ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 711-729
DOI 10.5755/j01.itc.55.2.43473

Basketball Action Recognition Based on Inv-HRNet-BiLSTM

Received 2025/10/31 | Accepted after revision 2026/03/19

HOW TO CITE: Sun, Y. (2026) Basketball Action Recognition Based on Inv-HRNet-BiLSTM. *Information Technology and Control*, 55(2), 711-729. https://doi.org/10.5755/j01.itc.55.2.43473

Basketball Action Recognition Based on Inv-HRNet-BiLSTM

Yanzhong Sun
School of Physical Education, Zhoukou Vocational and Technical College, Zhoukou, 466000, China

Corresponding author: sunsyanz@outlook.com

With the development of computer vision technology, the application of action recognition in sports is becoming increasingly widespread, and basketball has become a research focus due to its fast pace, drastic limb changes, and frequent occlusion. To address the issues of insufficient spatial modeling capabilities and limited temporal expression effects in existing models, this paper proposes a novel spatiotemporal joint modeling basketball action recognition model called Inv-HRNet-BiLSTM. This model adopts the core methodology of "dynamic spatial feature extraction, temporal dependency modeling, feature enhancement optimization", innovatively integrating the Involution spatial adaptation mechanism with High-Resolution Networks (HRNet) to construct the Inv-HRNet human pose estimation network. By dynamically generating exclusive convolution kernels for each spatial position, the modeling capabilities of joint displacement and motion deformation are enhanced. At the same time, the multi-scale parallel structure of HRNet is used to maintain high-resolution feature transmission and improve detail capture accuracy. On this basis, the model integrates a Bidirectional Long Short-Term Memory Network (BiLSTM) and introduces a dual encoding mechanism, combing position encoding and motion amplitude, to achieve bidirectional deep modeling of action temporal features. This enables the accurate capture of continuity and dynamic changes in action stages. The experimental results show that the accuracy of the Inv-HRNet pose recognition network reaches 94.94%, with a recall rate of 92.23%. The average recognition accuracy of the Inv-HRNet-BiLSTM model in complex scenes, such as occlusion and blur, is 78.58%, and the highest recognition accuracy for eight typical basketball movements, such as dribbling, shooting, and passing, is 95.28%. At the same time, the inference speed is 20.56 ms/frame, and the memory usage is only 723.90 MB, which is significantly better than existing comparative models. The model achieves efficient integration of recognition accuracy, robustness, and real-time performance through organic integration and collaborative optimization of multiple modules, providing a lightweight and high-performance system architecture reference for the field of sports intelligent recognition.

KEYWORDS: Basketball action recognition; Involution operator; HRNet; Feature extraction network; Human pose estimation

1. Introduction

Driven by the growing demand for intelligent sports, accurately recognizing complex basketball movements has become a hot topic in computer vision and human behavior understanding research. Basketball action has the characteristics of fast pace, violent limb changes, frequent occlusion, etc. Traditional methods based on the convolution of still images or single frame pose estimation generally have issues such as insufficient recognition accuracy and weak time series modeling in real scenes [21, 28]. In the last few years, a variety of advanced models have been proposed to improve the recognition performance. For example, the Graph Convolution Network (GCN) has made a breakthrough in the modeling of bone structure. The Transformer structure introduces the global attention mechanism to enhance the time series dependent modeling ability. The High-Resolution Network (HRNet) improves the positioning accuracy of key points through the multi-scale parallel structure [2, 22]. However, these methods still face challenges such as poor spatial adaptability and model redundancy. Fixed convolution is difficult to adapt to dynamic attitude changes, and the complexity of the model structure also limits the real-time and deployment performance [26]. For this reason, the study combines the advantages of spatial adaptive feature extraction and time series modeling, designs HRNet integrating the Involution (Inv-HRNet), and constructs a Basketball Action Recognition (BAR) model with Bidirectional Long Short-Term Memory Network (BiLSTM). The Involution operator can generate a dynamic convolution kernel for each spatial position to enhance the modeling ability of joint point displacement and motion deformation. HRNet multi-branch structure maintains high-resolution feature transfer and improves the ability of detail capture. BiLSTM conducts two-way modeling of the key point sequence to strengthen the understanding of the complete action phase. The model further improves the expression of temporal features by introducing the mechanism of position coding and motion amplitude coding. The study aims to achieve a BAR system with high accuracy and low delay, and to solve the problem that the accuracy and efficiency of the existing model are difficult to take into account in a complex environment. The novelty of the study lies in the combination of a dynamic convolution kernel structure and a bidirectional time series modeling mechanism to build a lightweight and modular recognition framework. The expected results can provide efficient and practical technical support for sports intelligence analysis, game decision support, and training monitoring.

In recent years, aiming at the issues of low accuracy and insufficient time series modeling in BAR, researchers have proposed a variety of improved methods of fusion detection and spatio-temporal modeling. Khobdeh et al. designed a model combining YOLO and Long Short-Term Memory Network (LSTM). The player position was detected by YOLO, and the classification was realized by fusing fuzzy logic and LSTM. The results demonstrated that the performance was better than that of the baseline model on SpaceJam and Basketball-51 datasets [8]. To enhance the recognition efficiency and accuracy of basketball action in video, Yan et al. designed a spatiotemporal neural network based on dual-stream fusion and attention mechanism. By introducing the multi-head spatial self-attention and spatial attention module, the speed and accuracy of BAR were greatly improved [25]. Xue et al. proposed a robot system called InVision based on deep trusted federated learning for ultra-high speed target recognition in complex electromagnetic environments. This method significantly improved the complex representation and description ability of neural networks by introducing geometric component transformers, and developed a fuzzy perception cooperative learning scheme to alleviate the noise labeling problem. In addition, decentralized federated training aimed to alleviate privacy protection issues, effectively search for similarity, and reduce representation redundancy. The results showed that InVision significantly outperformed excellent comparison methods in terms of accuracy, security, speed, and robustness. Moreover, efficient connections and extractions were established while providing security guarantees [24]. Cui proposed a continuous frame action recognition method based on a single multi-frame detection algorithm and a 3D Convolutional Neural Network (CNN) to address that existing BAR methods are difficult to use with continuous frame data. By fusing the features of the original frame and the clipped frame, the average recognition rate of 94.6% was achieved [4].

At the same time, as an efficient feature extraction operator, because of its stronger spatial adaptability, Involution is broadly utilized in the tasks of expression recognition, target detection, and small sample learning. To deal with the issues of poor fitting and weak generalization of small sample learning in new fields, L. Qi et al. proposed a ternary merging network with the Involution operator. The common and unique features were extracted by double encoders combined with the Involution operator, and the training speed and accuracy were greatly improved [11]. To suppress background interference and balance spatial information, Guo et al. proposed a 19-layer expression recognition network based on the visual geometry group improved by an attention mechanism and involution operator. Deep semantic feature extraction was achieved through joint normalization and dense connection, which greatly enhanced the accuracy of expression recognition [5]. To reduce the high pixel-level annotation cost of image instance segmentation in intelligent animal husbandry, H. Wang proposed a weakly supervised segmentation strategy. Through deformable convolution, spatial attention, and Involution to expand the receptive field, the accuracy of individual contour segmentation in stacked scenes was significantly improved [17].

In addition, the advantages of HRNet in key point positioning and multi-scale feature fusion are also paid attention to, which promotes the in-depth application of HRNet in medical image, remote sensing recognition, and target detection. To accurately and efficiently measure the morphological parameters related to rotator cuff injury and solve the disadvantages of manual measurement, Zheng et al. designed a cascaded HRNet model based on deep learning. Through the training and testing of a large number of X-ray photos, the automatic and reliable measurement of shoulder shape parameters was realized, which helped clinical screening and decision-making [29]. To improve the accuracy of rice remote sensing monitoring in cloudy and rainy areas, Cai et al. proposed an improved HRNet deep learning model based on the attention mechanism. By using channel attention and self-attention modules to extract multimodal features and designing a dual-attention fusion mechanism to optimize feature interaction, high-precision identification of rice planting areas was achieved [1]. To solve the problems of complex models and multi-scale target detection, Chen et al. designed a multi-scale feature fusion algorithm based on HRNet and adaptive spatial feature fusion. By improving the network structure, integrating multi-scale features, and introducing a new loss function, efficient detection performance with fewer parameters was achieved [3].

In summary, the current mainstream BAR can be divided into three categories: CNN-based methods, GCN-based methods, and Transformer-based methods. The CNN-based method excels at extracting spatial features, but has limitations in modeling the temporal dimension and spatial adaptability. The GCN-based method utilizes graph structure to capture the correlations between joints, but is sensitive to keypoint localization errors, has a single temporal modeling, and lacks generalization ability. The Transformer-based method performs outstandingly in modeling long sequence actions, but it has high computational overhead, high dependence on data volume, and insufficient ability to capture local features. The common bottlenecks of existing methods include the difficulty in balancing spatial adaptability and temporal modeling capabilities, insufficient robustness in complex scenarios, and the contradiction between model complexity and real-time performance. The Involution operator and HRNet are outstanding in feature extraction and multi-scale fusion, which provides a new idea for improving recognition accuracy. Based on this, a BAR model combining Involution feature extraction and HRNet is designed to strengthen the ability of spatial feature expression and temporal information modeling, realize more accurate and efficient BAR, and help the development of sports intelligent technology. The innovation of the study lies in: 1) A new human pose network (Inv-HRNet) is proposed by integrating Involution convolution and HRNet. The complex spatial structure in basketball can be better modeled through dynamically generated kernels, significantly improving the accuracy of key points; 2) A spatiotemporal framework is constructed by combining BiLSTM with dual position and motion encoding. This framework effectively captures temporal dynamics and improves the robustness of occluded and blurred scenes; 3) A lightweight system IS designed that significantly reduces inference latency and memory usage while maintaining high accuracy, achieving the optimal balance between accuracy and efficiency.

2. Research Design

2.1. Design of Human Pose Estimation Network Based on Inv-HRNet

To improve the modeling ability of complex spatial deformation in human pose estimation, the involution operator is introduced based on the HRNet structure. Compared with the traditional weight-sharing convolution method, Involution can dynamically generate a dedicated convolution kernel for each spatial position, better adapt to the flexible changes of joint points in basketball, and enhance the spatial adaptability of feature extraction [9]. The operation flow of the Involution convolution is shown in Figure 1.

In Figure 1, unlike ordinary convolution, Involution generates a unique convolution kernel for each spatial position in a spatially adaptive manner, thereby effectively enhancing the model's ability to perceive changes in spatial structure [10]. First, for the input feature map, each spatial position is encoded by a lightweight function to generate an intermediate feature whose size is the square of $1\times1\times R^2$. R denotes the size of receptive field. The kernel weight generation of the Involution operation on the location x can be expressed, as shown in Equation (1) [15].

$$K(x)=\phi(x)\in R^{R\times R\times G} \tag{1}$$

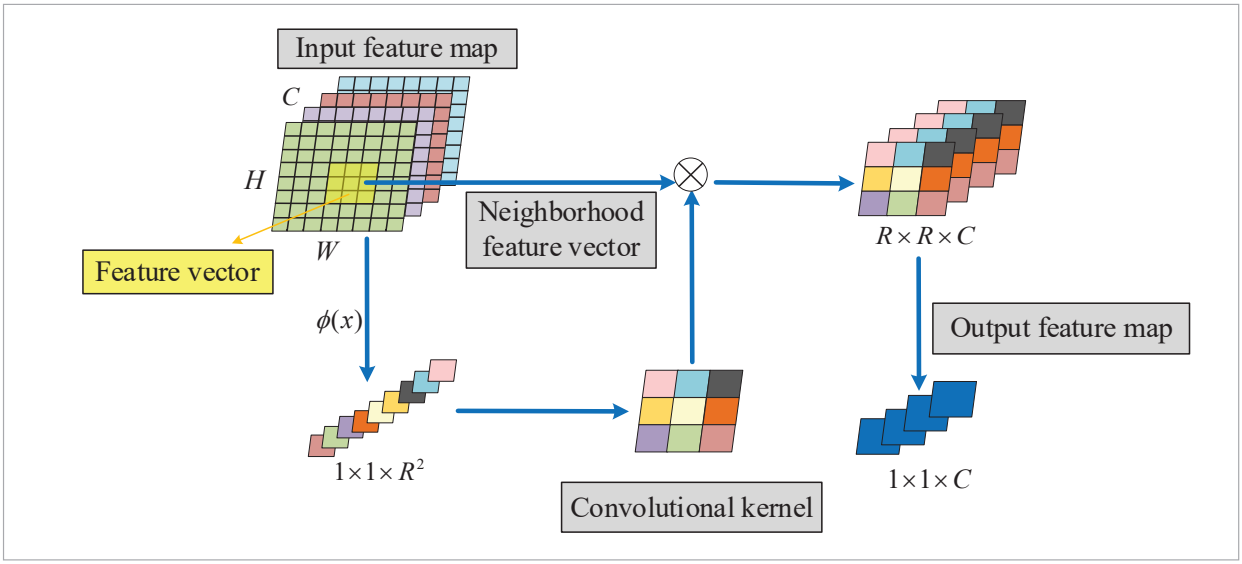
In Equation (1), $K(x)$ represents the weight of the Involution kernel generated at position x. $\phi(x)$ is the kernel generating function. R represents the set of real numbers. G indicates the number of channel packets. Next, this vector is reshaped as a convolution kernel of $R\times R\times1$, which is used for the spatial weighting operation of local regions. The convolution kernel then performs point by point multiplication with the C channel features around the current position to generate a set of response graphs with the size of $R\times R\times C$. The local neighborhood centered on the position x is shown in Equation (2).

$$X_{N(x)}=\{X(x+\delta)\mid\delta\in\Pi\} \tag{2}$$

$N(x)$ refers to the neighborhood area centered on the location x. $X_{N(x)}$ represents the local feature subset around the position x. X represents the original input characteristic diagram. δ is the spatial offset. Π represents the collection of offset positions in the receptive field. All response graphs are then aggregated by channel, and finally a characteristic graph with the same number of channels as the original input is output. The calculation of the Involution operation is shown in Equation (3) [12].

$$Y^g(x)=\sum_{\delta\in R}K^g(x,\delta)\cdot X^g(x+\delta) \tag{3}$$

Figure 1
Involution convolution operation flow.



In Equation (3), $Y^g(x)$ represents the value of the g th group output channel at position x. $K^g(x,\delta)$ represents the weight of δ at the offset of the g th group of Involution cores. $X^g(x+\delta)$ represents the value of the g th group in the input characteristic map at the neighborhood offset δ. In the whole process, the convolution kernel at each position is generated adaptively and is no longer shared with fixed weights, which greatly reduces the parameter redundancy and improves the diversity of feature expression [27]. Therefore, Involution shows great potential in modeling complex spatial structures, such as pose estimation, action recognition, and other tasks. To further strengthen the ability to express multi-scale spatial features, the HRNet structure of high-resolution preservation and multi-branch fusion is introduced. Figure 2 shows the schematic diagram of the HRNet structure.

Figure 2 shows the core design of the HRNet structure. The network uses a multi-branch parallel architecture, and feature maps with different resolutions are transmitted and updated in the whole network simultaneously. The network starts from the high-resolution path, and gradually introduces low- and medium-resolution branches. At each stage, it will carry out feature up and down sampling and cross-branch fusion [13, 23]. In this way, HRNet not only maintains high-resolution features but also integrates spatial semantic information of various scales to improve the recognition ability of key points, edges, and other details [7]. However, HRNet still uses fixed convolution in spatial modeling, and there is still room for improvement in spatial adaptability and redundancy suppression. Therefore, the study proposes to deeply integrate the Involution operator into the HRNet framework. This operator replaces the static convolution kernel through the dynamic kernel generation mechanism, making the feature extraction process spatially adaptive [14]. Figure 3 shows the basic structure of Inv-HRNet.

In Figure 3, the bottleneck structure of the original HRNet uses a three-layer convolution sequence: the first layer 1×1 convolution performs channel dimensionality reduction; the middle 3×3 convolution performs spatial feature extraction; the last layer 1×1 convolution restores channel dimensions, interspersed with batch normalization and ReLU activation functions. The core innovation of the study is to replace the intermediate standard 3×3 convolution with 3×3 Involution operator to form a new Involution bottleneck module. By dynamically generating convolution kernels for each spatial position, the Involution unit realizes adaptive modeling of key areas and effectively enhances the ability to capture spatial features such as joint point displacement and limb deformation in basketball. In the study, the Involution operator is deployed using a strategy of "full branch adaptation + critical stage reinforcement":

Figure 2
Schematic diagram of HRNet structure.

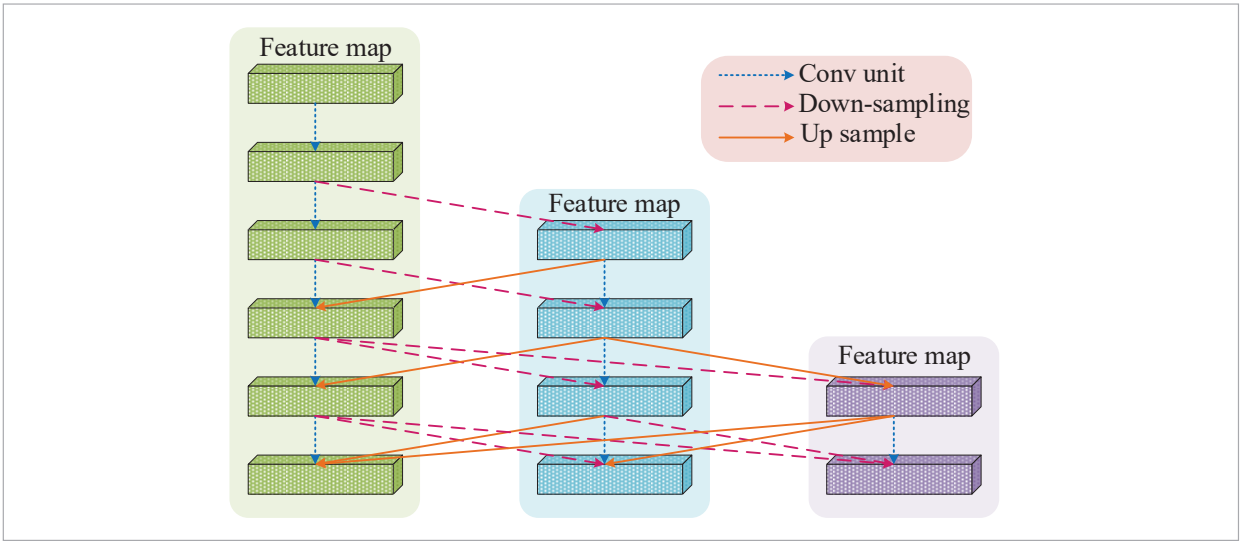
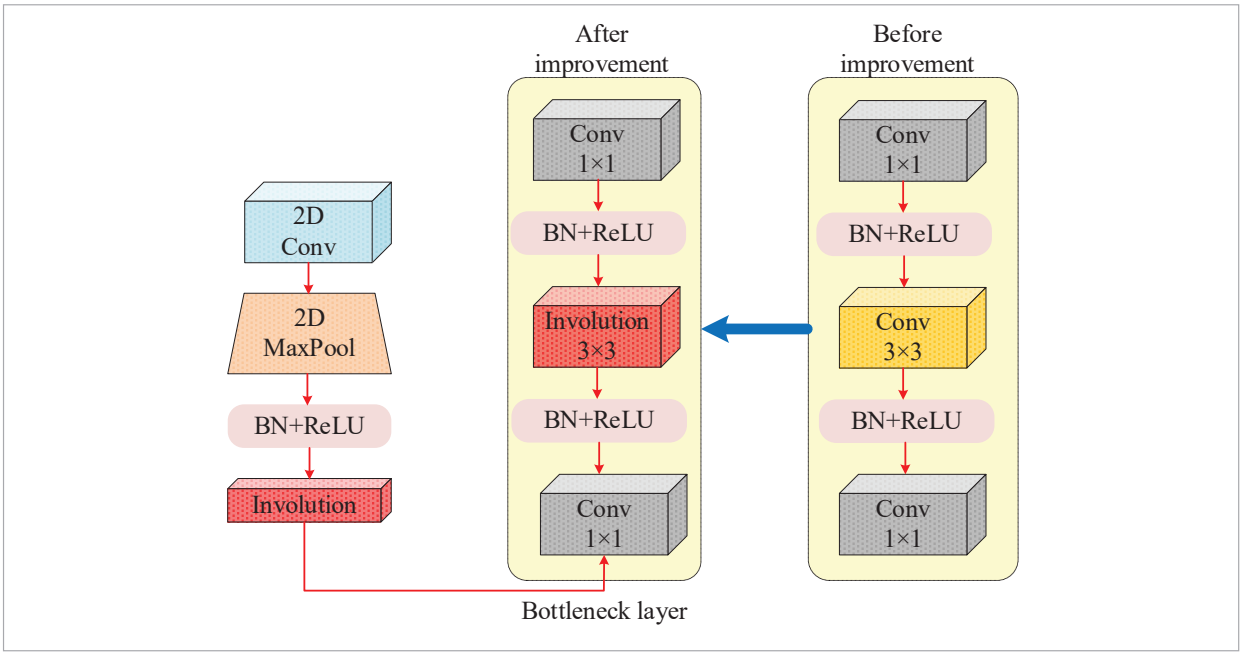
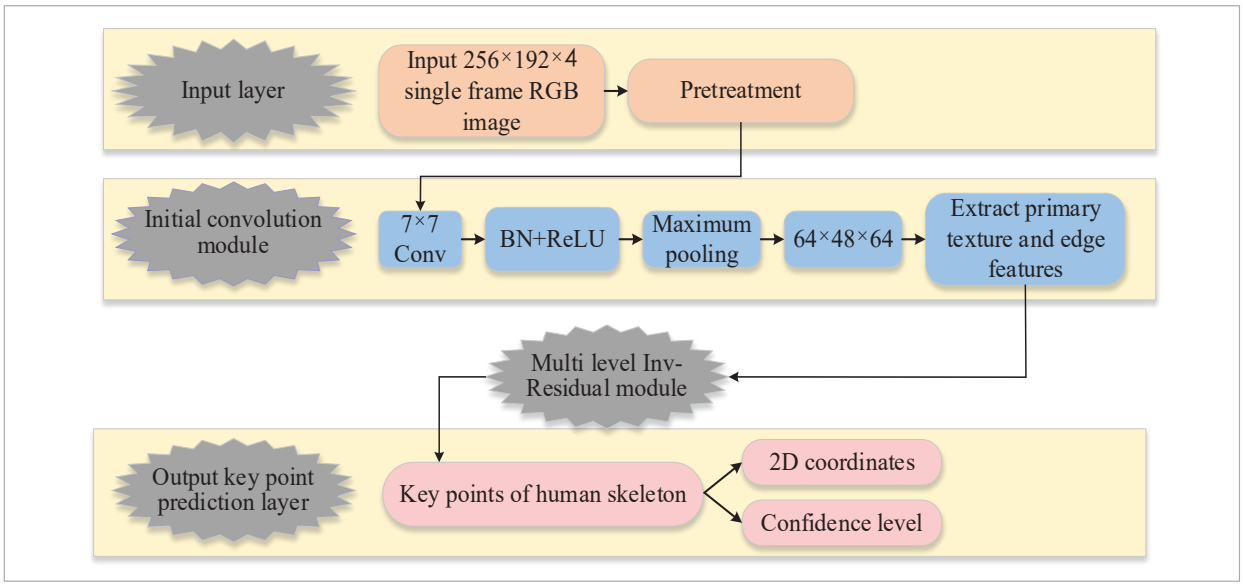


Figure 3
Inv-HRNet structure diagram.



the Involution operator is uniformly introduced in all resolution branches of HRNet to ensure that feature extraction at all scales has spatial adaptability. At the same time, the channel grouping density is strengthened in the middle and high resolution branches (corresponding to P2-P4 stages) to adapt to the precise modeling of core details such as joint points and limb contours in basketball sports. This deployment method not only avoids the uneven feature extraction caused by a single branch application but also balances the computational complexity and representation ability through phased optimization. Full branch deployment does not significantly increase additional overhead, and key stage reinforcement improves the accuracy of core feature capture, enabling the model to efficiently perform in complex pose scenarios. The reason why the Involution operator can significantly improve spatial modeling capabilities lies in its dynamic and position-specific kernel generation mechanism. This fundamentally differs from the fixed shared weights of standard convolutions. Standard convolution uses a globally unified weight matrix and applies the same feature extraction rules to all spatial positions, making it difficult to adapt to severe limb deformations such as wrist flips and torso twists in basketball, as well as complex joint configurations. The Involution adaptively generates exclusive convolution kernels for each pixel position, with kernel weights dynamically determined by the local contextual information of that position, such as the relative position of adjacent joints and the deformation trend of muscle movement. This mechanism enables the model to accurately focus on specific contextual features. For example, during shooting, a bent elbow will generate a kernel that focuses on "joint folding"; during a breakthrough, an extended leg will generate a kernel that focuses on "limb contour continuity", thereby effectively capturing the spatial details of complex poses. At the same time, dynamic kernel generation avoids the feature extraction redundancy caused by fixed weights. This enables the model to accurately lock core joint features when dealing with occlusion, motion blur, and other scenes, greatly enhancing its ability to adapt to complex spatial deformations in basketball. The module retains the 1×1 convolution layer at the beginning and end to realize the channel compression and expansion function, ensuring that the learning ability of the deep network is not affected. Among them, the fusion process of different resolution feature maps in HRNet at each layer is shown in Equation (4).

Figure 4
Schematic diagram of human body pose recognition network based on Inv-HRNet.



$$F_a^{(l)} = \sum_{S}^{b=1} \text{Upsample / Downsample}(F_b^{(l-1)}, r_a, r_b) \quad (4)$$

In Equation (4), $F_a^{(l)}$ denotes the output characteristic diagram of the a th branch of the l th layer. S means the total number of network layers. Upsample/Downsample indicates up sampling or down sampling operation. $F_b^{(l-1)}$ represents the characteristic diagram of the b th branch on the l–1 th layer. r_a and r_b represent the resolutions of the a and b branches respectively. The expressions of up sampling and down sampling operations are shown in Equation (5).

$$F_a^{\uparrow/\downarrow} = \begin{cases} \text{Conv}_{1\times1}(\text{Interpolate}(F_b)), r_b < r_a \\ \text{Conv}_{3\times3,stride-s}(F_b), r_b > r_a \\ F_b, r_b = r_a \end{cases} \quad (5)$$

In Equation (5), $F_a^{\uparrow/\downarrow}$ indicates the result obtained after upsampling or downsampling the feature map. Interpolate (F_b) means up sampling interpolation operation. $\text{Conv}_{1\times1}$ and $\text{Conv}_{3\times3,stride=s}$ indicate convolution operations. s is the downsampling step. Sampling operation is performed when $r_b<r_a$. The down sampling operation is performed at $r_b>r_a$ to ensure that the size of the feature map is consistent during cross-branch fusion. Based on this, a human pose estimation network based on Inv-HRNet is constructed, as shown in Figure 4.

From Figure 4, the overall structure of Inv-HRNet is composed of four parts: input layer, initial convolution module, multi-level Involution residual module, and output key point prediction layer. First, the input layer receives a single frame RGB image with the size of 256×192×3, and performs preprocessing such as normalization and edge filling. Then, the initial convolution module extracts low-level texture features through 7×7 convolution, batch normalization, ReLU activation, and maximum pooling operations, and reduces the image to the feature map of 64×48×64. The backbone structure is composed of several residual bottleneck modules. Each module completes 1×1 convolution dimensionality reduction, 3×3 Involution feature extraction, and 1×1 convolution dimensionality increase in turn, while retaining residual connections to enhance the deep feature transfer ability. By dynamically generating convolution kernels for each pixel position, the Involution unit effectively improves the network's ability to perceive local spatial deformation. The whole network integrates the multi-scale parallel architecture of HRNet to realize the continuous transmission and interaction of different resolution features. Finally, the output module maps the feature to the key points of the human skeleton and calculates the 2D coordi-

nates and confidence of each key point through the maximum response point regression, thereby realizing the accurate recognition of human posture.

2.2. Design of Sports Basketball Action Recognition Model Based on Inv-HRNet-BiLSTM

After completing the design of a human pose estimation network based on Inv-HRNet, a BAR model with time series information is further constructed to improve the accuracy of action recognition and time series understanding ability. The BAR model based on Inv-HRNet and BiLSTM (Inv-HRNet-BiLSTM) is shown in Figure 5.

From Figure 5, the sports BAR model based on Inv-HRNet-BiLSTM includes four parts: data processing module, human action recognition module based on Inv-HRNet, time series modeling module, and action classification module. First, the data processing module preprocesses the input basketball video. Then, the human action recognition module adopts the Inv-HRNet structure. The position information of basketball players' joint points is dynamically captured through the spatial adaptive convolution kernel to achieve accurate positioning of key points of the human body. Subsequently, the time series modeling module uses BiLSTM to deeply learn the timing data of key points, capture the time dynamic changes and stage characteristics of actions, and improve the ability to understand the details of continuous actions [20]. Finally, the action classification module inputs the extracted temporal features into the multi-layer perceptron and accurately identifies basketball actions through the full connection layer and classifier, covering shooting, passing, breakthrough, and other common actions. The basic structure of the timing modeling module is shown in Figure 6.

From Figure 6, the module first receives the key point coordinate sequence extracted from consecutive frames. The key point data contain the spatial position information to form a high-dimensional temporal feature input. To strengthen the sense of order in the time dimension, the module introduces the position coding mechanism, so that the model can accurately grasp the time series of key points. At the same time, the key point displacement between adjacent frames is quantified by motion amplitude coding to highlight the intensity and rhythm of motion changes. Then, the sequence data are input into BiLSTM. The network uses a two-way structure to model the forward and backward dependencies of the time series, respectively, to effectively capture the global timing information of the action. To prevent over-fitting, the module introduces the Dropout layer to strengthen the generalization ability [16]. Finally, the time series modeling module outputs an action timing feature vector to learn the structure and rhythm information of key points that change with time. Position encoding and motion amplitude encoding play different and complementary roles in temporal modeling. The two work together to construct a more comprehensive representation of temporal features. The core function of position encoding is to provide absolute time series information. By assigning unique encoding to keyframes at

Figure 5
A BAR model based on Inv-HRNet-BiLSTM.

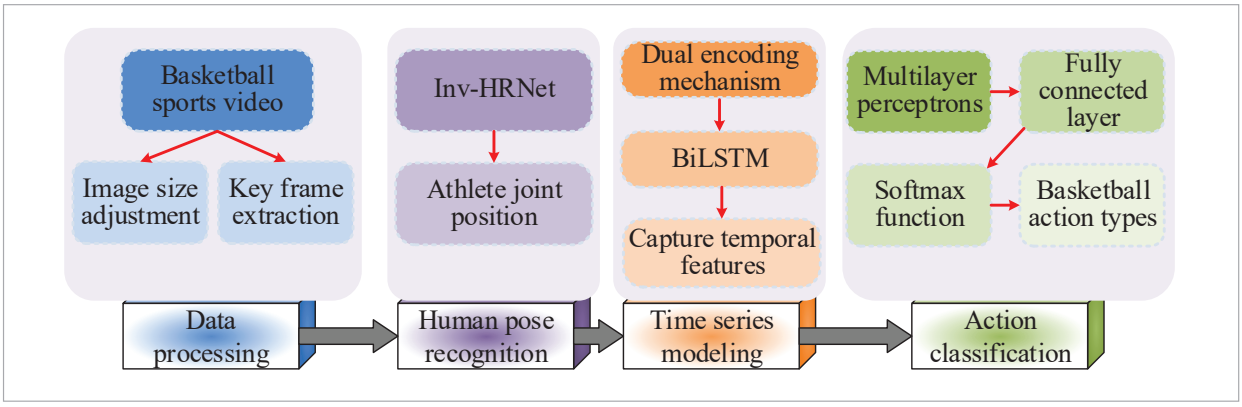
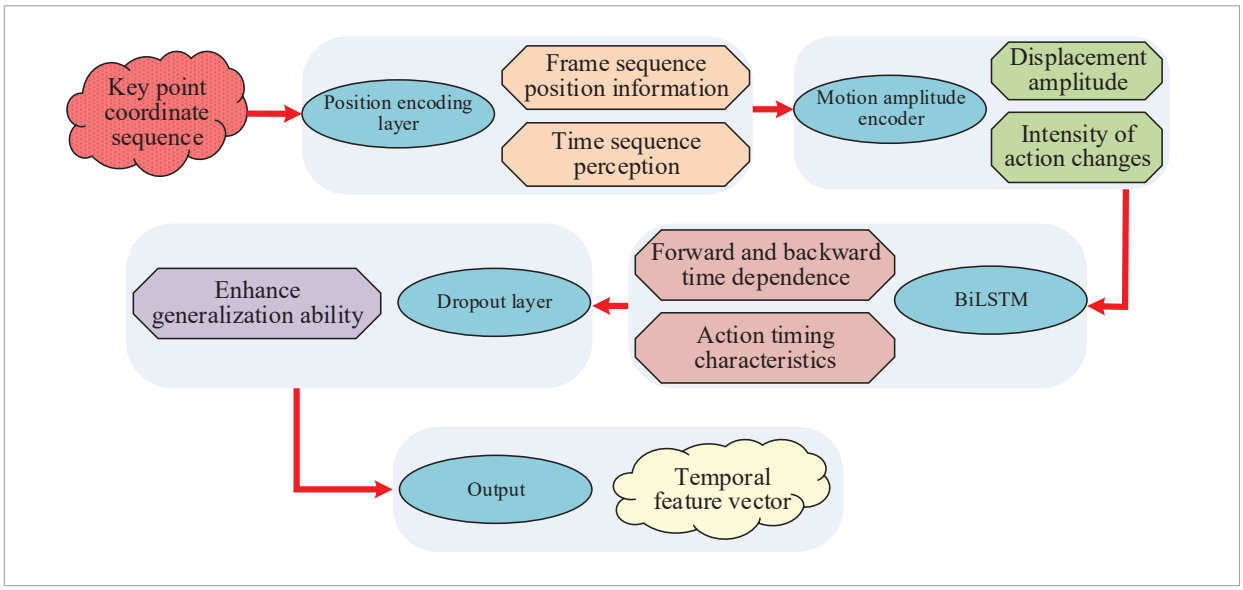


Figure 6
Schematic diagram of timing modeling module.



different time steps, it clarifies the sequential logic of action stages, such as the order relationship of "lift, jump, release" in shooting actions. This is the basis for understanding the integrity of actions and avoiding misjudgment of action categories due to frame sequence confusion. The Encoding of motion amplitude focuses on quantifying the relative displacement intensity of key joints between adjacent frames to accurately capture the dynamic rhythm and force characteristics of motion. For example, the quick swing of the wrist when dribbling and the kick of the leg when jumping. This dynamic information directly reflects the core attributes of actions and is the key to distinguishing similar actions (such as shooting and passing).

The combination of the two achieves a dual characterization of "order" and "potential" in the time dimension: position encoding ensures the accuracy of action timing, and motion amplitude encoding enhances the recognition of action dynamics. Compared with a single encoding method, it can more comprehensively cover the temporal characteristics of basketball actions, providing richer input information for BiLSTM to deeply mine the long-term dependencies of actions and significantly improving the accuracy of action recognition in complex scenes. Position encoding and motion amplitude encoding are aligned and fused with BiLSTM through "dimension concatenation + feature normalization". Firstly, the output features of both encodings are mapped to 256 dimensions (consistent with the input dimension of BiLSTM), then concatenated into 512-dimensional joint features according to the channel dimension. They are normalized through the LayerNorm layer to eliminate differences in feature distribution. The fused features contain both absolute temporal information and dynamic motion features, which are directly input into the hidden layer of BiLSTM. This enables bidirectional temporal modeling to simultaneously utilize two types of key information, ensuring efficient adaptation between encoding and network structure. The expression of the position coding is shown in Equation (6) [19].

$$\begin{cases} PE_{(pos,2i)} = \sin\left(\dfrac{pos}{10000^{2i/d}}\right) \\ PE_{(pos,2i+1)} = \cos\left(\dfrac{pos}{10000^{2i/d}}\right) \end{cases} \tag{6}$$

In Equation (6), i represents the coding dimension index. $PE_{(pos,2i)}$ and $PE_{(pos,2i+1)}$ represent the values of location code in location pos and dimensions $2i$ and $2i+1$. d indicates the total dimension of the code. The motion amplitude code is shown in Equation (7) [30].

$$M_t = \| P_t - P_{t-1} \|_2 \tag{7}$$

In Equation (7), M_t represents the movement amplitude of key points in time step t. P_t and P_{t-1} represent the key point coordinate vectors of time steps t and t–1. The expression of Dropout operation is shown in Equation (8).

$$\tilde{\mathbf{h}}_t = \mathbf{h}_t \odot \mathbf{m}, \quad \mathbf{m} \sim \text{Bernoulli}(p) \tag{8}$$

In Equation (8), $\tilde{\mathbf{h}}$ represents the eigenvector after dropout. $\mathbf{h}$ represents the original eigenvector. $\mathbf{m}$ represents the binary mask matrix with the same dimension as $\mathbf{h}$. $\odot$ means multiply by element. $\mathbf{m}$~Bernoulli(p) means that each element independently obeys Bernoulli distribution. Finally, the high-dimensional action features output by the time series modeling module are mapped to specific action categories. The basic structure of the action classification module is shown in Figure 7.

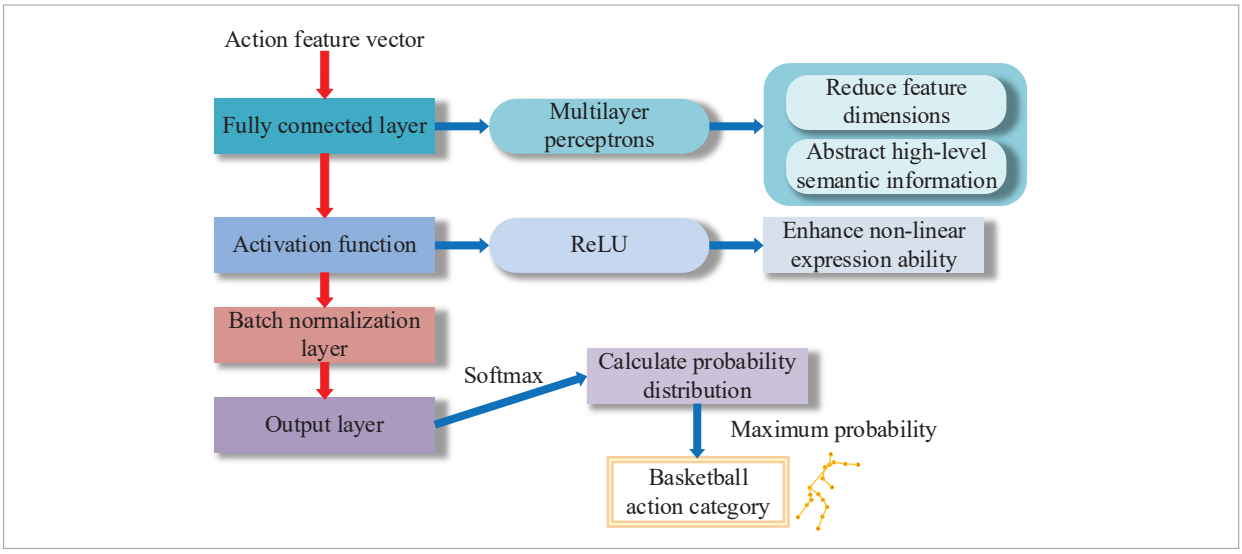
From Figure 7, the action classification module is the key part of the whole BAR system. The module first receives the action feature vector from the time series modeling module, and the input feature is nonlinearly mapped and abstracted through the multi-layer full connection layer. The expression ability and nonlinear fitting ability are improved through the ReLU activation function between layers [18]. To speed up the model training and stabilize the learning process, the module introduces a batch normalization layer to reduce the internal covariate offset. The output layer uses the Softmax function to convert the network output into the probability distribution of each action category. Common basketball action categories include shooting, passing, layup, breakthrough, and defensive sliding [6]. The softmax function expression is shown in Equation (9).

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{j=1}^{Q} \exp(z_j)} \tag{9}$$

In Equation (9), $\hat{y}_i$ is the probability predicted as category i. z_i and z_j respectively represent the unnormalized scores of the classifier for category i and category j. Q means the total amount of action categories. Finally, the model completes the action recognition according to the category label with the highest probability, and realizes the effective classification of various actions in basketball. This module is reasonably designed and has clear layers. It can make full use of time series feature information to ensure the accuracy and real-time performance of identification, and is suitable for actual sports intelligent analysis applications.

Figure 7

Schematic diagram of action classification module structure.



3. Results and Analyses

3.1. Verification of Human Pose Estimation Network Based on Inv-HRNet

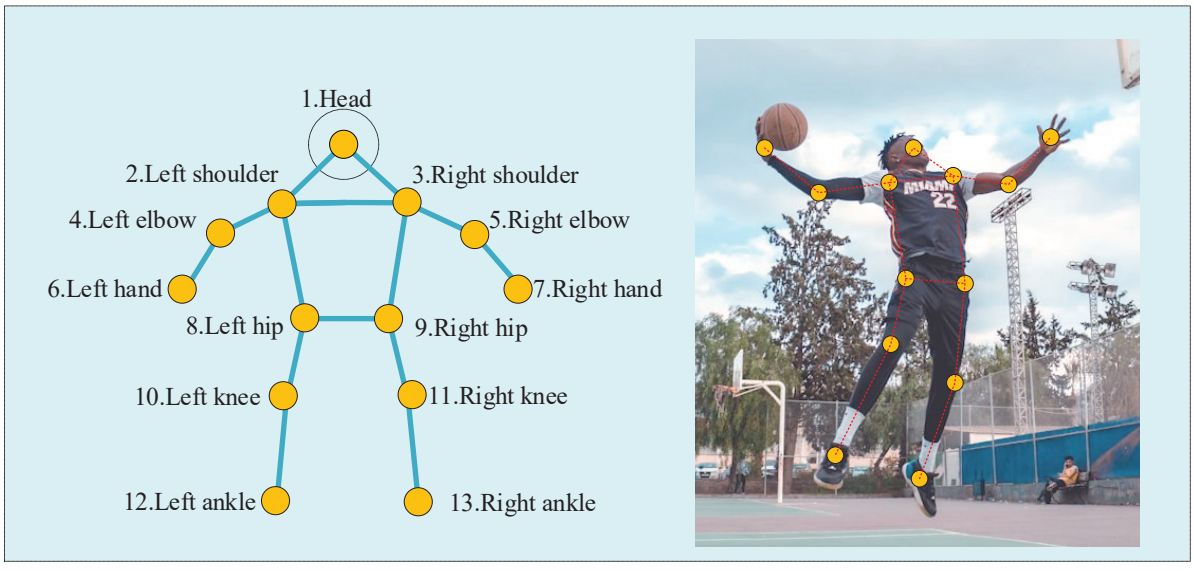
To ensure the reliability and reproducibility of the experimental results of the proposed Inv-HRNet model, the system experiments were carried out in a unified software and hardware environment. The specific experimental configuration and model training parameters are shown in Table 1. The experiment combined two open-source datasets, the SpaceCam Basketball Action dataset and the Kinetics Skeleton dataset, to evaluate the performance in BAR and daily sports scenarios. The SpaceCam dataset provides basketball game videos and action annotations, while the Kinetics Skeleton dataset contains 2D human keypoint data. The study used a mixed training set and independent test set method, with a total of 8,972 samples. The SpaceCam test set had 1,346 samples, and the Kinetics Skeleton test set had 1,568 samples.

According to the experimental environment in Table 1, to intuitively show the capture effect of the proposed Inv-HRNet pose recognition network on the key points of basketball players' bones, the 13 point key point labeling scheme was used to replace

Table 1
List of experimental equipment configuration.

Category	Name	Configuration
Hardware	GPU	NVIDIA GeForce RTX 3090 × 2
	CPU	Intel Core i9-13900K
	Memory	128 GB DDR5
	Storage	1TB NVMe SSD, read/write speed 3500 MB/s
Software	Operating System	Ubuntu 20.04
	Programming Language	Python 3.9
	Deep Learning Framework	PyTorch 2.0
	Image Processing Library	OpenCV 4.7.0
Parameters	Initial Learning Rate	0.001 (with StepLR learning rate scheduler)
	Batch Size	32
	Number of Epochs	100

Figure 8
Key point recognition effect of Inv-HRNet network skeleton.



(Image source: https://unsplash.com/photos/a-man-jumping-up-into-the-air-to-dunk-a-basketball-NvJOYJHseBY)

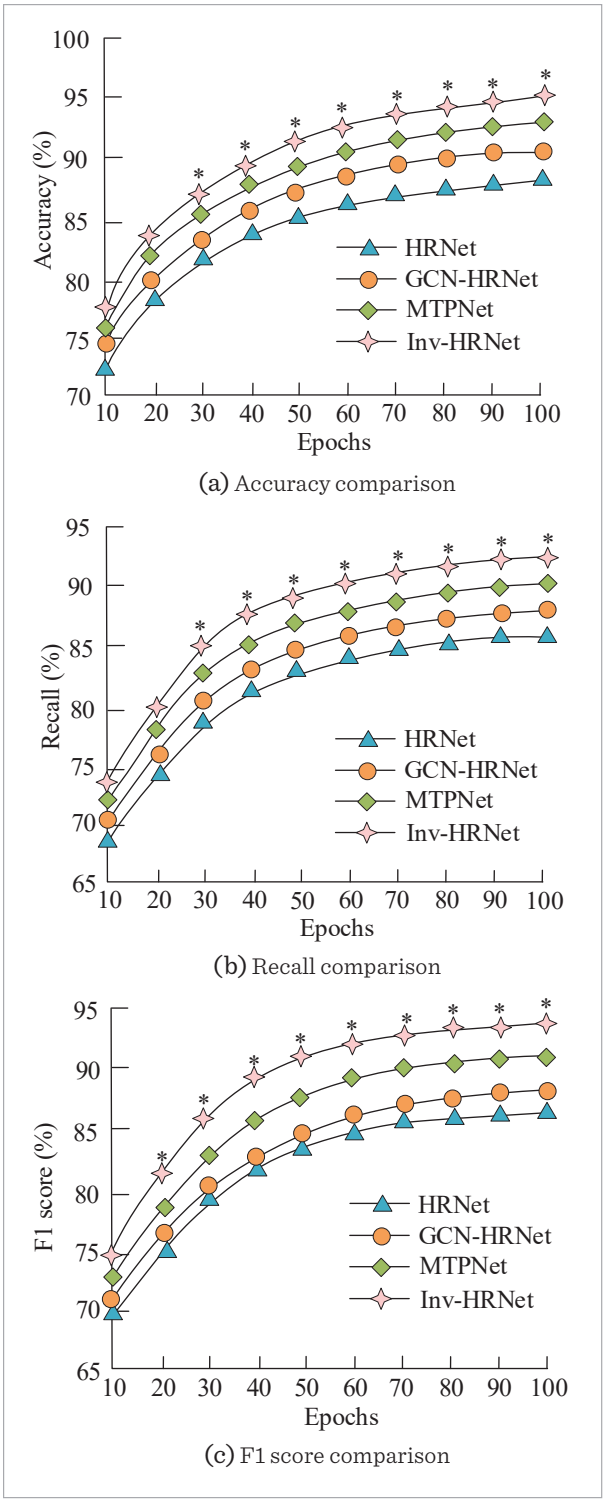
the commonly used 17 point labeling. The aim was to highlight the most representative joint parts in basketball movement. By decoding and visualizing the key points output from the network, the skeleton connection graph was generated. This directly reflected the positioning accuracy and spatial adaptability of the key points of the model in the complex dynamic environment. Figure 8 shows the visualization analysis findings.

From Figure 8, the Inv-HRNet showed excellent key point recognition ability in basketball scenes. The model accurately captured the position of the main joint points of the human body during the shooting process, including the head, shoulders, elbows, wrists, knees, ankles, and other key parts. The positioning of key points closely fitted the contour of the human body, and the skeletal structure formed by the connection between the joint points was clear, which fully reflected the dynamic posture characteristics of shooting. In addition, although the 13-point labeling simplified the bone structure, it could still cover the core posture characteristics of the shooting action. Compared with the conventional action point labeling method, it could effectively reduce the calculation of the model and improve the recognition efficiency of the model. To further verify the recognition effect of Inv-HRNet, the unimproved HRNet, GCN-based HRNet (GCN-HRNet), and Multi-modal Transformer-based Pose Network (MTPNet) were introduced and compared. The results are shown in Figure 9.

Figure 9(a) shows the accuracy curves of the four methods. The accuracy of Inv-HRNet was always ahead of other models, reaching 77.42% in the 10th round, and 94.94% in the 100th round, which was 6.79% higher than that of HRNet, indicating its stronger feature extraction ability and faster convergence speed. Combined with Figure 9(b), Inv-HRNet also showed obvious advantages. In the early stage of training, the recall rate was 73.91%, which was significantly better than 68.92% of HRNet, and it always kept the lead in the whole iteration process. At the 100th iteration, the recall rate of Inv-HRNet reached 92.23%, increased by 6.62% compared with HRNet, and increased by 5.19% and 2.43% compared with GCN-HRNet and MTPNet, respectively. It showed that the network not only had excellent feature recognition ability but also realized more complete key point capture and recognition in multi-pose complex

Figure 9

Comparison of recognition accuracy of different methods in different iterations.



Note: * indicates $p<0.05$.

scenes. From the F1 scores of the four methods in Figure 9(c), with the increase of training rounds, the F1 score rose rapidly, and finally converged to 94.12%, ahead of 91.24% of MTPNet and 87.65% of GCN-HRNet. It showed that Inv-HRNet not only maintained high accuracy but also ensured the recall of key actions, which reflects its advantages in action recognition accuracy and stability. At the same time, the reasoning speed and memory usage of the four methods were compared. Figure 10 shows the results.

From Figure 10(a), Inv-HRNet always maintained the fastest reasoning speed under different sample sizes, and the average reasoning time was about 11.93 ms/frame, which was significantly better than other models. When the sample size increased from 100 to 1,000, the speed remained highly stable, and the fluctuation range was controlled within 0.07 ms. It showed that Inv-HRNet not only had excellent real-time processing ability but also could maintain computational efficiency in the case of large-scale input, which is suitable for the actual basketball scene recognition task with strict delay requirements. Combined with Figure 10(b), Inv-HRNet also maintained the lowest memory consumption in the sample range of 100 to 1,000, and its usage increased from 392.74 MB to 433.79 MB, which was always lower than HRNet, GCN-HRNet, and MTPNet, indicating excellent resource control ability. Inv-HRNet's memory advantage benefited from its dynamic sparse computing strategy. Through the adaptive modeling of spatial structure, it effectively compressed redundant feature channels and improved the memory efficiency.

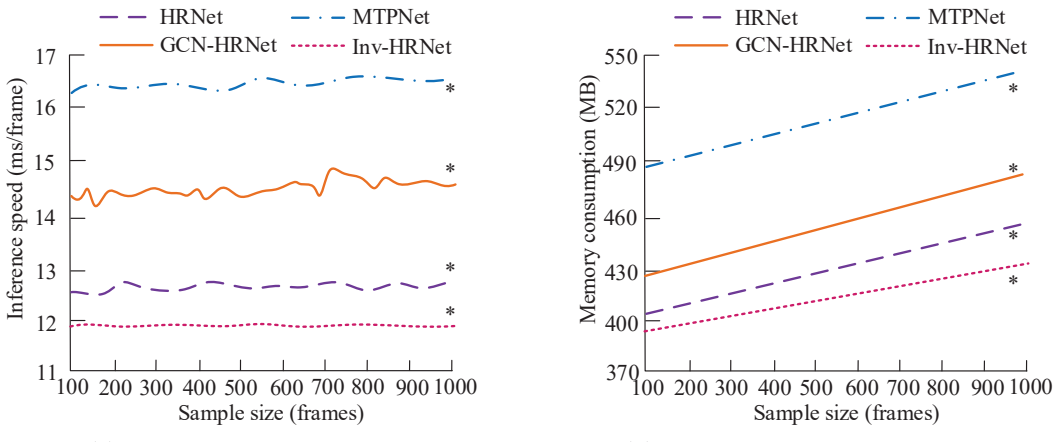
3.2. Verification of Basketball Action Recognition Effect Based on Inv-HRNet-BiLSTM

To verify the role of the key modules in the Inv-HRNet-BiLSTM model in BAR, ablation experiments were designed. Variants of ablation experiments include: basic model (HRNet-LSTM); Involution-enhanced spatial deformation modeling (Inv-HRNet-BiLSTM); BiLSTM integrating the action time series characteristics (Inv-HRNet-BiLSTM); Position coding adding to strengthen time series positioning capability (Inv-HRNet-BiLSTM-PE); Motion amplitude coding adding to quantify joint displacement intensity (Inv-HRNet-BiLSTM-ME); The complete model integrating position and motion double coding (Inv-HRNet-BiLSTM-Full). The specific detail is listed in Table 2.

From Table 2, the Inv-HRNet-BiLSTM model performed best under the complete structure with all key modules, with an accuracy rate of 85.92% and an F1 score of 84.50%, indicating that there was good synergy among the submodules. Among them, when the Involution operator was replaced by the ordinary convolution (w/O Involution), the accuracy rate decreased to 81.73%, which was the lowest among all

Figure 10

Comparison of reasoning speed and memory usage of each method under different sample sizes.



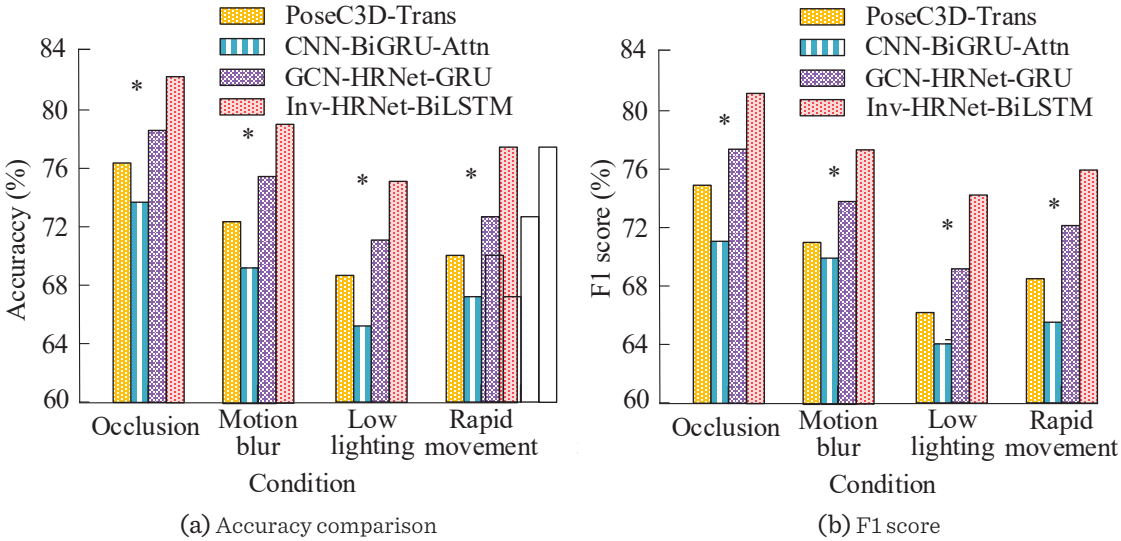
(a) Comparison of reasoning speed

(b) Comparison of memory consumption

Note: * indicates $p<0.05$.

Figure 11

Robustness comparison of four models under different interference scenarios.



Note: * indicates $p<0.05$.

Table 2

Results of key module ablation experiment on Inv-HRNet-BiLSTM model.

Model variant	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Parameters (M)	Inference delay (ms)
w/o Involution	81.73	80.24	79.85	80.04	65.21	28.36
w/o BiLSTM	83.12	81.67	81.09	81.38	66.44	27.91
w/o PosEnc	84.56	83.41	82.95	83.18	67.08	28.05
w/o MagEnc	84.21	83.06	82.57	82.81	67.08	28.02
w/o PosEnc & MagEnc	82.87	81.22	80.34	80.78	67.08	28.08
Inv-HRNet-BiLSTM (full)	85.92	84.74	84.26	84.5	67.89	28.12

Note: w/o represents "without", where the specific configuration of the "w/o BiLSTM" variant is to replace the BiLSTM with a unidirectional LSTM.

variants. This indicated that Involution had significant advantages in modeling spatial structure changes, especially in enhancing the adaptability to the changes of key points in complex basketball movements. After changing from BiLSTM to unidirectional LSTM (w/O BiLSTM), the F1 score decreased to 81.38%, indicating that BiLSTM's bidirectional modeling of the time dimension is helpful to capture the complete characteristics of action stages. When w/o PosEnc and w/o MagEnc were removed, the accuracy decreased to 84.56% and 84.21%, respectively, and further decreased to 82.87% when both were removed at the same time. This result verified their complementarity in improving the expression of temporal modeling.

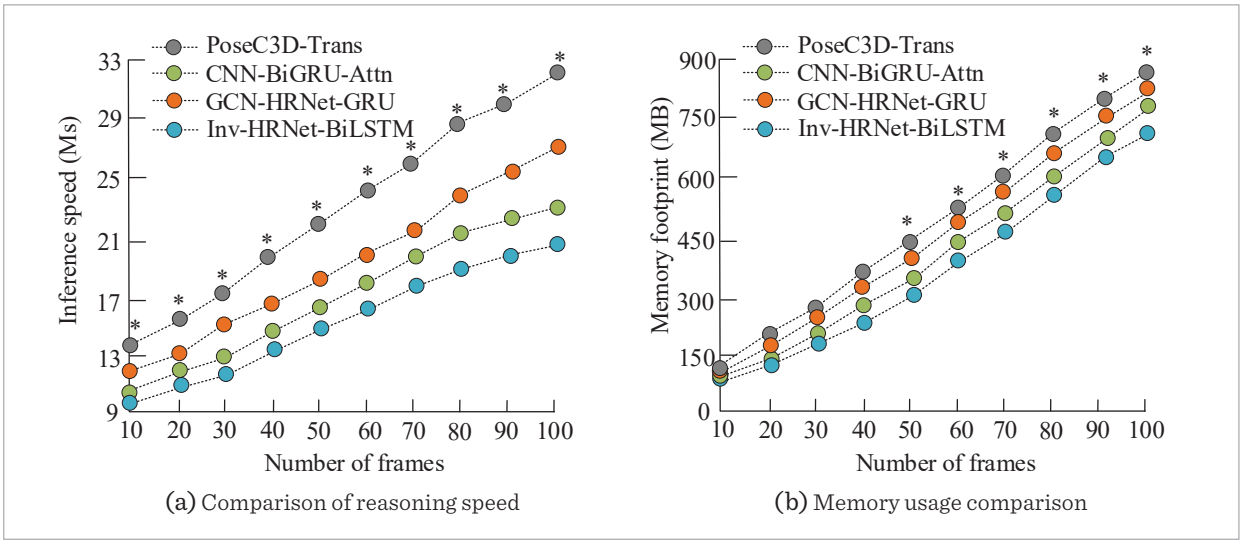
At the same time, the study introduced three mainstream BAR models for comparison: Pose-based Channel-wise 3D Convolutional Network with Temporal Transformer (PoseC3D-Trans), CNN-BI-GRU-Attn, and GCN-HRNet-GRU. Firstly, the robustness of each model was tested, and four typical scenes, including occlusion, motion blur, low illumination, and fast moving, were selected for testing. The results are shown in Figure 11.

Figure 11(a) shows the comparison of the recognition accuracy of the four models under different interference scenes. Inv-HRNet-BiLSTM showed excellent action recognition accuracy in four typical complex scenes, with an average of 78.58%, which was significantly higher than PoseC3D-Trans, CNN-BiGRU-Attn, and GCN-HRNet-GRU. In occluded scenes, the accuracy rate reached 82.61%, indicating that its Involution spatial adaptive mechanism could effectively model the spatial features of locally missing regions. In motion blur and fast motion scenes, the accuracy was 78.95% and 77.36%. This was attributed to BiLSTM's sensitive modeling ability of temporal dynamics, which allowed the model to maintain stable recognition even when frames are blurred or actions undergo sudden changes. Under low-light conditions, the accuracy rate was 75.41%, also leading other models, indicating that its keypoint-driven approach does not rely on overall image brightness and possesses excellent visual interference resistance. From Figure 11(b), Inv-HRNet-BiLSTM also showed the highest comprehensive robustness. Under occlusion and motion blur, the F1 score was 81.38% and 77.22%, respectively, showing its ability to maintain the integrity of the action phase under incomplete image information. In low-light and fast-moving environments, the F1 score was 74.63% and 76.05%, respectively, indicating that its temporal modeling mechanism and dual coding enhancement strategy can accurately extract spatio-temporal motion features in complex scenes. On the whole, Inv-HRNet-BiLSTM had strong generalization ability and adaptability in the robustness dimension. On this basis, the reasoning speed and memory usage of the four models were analyzed, as shown in Figure 12.

Figure 12(a) shows the comparison of the action recognition time of the four models. The Inv-HRNet-BiLSTM model always maintained the lowest delay under each frame number, only increasing from 9.76 ms to 20.56 ms, showing excellent real-time performance, especially suitable for the demand for rapid response in actual basketball games. The superior inference speed of Inv-HRNet was attributed to the efficient parameter design of the Involution operator. Unlike standard convolutions that obtain large receptive fields by increasing the number of channels or enlarging the kernel size, Involution focused on cross-channel spatially specific processing without the need for redundant shared weights. Its dynamic kernel generation mechanism used channel grouping and local context encoding to ensure receptive field coverage while significantly reducing the total number of parameters and computational complexity. This design avoided the dependence of standard convolution on large channels and kernels, improved spatial adaptability, and reduced computational overhead. This enabled the model to

Figure 12

Comparison of reasoning time and memory consumption of different model action recognition.



(a) Comparison of reasoning speed

(b) Memory usage comparison

Note: * indicates $p<0.05$.

Table 3
Comparison of accuracy of each model in typical basketball action recognition (%).

Action Class	PoseC3D-Trans	CNN-BiGRU-Attn	GCN-HRNet-GRU	Inv-HRNet-BiLSTM
Dribble	93.15	91.47	92.34	95.28
Pass	89.6	87.33	88.91	91.76
Shoot	92.82	90.45	91.78	94.12
Rebound	86.41	84.76	85.35	88.9
Block	88.67	86.02	87.15	90.08
Tackle	85.33	82.57	83.69	87.4
Layup	91.08	89.73	90.65	93.57
Dunk	90.95	89.1	89.88	93.12

maintain fast inference in complex pose modeling scenarios, demonstrating computational advantages that balance accuracy and efficiency. Combining with Figure 12(b), PoseC3D-Trans reached 875.50 mb at 100 frames, which was much higher than other models, indicating that it has high requirements for hardware resources. Although the resource overhead of CNN-BiGRU-Attn and GCN-HRNet-GRU was smaller than that of PoseC3D-Trans, they still did not have good scalability. In contrast, Inv-HRNet-BiLSTM showed the lowest memory consumption in all frames, and its memory consumption was only 723.90 MB at 100 frames, reflecting that it is more suitable for BAR applications in edge devices or resource-constrained scenes. Finally, the study tested the recognition accuracy of the four models for different basketball actions and analyzed eight typical actions, including dribbling, passing, shooting, rebounding, blocking, tackling, layups, and dunks. The findings are exhibited in Table 3.

From Table 3, Inv-HRNet-BiLSTM achieved the highest recognition accuracy in all action categories. Specifically, the recognition accuracy of the Inv-HRNet-BiLSTM for dribble movement reached 95.28%, which was significantly higher than 93.15% of PoseC3D-Trans, 92.34% of GCN-HRNet-GRU, and 91.47% of CNN-BiGRU-Attn. This advantage was mainly due to the efficient modeling ability of the Involution module for local spatial features, as well as the bidirectional perception characteristics of BiLSTM for action timing features. In the layup and shooting with strong movement continuity, the accuracy of Inv-HRNet-BiLSTM model was 93.57% and 94.12%, respectively, which reflected its accuracy in capturing the details of basketball technical movements. However, the recognition performance of 88.90% and 87.40% was still high in rebounds and tackles with more occlusion and interference, indicating that the model has strong robustness. In contrast, the modeling ability of the other three models in dynamic scenes was relatively limited, especially in the action categories with information masking or strong temporal dependence. In general, Inv-HRNet-BiLSTM provided higher accuracy and practical support for complex BAR with its advantages of multi-dimensional feature fusion and time series modeling. To verify the rationality of the 13-point keypoint configuration, an ablation experiment was added to compare the performance differences between the 13-point and 17-point annotation schemes. The experiment was conducted on the same dataset (SpaceCam+Kinetics Skeleton) and training parameters, and the results are shown in Table 4.

According to Table 4, compared to the 17-point standard configuration, the accuracy of the 13-point configuration only decreased by 0.18%, the recall rate decreased by 0.08%, and the F1 score remained basically the same (with a difference of only 0.08%), without any substantial negative impact on recognition performance. At the same time, the 13-point configuration simplifies non-core joint points (such as ear and hip secondary nodes), increasing inference speed by 12.1% and reducing memory usage by

Table 4
Performance comparison of different keypoint annotation schemes.

Index	17 points	13 points
Accuracy (%)	95.12	94.94
F1 score (%)	94.20	94.12
Inference speed (ms/frame)	13.58	11.93
Memory usage (MB)	476.32	433.79
Recall rate (%)	92.31	92.23

8.9%, significantly optimizing the real-time performance and resource efficiency. This result proved that the 13-point annotation scheme achieved a balance between performance and efficiency while retaining the core posture features of basketball movements, and was more suitable for deployment requirements in practical application scenarios. In addition, to further verify the lightweight advantage of Inv-HRNet, a comparison of its parameter counts and GFLOPs with the original HRNet was made: the original HRNet had a total parameter count of 24.3M and GFLOPs of 36.8G; Inv-HRNet compressed redundant parameters through a dynamic kernel generation mechanism, reducing the total parameter count to 18.7M (23.0% reduction) and GFLOPs to 29.2G (20.7% reduction). This comparison indicated that the introduction of the Involution operator enhanced spatial modeling capabilities without increasing computational burden, but instead reduced model complexity by optimizing parameter efficiency. This result provided more favorable hardware adaptation conditions for subsequent real-time deployment.

4. Conclusion

Aiming at the problems of insufficient spatial structure modeling and incomplete time series dynamic capture in the process of BAR, the study aimed to build a recognition framework with high accuracy and efficiency. The aim was to enhance the recognition ability and real-time performance in complex scenes. To achieve this goal, the pose recognition network based on Inv-HRNet was designed, and the dual-encoding mechanism of BiLSTM, position, and motion was further introduced to build the timing modeling module, forming a complete Inv-HRNet-BiLSTM BAR model. Experiments showed that the accuracy of Inv-HRNet in the skeleton key point recognition task was improved by 6.79%, and the recall rate was improved by 6.62% compared with the original HRNet. The accuracy of the Inv-HRNet-BiLSTM action recognition model in occluded scenes, motion blur, fast-moving, and low illumination was 82.61%, 78.95%, 77.36%, and 75.41%, respectively. At the same time, the model still had a high recognition performance of 88.90% and 87.40% in rebounds and tackles with more occlusion and interference. In addition, the average reasoning delay of the model was controlled within 20.56 ms, and the memory consumption was no more than 724 MB, which had good deployment adaptability and engineering practical value. The results showed that the joint optimization of dynamic space modeling and temporal features could effectively solve the modeling problems of deformation sensitivity and continuity in BAR. Although the model performs well in the experimental environment, it still needs to further optimize the lightweight structure and improve the adaptability to extreme light and abnormal posture in the actual deployment. In the future, multimodal information such as audio, acceleration data, and self-monitoring pre-training strategies can be combined to enhance the model's ability to understand and analyze complex behaviors in real game videos. Moreover, it can expand its practical value in sports teaching, referee assistance, and sports rehabilitation.

Conflicts of Interest

The author declares that there is no conflict of interest.

References

1. Cai, Y., Liu, Z., Meng, X., Wang, S., Gao, H. Rice Extraction from Multi-Source Remote Sensing Images Based on HRNet and Self-Attention Mechanism. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(4), 186-193. https://doi.org/10.11975/j.issn.1002-6819.202311117
2. Chen, D., Ni, Z. Action Recognition Method of Basketball Training Based on Big Data Technology. International Journal of Advanced Computer Science and Applications, 2024, 15(2), 340-350. https://doi.org/10.14569/IJACSA.2024.0150236
3. Chen, Z., Li, Z., Lv, C., Yue, H., Peng, Y. A Feature Fusion Object Detection Algorithm Based on HRNet and ASFF. Control and Decision, 2024, 39(10), 3207-3215. https://doi.org/10.13195/j.kzyjc.2023.0852
4. Cui, Z. 3D-CNN-Based Action Recognition Algorithm for Basketball Players. Informatica, 2024, 48(13), 97-110. https://doi.org/10.31449/inf.v48i13.6100
5. Guo, J., Dong, Y., Liu, X., Lu, S. Facial Expression Recognition Based on Attention Mechanism and Involution. Computer Engineering and Applications, 2023, 59(23), 95-103. https://doi.org/10.3778/j.issn.1002-8331.2207-0412
6. Hasanvand, M., Nooshyar, M., Moharamkhani, E., Selyari, A. Machine Learning Methodology for Identifying Vehicles Using Image Processing. Advances in Intelligent Applications, 2023, 1(3), 170-178. https://doi.org/10.47852/bonviewAIA3202833
7. Jayaweera, S. S., Regani, S. D., Hu, Y., Wang, B., Liu, K. J. R. HRNet: High-Resolution Neural Network for Human Imaging Using mmWave Radar. IEEE Internet of Things Journal, 2025, 12(1), 881-893. https://doi.org/10.1109/JIOT.2024.3470813
8. Khobdeh, S. B., Yamaghani, M. R., Sareshkeh, S. K. Basketball Action Recognition Based on the Combination of YOLO and a Deep Fuzzy LSTM Network. Journal of Supercomputing, 2024, 80(3), 3528-3553. https://doi.org/10.1007/s11227-023-05611-7
9. Kim, W., Cho, S., Jeong, J., Kim, Y., Kim, H. O., Choi, M. Evaluation of Deep Learning-Based Water Bodies and Flooded Area Detection with Nanosatellites: The PlanetScope Satellite Imageries and HRNet Model. Korean Journal of Remote Sensing, 2024, 40(5), 617-627. https://doi.org/10.7780/kjrs.2024.40.5.1.16
10. Li, X., Wu, D., Wang, Y., Chen, W., Gu, H. Improved HRNet Semantic Segmentation Model for Remote Sensing Images of Winter Wheat. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(3), 193-200. https://doi.org/10.11975/j.issn.1002-6819.202308134
11. Qi, L., Xu, R., Zhao, S., Zheng, M., Yu, W. TMNIO: Triplet Merged Network with Involution Operators for Improved Few-Shot Image Classification. IET Image Processing, 2024, 18(6), 1629-1641. https://doi.org/10.1049/ipr2.13055
12. Sadeghi, N., Mahdavi-Nasab, H., Zeinali, M., Pourghassem, H. Comparing the Semantic Segmentation of High-Resolution Images Using Deep Convolutional Networks: SegNet, HRNet, CSE-HRNet and RCA-FCN. Journal of Information Systems and Telecommunication, 2023, 11(4), 359-367. https://doi.org/10.61186/jist.39680.11.44.359
13. Saleem, A. L., Mamidisetti, G. Semantic Object Segmentation Using Modified HRNet Deep Learning Model. Journal of Information Systems Engineering and Management, 2025, 10(4), 375-390. https://doi.org/10.52783/jisem.v10i4s.530
14. Takagi, S., Xiong, L., Kato, F., Cao, Y., Yoshikawa, M. HRNet: Differentially Private Hierarchical and Multi-Resolution Network for Human Mobility Data Synthesization. Proceedings of the VLDB Endowment, 2024, 17(11), 3058-3071. https://doi.org/10.14778/3681954.3681983
15. Tang, Q., Jiang, Z., Pan, B., Guo, J., Jiang, W. Scene Text Detection Using HRNet and Spatial Attention Mechanism. Programming and Computer Software, 2023, 49(8), 954-965. https://doi.org/10.1134/S0361768823080212
16. Tong, J., Wang, F. Basketball Sports Posture Recognition Technology Based on Improved Graph Convolutional Neural Network. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2024, 28(3), 552-561. https://doi.org/10.20965/jaciii.2024.p0552
17. Wang, H., Jiang, Y., Li, X., Ma, Y., Liu, X. Pig Image Instance Segmentation Based on Weakly Supervised Dataset. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(10), 255-265. https://doi.org/10.6041/j.issn.1000-1298.2023.10.025

18. Wang, M., Zhou, L., Zhang, C. Upper Limb Movement Trajectory Recognition of Basketball Players Based on Machine Learning. International Journal of Information and Communication Technology, 2023, 23(1), 15-28. https://doi.org/10.1504/IJICT.2023.132170

19. Wang, X., Shen, Y., Xing, Q. Two-Branch Video Action Recognition Method Based on High-Efficiency Parameter Fine-Tuning. Journal of Henan Polytechnic University (Natural Science), 2025, 44(4), 21-28. https://doi.org/10.16186/j.cnki.1673-9787.2025020018

20. Wang, X., Tian, H., Shi, W., Wang, T., Chen, R. Research on the Enhancement Algorithm of Qing Dynasty Official Robes Detection Based on YOLOv5. Journal of Beijing Institute of Fashion Technology (Natural Science Edition), 2023, 43(1), 7-15. https://doi.org/10.16454/j.cnki.issn.1001-0564.2023.01.002

21. Wu, M., Peng, J. Basketball Players' Step Action Recognition Based on Two-Model Convolutional Neural Network. Journal of Computational Methods in Sciences and Engineering, 2025, 25(3), 2756-2769. https://doi.org/10.1177/14727978251321689

22. Xia, Z. A Recognition Method of Basketball Players' Shooting Action Based on Gaussian Mapping. International Journal of Reasoning-Based Intelligent Systems, 2023, 15(2), 105-110. https://doi.org/10.1504/IJRIS.2023.130192

23. Xu, Y., Gong, J., Huang, X., Hu, X., Li, J., Li, Q., Peng, M. Luojia-HSSR: A High Spatial-Spectral Resolution Remote Sensing Dataset for Land-Cover Classification with a New 3D-HRNet. Geo-Spatial Information Science, 2023, 26(3), 289-301. https://doi.org/10.1080/10095020.2022.2070555

24. Xue, B., Zheng, Q., Li, Z., Wang, J., Mu, C., Yang, J., Li, X. Perturbation Defense Ultra High-Speed Weak Target Recognition. Engineering Applications of Artificial Intelligence, 2024, 138, 109420. https://doi.org/10.1016/j.engappai.2024.109420

25. Yan, X. Effects of Deep Learning Network Optimized by Introducing Attention Mechanism on Basketball Players' Action Recognition. Informatica, 2024, 48(19), 179-194. https://doi.org/10.31449/inf.v48i19.6188

26. Zhang, C., Wang, M., Zhou, L. Recognition Method of Basketball Players' Throwing Action Based on Image Segmentation. International Journal of Biometrics, 2023, 15(2), 121-133. https://doi.org/10.1504/IJBM.2023.129216

27. Zhang, D., Tian, Y., Xu, P., Zhong, C. Road Surface Damage Area Identification Method Based on E-HRNet. Journal of Beijing Jiaotong University, 2023, 47(4), 110-119. https://doi.org/10.11860/j.issn.1673-0291.20220142

28. Zhang, H., Zhou, Y., Qiu, Y. Detection Method of Steel Surface Defects with Fusion of HGNetV2 and Attention Mechanism. Journal of Electronic Measurement and Instrumentation, 2025, 39(1), 36-49. https://doi.org/10.13382/j.jemi.B2407618

29. Zheng, Y., Wu, Y., Chen, X., Wang, P., Dong, F., He, L., Su, Q., Cheng, G., Ma, C., Yao, H., Zhou, S. Automatic Measurement of X-Ray Radiographic Parameters Based on Cascaded HRNet Model from the Supraspinatus Outlet Radiographs. Quantitative Imaging in Medicine and Surgery, 2025, 15(2), 1425-1438. https://doi.org/10.21037/qims-24-1373

30. Zhou, Y., Wang, R., Wang, Y., Sun, S., Chen, J., Zhang, X. A Swarm Intelligence Assisted IoT-Based Activity Recognition System for Basketball Rookies. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(1), 82-94. https://doi.org/10.1109/TETCI.2023.3319432