information

MDPI

Article

Basketball Action Recognition Method of Deep Neural Network Based on Dynamic Residual Attention Mechanism

Jiongen Xiao [1,2], Wenchun Tian [3,*] and Liping Ding [2]

[1] International Business School, Guangdong University of Finance and Economics, Guangzhou 510320, China
[2] Electronic Forensics Laboratory, Guangzhou Institute of Software Application Technology, Guangzhou 511458, China
[3] School of Electrical and Computer Engineering, Guangzhou Nanfang College, Guangzhou 510970, China
* Correspondence: tianwch@nfu.edu.cn

Abstract: Aiming at the problem that the features extracted from the original C3D (Convolutional 3D) convolutional neural network(C3D) were insufficient, and it was difficult to focus on keyframes, which led to the low accuracy of basketball players' action recognition; hence, a basketball action recognition method of deep neural network based on dynamic residual attention mechanism was proposed. Firstly, the traditional C3D is improved to a dynamic residual convolution network to extract sufficient feature information. Secondly, the extracted feature information is selected by the improved attention mechanism to obtain the key video frames. Finally, the algorithm is compared with the traditional C3D in order to demonstrate the advance and applicability of the algorithm. Experimental results show that this method can effectively recognize basketball posture, and the average accuracy of posture recognition is more than 97%.

Keywords: behavior recognition; deep learning; C3D convolution; residual network; attention mechanism

1. Introduction

In recent years, with the development of computer vision, more and more technical achievements have been receiving more and more attention, especially in sports. Indeed, 3D vision is used for human action recognition and behavior analysis [1]. Through intelligent detection of sports movements, it can provide athletes, coaches or analysts with guidance on movement techniques, or assist sports field referees to make more reasonable and effective judgments. Professional basketball is becoming more and more influential worldwide, and the corresponding sports science analysis industry is also booming. To analyze the athletic state of basketball players more scientifically and to improve the scientific nature of coaches' training plans, it is of great significance to improve the training effectiveness of athletes [2].

Basketball is now one of the most popular ball games. In basketball training and competitions, coaches make corresponding training plans according to different athletes to improve their basketball skills [3,4]. The traditional training method is to make a game plan according to the coaches' own experience and the athletes' technical level. This method is highly subjective and requires a lot of time to analyze the movements of the training movement, so it is difficult to evaluate the training effect [5]. The core of modern sports training is precision and efficiency. If the coach can accurately control the athletes' sports posture, the training effect can be greatly improved. Therefore, accurate identification of sports postures is important to improve the scientific nature of coaches' training plans and improve the training effects of athletes [6,7].

The methods for basketball gesture recognition mainly include two categories: inertial sensor-based gesture recognition [8] and image-acquisition-based gesture recognition [9]. The inertial sensor-based pose recognition method requires the athlete to wear the sensor and send the collected data to the data processing terminal for action recognition analysis, which is a large amount of equipment and not conducive to wide application. The pose

check for updates

Citation: Xiao, J.; Tian, W.; Ding, L. Basketball Action Recognition Method of Deep Neural Network Based on Dynamic Residual Attention Mechanism. *Information* **2023**, *14*, 13. https://doi.org/10.3390/info14010013

Academic Editor: Gholamreza Anbarjafari (Shahab)

Received: 28 November 2022
Revised: 22 December 2022
Accepted: 23 December 2022
Published: 27 December 2022

recognition method for image acquisition firstly adopts video or image captured by the camera, then carries on feature extraction to hidden features in video or image, and finally uses a classifier to carry on motion recognition. Feature extraction is the most important method for attitude recognition of image acquisition. In recent years, deep learning has advanced performance in adaptive dynamic extraction of important features, such as C3D [10], I3D [11], P3D [12], TSN [13] and dual-stream networks [14]. At present, the behavior recognition algorithms based on deep learning usually use C3D to extract behavioral features from the input video and continuously train the models [15–19], after which the trained models are used for classification recognition. Tran et al. [20] found the most suitable C3D convolutional kernel size for behavior recognition through systematic experimental studies and proposed C3D networks for direct extraction of spatiotemporal features for behavior recognition. In order to improve the generalization ability of the 3D convolution network, the 3D residual convolution network [21] and pseudo-3D residual network [22] are proposed, one after another. Three-dimensional convolution has high computational and memory costs, so it can train videos by using mixed convolution tubes [23] and generative adversarial networks, which improve recognition efficiency and performance. A single-stream network can simply and directly capture the time dynamics in a short time, but it cannot capture the long-term time information. To solve this problem, Simonyan et al. [24] put forward a double-stream convolution neural network method. Feichtenhofer et al. [25] proposed a series of spatial fusion functions, which solved the problem of information interaction between convolution streams. To solve the time-consuming problem of the optical flow input calculation used in the double-stream convolutional neural network, Zhang et al. [26] proposed a real-time motion recognition framework to extract RGB images and motion vectors from compressed video.

Although the above algorithm based on deep learning can complete the recognition of basketball posture. However, there is a problem of insufficient robustness. For example, in practical applications, it is easily affected by complex environmental factors such as light, background clutter, and camera angle of view [6]. At the same time, due to the simple structure of the traditional C3D network, it is difficult to recognize the basketball posture, resulting in low efficiency and a high error rate. To solve these problems, we propose a deep neural network basketball action recognition method based on an efficient dynamic residual attention mechanism to address the problems that the original C3D convolutional neural network (C3D) extracts insufficient features and has difficulty focusing on keyframes, resulting in low accuracy of basketball player action recognition. The main contributions of our paper are as follows:

1. The basketball motion image obtained from the video contains a large amount of noise, in order to achieve a better suppression of the noise in the image, this paper uses a median filter to pre-process the image, so that the extraction of the basketball motion in the video image is the least interference.
2. In order to fully extract the effective feature information from the video image basketball, this paper improves the convolution layer to dynamic residual convolution on the basis of the original C3D network.
3. In order to efficiently recognize basketball action in video images, the extracted feature information is focused on the important features by improving the attention mechanism and eliminating the unimportant feature information, so as to delete the feature information that is beneficial to the basketball player's pose recognition and improve the accuracy of basketball pose recognition.

The experiments show that the experimental method designed in this paper for eight kinds of postures in basketball can obtain the basketball action information of the detected person in real time, realize the accurate extraction of individual action data, and complete the recognition of basketball postures, and its average accuracy can reach 98%, which has certain practical value in the recognition of basketball postures.

The rest of the paper is organized as follows. Section 2 introduces the related basic methods. Section 3 introduces the dynamic residual convolutional network, the improved attention mechanism network and the process of basketball motion pose recognition. Section 4 introduces the experiments to verify the effectiveness of the methods in this paper. Section 5 introduces the conclusions and discussion of the study.

2. Materials and Methods

2.1. C3D Neural Network

2D convolution is mainly applied in spatial feature extraction of static images, and 2D convolution neural network has made remarkable achievements in image classification, target segmentation, detection and other tasks. However, for motion behavior in dynamic videos, 2D convolution can only extract spatial features, and cannot extract motion feature information in the temporal dimension [27]. Therefore, 2D convolution cannot be applied to tasks with temporal dimensional information.

In order to be able to extract feature information in the temporal dimension of video data or multi-frame images, Simonyan et al. [28] proposed a dual-stream convolutional neural network, but this method requires a lot of time to calculate the optical flow map in advance, and the optical flow map will contain many invalid motion features. In contrast, C3D in deep learning algorithms can learn end-to-end to achieve the extraction of motion features. In particular, the C3D proposed by Tran et al. [20] adds convolution to the temporal dimension of video data and can extract spatial features and motion features simultaneously. Compared with a 2D convolution network, C3D adds an extra time dimension to the input data and convolution kernel, and multiple consecutive video frames form a cube as the input. Then, the C3D convolution kernel is used in the cube, and each feature graph in the convolution layer is extracted from multiple consecutive frames in the previous layer. Therefore, C3D convolution captures motion information and is suitable for behavioral recognition tasks. The 2D convolution and C3D operations are shown in Figure 1.

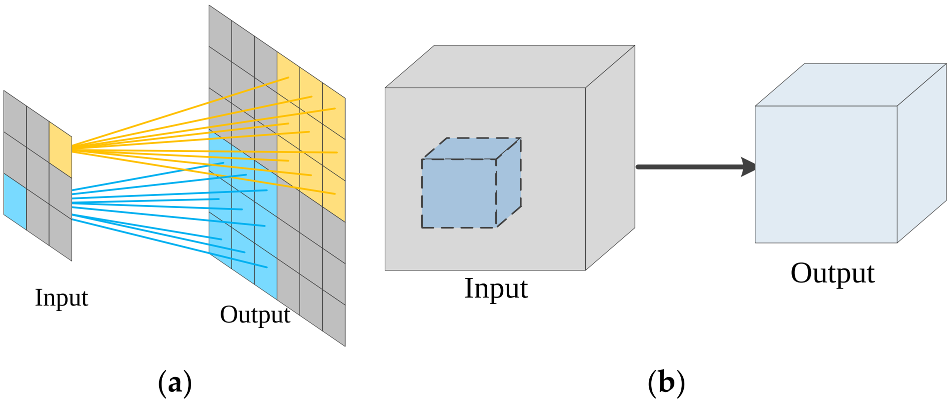


Figure 1. Convolutional network diagram: (**a**) 2D Convolution and (**b**) C3D Convolution.

The C3D network is a deep convolutional neural network with a total of 10 layers, including 8 convolutional layers and 2 layers of fully connected layers. Firstly, the network starts with a 3-channel, 16-frame video image with an image size adjusted to (112 × 112) as input. Then the key feature information in the video image is extracted by network training. Finally, the extracted features are input to the later two-layer fully connected layer for feature classification, and the classification probabilities of each human action category in the video are output by a Softmax classifier. The overall structure of the C3D neural network is shown in Figure 2.

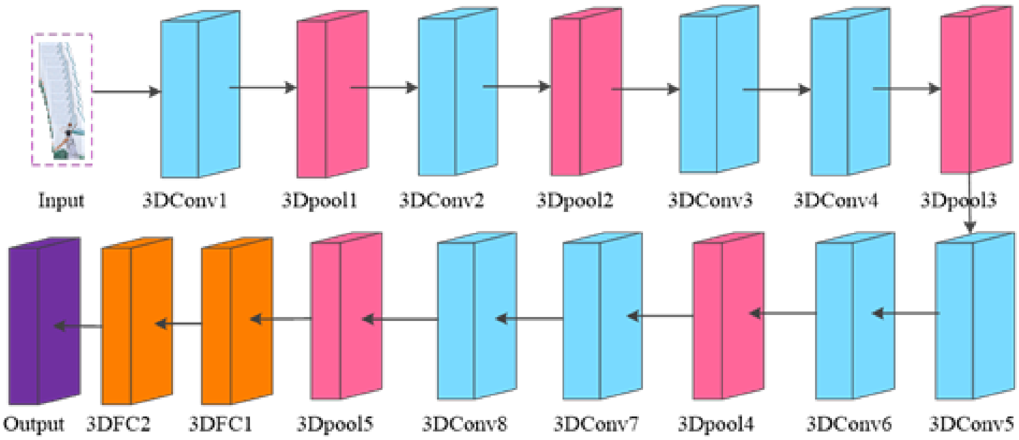


Figure 2. The overall structure of C3D neural network.

2.2. Residual Network

In deep learning networks, the more layers there are, the more detailed and comprehensive data feature extraction. However, the increase in network layers will easily lead to over-fitting and network degradation. In order to avoid the above problems, He et al. [29] proposed a structure that can optimize network training—Residual Networks. The core idea of Residual Networks is Residual connection and identity connection so that the input signals of deep networks learn new features by the Residual connection structure. The identity connection structure is used to preserve the original input characteristics so that the performance of the deep neural network will not be degraded. The structure of the residual block is shown in Figure 3. The mapping relationship is mathematically represented as:

$$H(x) = F(x) + x \tag{1}$$

where x represents the input value, $F(x)$ represents the residual mapping value, and $H(x)$ represents the output value of the residual structure.

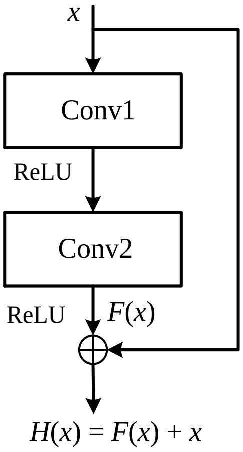


Figure 3. The overall structure of C3D neural network.

The residual network uses constant transformation and jump connection to solve the problem of layer degradation in deep learning networks. The jump connection protects the integrity of information, and the network only needs to learn the difference between input and output, which reduces the complexity [30]. The constant transformation effectively expands the depth of the network, which can avoid the gradient disappearance and training difficulty due to the increase in depth. The performance of the residual network will not degrade but improve to a certain extent as the number of layers increases and improve the network performance and mobility to a certain extent.

2.3. Attention Mechanism

In feature extraction, treating key information and noise data equally will not only waste computing resources but also affect the classification performance of the model [31]. In the model of attention mechanism, firstly, the global image is scanned quickly to find the focus target area, and then the attention resources are focused on this area to obtain more details of the focus target and suppress irrelevant information around, which greatly improves the efficiency and accuracy of visual information processing. Therefore, in the field of computer vision research, deep neural networks are usually also combined with attention mechanisms [32], which can not only guide the model to achieve a reasonable allocation of computational resources but also effectively improve the performance of deep learning tasks. Attention mechanisms mainly include the spatial attention mechanism and channel attention mechanism. The following two attention mechanisms are introduced in detail:

(1) The spatial attention mechanism acts on the two-dimensional spatial plane and uses the learned attention mask to determine the corresponding attention weights of each element on the feature plane, so as to evaluate the correlation between different spatial positions and the target object and highlight the significant areas in space. The spatial attention mechanism helps the model to search for regions with a high concentration of target objects in the input feature plane and avoids the interference of chaotic background information to a certain extent, which is of great significance for the HAR task.

(2) The channel attention mechanism acts on different convolutional channels of the input features and adaptively adjusts the feature response values of each channel using the learned attention masks. The purpose is to filter out the features that contribute more to the recognition results in the feature grasping process, through which the channel attention is weighted to the different channels of the input features, thus assisting the model to learn more meaningful features.

3. Approach

3.1. Overview

The key problem with behavior recognition is extracting the behavioral features of interest accurately. Current behavior recognition methods extract the overall features of images without distinguishing between body movement area and other areas. In this paper, the attention mechanism and 3D convolutional network are combined to focus on the features of body movement parts. A basketball action recognition method of deep neural network based on a dynamic residual attention mechanism is proposed to enhance the accuracy of basketball motion recognition. The frame diagram of the basketball movement recognition method based on an improved residual attention mechanism is shown in Figure 4. The steps of the method are as follows:

Step 1 Data Preprocessing: Basketball images contain a large amount of noise. In order to achieve a better suppression of noise in the image, this paper uses a median filter to preprocess the image, so that the extraction of basketball action in the video image minimizes interference.

Step 2 The Network Training: In order to be able to recognize basketball actions in video images, the attention mechanism is deployed in each convolutional block to focus on important features and eliminate unimportant feature information, so as to delete feature information that is beneficial to basketball player pose recognition and improve the accuracy of basketball pose recognition.

Step 3 Network Optimization: As the network progresses, a dynamic residual network is used to replace the ordinary C3D convolutional layer in order to improve the information flow and speed up the network training.

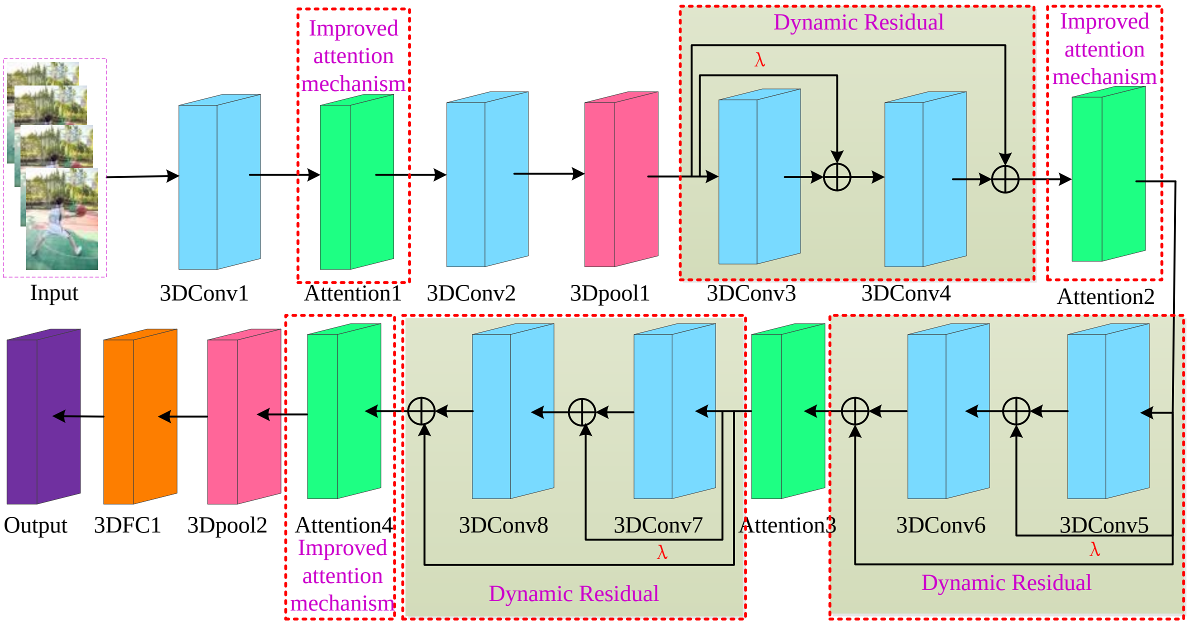


Figure 4. Network framework.

3.2. Dynamic Residual Network

In order to avoid the problem of network degradation caused by increasing the convolutional layers in extracting depth features, this paper does not adopt the traditional form of residual network structure but adopts an improved residual network, which not only can effectively solve the phenomenon of overfitting due to the increase of network depth but also can adaptively extract the feature information of the deep spatiotemporal domain. Because the second convolution in the traditional residual network only convolves the feature vector of the first convolution and does not effectively use the correlation between the input vector of the residual block and the second convolution layer, this paper designs a dynamic residual block with the structure shown in Figure 5.

In Figure 5, a jump connection line with the jump coefficient is added to the input and the second convolution layer to make the input and the first convolution layer form a sub-residual block. In this way, the second convolution layer can not only elucidate the feature vector of the first convolution layer but also learn the input vector of the residual block. Therefore, residual blocks with jump connections can extract the spatiotemporal characteristic information of basketball images more effectively and have higher learning efficiency. Although the introduction of the jump connection line can improve the ability of the residual block to extract feature information from basketball images, the choice of parameters has a great impact on the overall performance. If the parameters are too large, the second convolutional input will have more feature information and increase the burden of the network, and if the parameters are too small, the network feature information will be lost. Therefore, in this paper, we obtain a scaling parameter by two FC layers, and then use the Sigmoid function to scale the one-dimensional parameters in the range of 0–1, which is equivalent to adaptively learning the parameters according to the process of updating the network parameters. This adaptively jumps the residual blocks of the connection line more conducive to the flow of information and accelerates the network training to improve the final recognition effect of the model.

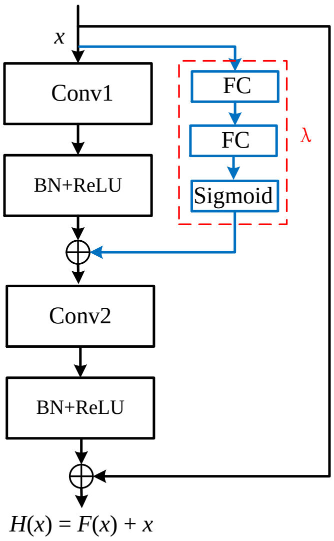


Figure 5. Dynamic residual block.

3.3. Improved Attention Mechanism

Basketball posture recognition requires that video clips are first processed into time-series video frames and then sent to the network for classification and recognition. However, the keyframes in a video clip that can accurately identify the action are often contained in a large number of redundant frames, so attention modules that can generate key frame information are needed in the network.

For the basketball posture recognition task, it is necessary not only to focus on specific key parts of given input images but also to find which are the keyframes. However, the traditional attention mechanism can only focus on some specific features of the image through channel attention and spatial attention, but cannot focus on the keyframes in time. Therefore, this paper adopts the improved attention mechanism to generate the attention feature map along the three dimensions of channel, space and time. In this process, the output features of each dimension are multiplied by the input features of this dimension to carry out adaptive feature refinement to produce the final attentional feature map, as shown in Figure 6.

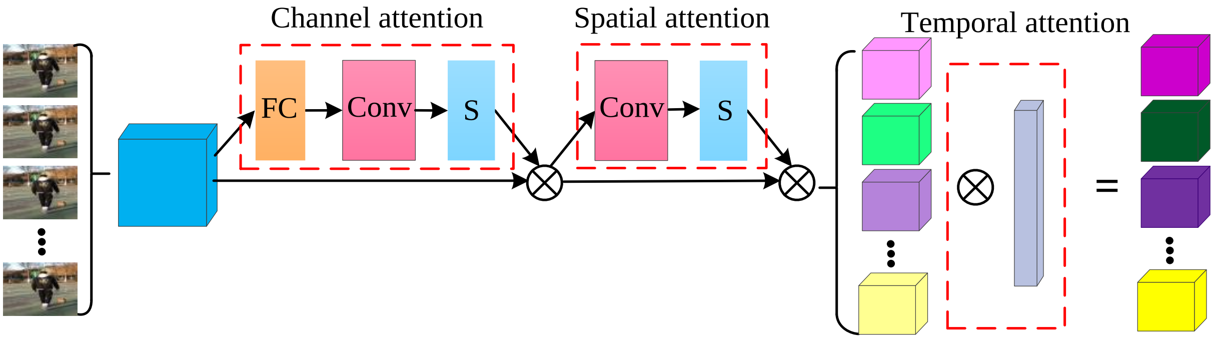


Figure 6. Diagram of improved attention mechanism.

Firstly, the obtained 3D convolutional feature map is passed through the channel attention mechanism (as shown in Equation (2)) to obtain the channel attention feature map F'. Due to the channel attention mechanism, the channel modulation weights are generated by exploring the interdependencies among the channels, and then the channel-level attention distribution is obtained. So global averaging pooling with focusing the signal association information of each channel and avoiding the interference of local spatial information is used. Two-dimensional convolution is used to obtain the non-normalized channel attention mapping. To take full advantage of the feature interdependence, a sigmoid function is used as a gating mechanism to obtain channel attention weights between 0 and 1.

Then, the new adaptive feature refinement feature map F' is obtained by calculating the spatial attention mechanism as in Equation (3). To obtain the correlation between different spatial locations in the feature map F' and the target action, a 2D convolution operation is taken to calculate its spatial attention distribution.

Finally, in order to find out the key frames from the video frames of the feature map after feature refinement, a temporal attention mechanism is used to distinguish the key video frames as shown in Equation (4). The final spatiotemporal channel attention feature map F''' was obtained from the original feature map after adaptive feature refinement.

$$F' = F \otimes M_c(F) \tag{2}$$

$$F'' = F' \otimes M_s(F') \tag{3}$$

$$F''' = F'' \otimes M_m(F'') \tag{4}$$

where $\otimes$ represents the element-level multiplication operator.

Specific parameters of the basketball high-efficiency recognition network based on improved residual attention mechanism are shown in Table 1.

Table 1. Network-specific parameters.

Layer	Kernel Name	Kernel Size	Stride	Output Size
Input	-	-	-	[16,112,112,1]
Conv1	64	[3,3,3]	[2,2,2]	[8,56,56,64]
Attention1	128	[1,1,1]	[1,1,1]	[8,56,56,128]
Conv2	128	[3,3,3]	[2,2,2]	[4,28,28,128]
Pool1	-	[2,2,2]	[2,2,2]	[4,14,14,128]
Residual1	256	[3,3,3]	[1,1,1]	[4,14,14,256]
Attention2	256	[1,1,1]	[1,1,1]	[4,14,14,256]
Residual2	512	[3,3,3]	[2,2,2]	[2,7,7,512]
Attention3	512	[1,1,1]	[1,1,1]	[2,7,7,512]
Residual3	512	[3,3,3]	[1,1,1]	[1,7,7,512]
Attention4	512	[1,1,1]	[1,1,1]	[1,7,7,512]
Pool2	-	[2,2,2]	[2,2,2]	[1,4,4,512]
FC	4096	-	-	4096

4. Experiments

4.1. Experimental Data

In the experimental phase of this section, the original frames of the basketball technical action dataset contain a total of eight technical actions: dribbling, running, shooting, passing, jumping, catching, walking, and no action. Each action has 200 segments, totaling 1600 videos. This class randomly selects 50 videos from each technical action, a total of 400 videos as the test set, and the rest of the videos as the training set.

To prevent a large loss of long-term timing information, a sparse sampling strategy [32] is used for video frame acquisition. First, the basketball video sequence is edited into n clip videos. Then one frame is randomly captured from each clip video. Finally, n-frame images are preprocessed and stacked as the model input tensor. In the next experiment, feature information is extracted from the input tensor, and the extracted feature information is optimized in small batches using the optimizer Adam with the loss function of the cross-entropy loss function. In this paper, the experiments are conducted using Python 3.7, based on the Keras deep learning framework. The experimental parameters are set as shown in Table 2.

Table 2. Network-specific parameters.

Parameter Type	Parameter Values
Batch-size	64
Learning rate	0.001
decay	0.9
Dropout	0.5
epoch	50
classifier	Softmax

4.2. Dynamic Residual Network Impact

In order to verify the influence of the dynamic residual network on posture recognition of basketball players, a dynamic value obtained by dynamic learning of jump connection line in the dynamic residual network proposed in this paper was used for network training with the given value to obtain experimental results, as shown in Figure 7. As can be seen from Figure 7, the model of dynamic value training is more stable and the curve is more stable, mainly because in the process of network training, the optimal parameters are constantly updated according to the needs of the network so that the posture of basketball movement can be better identified. These results demonstrate that the dynamic residual network can improve the information flow, speed up the network training, and enhance the final recognition effect of basketball actions.

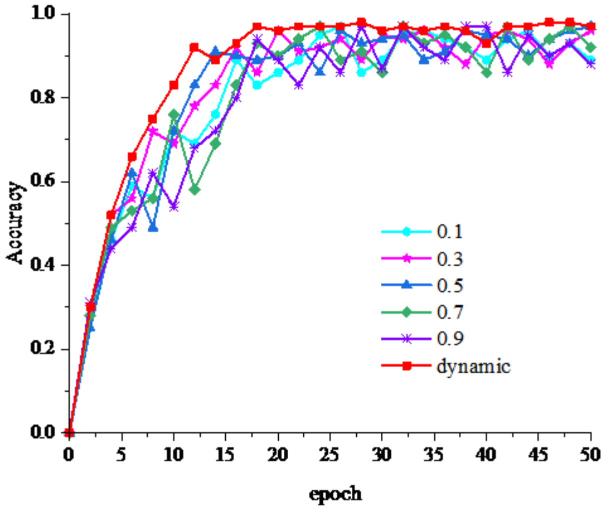


Figure 7. Graph of dynamic residual experimental results.

4.3. Analysis of Experimental Results

The effect of dynamic residual attention networks on video image behavior feature extraction models is observed experimentally. In this experiment, only the original RGB frames of the video are used as input to calculate the video behavior recognition accuracy of each model on the dataset. The experimental results are shown in Table 3. As we can see, our proposed method achieves the best accuracy for basketball action recognition. Specifically, the recognition accuracy of our proposed method, respectively, achieves 8.92%, 9.52%, 11.32% and 17.52% more improvements than the ShuffleNetV2, EfficientNet-B0, ResC3D Network and Traditional C3D Network. It can be found that the performance of the proposed network outperforms traditional C3D networks as well as the existing network model's ResC3D network. This is because the traditional C3D network and ResC3D network have simple network structures, so the features that can be extracted are limited, and better recognition results cannot be achieved. Compared with the traditional C3D network and ResC3D network, the dynamic residual network and attention mechanism network can extract beneficial action features more effectively, reduce over-fitting, and improve network training efficiency. To sum up, the method proposed in this paper can better extract and save enough space-time information features and achieve a better basketball action recognition effect.

Table 3. Comparison of experimental results on basketball action dataset.

Methods	Accuracy
Traditional C3D Network	80.3%
ResC3D Network	86.5%
Only dynamic residual networks	87.6%
EfficientNet-B0	88.3%
ShuffleNetV2	88.9%
Only the attention mechanism network	89.4%
Our proposed method	97.82%

In order to verify the effectiveness of the proposed method in this paper, the method is used to recognize different basketball actions: dribbling, running, shooting, passing, jumping, catching, walking, and no action. The experiments show that the proposed method in this paper is more efficient in recognizing basketball actions. As can be seen from Figure 8 the recognition accuracy rates of eight basketball actions are over 96%, and the recognition rate of the whole basketball action is as high as 98.26%. Specifically, the recognition accuracy rate of jumping basketball action is up to 99.27%. The recognition accuracy rate of running basketball action is 96.74%. This result is mainly because the proposed method can fully extract the feature information for action recognition mainly through the improved residual network, and then the improved attention mechanism can efficiently cull the feature information that is beneficial for action recognition. In summary, the method proposed in this paper improves the accuracy of different basketball action recognition through a dynamic residual network and improved attention mechanism.

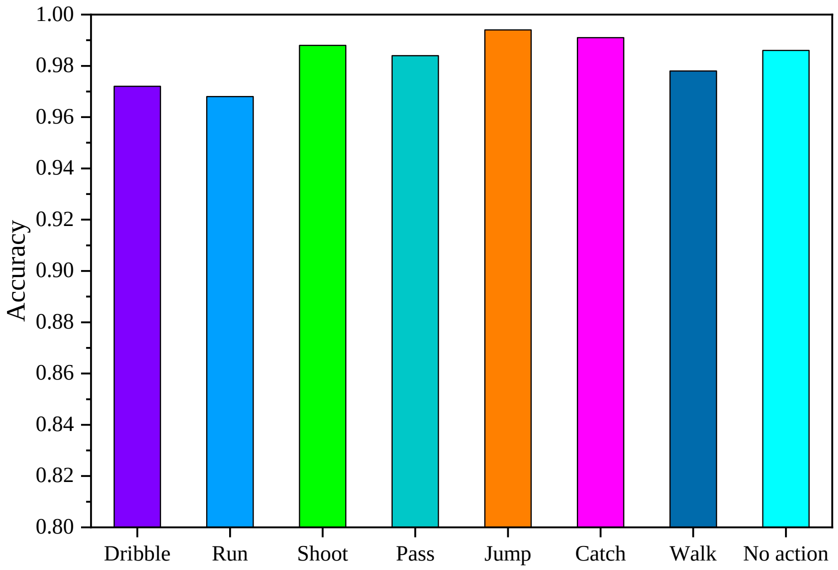
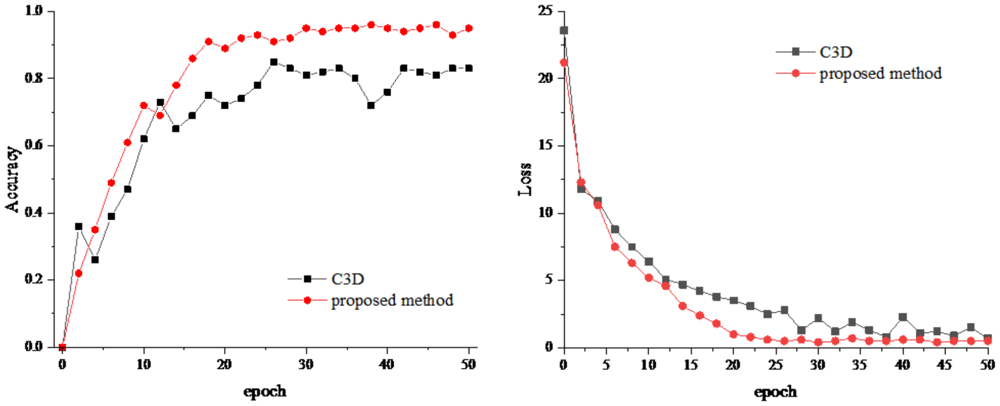


Figure 8. The results of basketball action recognition by this method.

To evaluate the performance of the proposed method in this paper, experimental comparisons with traditional C3D convolution are performed, and the accuracy and loss function results of the experimental training are shown in Figure 9.

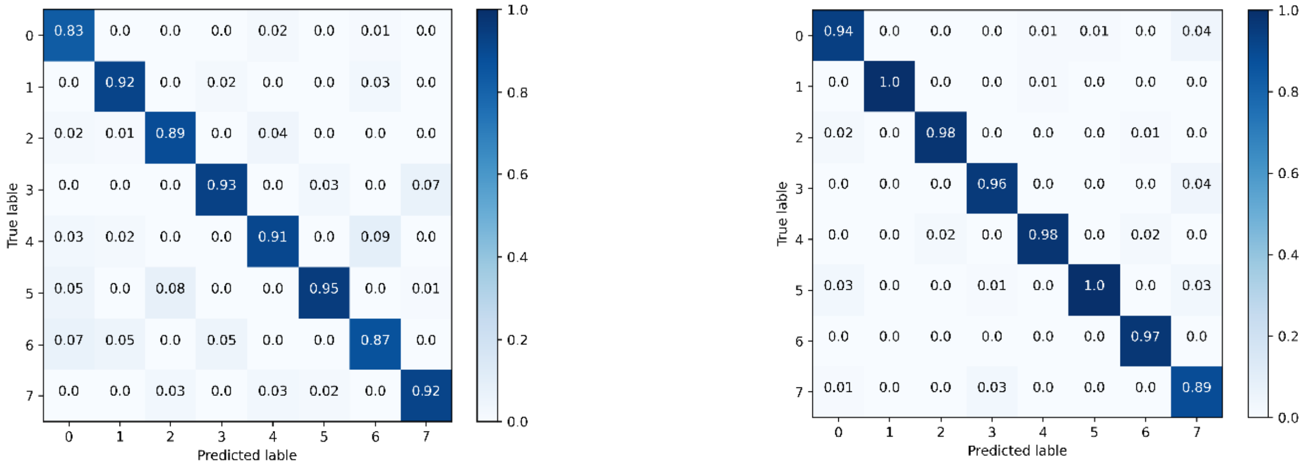


(**a**) (**b**)

Figure 9. Network training result graph. (**a**) Accuracy result graph. (**b**) Loss result graph.

The accuracy and loss results of training in Figure 9 show that the curve tends to be stable as the network is continuously trained. The accuracy of the proposed method is higher than that of the traditional C3D method, and the curve is more stable. The loss curve of the proposed method is not only stable but also low, which indicates that the proposed method extracts sufficient feature information by improving the residual module and then uses the improved attention mechanism for feature culling, and the extracted features are easier for basketball action recognition.

Figure 10 shows the graph of the experimental confusion matrix results. From Figure 10, we can see that the proposed method adopts an improved residual network and attention mechanism, and the proposed feature information is more sufficient and effective, which makes the recognition of basketball action higher, and the recognition result is above 96.5%. In contrast, the traditional C3D method has only 90.25% of final recognition due to the insufficient extracted feature information, especially the dribbling recognition is only 83%. The experimental results verify the effectiveness of the basketball motion recognition method based on the improved residual attention mechanism proposed in the study.



(**a**) (**b**)

Figure 10. Experimental comparison of confusion matrix. (**a**) C3D method. (**b**) Proposed method in this paper.

Figure 11 shows the recognition results of jumping action of basketball training actions by this paper's algorithm, which proves the effectiveness of this paper's algorithm.

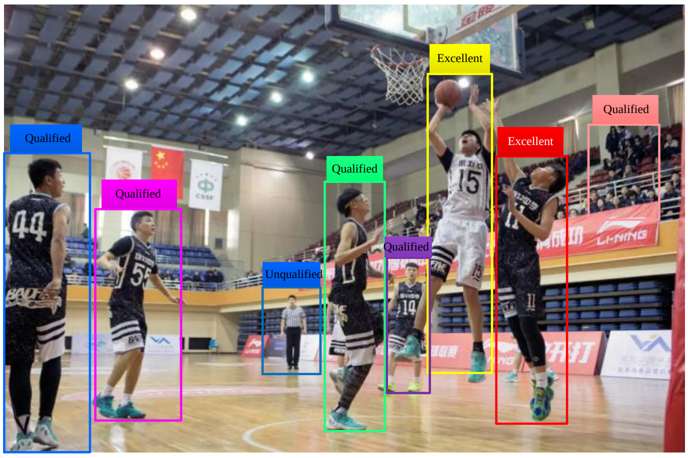


Figure 11. Recognition results of basketball training actions.

4.4. Ablation Study

In order to verify the performance improvement of the improved residual network in Figure 6 with the original residual network. In this paper, experiments are set up to verify the performance improvement of the improved residual network on the original residual structure, and the experimental results are shown in Table 4.

Table 4. Impact of improved residual structure on performance.

Method	Training Time/Min	Accuracy
Residual block	39.4	92.45%
Improve residual block	36.1	97.82%

Meanwhile, to verify the extent to which the two different attention mechanisms in Figure 7 improve the performance of the model. In this paper, channel attention and spatial attention were added to the network separately to verify their performance, and the experimental results are shown in Table 5.

Table 5. Impact of different attention mechanisms on model performance.

Method	Training Time/Min	Accuracy
Channel attention	38.6	91.69%
Spatial attention	39.1	93.58
Channel attention + Spatial attention	36.1	97.82%

From Table 4 it can be demonstrated that the improved residual network in Figure 6 outperforms the original residual structure. As can be seen in Table 5, using one attention mechanism alone is not as effective as using two attention mechanisms at the same time.

5. Discussion

In previous approaches, there is the problem that keyframes are difficult to be focused. To solve this problem, we designed a deep neural network based on a dynamic attention mechanism from several perspectives. First, we used a median filtering method to preprocess the images to obtain basketball motion images with less noise. In addition, we also modified the original convolutional layer into a dynamic residual convolution, which not only improved the correct rate but also made the final network extraction more efficient. Most importantly, we further improved the attention mechanism to be able to extract the most important features in basketball sports images, further improving the accuracy of basketball sports recognition.

The contribution of our research to computer vision and its related fields can be summarized in three main parts. First, we propose a paradigm for solving sports-like behavior recognition, and this problem-solving paradigm can be easily extended to other fields, not only for basketball sports behavior recognition. Second, we bring inspiration for other complex behavior recognition. Importantly, we introduce attention mechanisms and improve them for specific problems, and the experimental part illustrates the effectiveness of our approach. Third, we bring some proven ideas for network design. In this paper, to circumvent the problem of network degradation due to the deepening of network layers, we use an improved residual network structure, which is different from the traditional residual network structure. We discuss this structure in detail in the methods section.

6. Conclusions

In this paper, we propose a deep neural network basketball action recognition method based on an efficient dynamic residual attention mechanism to address the problem of the low recognition rate of basketball sports in the traditional C3D network. To be able to extract video frames adequately, the improved dynamic residual network is used to extract video frames, and the improved dynamic residual convolution is the traditional residual to a weight value through dynamic learning. Then this weight is connected by a jump connection line to extract sufficient feature information for dynamic residuals. Finally, the extracted feature information is censored by the improved attention mechanism to select the key video frames. The experimental results show that the method in this paper has good recognition ability while improving the training speed of the network without loss of performance, and the average accuracy of posture recognition is more than 98%.

Although this method solves the problems of low efficiency and high error rate in the existing basketball action recognition technology to a certain extent, it is mainly effective in recognizing the most common eight kinds of basketball actions, and there is a problem of low efficiency in recognizing some other actions similar to these common movements. Therefore, in future research, we will focus on the recognition of similar basketball actions. In addition, we will also try to apply the method to other related fields.

Author Contributions: W.T.: Conceptualization, Methodology, Writing—Reviewing and Editing. J.X.: Data curation, Writing—Original draft preparation. L.D.: Visualization, Investigation, Software, Validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Project of Guangdong Science and Technology Department (CN), Grant No. 2020B010166005, the Fund project of Department of Science and Technology of Guangdong Province (GDKTP2021032500), Ministry of Education Social Science Fund (22YJC630167) and the Post-Doctoral Research Project (Z000158).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ning, X.; Tian, W.; He, F.; Bai, X.; Sun, L.; Li, W. Hyper-sausage coverage function neuron model and learning algorithm for image classification. *Pattern Recognit.* **2022**, *136*, 109216. [CrossRef]
2. Hou, X.; Ji, Q. Research on the Recognition Algorithm of Basketball Technical Action Based on BP Neural System. *Sci. Program.* **2022**, *2022*, 7668425. [CrossRef]
3. Fan, J.; Bi, S.; Xu, R.; Wang, L.; Zhang, L. Hybrid lightweight Deep-learning model for Sensor-fusion basketball Shooting-posture recognition. *Measurement* **2022**, *189*, 110595. [CrossRef]
4. Yuan, B.; Kamruzzaman, M.; Shan, S. Application of motion sensor based on neural network in basketball technology and physical fitness evaluation system. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 5562954. [CrossRef]
5. Wei, Y.; Jiao, L.; Wang, S.; Bie, R.; Chen, Y.; Liu, D. Sports motion recognition using MCMR features based on interclass symbolic distance. *Int. J. Distrib. Sens. Netw.* **2016**, *12*, 7483536. [CrossRef]

6. Li, G.; Zhang, C. Automatic detection technology of sports athletes based on image recognition technology. *EURASIP J. Image Video Process.* **2019**, *2019*, 15. [CrossRef]
7. Wu, G.; He, F.; Zhou, Y.; Jing, Y.; Ning, X.; Wang, C.; Jin, B. ACGAN: Age-compensated makeup transfer based on homologous continuity generative adversarial network model. *IET Comput. Vis.* **2022**. [CrossRef]
8. Song, Z.; Zhao, X.; Hui, Y.; Jiang, H. Fusing Attention Network based on Dilated Convolution for Super Resolution. *IEEE Trans. Cogn. Dev. Syst.* **2022**. [CrossRef]
9. Zhao, W.; Wang, S.; Wang, X.; Zhao, Y.; Li, T.; Lin, J.; Wei, J. CZ-Base: A Database for Hand Gesture Recognition in Chinese Zither Intelligence Education. In Proceedings of the International Forum on Digital TV and Wireless Multimedia Communications, Shanghai, China, 2 December 2020; Springer: Singapore, 2020; pp. 282–292.
10. Qu, W.; Zhu, T.; Liu, J.; Li, J. A time sequence location method of long video violence based on improved C3D network. *J. Supercomput.* **2022**, *78*, 19545–19565. [CrossRef]
11. Zhang, Y.H.; Wen, C.; Zhang, M.; Xie, K.; He, J.B. Fast 3D Visualization of Massive Geological Data Based on Clustering Index Fusion. *IEEE Access* **2022**, *10*, 28821–28831. [CrossRef]
12. Lin, J.; Mou, L.; Zhu, X.X.; Ji, X.; Wang, Z.J. Attention-aware pseudo-3-D convolutional neural network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7790–7802. [CrossRef]
13. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 20–36.
14. Zhao, Y.; Man, K.L.; Smith, J.; Siddique, K.; Guan, S.-U. Improved two-stream model for human action recognition. *EURASIP J. Image Video Process.* **2020**, *2020*, 1–9. [CrossRef]
15. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 445–450.
16. Li, Y.; Miao, Q.; Tian, K.; Fan, Y.; Xu, X.; Li, R.; Song, J. Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2956–2964. [CrossRef]
17. Yang, J.; Wang, F.; Jieru, Y. A review of action recognition based on convolutional neural network. *J. Phys. Conf. Ser. IOP Publ.* **2021**, *1827*, 012138. [CrossRef]
18. Xu, H.; Das, A.; Saenko, K. R-c3d: Region convolutional 3d network for temporal activity detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5783–5792.
19. De Melo, W.C.; Granger, E.; Hadid, A. Combining global and local convolutional 3d networks for detecting depression from facial expressions. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–8.
20. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
21. Tran, D.; Ray, J.; Shou, Z.; Chang, S.F.; Paluri, M. Convnet architecture search for spatiotemporal feature learning. *arXiv* **2017**, arXiv:1708.05038.
22. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October; pp. 5533–5541.
23. Zhou, Y.; Sun, X.; Zha, Z.J.; Zeng, W. Mict: Mixed 3d/2d convolutional tube for human action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 449–458.
24. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 October 2014; pp. 27–42.
25. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
26. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with enhanced motion vector CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2718–2726.
27. Yao, G.; Lei, T.; Zhong, J. A review of convolutional-neural-network-based action recognition. *Pattern Recognit. Lett.* **2019**, *118*, 14–22. [CrossRef]
28. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition. In Proceedings of the Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015.
29. He, K.; Zhang, X.; Ren, S.; Su, J.N. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Zhang, Z.; Lv, Z.; Gan, C.; Zhu, Q. Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions. *Neurocomputing* **2020**, *410*, 304–316. [CrossRef]

31. Zhao, D. Injuries in college basketball sports based on machine learning from the perspective of the integration of sports and medicine. *Comput. Intell. Neurosci.* **2022**, *2022*, 1429042. [CrossRef] [PubMed]
32. Wang, J.; Chen, Y.; Chakraborty, R.; Yu, S.X. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13–19.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.