scales and extracts multi-scale information to improve the robustness of the model.

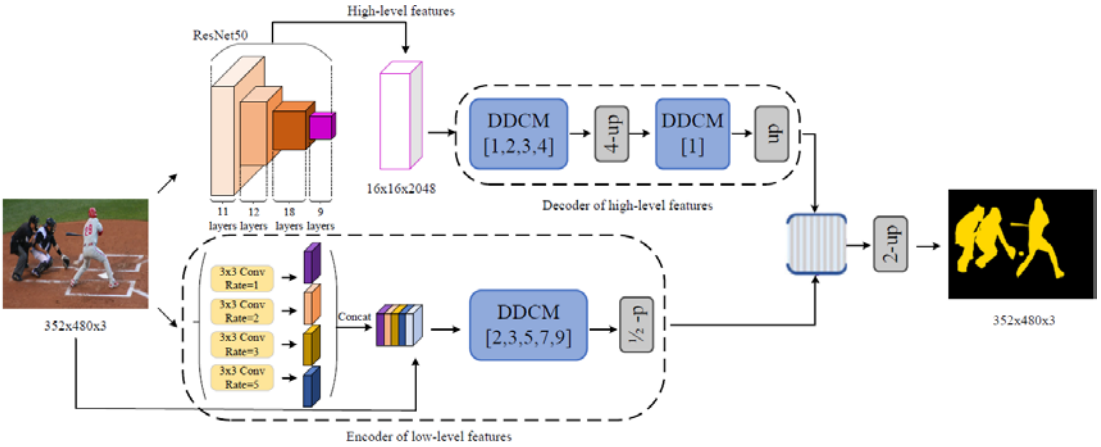

Fig. 1. Architecture of the proposed PDDCMNet. The encoder of low-level features module is composed of parallel dilated convolution, feature maps concatenate operation and a DDCM module (3×3 convolution operation with different dilated rates). The decoder of high-level features module consists of two DDCM modules. The low-level features of the decoder of low-level features module and the high-level features of the encoder of high-level features module are fused as the final output result. Here, “up” and “p” represent up-sampling and pooling operations, respectively.

3. Proposed Method

3.1 Architecture of PDDCMNet

Fig. 1 shows the parallel dense dilated convolution merging Network (PDDCM-Net) combined with a pre-trained model for movement scene semantic segmentation. Compared with other codec structure models, the PDDCM model only fuses low-level features and high-level features once in the final output, and the intermediate feature map in the encoding stage does not participate in the fusion operation, which reduces the model complexity. Particularly, the PDDCM-Net employs parallel convolution operations with different dilation rates in the encoding stage of low-level features to increase the receptive field, which enriches the global information, and this model uses convolution operations with different dilation rates to extract deep semantic features.

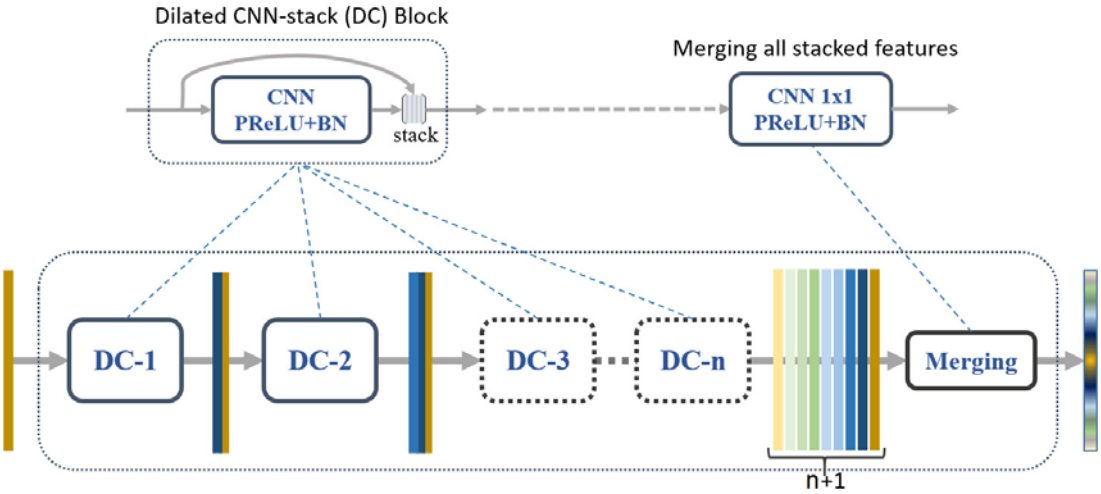

Fig. 2. The DDCM module consists of $n\{1,2,3,...,n\}$ various dilation rates DC blocks.

Specifically, PDDCMNet is divided into two parallel branches, one of which takes the final output of ResNet50 as the high-level feature to be enhanced, and then passes the advanced feature through the decoder of high-level features; the other branch is to concatenate the feature maps after the parallel dilation convolution and the input image, and then go through the encoder of low-level features. Finally, the results of the two branches are concatenated and up-sampled as the final prediction result map. We take the final output of resnet50 as the high-level feature to be processed, and its size and channel are 16×16 and 2048, respectively. In order to ensure that the size of the feature map is 16×16, the stride of the convolution operation in the 43rd layer of ResNet50 model is set to 1.

3.2 DDCM

The dense dilated convolution merging module (DDCM-Module) is composed of different numbers of (DC) blocks with dilated CNN-stack. **Fig. 2** illustrates the structure of DDCM, where a DC block usually consists of a dilated convolution operation followed by a nonlinear activation function and batch normalization (BN). After that, the output is concatenated with input and both of them are used as the input of the next layer. Finally, the output of the model is a 1×1 convolution operation and with the activation function and BN layer. This method can effectively extract the salient features in the intermediate stacked features. Convolution operations with different dilation rates are continuously used in DDCM, and the receptive field of the model can be greatly increased by using only a few DC blocks. It should be noted that in order to ensure that the input feature map has the same resolution as the output feature map, zero-padding is used during the convolution operation. It should be noted that in order to ensure that the input feature map has the same size as the output feature map resolution, o padding is used during the convolution operation. Besides, a 2-D dilated convolution operation is defined as:

$$h(m_n) = \sum_{c \in C_n} \alpha_{s,d}^{c} \bullet m_n^{c} \tag{1}$$

where $\bullet$ represents a convolution operation, and $h : F^{H_n \times W_n \times C_n} \to F^{H_{n+1} \times W_{n+1}}$ convolves the maps $m_n \in F^{H_n \times W_n \times C_n}$ of input features. In addition, $\alpha_{s,d}$ is a dilated convolution where s is the

filter size and $d \in Z^+$ is the dilation rate that is nonzero for multiplication of r pixels from the center. Moreover, a kernel size s in a dilated convolution can be effectively extended to $s+(s-1)(d-1)$ with dilation factor d, when $d=1$, a dilated convolution degrades to a standard convolution.

4. Experiments and Results

4.1 Data Description

COCO-Stuff is the data set used in this experiment. COCO-Stuff [30] is a recently labeled dataset based on MS-COCO [31] for stuff segmentation in context. In the dataset, 9,000 images are chosen for training and the rest 1,000 images are chosen for testing, including 91 categories. We select 400 and 50 pictures from the corresponding data set as the training and testing sets in our experiment, and process the image labels so that the labels only contain two categories: person and background. **Fig. 3** shows four examples selected from the processed COCO-Stuff data set where rows (a) and (c) are the four sports movement scenes, and rows (b) and (d) are the corresponding segmentation results.

Fig. 3. The above movement scene images and the corresponding ground truth are selected from the COCO-Stuff data set. Lines (a) and (c) are images (b) and (d) are their corresponding ground truth.

4.2 Training Details and Evaluation Methods

We implemented the proposed PDDCMNet based on Pytorch. The learning rate was initialized to 0.0001 for COCO-Stuff. After every 50 epochs, the learning rate was updated to be 0.5 times of the original rate to guarantee the convergence speed. Adam optimizer was utilized to train the network with batch size as 4. We trained 500 epochs on different models. To augment the data, we just resized the image size of the original data set to 352×480.

Overall accuracy (OA), per-class accuracy and Mean IoU (mIou) are commonly employed as quantitative assessment indicators for scene segmentation. Specifically, OA is the ratio between the number of correctly predicted pixels and the number of all pixels, which can only reflect the performance of each method. MIOU is the average of the ratio of the intersection to the union of each class in the dataset, and it is an important indicator of model performance in semantic segmentation.

Table 1. Display of test indexes of competing models (on COCO-Stuff data set).

Model	Per-class accuracy (%)		OA (%)	mIou (%)
	Person	Background		
Deeplabv3+ [27]	59.26	94.86	77.06	67.41
SENet [29]	58.64	95.23	76.93	67.42
DANet [30]	81.28	96.96	89.12	82.33
PDDCMNet	82.52	97.15	89.84	83.12

4.3 Parameter setting

We set up experiments with different parameters to ensure that the selected parameters are optimal. First, in the case of keeping other parameters unchanged, the batch size is set to 4, which only changes the number of convolution operations of the DDCM module for encoder of low-level features. A small number of convolution times cannot extract deep-level features. Multiple convolution operations will increase the complexity of the model and may also inhibit model performance. In order to have a trade-off between model performance and model complexity, we separately set [2, 3, 5], [2, 3, 5, 7], [2, 3, 5, 7, 9] and [2, 3, 5, 7, 9, 11] to conduct experiments, experimental results as shown in **Fig. 4**(a), when setting [2, 3, 5, 7, 9] (representing the dilation rate per convolution) the model achieves the best segmentation performance. The setting of the batch size will also affect the segmentation performance of the model. When it is set to [2, 3, 5, 7, 9] and all other parameters remain unchanged, we set the batch size to 1, 2, 4, and 8. The experimental results in **Fig. 4(**b) show that when the batch size is set to 4, the segmentation performance of the model is the best.

In summary, when the convolution operations with different dilation rates are set to [2, 3, 5, 7, 9] and the batch size is set to 4, the model achieves the best segmentation performance.

4.4 Quantitative Evaluation

We compare the experimental results of PDDCMNet with other models. For other models, we also use ResNet50 as the backbone to extract high-level features. **Table 1** lists the results of different models applied to the COCO-Stuff dataset. From the results, we can see that for the COCO-Stuff dataset, the Deeplabv3+ network obtained higher OA than the SENet [29] model,

that is, Deeplabv3+ obtained an OA value of 77.06% while SENet obtained an OA value of 76.93%. The former deeplabv3+ model uses hole convolution operations and spatial pyramid pooling to exploit multi-scale information to ensure the segmentation precision of object boundaries. The latter SENet only uses one pooling operator to extract deep-level features, which is obviously not sufficient, but it costs less memory than the deeplabv3+ model. Compared with SENet and Deeplabv3+, DANet [30] has achieved better results. DANet consists of a feature extractor, a channel attention module and a position attention module. The position attention module is used to extract the associated features in the same channel and the channel attention module integrates the correlation of features between different channels, and these two modules enhance the feature representation so as to improve model performance. However, the receptive field accepted by the model in DANet is still insufficient. In the proposed PDDCMNet, parallel convolution kernels with different dilated rates are used to increase the model's receptive field, while the DDCM module can also extract richer and more detailed semantic features. Compared with SENet and Deeplabv3+, PDDCMNet has a significant improvement in OA and mIou, which is 0.72% and 0.79% higher than DANet in terms of OA and mIou, respectively.

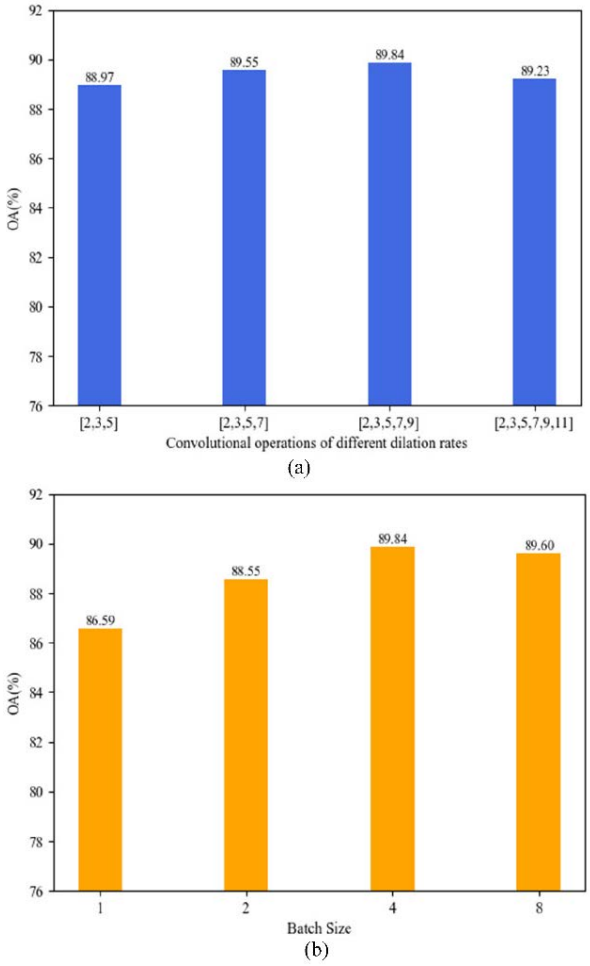

Fig. 4. (a) plots the OA values as a function of different dilation rate convolution operations, (b) plots the OA values as a function of different batch size.

Table 2. Comparison results of inference time (measured on the input image size of 3 × 352 × 480) for different models.

Model	Deeplabv3+	SENet	DANet	PDDCMNet
Inference time(ms-CPU)	2716.96	298.45	291.59	687.03
Inference time(ms-GPU)	182.78	88.17	89.06	69.98

To verify the computational efficiency of the proposed PDDCMNet model, we tested the running time of all comparison methods. The running environment is PyThon3.6 and PyTorch1.3, the CPU is Intel (R) Xeon (R) Silver 4210, and the GPU is NVIDIA Geforce GTX 2080Ti. **Table 2** shows the forward propagation times on CPU and GPU for all comparison models, where the higher the computational complexity, the more time is consumed. Deeplabv3+ employed a multi-branch dilated convolution operation, thus spending a lot of memory so that the forward propagation time of the model increased. For SENet and DANet, there is little difference in the time consumption of forward propagation on CPU or GPU. Compared with the Deeplabv3+ model, they are more lightweight models so the forward propagation time is shorter than Deeplabv3+. From **Table 2**, it can be shown that the PDDCMNet has the shortest forward propagation time on the GPU, the reason is that compared with the Deeplabv3+ model, we also use dilated convolution but choose suitable dilated rates, in addition, the PDDCMNet structure is simple thus reduced computational cost, so the model spends less time in forward propagation.

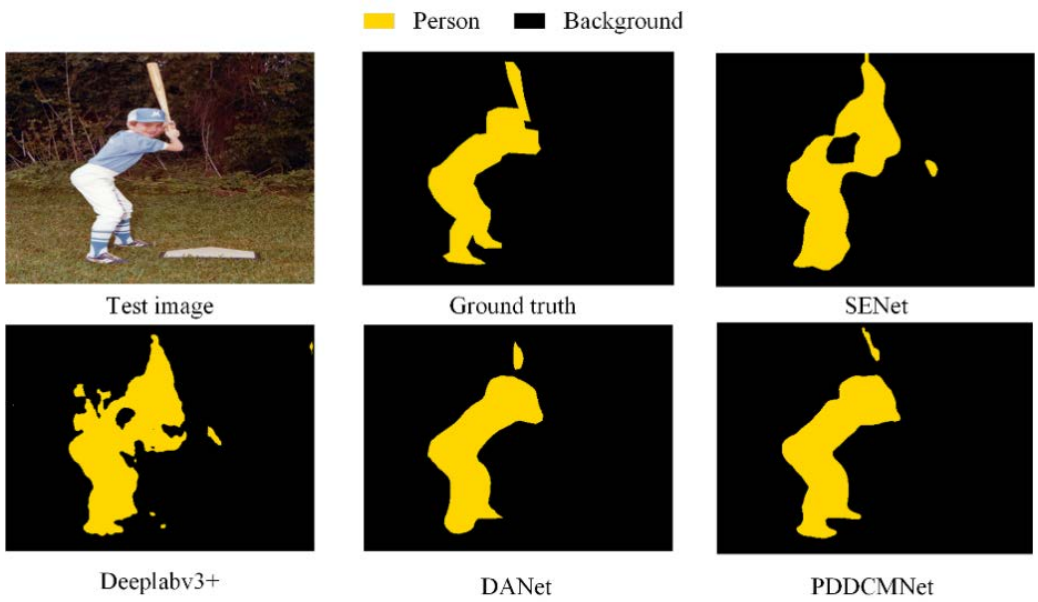

Fig. 5. Semantic segmentation maps of all competing methods on COCO-Stuff data set.

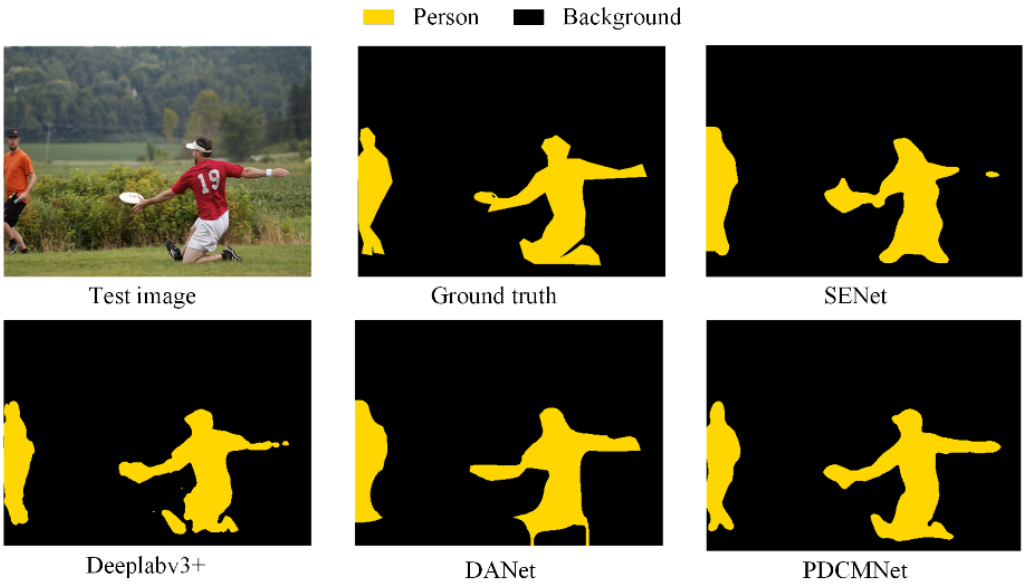

Fig. 6. Semantic segmentation maps of all competing methods on COCO-Stuff data set.

4.5 Qualitative Evaluation

Figs. 5 to **8** are the comparison results of the visualization effect of PDDCMNet and several representative methods. Due to the limitation of the size of the receptive field, the segmentation results of the SENet and Deeplabv3+ models are very confusing, and even the contours of the object positions in the image are ambiguous. Compared with the previous two methods, DANet has a much better prediction result. The position of the object is segmented, but the segmentation of object edge details is still very rough. However, in PDDCMNet, whether it is the segmentation of a single object or the segmentation of multiple objects in **Figs. 5** to **8**, this model is much better than other methods in the overall position of the object and the edge detail segmentation.

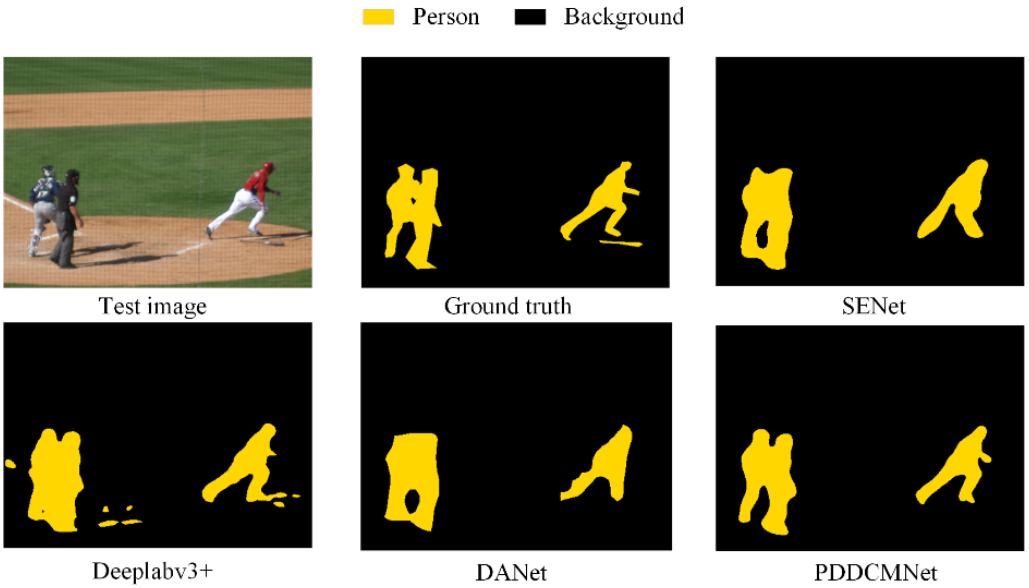

Fig. 7. Semantic segmentation maps of all competing methods on COCO-Stuff data set.

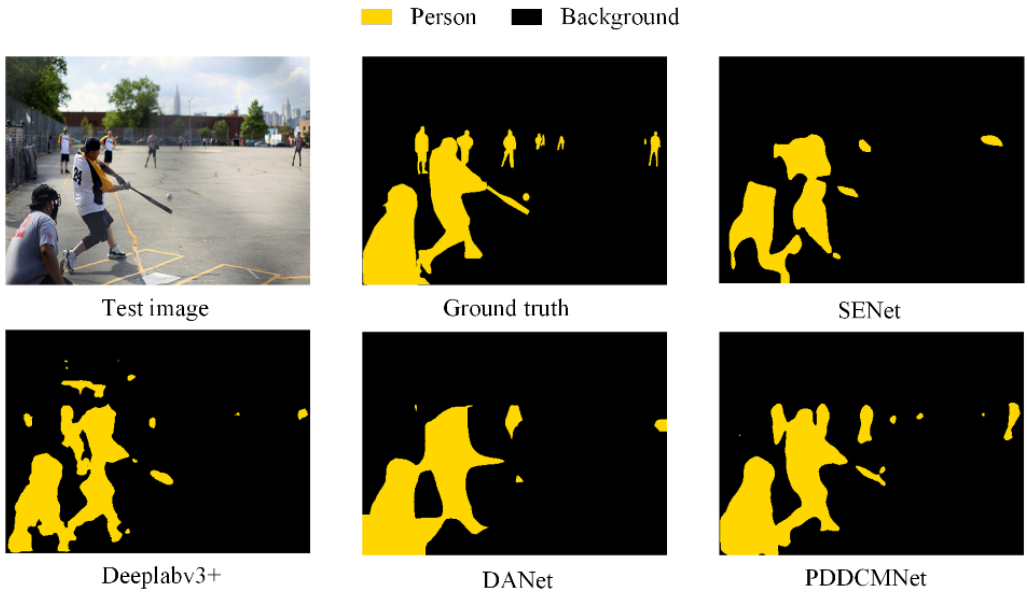

Fig. 8. Semantic segmentation maps of all competing methods on COCO-Stuff data set.

5. Conclusion

The PDDCMNet we proposed is composed of backbone, parallel dilated convolution and DDCM modules. The model extracts feature with rich semantic information through backbone, but using parallel empty convolution operation can expand the receptive field of the moving scene image and then provide a great contribution to the precise location of the object, making the location of the object more accurately obtained. In addition, the DDCM module enhances the expression power of the features. Finally, the fusion of low level features and high-level features is the realization of compensation for geometric (such as edges) information and spatial or context information, and alleviate the problem of blurred object edge segmentation. Experimental results of the proposed PDDCMNet show enhanced feature representation to promote the model segmentation performance on a sports movement scene dataset. Moreover, compared with several representative methods, better segmentation prediction maps can be obtained, especially the edge segmentation of objects is more accurate. From the segmentation accuracies in **Table 1** and the segmentation maps from **Fig. 5** to **Fig. 8**, it is clear to validate the effectiveness of our model.

However, the poor segmentation accuracy of small target objects is still a problem. In order to solve this problem, we plan to combine dilated convolution with the transformer module to enhance the feature expression ability of the model so as to further improve the segmentation performance of the model.

References

[1] A. Bar, J. Lohdefink, N. Kapoor, S. J. Varghese, F. Huger, P. Schlicht, and T. Fingscheidt, "The vulnerability of semantic segmentation networks to adversarial attacks in autonomous driving: Enhancing extensive environment sensing," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 42-52, Jan. 2021. Article (CrossRef Link)

[2] X. Cao and Y. Lin, "CAggNet: Crossing Aggregation Network for Medical Image Segmentation," in *Proc. of 25th International Conference on Pattern Recognition (ICPR)*, pp. 1744-1750, 2021. Article (CrossRef Link)
[3] J. Ren, M. Chai, S. Tulyakov, C. Fang, X. Shen, and J. Yang, "Human motion transfer from poses in the wild," in *Proc. of the European Conference on Computer Vision*, pp. 262-279, 2020. Article (CrossRef Link)
[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 3431-3440, 2015. Article (CrossRef Link)
[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234-241, 2015. Article (CrossRef Link)
[6] 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, 1 Dec. 2017. Article (CrossRef Link)
[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881-2890, 2017. Article (CrossRef Link)
[8] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834-848, 1 April 2018. Article (CrossRef Link)
[9] Z. Wu, C. Shen, and A. Hengel, "Real-time semantic image segmentation via spatial sparsity," *arXiv preprint*, 2017. Article (CrossRef Link)
[10] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: a deep neural network architecture for real-time semantic segmentation," *arXiv preprint*, 2016. Article (CrossRef Link)
[11] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 418-434, 2018. Article (CrossRef Link)
[12] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263-272, Jan. 2018. Article (CrossRef Link)
[13] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: deep feature aggregation for real-time semantic segmentation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9522-9531, 2019. Article (CrossRef Link)
[14] A. Cioppa, A. Deliege, M. Istasse, C.D. Vleeschouwer, and M.V.Droogenb, "ARTHuS: Adaptive Real-Time Human Segmentation in Sports Through Online Distillation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pp. 1-10, 2019. Article (CrossRef Link)
[15] A. Mustafa, C. Russell, and A. Hilton, "4D Temporally Coherent Multi-Person Semantic Reconstruction and Segmentation," *International journal of computer vision*, vol. 130, no. 6, pp. 1583-1606, 2022. Article (CrossRef Link)
[16] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 49-59, 2018. Article (CrossRef Link)
[17] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010. Article (CrossRef Link)
[18] A. Cioppa, A. Deliege, M. Istasse, C. De Vleeschouwer, and M. Van Droogenbroeck, "Arthus: adaptive real-time human segmentation in sports through online distillation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-10, 2019. Article (CrossRef Link)
[19] J. Xie, B. Shuai, JF. Hu, J. Lin, and WS. Zheng, "Improving fast segmentation with teacher-student learning," *arXiv preprint*, 2018. Article (CrossRef Link)

[20] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4083-4090, 2015. Article (CrossRef Link)
[21] P. F. Alcantarilla, J. J. Yebes, J. Almazán, and L. M. Bergasa, "On combining visual slam and dense scene flow to increase the robustness of localization and mapping in dynamic environments," in *Proc. of the IEEE International Conference on Robotics and Automation*, pp. 1290-1297, 2012. Article (CrossRef Link)
[22] P. Bideau, A. RoyChowdhury, R. R. Menon, and E. Learned-Miller, "The best of both worlds: combining cnns and geometric constraints for hierarchical motion segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 508-517, 2018. Article (CrossRef Link)
[23] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1857-186, 2018. Article (CrossRef Link)
[24] L. Ding, H. Tang and L. Bruzzone, "LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426-435, Jan. 2021. Article (CrossRef Link)
[25] Q. Liu, M. Kampffmeyer, R. Jenssen and A. B. Salberg, "Dense Dilated Convolutions' Merging Network for Land Cover Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6309-6320, Sept. 2020. Article (CrossRef Link)
[26] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint*, 2014. Article (CrossRef Link)
[27] B. Yu, L. Yang and F. Chen, "Semantic Segmentation for High Spatial Resolution Remote Sensing Images Based on Convolution Neural Network and Pyramid Pooling Module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252-3261, Sept. 2018. Article (CrossRef Link)
[28] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011-2023, 1 Aug. 2020. Article (CrossRef Link)
[29] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146-3154, 2019. Article (CrossRef Link)
[30] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: thing and stuff classes in context," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209-1218, 2018. Article (CrossRef Link)
[31] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: common objects in context," in *Proc. of European Conference on Computer Vision*, pp. 740-755, 2014. Article (CrossRef Link)

Dongya Huang received a bachelor's degree in physical education from China University of Mining and Technology, Jiangsu, China, in 2009, and a master's degree in physical education and training from Nanjing University of Science and Technology, Jiangsu, China, in 2012. Since He is now an associate professor in the Physical Education Department of Nanjing Vocational Institute of Railway Technology. His research interests include school sports, smart sports, social sports, sports training and other fields.

Li Zhang received a bachelor's degree in physical education from Xinyang Normal University, Xinyang, Henan, China, in 2011 and a master's degree in physical education and training from Nanjing University of Science and Technology, Nanjing, Jiangsu, China, in 2014, with research interests in school sports, social sports, leisure sports, and smart sports.