Hindawi
Scientific Programming
Volume 2022, Article ID 3311931, 7 pages
https://doi.org/10.1155/2022/3311931

Research Article

Study on the Segmentation Method of the Improved DeepLabv3+ Algorithm in the Basketball Scene

Kai Li

Department of Physical Education, Jiangxi University of Science and Technology, Jiangxi, Ganzhou 341000, China

Correspondence should be addressed to Kai Li; 9120110080@jxust.edu.cn

Received 25 April 2022; Revised 20 May 2022; Accepted 25 May 2022; Published 17 August 2022

Academic Editor: Punit Gupta

Aiming at the weak robustness and low segmentation accuracy of traditional basketball scene segmentation methods, this study proposed an improved basketball scene semantic segmentation model based on DeepLabv3+ for the purpose of basketball scene segmentation and accurate positioning of players. In this model, a relatively complex decoder is designed based on the DeepLabv3+ network. Multiple feature fusion is used to restore the semantic information of the image, and the convolutional block attention mechanism is introduced to optimize the channel weight and position information, reduce the computational complexity of the model, and improve the edge sensitivity. Experimental results show that the proposed model is 21.8% better than FCN's full convolution model and 1.9% better than DeepLabv3+. At the speed of segmentation, it can process 6 pictures per second, greatly improving the accuracy of semantic segmentation for basketball scenes. In the future, real-time detection of sports such as basketball using computer vision methods will become more and more important.

1. Introduction

In recent years, basketball has become popular both at home and abroad. However, players in professional leagues fight fiercely, so it is hard to avoid missing or misjudging [1]. Fair refereeing is very important to the basketball game. The refereeing decisions often decide the trend of the game. Therefore, how to improve the situation has become a big problem. In CBA, there are frontcourt referees, backcourt referees, and video replay. Although the video replay is very clear, if every shot is judged by replay, the game will become extremely complicated and time-consuming, without timeliness [2]. Therefore, it is necessary to study the behavior identification of athletes. At present, the identification method of athlete attributes still studies the existence of attributes without obtaining the position information of the attributes of man and the ball. Obtaining the position of the player and the ball, that is, accurate positioning, is the premise of attribute judgment, which is very meaningful for semantic segmentation of basketball scenes [3].

Deep learning is widely used in various fields because of its ability to extract image features and fit complex problems, and semantic segmentation is one of the key tasks of deep learning. Some researchers replaced the fully connected layer with the full convolution layer and proposed FCN, which realized end-to-end and pixel-to-pixel image segmentation for the first time, thus opening the door to semantic segmentation [4]. In the same year, one researcher proposed DeepLabv1, which introduced the dilated convolution of the 1990s into the field of semantic segmentation and increased the receptive field without increasing parameters. Influenced by the success of SPP in the target detection algorithm R-CNN, the model PSPNet combined with spatial pyramid was proposed. Some researchers also proposed DeepLabv2, which combined SPP and dilated convolution to form spatial pyramid structures with different dilation rates and realized multiscale feature extraction [5]. Before long, DeepLabv3 was proposed, and v3 adopted Xception as a feature extraction network, which greatly reduced the parameter calculation, and at the same time, the removed condition was postprocessed with the field, which realized the deep learning semantic segmentation model in the real sense. Some researchers were influenced by the idea of SegNet encoding and decoding structure and proposed a model

DeepLabv3+ containing a decoder. This model has achieved amazing results on multiple datasets, showing strong generalization ability [6].

As for the traditional methods, video shot segmentation has many characteristics, such as high complexity, high cost, great variation, and difficulty in automatic extraction. In this study, deep learning semantic segmentation is introduced into the basketball scene, and combined with the convolutional block attention mechanism, an improved semantic segmentation model based on DeepLabv3+ is proposed for real-time segmentation and accurate positioning of players [7]. Real-time basketball segmentation using the DeepLabv3+ algorithm will become an integral part of the athlete's court.

2. Traditional DeepLabv3+ Model

The original model of DeepLabv3+ is shown in Figure 1. The model is mainly composed of an encoder and a decoder. The encoder is divided into two parts: the DCNNS extraction network and ASPP spatial pyramid structure [8]. The decoder includes one feature fusion and two upsampling. During model training, the initial image is first entered into the coding module, and after DCNNS extraction, the network reduces the resolution of the image to 1/16 of the original [9]. The extracted feature tensor is then imported into the ASPP structure, which is a spatial pyramid structure with different dilution rates. Then, channel compression is realized through 1×1 convolution to prevent the prediction result from skewing to the underlying features [10]. In the decoder, images restored by the quadruple bilinear interpolation and feature extraction network are used for a splicing feature fusion, and then, a quadruple bilinear interpolation is used to realize the image output [11]. The construction of spatial pyramid structures with different dilation rates improves the extraction of multiscale features and achieves the balance of receptive field and resolution.

3. Improved DeepLabv3+ Model

For the segmentation of the basketball scene, if the traditional DeepLabv3+ model is used for segmentation, there is the problem of fuzzy edge segmentation. As shown in Figure 2, this study takes the DeepLabv3+ original model as the main body and improves the model from the following two aspects. First, inspired by the SegNet model, a more complex decoder is designed to supplement the underlying features and restore image information better. Second, because a large number of channels and parameters will affect the training of the network, ASPP connects a convolutional attentional mechanism module (CBAM) to strengthen the position and channel features of the feature graph and enhance the generalization ability of the network [12].

3.1. Improvement of Decoder. This study designed a relatively complex decoder [13]. The resolution of the basketball scene dataset in this study is $512 * 512$, which is relatively low. After four downsamplings, the resolution is only 1/16 of the original, that is, $32 * 32$. Although advanced features are extracted from the network, a large amount of low-level feature information is discarded, which makes the image blurred and difficult to recover in upsampling [14]. Therefore, a relatively complex decoder is designed in this study. The decoder is changed from one-time feature fusion to two-time feature fusion to extract the bottom layer reflected by the feature graph of the first several layers in the feature network, and local information complements the high-level semantics of the output layer, so that the upsampling process can better restore the semantic information of the image. In addition, in the decoder, the original two 4× upsamplings are changed to four 2× upsamplings. Compared with two 4× upsamplings, four 2× upsamplings can restore semantic information more accurately [15].

3.2. Attention Mechanism Module. The human visual system tends to automatically ignore some unimportant things when judging things and puts limited energy on things we need to pay attention to, which greatly reduces the time needed to process things. For example, when typing, we pay more attention to the screen of the computer than to the outline of the computer. This is the attention mechanism [16]. The mechanism model was first used in natural language processing and has gained great success, so it is favored in computer vision.

CBAM represents the attention mechanism module of the convolution module, which is a kind of attention mechanism module combining space and channel [17]. As shown in Figure 3, input feature graph F and obtain feature graph $F1$ through the channel attention mechanism. $F2$ is obtained through the spatial attention mechanism.

The traditional DeepLabv3+ model has too few modified edges, resulting in rough edge segmentation and making no difference in the category importance of segmented pixels [7]. As the original model network contains a large number of channel integration, the deeper the operation network, the more difficult it is to describe and extract the features, which is not conducive to the expansion of learning. Therefore, the CBAM dual-attention mechanism is introduced in this study to assign weight coefficients to channels and give higher weight to edge channels, to optimize the training network and enhance the generalization ability of the network.

4. Experiment

4.1. Dataset and Preprocessing. The experimental data were obtained in the continuous shooting mode of the mobile phone. The shooting was divided into 24 groups of 100 images each. The experimental data were divided into three parts: training set (2,000 images), validation set (200 images), and test set (200 images). After uniformly and continuously naming the collected images, Python code was used to cut the images with the same specifications, and the images with a resolution of 512×512 were produced. Finally, the processed images were classified and integrated into the initial dataset [18].

The annotation tool LabelMe in Anaconda was used to label the images in the initial dataset one by one and

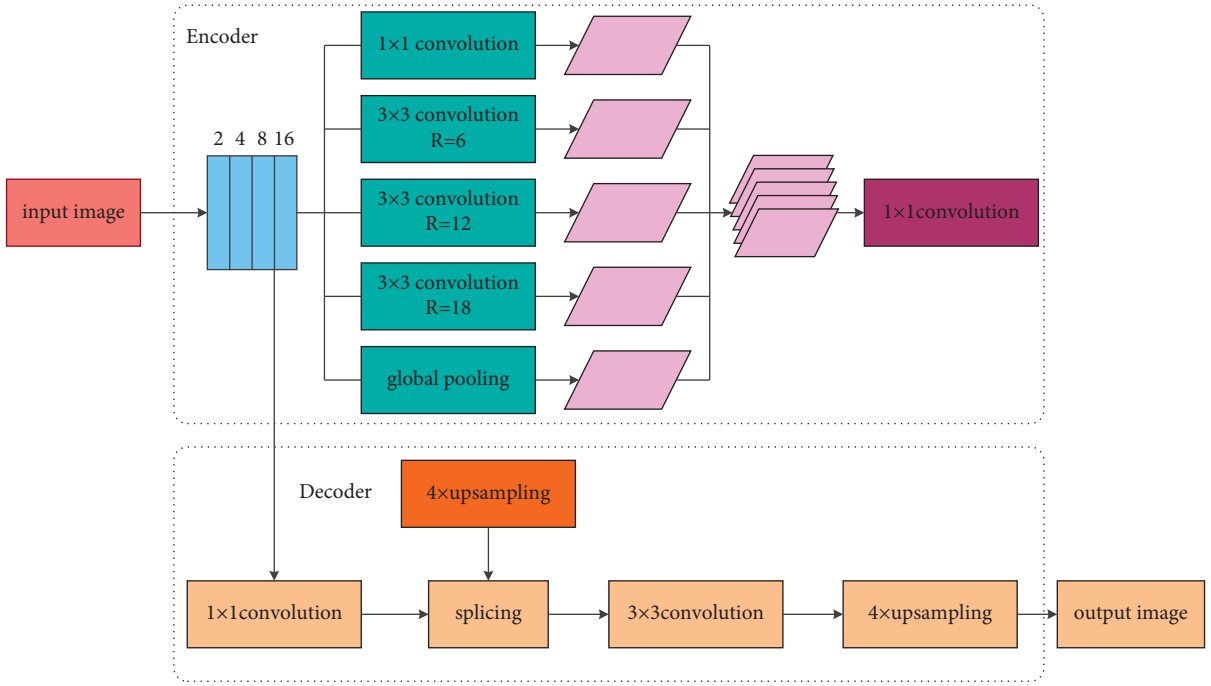

Figure 1: DeepLabv3+ original model.

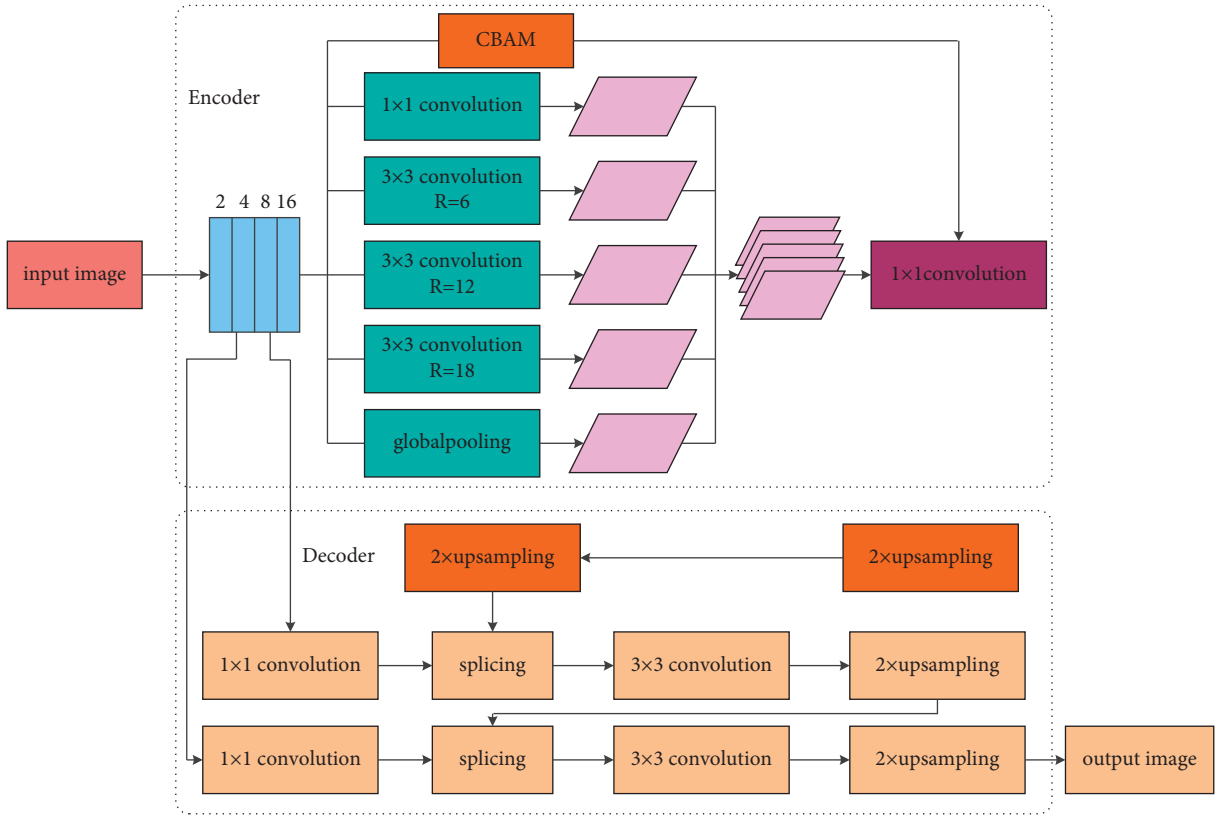

Figure 2: Improved model.

generate JSON files at the same time. Finally, the dataset was used to batch transform the images into gray maps with a bit depth of 24° [19]. The dataset used by the model in this study is a combination of the original dataset and the transformed gray map. The sample dataset is shown in Figure 4:

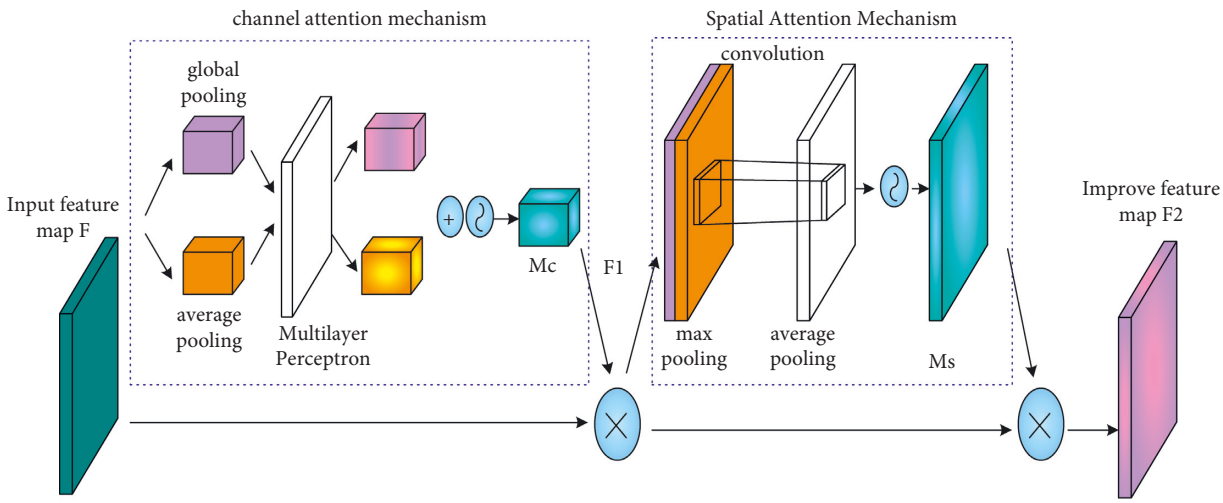

Figure 3: Convolutional block attention mechanism.

Figure 4: DeepLabv3+ data sample diagram.

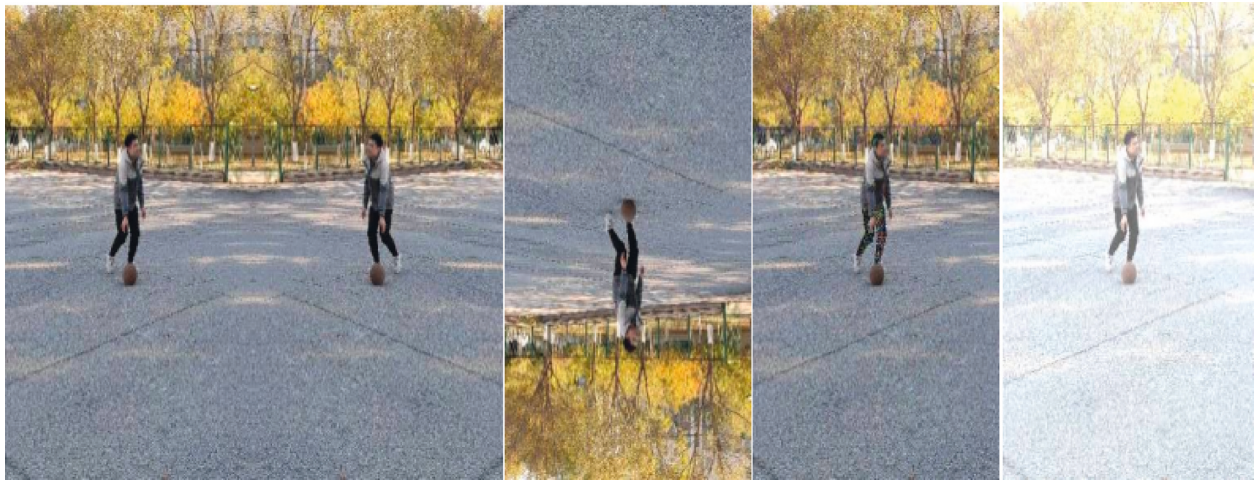

Figure 5: Data enhancement image.

In Figure 4, the left image represents the shooting image and the right image represents the annotated image. Then, according to the original image and annotated image, the single channel grayscale image is trained and combined with the shooting image to form a basketball dataset. The dataset contains 2,400 images, which are divided into three categories: people, basketball, and background images. In the

Table 1: Training and test environment.

Item	Parameter
Operating system	Linux
Algorithm framework	PyTorch
GPU	1660 Ti
Memory	8G

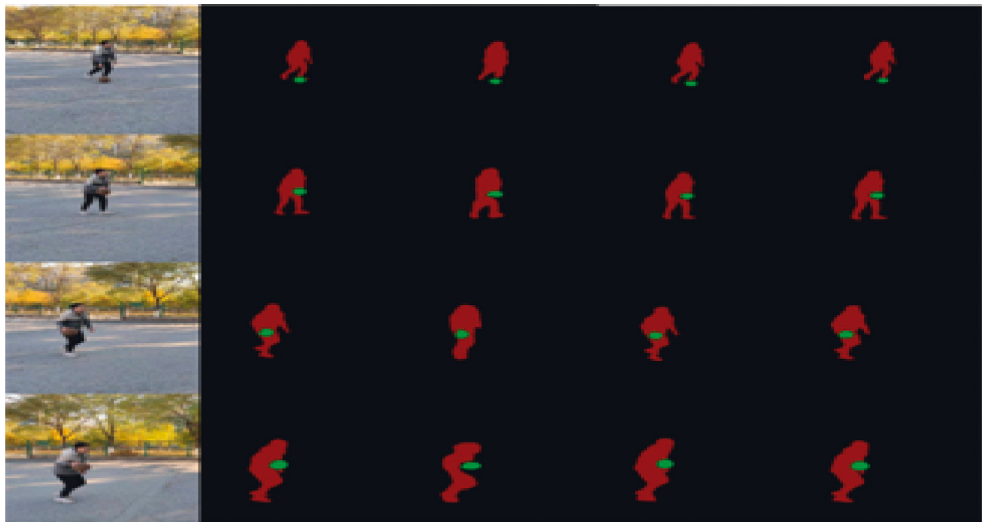

Figure 6: Comparison of segmentation effects of different models.

segmentation effect, the color of the image is divided into 3 categories, and each color corresponds to the corresponding category. In semantic segmentation datasets, common datasets contain tens of thousands or even hundreds of thousands of data, so as to meet the needs of model learning, improve segmentation accuracy, and reduce the overfitting phenomenon caused by too few images. The dataset used in this article is relatively small. To improve the diversity of training samples and reduce the underfitting problem caused by insufficient samples in the training process, this study uses horizontal flip, vertical flip, noise, and brightening to expand the dataset, so that the training set contains 10,000 images.

Figure 5 shows the effect of horizontal flip, vertical flip, noise, and brightening.

4.2. Experimental Configuration and Model Training. The experiment used a Linux system and PyTorch deep learning framework to train the semantic segmentation model in this study. Table 1 provides the specific configuration and training environment for specific experiments.

The number of iterations of the experiment was set to 8,000. ASPP dilated convolution rate was set to [6, 7, 12] and batch size to 4. To avoid stagnation in the training process, SGD was adopted to update the learning rate. The cross-entropy loss function was used as the loss function. At the beginning of the iteration, the decline of the loss function was very rapid, and after 400 iterations, the decline of the loss function slowed down significantly, and the final loss function approached about 1.0 [20].

4.3. Comparison and Analysis of Models. The evaluation index of the model is mainly divided into two parts. One is the pixel accuracy PA, which is the percentage of predicted pixels divided by original correct pixels. The value of PA is proportional to the image segmentation effect. The average value of PA of different kinds is calculated to obtain MPA.

Table 2: Segmentation accuracy comparison of different models on the basketball dataset.

Network model	MPA	MIoU/%	Time/FPS
FCN	92.1	62.7	7.6
SegNet	92.6	69.2	9.1
DeepLabv3+	92.7	82.6	5.8
Serial CBAM	93.6	83.7	5.6
Parallel CBAM	94.3	84.5	5.6

Table 3: Comparison of the segmentation accuracy of improved models with different attention mechanisms.

|Network model	CAM	CBAM	MIoU
DeepLabv3+			82.6
Serialv3+		✓	84.1
Parallelv3+	✓		83.9
Parallelv3+		✓	84.5

The other is IoU, that is, the ratio of the intersection of correctly segmented pixels and the union of two categories of pixels, known as the intersection over union (IoU). MIoU is the average value of IoU, that is, the average IoU.

To ensure the rigor and scientific nature of the experiment, the segmentation effect diagrams of the FCN model, the improved DeepLabv3+ model, and the basic DeepLabv3+ model were compared in the experiment. The segmentation comparison is shown in Figure 6.

It can be concluded from Figure 6 that FCN, the original semantic segmentation model, can roughly segment images of man and ball, but it is not ideal for edge and detail segmentation. DeepLabv3+ shows strong segmentation capability in this respect but still needs to be improved in position. The CBAM module of the parallel attention mechanism is added, which not only improves the edge segmentation considerably but also has a strong sensitivity to position, thus achieving a good segmentation effect [4].

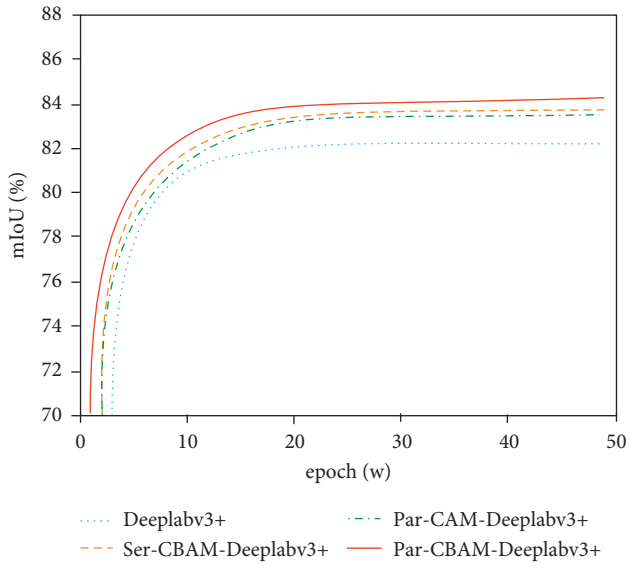

Figure 7: Effect comparison of the four networks on the test dataset.

FCN can achieve 62.7% MIoU in the basketball dataset, while DeepLabv3+ can achieve 82.6%. In the experiment, two different improvement methods, serial CBAM and parallel CBAM, were compared. As given in Table 2, MIoU of serial CBAM increased by 1.1% on the basis of traditional DeepLabv3+ and that of parallel CBAM increased by 1.9% based on DeepLabv3+, so parallel CBAM had a better segmentation effect [21].

To further discuss the details of serial and parallel experiments, this study compared the traditional DeepLabv3+, serial CBAM, parallel CAM, and parallel CBAM. Its segmentation index and convergence rate are given in Table 3.

It can be concluded from Table 3 that, compared with the traditional DeepLabv3+ model, whether in parallel or series, the segmentation accuracy of the networks with attention mechanisms has been improved. Among them, the networks with parallel CBAM have the best MIoU.

MIoU effects were compared on the test dataset in four cases, and the results are shown in Figure 7.

It can be concluded from Figure 7 that, compared with the traditional DeepLabv3+ model, the networks with the attention mechanism have faster convergence speed, and the networks with parallel CBAM have the best convergence speed.

The experimental results show that the parallel attention module can effectively improve the convergence speed and segmentation accuracy of the model.

5. Conclusion

In this study, the semantic segmentation of deep learning is introduced into the basketball scene, and the precise positioning of man and basketball is realized by segmentation, which improves the insufficient accuracy of traditional basketball segmentation. In this method, the SegNet decoder is improved to improve the segmentation accuracy, and the CBAM attention mechanism module is introduced, which is connected with the spatial pyramid structure in parallel to realize the optimization of redundant channels to improve the sensitivity of edge position. The experimental results show that the MIoU of the proposed model is 1.9% higher than that of DeepLabv3+, which improves the semantic segmentation accuracy of basketball scenes and lays a foundation for the development of referee robots in basketball scenes. At the same time, improving the speed of semantic segmentation will become an important direction for subsequent development.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] J. L. Russell, B. D. McLean, D. S. Impellizzeri, A. J. Strack, and A. J. Coutts, "Measuring physical demands in basketball: an explorative systematic review of practices," *Sports Medicine*, vol. 51, no. 1, pp. 81–112, 2021.

[2] M. S. Palmer, C. Wang, R. M. Plucinski, and R. M. Pringle, "BoomBox: An Automated Behavioural Response (ABR) camera trap module for wildlife playback experiments," *Methods in Ecology and Evolution*, vol. 13, no. 3, pp. 611–618, 2022.

[3] J. Liu, "Motion Action Analysis at Basketball Sports Scene Based on Image Processing," *Scientific Programming*, vol. 3, pp. 1–11, 2022.

[4] A. A. Adegun and S. Viriri, "FCN-based DenseNet framework for automated detection and classification of skin lesions in dermoscopy images," *IEEE Access*, vol. 8, pp. 150377–150396, 2020.

[5] L. Yan, D. Liu, Q Xiang et al., "PSP net-based automatic segmentation network model for prostate magnetic resonance imaging," *Computer Methods and Programs in Biomedicine*, vol. 207, Article ID 106211, 2021.

[6] S. Alqazzaz, X. Sun, X Yang, and L Nokes, "Automated brain tumor segmentation on multi-modal MR image using SegNet," *Computational Visual Media*, vol. 5, no. 2, pp. 209–219, 2019.

[7] S. C. Yurtkulu, Y. H. Şahin, and G. Unal, "Semantic segmentation with extended DeepLabv3 architecture," in *Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, Sivas, Turkey, April 2019.

[8] W. Sang, S. Yuan, X Yong, X Jiao, and S Wang, "DCNNs-based denoising with a novel data generation for multidimensional geological structures learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 10, pp. 1861–1865, 2021.

[9] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11030–11039, Seattle, Washington, USA, June 2020.

[10] Bo Jin, Leandro Cruz, and Nuno Gonçalves, "Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis," *IEEE Access*, vol. 8, pp. 123649–123661, 2020.
[11] T. Fetzer and G. Reis, "Fast projector-driven structured light matching in sub-pixel accuracy using bilinear interpolation assumption," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pp. 26–36, Springer, Lecce, Italy, September 2021.
[12] W. Deng, Y. Mou, T Kashiwa et al., "Vision based pixel-level bridge structural damage detection using a link ASPP network," *Automation in Construction*, vol. 110, Article ID 102973, 2020.
[13] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3907–3916, Long Beach, CA, USA, January 2019.
[14] X. Song and L. Fan, "Human posture recognition and estimation method based on 3D multiview basketball sports dataset," *Complexity*, vol. 2021, Article ID 6697697, 10 pages, 2021.
[15] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 8–15, IEEE, Sentosa, Singapore, December, 2019.
[16] H. Fukui and T. Hirakawa, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10705–10714, Seattle, WA, USA, June 2019.
[17] M. Canayaz, "C+EffxNet: A novel hybrid approach for COVID-19 diagnosis on CT images based on CBAM and EfficientNet," *Chaos, Solitons & Fractals*, vol. 151, Article ID 111310, 2021.
[18] J. Blank and K. Deb, "Pymoo: Multi-objective optimization in python," *IEEE Access*, vol. 8, pp. 89497–89509, 2020.
[19] L. Mikkelsen, K. Moesgaard, M Hegnauer, and AD Lopez, "ANACONDA: a new tool to improve mortality and cause of death data," *BMC medicine*, vol. 18, no. 1, p. 61, 2020.
[20] B. E. Woodworth, K. K. Patel, and N. Srebro, "Minibatch vs. local sgd for heterogeneous distributed learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6281–6292, 2020.
[21] J. Zhang, C. Lu, J. Wang, L Wang, and XG Yue, "Concrete cracks detection based on FCN with dilated convolution," *Applied Sciences*, vol. 9, no. 13, p. 2686, 2019.