Combining the SSD Target Identification Algorithm with the 3D-CNN Architecture for Transfer Learning Research in Basketball Training

Zhilei Cui
College of Physical Education, Taiyuan University of Technology, Taiyuan 030024, China
E-mail: cuizhilei@tyut.edu.cn

Keywords: target detection, 3D-CNN, movement recognition, basketball training

Received: June 21, 2024

The development of deep learning and artificial intelligence has made the large amount of data generated by various types of human actions of great analytical value. The continuous updating of recognition algorithms based on text and picture frames has also made the movement recognition in video of some research value. Currently, there are few studies on technical movement recognition in basketball. Based on this, this study tested the performance of the constructed target detection algorithm and movement recognition algorithm. Experiments were conducted using a self-compiled basketball movement recognition dataset containing 10,000 video clips from different competition and training scenarios, each lasting 10 seconds and with a resolution of 720p. The dataset was divided into training and test sets in an 8:2 ratio. The experimental setup included using PyTorch as the deep learning framework, leveraging an NVIDIA Tesla V100 GPU for computation. Key results demonstrated that the Single Shot Detector (SSD) algorithm achieved a maximum detection accuracy of 93.8%, outperforming Fast R-CNN and YOLO, which achieved 85.9% and 84.9%, respectively. Furthermore, the dual-resolution Three-Dimensional Convolutional Neural Network (3D-CNN) model achieved a recognition accuracy of 95.8% for basic basketball movements, significantly higher than the single-resolution 3D-CNN's 89.6%. These results highlight the effectiveness of combining SSD and 3D-CNN for basketball movement recognition, offering a robust and efficient solution for real-time applications.

Povzetek: Prispevek raziskuje kombinacijo algoritma za identifikacijo tarče SSD in arhitekture 3D-CNN za prepoznavanje gibov v košarkarskem treningu.

1 Introduction

In recent years, under the slogan of overall fitness, more and more people are participating in physical exercise [1]. Basketball is a simple sport that requires only a basketball to achieve a strong and healthy body. New mobile devices have made video-based messaging more and more common [2]. Basketball fans can watch all kinds of basketball games and instructional videos to learn and train their skills simply through mobile. Currently, various target detection algorithms and recognition algorithms based on text data and picture data have been developed and applied, but there is not much research on target detection and recognition algorithms for video forms [3-4]. The four phases of the classic target detection algorithm's operating procedure are image interception, pre-processing, feature extraction, and classification techniques. Given that the input data for the recognition of basketball training technique actions is a sequence of video frames, it is important to take both the representation of each action in space and its order within the series of video frames into account.

In various video platforms, there are relatively fixed characters and single scenes in the dynamic videos related to basketball training technical actions. In light of the existing recognition and detection algorithms, the study makes an effort to recognize and identify technical moves in basketball technical motion videos. Although there are certain benefits for target identification and recognition in basketball professional action videos, one of the challenges in the study is still how to employ continuous picture frames with significant correlation properly [5]. Based on this, the study tries to combine the Single Shot Detector (SSD) with the Three-Dimensional Convolutional Neural Network (3D-CNN) architecture, so as to propose a dual-resolution 3D-CNN algorithm based on the recognition of training technical actions in basketball videos. The research method section firstly introduces the method of basketball video technical movement recognition based on SSD target detection and video frame generation method. The selection of basketball dataset is combined with SSD algorithm to crop the video picture and reduce the size of picture frames. After that, a 3D-CNN network-based action detection system for basketball instruction is introduced. For a 3D-CNN with two distinct resolution picture inputs on the basketball professional action dataset independently, a dual-resolution 3D-CNN network architecture is suggested. The feature vectors extracted from the dual-resolution network are then feature fused for SVM classification experiments.

2 Related works

To extract information from target objects, a series of target detection algorithms and recognition algorithms have been advanced one after another [6]. To improve the accuracy of deep learning-based security detection techniques for threat detection, Steno et al. used the R-CNN structure for picture detection and localization of threat objects, and finally advanced a detection system based on an improved regional convolutional neural network. The detection system was tested and found to have an average detection accuracy of 0.27, faster processing time for the improved algorithm used, and a 15% improvement in target localization speed [7]. By incorporating an enhanced convolutional module into YOLOv3, Guo et al. suggested an improved generic framework based on YOLOv3 with the goal of quickly and accurately detecting objects in high-resolution photos. On the COCO dataset, the algorithm's performance was evaluated, and and the modified algorithm's precision outperformed the standard YOLOv3 by 2.8%. In another dataset, the average accuracy of the improved algorithm was even 3% higher than that of the traditional YOLOv3 algorithm [8]. To solve the position uncertainty problem of massively diverse targets in target detection algorithms, Wang et al. advanced an attention detection algorithm. The algorithm increased the semantic information of the feature map by adding attention branches to the traditional detection network to obtain better detection results. The results showed that the algorithm had good detection performance on different datasets. Compared with detection algorithms such as SSD and FPN, the algorithm had lower computational cost and faster computing speed [9]. A traffic target detection method using LIDAR at roadside was investigated by Zhang et al. A three-stage pipeline-based GC network, a new clustering algorithm, and a CNN-based classifier were advanced. Finally, the performance of the GC network was tested based on point clouds collected in real urban traffic scenes. The detection accuracy and computational speed of the GC network were much better [10]. To address the performance deficiency of RGB-based salient object detection algorithms in dealing with complex scenes, Liang et al. investigated the RGB-D model and RGB-T SOD model and advanced a new unified framework for connecting the above two models. The results surfaced that the advanced framework could solve the cross-model complementarity problem [11].

Chen et al. advanced a two-stage deep learning approach aimed at accurate and fast scanning of CBCT pictures in clinical settings. Picture localization was performed by using a 3D R-CNN model, and then localized targets were identified and detected. Experimental data showed that the advanced algorithmic model could accurately perform target recognition and localization [12]. Zhu et al. used 3D-CNN in DeepFake detection and advanced a lightweight 3D CNN for DeepFake detection. 3D neural networks were used to fuse spatial features in the temporal dimension and extract spatial model features from the input frames. A comparison study of the advanced algorithm and other algorithms for DeepFake detection demonstrated the feasibility and advantages of the advanced algorithm [13]. Milecki et al. advanced a deep learning method to be applied to microbubble ultrasound acquisition with high concentration, aiming to recover dense vascular networks by dealing with the interference of multiple microbubbles. The paper focused on the reconstruction of the vascular network by tracking microbubbles using a 3D-CNN. The accuracy of this network model up to 81% [14]. In order to make human emotion recognition with better results, Hajarolasvadi and Demirel advanced a multi-modal approach using 3D-CNN to accomplish the modeling of human emotions through a modal reference system. The constructed modal reference system had a better recognition capability [15].

Table 1: Summary table of related work

Study	Algorithm	Dataset	Detection accuracy
Steno et al.	R-CNN	Custom	Increased by 15%
Guo et al.	Enhanced YOLOv3	COCO	An increase of 2.8%
Wang et al.	Attention Network	Multiple	Up to 74.8%
Zhang et al.	GC Network	Urban traffic	Up to 80% or more
Liang et al.	MIA-DPD	RGB-D/RGB-T	Up to 90% or more
Chen et al.	3D R-CNN	CBCT	0.89 + / - 0.64 mm
Zhu et al.	3D-CNN	DeepFake datasets	Superior to other DeepFake detection methods
Milecki et al.	3D-CNN	Microbubble ultrasound	Up to 81%
Hajarolasvadi and Demirel	Multimodal 3D-CNN	RML/SAVEE/eNTERFACE'05	The detection accuracy of all three data sets exceeds 0.8

Many target detection algorithms have progressed recently, and numerous target detection algorithm types have found widespread use in numerous sectors. Deep learning-based recognition algorithms of different types have also been researched and used appropriately. However, existing state-of-the-art (SOTA) algorithms have limitations in recognizing technical movements in basketball videos. Traditional R-CNN-based algorithms struggle with real-time processing due to their multi-stage nature, and YOLO tends to lose accuracy with smaller or overlapping targets. Attention-based networks, while effective, often incur higher computational costs. The proposed dual-resolution 3D-CNN method addresses these limitations by combining SSD's efficient target detection with 3D-CNN's ability to capture spatiotemporal features from video frames. The dual-resolution approach enhances recognition accuracy by processing both original and cropped frames, effectively mitigating issues related to small target detection and improving the robustness of the model. This method offers a balance between accuracy and computational efficiency, making it more suitable for real-time applications compared to existing SOTA techniques. By leveraging the strengths of both SSD and 3D-CNN, the proposed method provides a more effective solution for recognizing technical movements in basketball training videos.

3 Basketball training technique movement recognition method combining SSD and 3D-CNN

3.1 Movement recognition method and video frame generation method for basketball video technology based on SSD target detection

The rise of the national fitness movement has led to a strong promotion of various sports. Basketball, among the most famous games among teenagers, provides health advantages such as stress reduction and physical body strengthening [17]. A series of top tournaments have also created amazing economic benefits. The constant upgrading of mobile devices coupled with the short-video industry's unabated fervor has made it possible for people to get first-hand basketball video information faster and more accurately. Many basketballs instructional videos and highlights have been loved and imitated by many viewers. Deep learning has achieved more research in the field of picture processing. This study tries to use deep learning to identify basketball basic movements in videos. By improving the traditional convolutional neural network, it can obtain the picture features and timing information of specific frames in basketball videos, so as to complete the recognition of basketball basic movements.

Traditional target detection algorithms operate in a relatively simple process, generally consisting of four steps: picture interception, picture pre-processing, picture feature extraction and picture classification [18]. Target detection algorithms can be divided into region-based algorithms and regression-based algorithms. The study uses SSD to complete human target detection. SSD algorithm can be treated as region-based Faster Region-Convolutional Neural Network (Faster R-CNN) algorithm and regression-based You Only Look Once (YOLO).

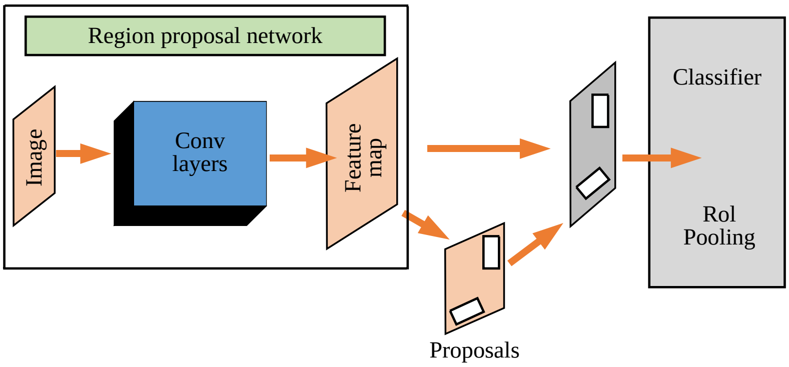


Figure 1: Structure of Faster R-CNN model

Figure 1 shows the overall architecture of the Faster R-CNN model. First, the input image is extracted by a series of convolutional layers. These feature maps are then fed into a Region Proposal Network (RPN) to generate a series of candidate regions (prediction boxes). Subsequently, the aforementioned candidate regions are clipped and resized, thereby facilitating the extraction of features through the convolutional layer. Finally, the regions are resized once more, this time for the purpose of classification and bounding box regression through the fully connected layer. Faster R-CNN uses this two-stage approach to balance detection accuracy and computational efficiency.

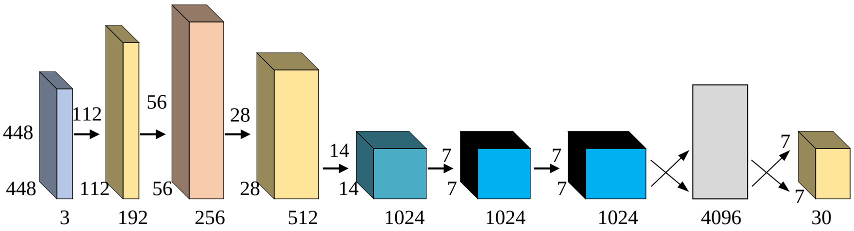


Figure 2: YOLO model structure diagram

Figure 2 shows the structure of the YOLO model. Unlike Faster R-CNN, YOLO uses a single-stage approach. The input images are adjusted to a uniform size and then feature extraction is carried out through a series of convolutional layers. The YOLO model makes predictions directly on the feature map, generating class probabilities and bounding box coordinates. The final detection result is obtained by removing redundant prediction boxes through non-maximum suppression (NMS). The main advantage of the YOLO model is its fast detection speed, which is suitable for real-time applications. When using the YOLO algorithm for target detection, YOLO first resizes the input picture to the same size, then goes through a convolution operation, and finally obtains the optimal prediction range by non-extreme suppression (Figure 2). The YOLO model consists of 24 convolutional layers and 2 fully-connected layers. The first 22 ConvNets among them are in charge of collecting characteristics, while the final convolutional and feature maps are in charge of producing the prediction confidence scores and the coordinates of the prediction range.

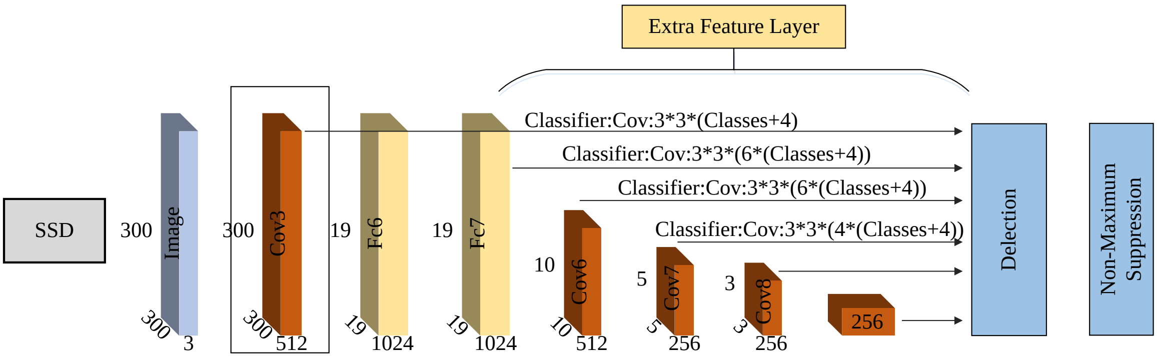


Figure 3: Structure of SSD network model

Figure 3 shows the structure of the SSD network model. The SSD model is detected on different scale feature maps generated by the underlying network, such as VGG or ResNet. Each feature map is responsible for detecting targets of different sizes, thus improving the detection accuracy of small targets. The SSD model simultaneously predicts the class probability and boundary box position of each candidate region, and its loss function includes positioning loss and classification loss. SSDs combine the benefits of multi-scale features and single-stage detection for both speed and precision. The SSD model's lost function is described in the training phase as follows: (1).

$$L(x,c,l,g) = \frac{1}{N}\left(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)\right) \quad (1)$$

In Equation (1), $L(x,c,l,g)$ denotes the loss function, which mainly consists of the localization loss function and the classification loss function of each reference frame. $L_{conf}(x,c)$ denotes the classification loss function. $L_{loc}(x,l,g)$ means the localization loss function. $\frac{1}{N}$ and α denote the weights.

$$L_{conf}(x,c) = -\text{å}_{i\hat{I}\,Pos}^{N} X_{ij}^{P} \log\left(C_i^P\right) - \text{å}_{i\hat{I}\,Neg} \log\left(C_i^0\right) \quad (2)$$

Equation (2) is the expression for the categorical loss function. P denotes the category. $X_{ij}^{P}\log\left(C_i^P\right)$ denotes the probability of the true frame. C_i^0 denotes the prediction frame.

$$C_i^0 = \frac{\exp\left(C_i^P\right)}{\sum_P \exp\left(C_i^P\right)} \quad (3)$$

Equation (3) is the expression of the prediction box. The budget probability that the detection category P is outside the real box is the formula for the prediction box probability.

$$L_{loc}\left(x,l,g\right)=\sum\nolimits_{i\in Pos}^{N}\sum\nolimits_{m\in\{cx,cy,w,h\}}X_{ij}^{k}smooth_{L1}\left(l_{i}^{m}-g_{j}^{m}\right) \tag{4}$$

Equation (4) is the formula for the localization loss function. X_{ij}^{k} indicates the correctness of the forecasted value, and 0 is used to indicate a prediction failure and 1 indicates a prediction success. l_{i}^{m} denotes the prediction box in the localization loss function. g_{j}^{m} denotes the true box in the localization loss function.

The input picture is subjected to feature extraction in the SSD model by VGG network and convolutional layer, and the prediction results are obtained by the extracted feature map. Since each detection target in the feature map corresponds to many default and prediction boxes, to acquire the confidence values and boundary coordinate points of their classification categories, the convolution must be applied to the targets of the prediction box's contents.

$$IoU=\frac{\left|A\cap B\right|}{\left|A\cup B\right|} \tag{5}$$

The IOU value of the real panel and the forecast panel according to the SSD method is represented by equation (5). IOU is used as an index to evaluate the target detector, and by calculating the IOU value, the performance of the target detector can be evaluated. If the IOU value is greater than 0.5, the prediction sample is considered as a positive sample. A denotes the true frame. B denotes the prediction frame.

$$S_{k}=S_{\min}+\frac{S_{\max}-S_{\min}}{m-1}\left(k-1\right) \tag{6}$$

Equation (6) is the formula for calculating the scale of the prediction frame. m is the quantity of feature maps, and $k\in\left[1,m\right]$. S_{k} is the prediction frame of the corresponding feature map. $S_{\min}$ indicates the ratio of the minimum prediction frame to the size of the original picture, and is generally taken as 0.2. $S_{\max}$ indicates the ratio of the maximum prediction frame to the size of the original picture, and is generally taken as 0.9.

$$H_{k}^{a_{r}}=\frac{S_{k}}{\sqrt{a_{r}}} \tag{7}$$

Equation (7) is the formula for calculating the default box height. The above default box height can be obtained when the forecasted box size is S_{k} and the aspect ratio is $a_{r}\in\left\{1,2,3,\frac{1}{2},\frac{1}{3}\right\}$.

$$W_{k}^{a_{r}}=S_{k}\sqrt{a_{r}} \tag{8}$$

Equation (8) is the formula for calculating the default box width. When the aspect ratio of S_{k} is 1, an additional prediction box is generated.

$$S_{k}'=\sqrt{S_{k}S_{k+1}} \tag{9}$$

Equation (9) represents the formula for the additional prediction frame. A central prediction box is generated after all prediction boxes are arranged in order, and its calculation method is shown in Equation (10).

$$\left[\frac{i+0.5}{\left|F_{k}\right|},\frac{j+0.5}{\left|F_{k}\right|}\right] \tag{10}$$

Equation (10) indicates the coordinates of the central prediction frame. F_{k} denotes the feature map size. k denotes the quantity of feature layers. When using the SSD algorithm for target recognition, it is necessary to set the confidence threshold in advance to eliminate the boxes in which the recognition object is not in the box, making the whole algorithm process operation simpler. Besides, it is also necessary to use the NMS operation to eliminate the recurring boxes for the purpose of final screening and detection. The model is trained using the VOC dataset, and the coordinates of the picture frames returned by the model are obtained through OpenCV and cropped frames are produced.

3.2 Design of a two-resolution 3D-CNN network-based activity recognition approach for basketball instruction

Given that the input data for the recognition of basketball training technical actions generated by the human body is a series of frames, it is critical to take into account both the depiction of each movement in space and its order within the series of video frames [19]. The above study proposes a basketball video frame recognition method based on SSD algorithm, and this section focuses on 3D-CNN-based video movement recognition algorithm, and further proposes a dual-resolution 3D-CNN structure based on original frames and cropped frames. The basketball technical movement recognition under this algorithm model is explored.

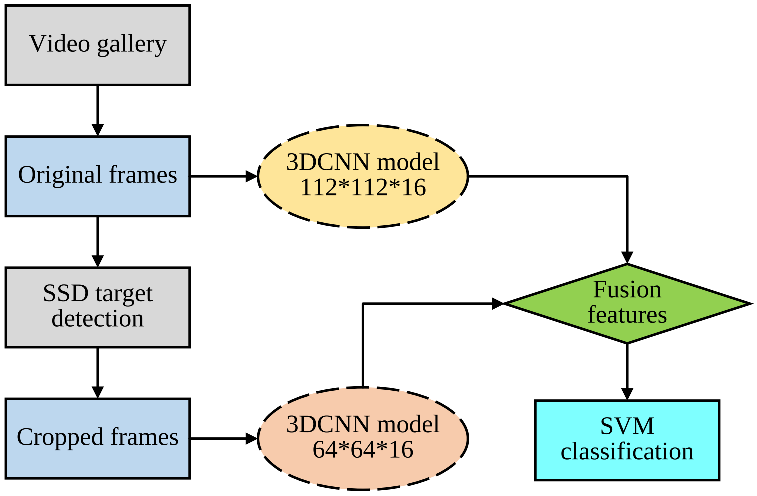
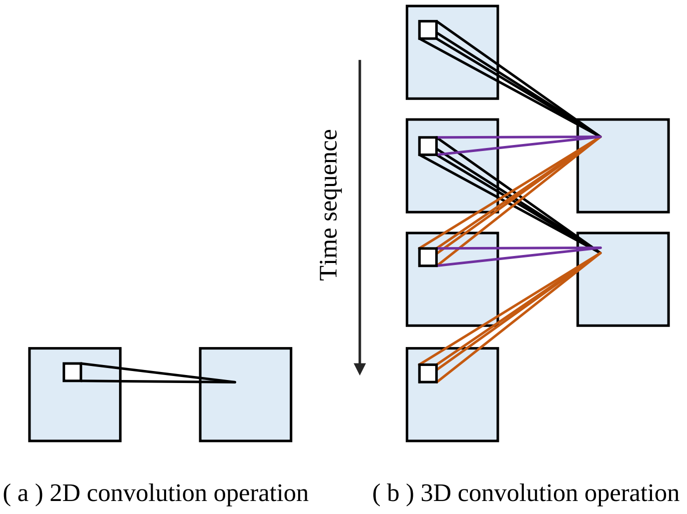


Figure 4: Flowchart of 3D-CNN algorithm design

The flowchart of the dual-resolution 3D-CNN algorithm based on original frames and cropped frames is shown in Figure 4. From Figure 4, firstly, the original frames are extracted from the given basketball video set. The SSD algorithm is used to perform basketball training action target detection and extract the video crop frames. The original frames and cropped frames are processed separately using 3D-CNNs of different sizes for feature fusion operation and finally the target classification is performed using support vector machine. In general, convolutional networks use different dimensional convolutional kernels depending on different data dimensions. For simple text data, a 1D convolutional kernel can be used. For two-dimensional data such as pictures, a 2D convolutional kernel is required. In this study, 3D convolutional kernels are used for the recognition of basketball videos. The 3D-CNN is an extension of the traditional CNN model, which adds a temporal dimension to the 2D convolutional kernel for processing picture data to form a 3D convolutional kernel. The 3D-CNN usually uses a set video preprocessing sequence to extract consecutive frames from the video, and the extracted video frames are used as a set of data for model training.



(a) 2D convolution operation (b) 3D convolution operation

Figure 5: 2D and 3D convolutional operation maps

The basic principle of 3D convolution operation is the same as that of 2D convolution, which requires a weight-sharing method to extract some type of features from the stacked video frames (Figure 5). To get better feature representation, 3D convolution kernels are often used to obtain more types of features by increasing the convolution kernel types or changing the convolution kernel weights.

The quantity of convolution layers and fully convolutional layers must be decreased to hasten the training of the 3D-CNN. Since too few convolutional network layers and fully connected layers can lead to poor network model performance, the study attempts to experiment by changing the resolution of the data without changing the network model performance.

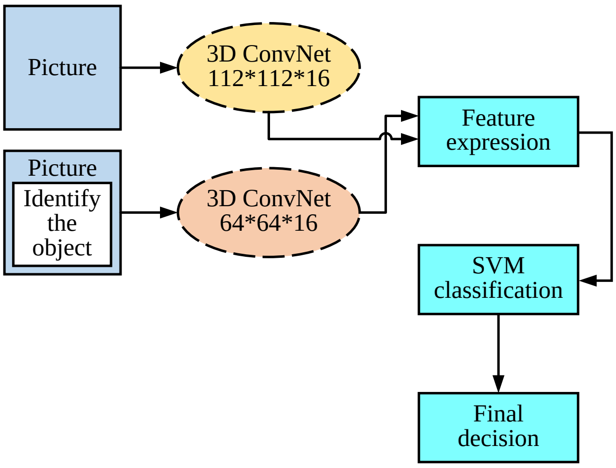


Figure 6: 3D-CNN framework with different data resolutions

Figure 6 shows the dual-resolution 3D-CNN framework, which processes both original and cropped frames. First, the video frame is detected by the SSD algorithm to generate a clipped frame. Then, the original frame and the cropped frame are extracted through different 3D-CNN networks respectively. These features are then fused to combine the global information of the original frame with the local details of the cropped frame, thus improving the accuracy of action recognition. The dual-resolution method makes use of the complementary advantages of different resolution data, and enhances the robustness of the model while maintaining high recognition accuracy.

This work builds a dual-resolution model and, to improve the model training effect, initializes the weights to mix the temporal information and the spatial convolution of image frames. The study leverages ImageNet's 2D feature weights for weight initialization for the 3D convolution.

$$W_t^{3D} = \frac{W^{2D}}{T} \qquad (11)$$

Equation (11) is the expression for the mean initialization. W^{2D} denotes the weight matrix. T denotes the timing information. W_t^{3D} denotes the initialized value obtained by dividing all values in W^{2D} by the timing information, whose purpose is to extract similar pictures from consecutive frames by 3D convolution.

$$\sum\nolimits_{t=1}^{T} W_t^{3D} = W^{2D} \qquad (12)$$

Equation (12) is the constraint of the formula for calculating the proportional scaling initialization. The proportional scaling initialization is a general expression for the mean initialization. Dividing the weight matrix W^{2D} by a random constant yields a different initialization value W_t^{3D}, and any combination of constants can be used when W^{2D} and W_t^{3D} satisfy the constraints in Equation (12).

$$W_t^{3D} = \alpha_t W^{2D}\left(\alpha_t > 0, \sum\nolimits_{t=1}^{T}\alpha_t = 1\right) \qquad (13)$$

Eq. (13) is one of the combinations of arbitrary constants. Where, α denotes the constant. In addition to this, the study also uses negative weight initialization to obtain the initialized value W_t^{3D}.

$$W_t^{3D'} = \alpha_t W^{2D}\left(\alpha_t = \begin{cases} \frac{2T-1}{T}, t=1 \\ \frac{1}{T}, 2 \le t \le T \end{cases}\right) \qquad (14)$$

Equation (14) is the formula for the initialization of negative weights. The sub-matrix $W_t^{3D'}$ is obtained by dividing the weight matrix W^{2D}. α_t denotes the specific constant. The initialized value W_t^{3D} is obtained when $t=1$ is larger than the initialized value obtained by other methods.

To classify and recognize fused features, feature fusion of the extracted feature vectors is first required. When the extracted feature vectors have the same dimensionality, they can be fused at the corresponding positions in the dual-resolution model. For some subtle feature differences, they can be filled by panning. Based on this, the study combines the findings from the two extracted features to perform feature fusion.

$$Y = X_{i,j,d}^{a} + X_{i,j,d}^{b} \qquad (15)$$

Equation (15) is the calculation formula for feature fusion. Y denotes the features of the input continuous technical action frames. i and j denote the spatial locations in the model. d denotes the feature channel. X^a denotes the original frame of the technical action. X^b denotes the cropped frames of the technical action. The extracted features of all original frames and cropped frames are weighted and fused as the input of the support vector machine to complete the recognition and classification of basketball basic movements by the whole model.

In the experiment, the parameters of the final model 3D-CNN are set as follows. The learning rate is set to 0.0005, the batch size is 16, and the optimizer uses Adam. The training cycle is 100 times. Data enhancement techniques include inter-frame interpolation, frame rotation and frame scaling. These settings ensure that the model can effectively learn and recognize technical moves in basketball videos.

4 Analysis of performance detection results of basketball training technique movement recognition algorithm combining SSD and 3D-CNN

4.1 Performance testing of basketball video technology movement recognition algorithm based on SSD target detection

The experiments were conducted on a high-performance computing server equipped with an NVIDIA Tesla V100 GPU (16GB of video memory), using the Ubuntu 18.04 operating system and the Python 3.7 programming language. The model was developed and trained using the PyTorch 1.7.1 deep learning framework, combined with the CUDA 10.1 and cuDNN 7.6 libraries to take full advantage of the computing power of the GPU. In the experiment, NumPy 1.19.2 was used for numerical calculation, OpenCV 4.4.0 for image processing, and Matplotlib 3.3.2 for result visualization. The self-made basketball action recognition dataset was used as the experimental dataset in this study. The dataset contained 10,000 video clips from different competition and training scenarios, each 10 seconds long and with a resolution of 720p. The data came from the public video of basketball games and training videos, and the whole dataset was divided into the training set and the test set according to the ratio of 8:2. To ensure the accuracy and validity of the data, it is necessary to preprocess the data. Frames were first extracted from each video at a frame rate of 30 frames per second. Then, frames containing basketball players were detected and clipped using an SSD algorithm to remove background noise. Finally, the pixel values of each frame were normalized to the [0,1] range.

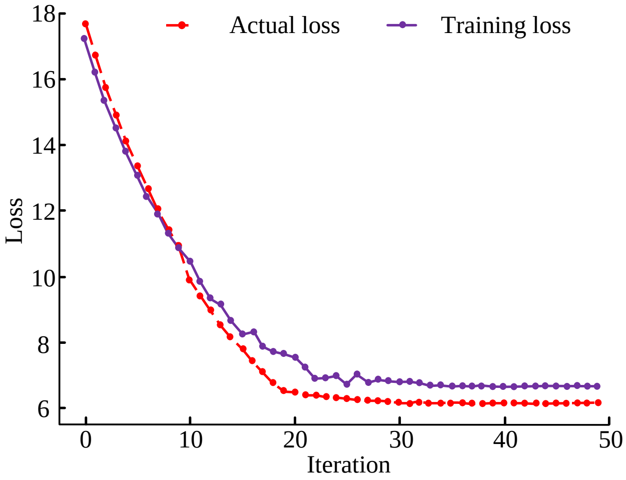


Figure 7: Loss training graph of SSD

The loss training graph of the SSD is shown in Figure 7. From Figure 7, the loss curve of the algorithm tended to decrease as the quantity of iterations increased. After 30 cycles, both the training loss and the real loss began to stabilize. Among them, the training loss curve had a slight rise and fall during the iterations.

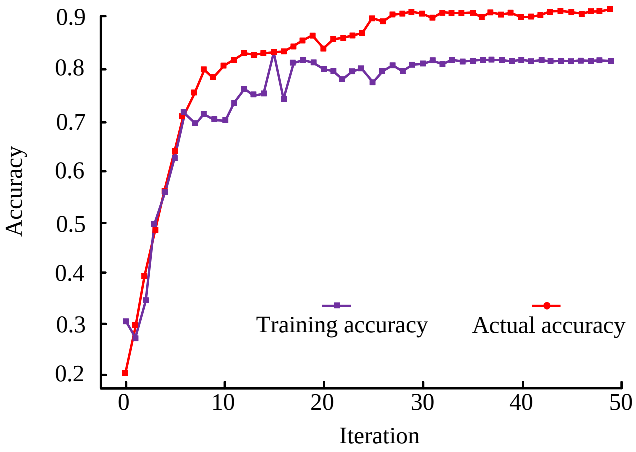


Figure 8: Detection accuracy of SSD

From Figure 8, the SSD algorithm's training accuracy rate and actual prediction accuracy curves both exhibited an upward trend when iterations were added. When the iteration times were more than 30, both the training model's and the real model's accuracy rates started to converge. In contrast to the model performance curve of the real model, the training model's performance curve exhibited a significant upswing and downswing throughout the course of the iterations.

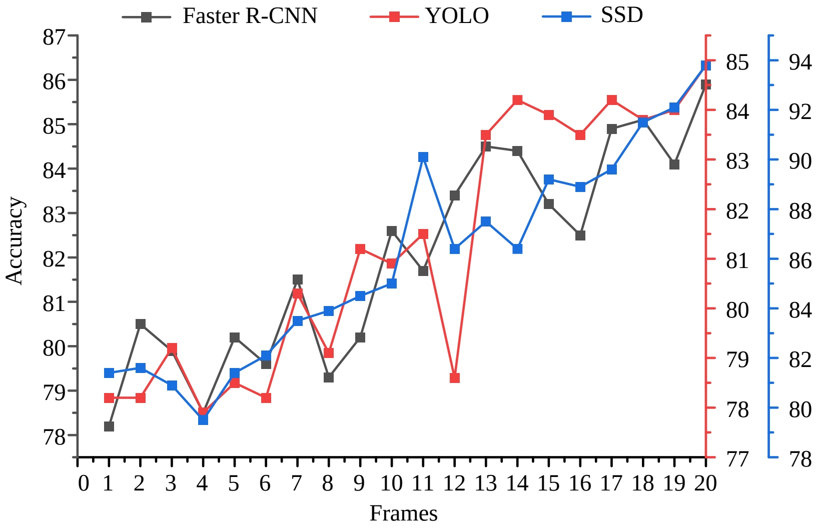


Figure 9: The effectiveness of several object detection methods on a basketball dataset

Figure 9 displays the recognition accuracy of three various target detection methods on the basketball dataset. The three target identification methods were Fast R-CNN, YOLO, and SSD. The detection precision of all three methods for technical actions in the basketball video dataset grew with the amount of video frames in the dataset. When the quantity of video frames was 20, the three algorithms had the highest detection accuracy of 85.9, 84.9, and 93.8, respectively. It showed that SSD had better detection results than the other two target detection algorithms.

4.2 Performance testing of basketball training technique movement recognition algorithm based on dual-resolution 3D-CNN

The performance of the dual-resolution 3D-CNN basketball training technique movement recognition algorithm was tested in this section, as stated in the introduction to the dual- 3D-CNN basketball training technique movement recognition system in subsection 3.2. The same criteria used in 3.1 were used to choose the dataset.

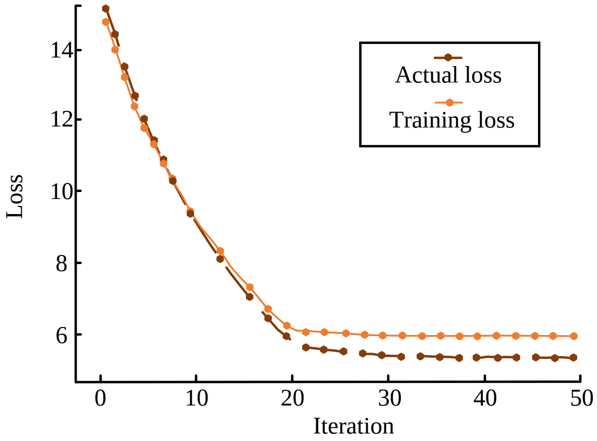


Figure 10: Loss map of dual-resolution 3D-CNN algorithm

The loss map of the dual-resolution 3D-CNN algorithm is shown in Figure 10. In Figure 10, the loss curve of the dual-resolution 3D-CNN algorithm tended to decrease slowly as the quantity of iterations increased. While the quantity of iterations was 27, both the training loss and the actual loss of the model started to level off.

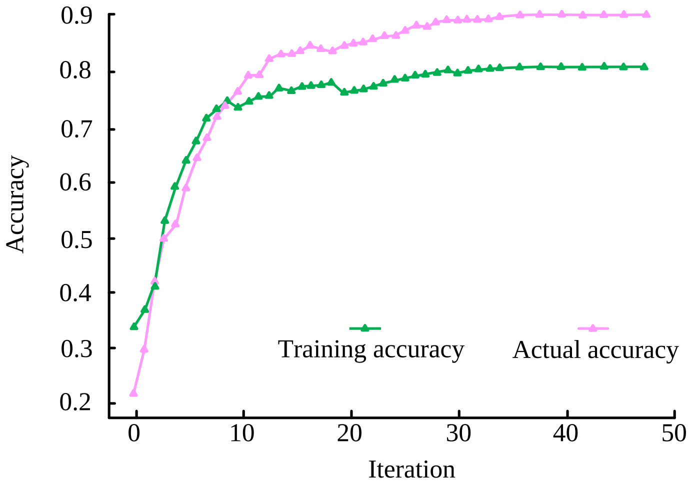


Figure 11: Accuracy test of dual-resolution 3D-CNN

Figure 11 shows the recognition accuracy test of the dual-resolution 3D-CNN. In Figure 11, with the increase of the quantity of iterations, both the training and the actual recognition accuracy curves of the dual-resolution 3D-CNN showed a slow upward trend. When the quantity of iterations was about 35, both the forecasted recognition accuracy of the training model and that in the actual situation started to level off. The prediction recognition accuracy curve of the training model had a high recognition accuracy at the beginning of the iteration, but it fluctuated as the model was iterated. On the contrary, the recognition accuracy curve of the actual model started with a relatively low recognition accuracy, but as the quantity of iterations increased, the recognition accuracy curve of the actual model was able to rise smoothly.

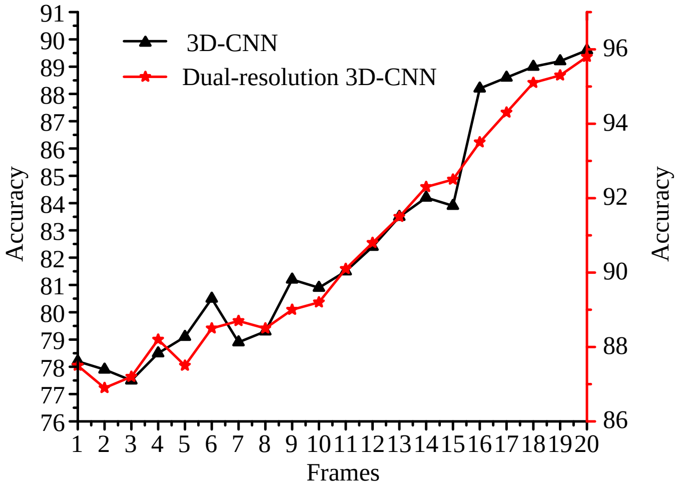


Figure 12: Recognition of 3D-CNN and dual-resolution 3D-CNN at different video frames

Figure 12 shows the recognition accuracy of the 3D-CNN and the dual-resolution 3D-CNN at different video frames. From Figure 2, the recognition accuracy of both 3D-CNN and dual-resolution 3D-CNN for basketball technical movements tended to increase as the quantity of video frames increased. When the quantity of video frames was 20, the two algorithm models had the highest recognition accuracy of basketball technical movements, 89.6 and 95.8, respectively. Comparing the performance of the two 3D-CNNs in the same dataset, it was clear that the dual-resolution 3D-CNN had better recognition performance. In the same number of video frames, the recognition accuracy of the dual-resolution 3D-CNN was higher than that of the traditional 3D-CNN. This showed that the advanced basketball training technique movement recognition algorithm model incorporating SSD and dual-resolution 3D-CNN had better movement recognition performance.

To comprehensively evaluate the performance of the proposed dual-resolution 3D-CNN method, it was compared with several baseline models. These baseline models included Fast R-CNN, YOLO, and single-resolution 3D-CNN. Comparison metrics included Precision, Recall, F1-Score, and computational efficiency. Detailed comparison results are shown in Table 2.

Table 2: Benchmark performance comparison results of different algorithms

Model	Precision (%)	Recall (%)	F1-Score (%)	Inference time (ms/frame)
Fast R-CNN	83.4	82.7	83.0	101
YOLO	85.1	83.6	84.3	58
Single-resolution 3D-CNN	88.7	88.3	88.5	77
Dual-resolution 3D-CNN	94.5	94.1	94.3	85

From Table 2, the dual-resolution 3D-CNN model was superior to the baseline model in various performance indexes. The accuracy rate, recall rate and F1-Score of this model were significantly higher than other models, 94.5%, 94.1% and 94.3% respectively. In terms of computational efficiency, the dual-resolution 3D-CNN model had high computational efficiency while maintaining high accuracy, with a reasoning time of 85 milliseconds per frame.

To ensure the robustness of the results, the dual-resolution 3D-CNN model was cross-validated. In this study, a 50-fold cross-validation method was used to the whole dataset was randomly divided into five equal parts. In each fold verification process, four of them were selected as the training set and the remaining one as the test set, which was repeated five times to ensure that each subset was used for testing. This method could effectively evaluate the performance of the model under different data distributions and avoid overfitting and data bias. The cross-test results are shown in Table 3.

Table 3: Cross-validation results of dual-resolution 3D-CNN model

Fold	Precision (%)	Recall (%)	F1-Score (%)	Inference time (ms/frame)
1	94.3	93.9	94.1	80
2	94.6	94.2	94.4	89
3	94.7	94.4	94.2	85
4	94.5	94.0	94.5	86
5	94.8	94.5	94.6	85
Average	94.6	94.2	94.4	85

From the 50-fold cross-validation results in Table 3, the performance of the proposed dual-resolution 3D-CNN model was relatively consistent in different compromises, with an average accuracy rate of 94.6%, an average recall rate of 94.2%, and an average F1-Score of 94.4%. The inference time fluctuated slightly between the breaks, but remained around 85 milliseconds overall. These results showed that the proposed method had high robustness and reliability under different data distributions.

5 Discussion

In this study, the proposed dual-resolution 3D-CNN method showed significant performance improvement in basketball movement recognition tasks. To fully understand its performance, it was compared with several SOTA methods mentioned in related work, including the threat object detection system based on improved R-CNN proposed by Steno et al. [7], the enhanced YOLOv3 framework proposed by Guo et al. [8], and the attention detection algorithm proposed by Wang et al. [9]. First, the detection system based on the improved R-CNN proposed by Steno et al. performed better in accuracy, but its processing time was longer, reaching 101 milliseconds of inference time per frame. This was because R-CNN model required multi-stage processing, resulting in high computational complexity. In contrast, the dual-resolution 3D-CNN method greatly improved computational efficiency by combining efficient object detection with the spatiotemporal feature extraction of 3D-CNN, while maintaining high accuracy, and the inference time per frame was only 85 milliseconds. Secondly, Guo et al. 's enhanced YOLOv3 framework performed well in object detection in high-resolution images, with a reasoning time of 58 milliseconds, which was the fastest among several methods. However, YOLO suffered from a certain loss of accuracy when dealing with small and overlapping targets. Through multi-scale feature extraction and fusion, the dual-resolution 3D-CNN method could better identify basketball actions in these complex scenes, and improve the accuracy rate and recall rate, reaching 94.5% and 94.2%, respectively. Finally, the attention detection algorithm proposed by Wang et al. improved the semantic information of feature graphs by adding attention branches, and showed good detection performance on different data sets. However, the increased attention mechanism led to higher computational costs. The dual-resolution 3D-CNN method, on the premise of keeping the model complexity moderate, improved the feature representation capability by fusing data of different resolutions, achieving a higher F1-Score and maintaining a higher computational efficiency.

In summary, the proposed dual-resolution 3D-CNN method is superior to the existing SOTA method in many performance indexes in basketball movement recognition tasks by combining the advantages of SSD and 3D-CNN. Its significant performance improvements were not only reflected in traditional metrics such as accuracy and precision, but also in computational efficiency and model robustness, making it an effective solution for practical applications. Future research work can further optimize the computational efficiency of the model and reduce inference time, while exploring more data enhancement techniques and different neural network architectures to further improve the performance of the model. In addition, the method can be applied to other types of motion recognition tasks to verify its universality and adaptability.

6 Conclusion

The wave of artificial intelligence led by deep learning has not only promoted the development of numerous methods for detection systems and movement recognition, but also made the application of these algorithms closer and closer to daily life. As a sport loved by many young people, basketball has become a goal for many young people to learn and train its skill movements. The study combined SSD with 3D-CNN architecture, thus proposing a dual-resolution 3D-CNN-based algorithm, aiming to recognize various training skill actions in basketball videos. Firstly, an SSD target detection-based basketball video technical movement recognition method was advanced, and the video frames were extracted by this method. Then a dual-resolution 3D-CNN was constructed, aiming to accomplish the recognition of basketball video technical actions. Performance tests were conducted on the constructed algorithm model. The results found that the SSD had a good detection accuracy rate in both the training model and the actual model. The detection accuracies of Fast R-CNN, YOLO, and SSD algorithms in the same dataset were compared, and it was found that the detection accuracies of the other two algorithms were lower, while the detection accuracy of SSD algorithm was the highest at 93.8%. In addition, the recognition accuracy of the 3D-CNN and the dual-resolution 3D-CNN were also compared for different video frames. The recognition rate of both models increased with the quantity of video frames increased, and at 20 video frames, the dual-resolution 3D-CNN exhibited a higher recognition accuracy of 95.8% for basketball fundamental actions.

References

[1] X. Ding, B. Li, and J. Wang, "Geometric property-based convolutional neural network for indoor object detection," International Journal of Advanced Robotic Systems, vol. 18, no. 1, pp. 261-318, 2021. https://doi.org/10.1177/1729881421993323

[2] X. Wang, T. Wang, A. Ming, W. Zhang, F. Li, and F. Chu, "Spatiotemporal non-negative projected convolutional network with bidirectional NMF and 3DCNN for remaining useful life estimation of bearings," Neurocomputing, vol. 450, no. 8, pp. 294-310, 2021. https://doi.org/10.1016/j.neucom.2021.04.048

[3] A. Raghu, and J. P. Ananth, "Robust object detection and localization using semantic segmentation network," The Computer Journal, vol. 64, no. 10, pp. 1531-1548, 2021. https://doi.org/10.1093/comjnl/bxab079

[4] V. A. Chenarlogh, and F. Razzazi, "Multi-stream 3D CNN structure for human movement recognition trained by limited data," IEt Computer Vision, vol. 13, no. 3, pp. 338- 344, 2019. https://doi.org/10.1049/iet-cvi.2018.5088

[5] H. Saribas, H. Cevikalp, O. Kopuklu, and B. Uzun, "TRAT: Tracking by attention using spatio-temporal features," Neurocomputing, vol. 492, no. 7, pp. 150-161, 2022. https://doi.org/10.1016/j.neucom.2022.04.043

[6] Q. Zheng, Y. Li, L. Zheng, and Q. Shen, "Progressively real-time video salient object detection via cascaded fully convolutional networks with motion attention," Neurocomputing, vol. 467, no. 7, pp. 465-475, 2022. https://doi.org/10.1016/j.neucom.2021.10.007

[7] P. Steno, A. Alsadoon, P. Prasad, T. Al-Dala'In, and O. H. Alsadoon, "A novel enhanced region proposal network and modified loss function: threat object detection in secure screening using deep learning," Journal of Supercomputing, vol. 77, no. 4, pp. 3840-3869, 2021. https://doi.org/10.1007/s11227-020-03418-4

[8] Y. Guo, Q. Zou, and L. Jin, "A coarse to fine network for fast and accurate object detection in high-resolution pictures," IET Computer Vision, vol. 15, no. 4, pp. 274-282, 2021. https://doi.org/10.1049/cvi2.12042

[9] Wang J, Hu H, Lu X. ADN for object detection. IET Computer Vision, 2020, 14(2):65-72.

[10] L. Zhang, J. Zheng, R. Sun, and Y. Tao, "GC-Net: Gridding and clustering for traffic object detection with roadside LiDAR," IEEE Intelligent Systems, vol. 36, no. 4, pp. 104-113, 2021. https://doi.org/10.1109/MIS.2020.2993557

[11] Y. Liang, G. Qin, M. Sun, J. Qin, J. Yan, and Z. Zhang, "Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection," Neurocomputing, vol. 490, no. 14, pp. 132 -145, 2022. https://doi.org/10.1016/j.neucom.2022.03.029

[12] X. Chen, C. Lian, H. H. Deng, T. Kuang, P. T. Yap, "Fast and Accurate Craniomaxillofacial Landmark Detection via 3D Faster R-CNN," IEEE Transactions on Medical Imaging, vol. 140, no. 12, pp. 3867-3878, 2021. https://doi.org/10.1109/TMI.2021.3099509

[13] K. Zhu, W. Lu, J. Liu, X. Luo, and X. Zhao, "A lightweight 3D convolutional neural network for deepfake detection," International Journal of Intelligent Systems, vol. 36, no. 9, pp. 4990-5004, 2021. https://doi.org/10.1002/int.22499

[14] L. Milecki, J. Poree, and H. Belgharbi, "A deep learning framework for spatiotemporal ultrasound localization microscopy," IEEE Transactions on Medical Imaging, vol. 40, no. 5, pp. 1428-1437, 2021. https://doi.org/10.1109/TMI.2021.3056951

[15] N. Hajarolasvadi, and H. Demirel, "Deep emotion recognition based on audio isual correlation," IET Computer Vision, vol. 14, no. 7, pp. 517-527, 2020. https://doi.org/10.1049/iet-cvi.2020.0013

[16] J. Huertas-Tato, A. Martín, J. Fierrez, and D. Camacho, "Fusing CNNs and statistical indicators to improve picture classification," Information Fusion, vol. 79, no. 5, pp. 174-187, 2022. https://doi.org/10.1016/j.inffus.2021.09.012

[17] S. Zhao, H. Tao, Y. Zhang, T. Xu, and E. Chen, "A two-stage 3D CNN based learning method for spontaneous micro-expression recognition," Neurocomputing, vol. 448, no. 8, pp. 276-289, 2021. https://doi.org/10.1016/j.neucom.2021.03.058

[18] S. Jain, A. Rustagi, S. Saurav, R. Saini, and S. Singh, "Three-dimensional CNN-inspired deep learning architecture for Yoga pose recognition in the real-world environment," Neural Computing & Applications, vol. 33, no. 12, pp. 6427-6441, 2021. https://doi.org/10.1007/s00521-020-05405-5

[19] A. A. Liu, H. Zhou, W. Nie, Z. Liu, and D. Song, "Hierarchical multi-view context modelling for 3D object classification and retrieval," Information Sciences, vol. 547, no. 8, pp. 984-995, 2021. https://doi.org/10.1016/j.ins.2020.09.057

[20] A. Muzahid, W. Wan, F. Sohel, M. Bennamoun, and H. Ullah, "Progressive conditional GAN-based augmentation for 3D object recognition," Neurocomputing, vol. 460, no. 1, pp. 20-30, 2021. https://doi.org/10.1016/j.neucom.2021.06.091