Exploring the Application of Neural Networks in the Learning and Optimization of Sports Skills Training

Dazheng Liu*
Yangzhou Polytechnic Institute, Jiangsu 225127, China

***Abstract*—Sports skills training is a crucial component of sports education, significantly contributing to the development of athletic abilities and overall physical literacy. It is essential to utilized neural networks to optimize traditional training methods that are inefficient and rely on subjective assessments. This paper develops methods for sports action recognition and athlete pose estimation and prediction based on deep neural networks. Given the complexity and rapid changes in sports skills, we propose a multi-task framework-based HICNN-PSTA model for jointly recognizing sports actions and estimating human poses. This method leverages the advantages of Convolution and Involution operators in computing channel and spatial information to extract sports skill features and uses a decoupled multi-head attention mechanism to fully capture spatio-temporal information. Furthermore, to accurately predict human poses to avoid potential sports injuries, this paper introduces an MS-GCN prediction model based on the multi-scale graph. This method utilizes the constraints between human body key points and parts, dividing the 2D human pose into different levels, significantly enhancing the modeling capability of human pose sequences. The proposed algorithms have been thoroughly validated on a basketball skills dataset and compared with various advanced algorithms. Experimental results sufficiently demonstrate the effectiveness of the proposed methods in sports action recognition and human pose estimation and prediction. This research advances the application of deep neural networks in the field of sports training, providing significant reference value for related studies.**

***Keywords*—Deep neural network; action recognition; 2D pose prediction; pose estimation; sports skill training; attention mechanism**

I. Introduction

Sports play an indispensable role in the cultural development of nations, serving not only as a key factor of citizen welfare but also as an important vessel for cultural identity. Recently, numerous policies have been published to encourage public participation in sports activities, with an increasing number of individuals seeking to alleviate stress and release emotions through sports [1]. As societal enthusiasm for sports activities grows, the learning and optimization methods of sports skills have gained more attention. Traditionally, this process has been predominantly governed by the professional capabilities and personal experiences of coaches, considering as an individual-dependent method that lacks objectivity and is resource-intensive. Therefore, it is essential to explore how advanced artificial intelligence algorithms can be utilized to enhance the efficiency of the sports skill learning and optimization process. In recent years, deep neural networks have been widely applied in various fields, such as speech recognition, fault monitoring, and text analysis. Notably, in the field of image recognition, deep convolutional neural networks (DCNNs) [2] have demonstrated the ability to effectively process unstructured image inputs and uncover latent features within massive datasets, providing a novel approach for sports training.

Employing neural network algorithms to identify sports skills presents an intriguing research problem. Such methods leverage the powerful image recognition capabilities inherent in deep learning algorithms to analyze the types of movements performed by athletes, detect key points in the human body and postural information, and thereby aid athletes in enhancing their motor skills and improving the quality of their movements. Additionally, by extracting temporal information from continuous inputs, deep neural networks can effectively predict future movements, thereby preventing potential risks and avoiding injuries resulting from improper actions. Therefore, the accurate recognition of sports actions and estimation of human poses can not only enhance the efficiency and quality of motor skill learning but also provide sports enthusiasts with more effective training methods.

In sports training, accurately identifying and predicting sequences of athletic movements poses a significant challenge. This challenge arises from the inherent complexity of human posture, the diversity of athletic skills, and the uncertainty in the execution of movements. To address the aforementioned challenges and enhance the feature extraction capability possessed by neural networks, it is imperative to comprehensively capture the temporal-spatial relationships inherent in sports movements. To this end, this paper proposes a novel multitask framework to jointly recognize sports actions and estimate human poses based on the hybrid deep neural network that integrates the Involution operator and Convolution operator. This approach significantly enhances the model's ability to capture spatial information, surpassing the performance of traditional convolutional neural networks. A parallel spatial-temporal attention mechanism is further designed to operate in a decoupled manner and focused separately on temporal and spatial dimensions. It facilitates the neural network's ability to identify crucial movements and detect subtle variations across different frames. Finally, a sports pose prediction method is proposed based on the multiscale graph convolutional networks, thereby optimizing the effectiveness and practicability when applied to sports skill training.

*Corresponding Author

This work is supported by University-level scientific research project (Social science) in 2023: Research on the construction of National fitness public service system embedded in digital technology in Yangzhou (Project Number: 2023xjsk003)

II. Literature Review

This section will present existing works that are the most relevant methods related to our work, including human action recognition, human pose estimation, and human motion prediction.

A. Human Action Recognition Methods

Human action recognition is a vision task centered on humans, aiming to identify the classification results corresponding to the input action sequences. It has extensive applications in fields such as human-computer interaction, sports training, and smart security [3]. Human action recognition has evolved from the use of hand-crafted features to features automatically obtained via deep neural networks. Early research primarily focused on extracting shallow information from input images, such as angles, edges, or contours [4]. To effectively recognize the motion information contained in human actions, optical flow and Histograms of Oriented Gradients (HOG) are often adopted as part of the feature set. For instance, FarajiDavar et al. [5] utilized HOG3D features to describe tennis actions and explored feature re-weighting and feature translation methods based on these features. Calandre et al. [6] employed optical flow information to detect table tennis stroke actions, thereby identifying the most relevant frames in the input videos. However, the use of manual features and machine learning methods suffers from poor generalization, complex feature extraction processes, and reliance on shallow features that inadequately describe the action information reflected in the original inputs, especially in the domain of sports action recognition.

Currently, deep learning techniques, represented by deep neural networks, have become the predominant method for human action recognition. By establishing methods for human action recognition based on deep learning, it is possible to construct more efficient and comprehensive sports training systems, pushing the development of sports skills learning in a more intelligent direction. Simonyan et al. [7] proposed a two-stream convolutional network for action recognition, which utilizes both input RGB images and optical flow images to extract spatial and temporal features respectively, with the recognition results obtained through the fusion at the decision layer. Moreover, utilizing human posture information to enhance action recognition results is also considered an effective means. Nie et al. [8] proposed a hierarchical structure to capture the geometric and appearance variations in posture. Notably, the lateral connections between adjacent frames were considered to describe the action-specific information. Furthermore, Lin et al. [9] developed a Temporal Shift Network (TSN), which can switch feature channels along the temporal dimension to exchange temporal information between adjacent frames. This module can also be embedded as an independent structure into any deep convolutional network model and can significantly improve recognition performance while maintaining lower FLOPs.

On the other hand, as the primary data for action recognition often comprises video data, deep models based on 3D convolution have also received considerable attention. Cao et al. [10] proposed a dual-stream bilinear 3D-CNN model that utilizes selective convolutional layer activations to form discriminative descriptors for videos, ultimately achieving a recognition accuracy of 95.3% on the PENN Dataset. Additionally, Baradel et al. [11] introduced a novel attention mechanism known as Glimpse Clouds, which learns to focus on specific image patches in space and time, aggregating the patterns and softly assigning each feature. Overall, action recognition methods based on deep learning, allowing for fully automated feature extraction and action classification, avoiding the influence of subjective factors, and having high accuracy and generalizability, becoming the mainstream approach in action recognition tasks.

B. Human Pose Estimation Methods

Human pose estimation involves determining the location of body joints and the connections between various body parts in 2D/3D spaces. This field has been actively researched over the past few years, evolving from conceptual frameworks such as Pictorial Structures [12] to recent deep neural network-based approaches. An effective approach considers human pose estimation as a detection task, specifically by obtaining heatmaps at body joints based on detection scores. Newell et al. [13] proposed a novel CNN architecture that employs repeated bottom-up and top-down processing. The proposed Stacked Hourglass networks were evaluated on the FLIC and MPII benchmarks, demonstrating improvements in 2D pose estimation. Pishchulin et al. [14] introduced the DeepCut method, which initially detects regions potentially containing human key joints, followed by creating a connection graph encompassing all regions. However, detection-based methods do not directly provide coordinates of human keypoints and instead infer them indirectly by maximizing the posterior probability.

Regression-based approaches involve projecting input actions onto desired keypoint coordinates through nonlinear functions. Toshev et al. [12] were pioneers in proposing a method for human pose estimation based on DNNs and cascade regression, which avoids the need for explicitly designing feature representations or detectors for body parts. Cheng et al. [15] addressed the issue of scale variation in multi-person pose estimation by proposing a method that utilizes a high-resolution feature pyramid to learn scale-aware representations, thereby achieving more precise keypoint localization in multi-person pose estimation. The proposed method achieved an Average Precision (AP) of 67.6% in the CrowdPose test.

In recent years, a key research focus has been on calculating the spatial information of each keypoint based on their 2D coordinates to obtain the 3D position of human posture. With the release of more high-accuracy 3D data, it has become feasible to train 3D human pose estimation models using deep neural network algorithms. Chen et al. [16] innovatively decomposed the 3D pose estimation problem into a 2D estimation based on camera coordinates and a 2D-to-3D matching using a non-parametric shape model. Pavllo et al. [17] proposed a multi-view fusion 3D pose estimation algorithm based on 2D keypoint trajectories, utilizing 2D keypoints to estimate 3D poses and back-projecting to 2D space to enable semi-supervised training. The proposed method reduced the error by 11% compared to the previous state-of-the-art on the Human3.6M dataset.

C. Research Gaps

Although deep learning has achieved significant success in human-centered fields such as action recognition, its application in sports skill training still presents numerous challenges, particularly when dealing with rapid pose movements and the diversity of actions under the same sports skill. Based on these considerations, this paper aims to address the following key issues:

1) Limitations of traditional CNN models in sports skills training: The utilization of traditional Convolutional Neural Networks (CNNs) in the domain of sports skill training, particularly for sports action recognition and human pose estimation, has exposed specific deficiencies. Standard CNN architectures are characterized by fixed, limited receptive fields and inherent spatial invariance, which compromise their capacity to effectively model the contextual nuances of complex athletic movements and limit their sensitivity to variations in spatial configurations. Furthermore, the generalization of conventional CNNs is predominantly contingent upon the original training dataset, typically necessitating substantial retraining or fine-tuning to adapt these models for diverse sports training applications.

2) Lack of Multitask Framework for Sports Action Recognition and Pose Estimation: Contemporary studies in implementing action recognition and human pose estimation for sports skill training typically utilize independent operational frameworks. While this method permits the tailored algorithmic development specific to each task, it frequently neglects the potential synergistic interactions between these intimately connected tasks. Furthermore, both pose estimation and action recognition generally share analogous feature extraction phases and operating these models independently leads to repetitive processing steps, thereby diminishing computational efficiency and increasing the complexity of real-time applications. Employing a multitask framework to concurrently learn shared features from pose estimation and action recognition can facilitate the acquisition of more robust and extensible features, offering a more holistic comprehension of sports actions and markedly enhancing the support for the learning and optimization of sports skills.

3) Limitations of CNN models in sports pose prediction: Although CNN backbone networks are widely used in action recognition and pose estimation, the inherent non-Euclidean nature of human keypoints makes it challenging for CNN models to achieve satisfactory results in human pose prediction. Particularly when dealing with the relationships or constraints between human keypoints and body parts, models based on CNN backbones struggle to incorporate such information in a priori manner, which is crucial for accurate human pose prediction. Representing the human pose in the form of a graph allows for a more precise reflection of the structural and functional relationships between different body parts. Thus, constructing a backbone model based on GCNs (Graph Convolutional Networks) to model the spatiotemporal relationships of human poses better meets the requirements of motion pose prediction.

In summary, future studies should concentrate on creating innovative models and approaches that tackle the existing challenges in sports skill training, aiming to not only boost the practicability but also enhance the efficacy and accuracy of sports training.

III. Research on Recognition Methods of Sports Action and Human Pose based on Deep Neural Networks and Attention Mechanism

In this section, a novel multitask framework based on deep neural networks and attention mechanisms is proposed for sports action recognition and human pose estimation. Furthermore, a novel multiscale model is proposed for human pose prediction based on the estimated body keypoints. The preliminary knowledge of Hybrid Involution and Convolution Neural Networks (HICNN) is first introduced, followed by the proposed multitask framework and parallel spatial-temporal attention (PSTA). Finally, the multiscale Graph Convolutional Network (MS-GCN) is designed to predict human poses for sports skill training.

A. Hybrid Involution and Convolution Neural Network

As a primary component of deep neural networks, Convolutional Neural Networks (CNNs) utilize the spatial invariance and channel specificity of convolution kernels to enhance computational efficiency and the ability to interpret translation equivalency. However, these characteristics hinder the adaptability of convolution kernels to different spatial positions, and their limited receptive fields pose challenges in modeling long-distance relationships. To address these issues, the involution operator, which possesses symmetrically inverse inherent characteristics, has been proposed. Specifically, the involution operator shares weights across different channels while varying spatially, thereby compensating for the deficiencies of convolution kernels in capturing long-distance relationships [18].

Given the input feature map as $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$, where H,W,C represent its height, width, and channels. By applying multiply-add operations in a sliding-window manner, the output feature map can be expressed as

$$\boldsymbol{Y}_{i,j,k} = \sum_{c=1}^{C} \sum_{(u,v)\in\Delta_K} \Theta_{k,c,u+\lfloor K/2\rfloor,v+\lfloor K/2\rfloor} \boldsymbol{X}_{i+u,j+v,c} \tag{1}$$

where $\Theta \in \mathbb{R}^{C_0\times C_i\times K\times K}$ represents the convolution filters with the fixed kernel size of $K\times K$, and $\Delta_K \in \mathbb{R}^2$ refers to the set of offsets in the neighborhood considering convolution conducted on the center pixel, written as

$$\Delta_K = [-\lfloor K/2\rfloor,\cdots,\lfloor K/2\rfloor]\times[-\lfloor K\times 2\rfloor,\cdots\lfloor K/2\rfloor] \tag{2}$$

Compared to the standard convolution kernel, involution kernels are devised with the inverse characteristics in the spatial and channel dimension, expressed as $\Psi \in \mathbb{R}^{H\times W\times K\times K\times G}$. Specifically, an involution kernel is tailored for the pixel $\boldsymbol{X}_{i,j} \in \mathbb{R}^{C}$ but shared over different channels. The output feature

map of involution can be obtained by applying multiply-add operations with involution kernels, that is

$$\boldsymbol{Y}_{i,j,k} = \sum_{(u,v)\in\Delta_K} \Psi_{i,j,u+\lfloor K/2\rfloor,v+\lfloor K/2\rfloor,\lceil KG/C\rceil} \boldsymbol{X}_{i+u,j+v,k} \tag{3}$$

To fully leverage the capabilities of convolution and involution, this section explores the alternating stacking of these two types of operators, constructing the Hybrid Incolution and Convolution Neural Networks (HICNN). This hybrid architec- - ture serves as the backbone extractor aiming to enhance the model's ability to discern complex spatial relationships while maintaining computational efficiency.

B. HICNN-PSTA for Sports Action Recognition and Pose Estimation

In this section, we introduced the HICNN-PSTA (Hybrid Involution and Convolution Neural Network with Parallel Spatial-Temporal Attention (PSTA), which is specially designed for joint sports action recognition and human pose estimation, as shown in Fig. 1.

Different from the previous work, this section attempts to establish a multitasking framework by predicting human poses and recognizing sports actions in parallel. The input RGB frames are first fed into the HICNN model to extract low-level visual features. Besides, a novel two-pathway attention mechanism, namely PSTA, is proposed to model spatial and temporal information in parallel. The PSTA mechanism significantly enhances the processing capabilities for both single-frame image and image sequences, which is crucial for multitask frameworks. Specifically, spatial attention aids in focusing on critical human-related information within individual frames, thereby increasing the accuracy of human pose estimation. Temporal attention, on the other hand, concentrates on the continuity of actions, which is essential for understanding action sequences and patterns.

The proposed PSTA is illuminated in Fig. 2 that originates from the vanilla multi-head self attention module, defined as

$$MSA(Q,K,V) = \text{softmax}(\frac{Q\cdot K^T}{\sqrt{C}})\cdot V \tag{4}$$

where $Q,K,V \in \mathbb{R}^{N\times C}$ denote the queies, keys and values. In PSTA module, the input embedding $E \in \mathbb{R}^{T\times N\times C}$ are firstly mapped into queries, keys and values with the same dimensions. Then, the mapped tensors are evenly divided into two groups along the channel dimension, results in time group $\{Q_T,K_T,V_T\}$ and space group $\{Q_S,K_S,V_S\}$. To model the spatial-temporal dependencies between joints avoiding the quadratic computation, the temporal and spatial correlations are calculated in two separate self-attention modules, which can be expressed as,

$$\begin{aligned} H_T &= MSA(Q_T,K_T,V_T) \\ H_S &= MSA(Q_S,K_S,V_S) \\ \boldsymbol{H} &= cat(H_T,H_S) \end{aligned} \tag{5}$$

Based on the latent representation, the multi-task prediction block produces single frame features, multi-task features, and image sequence features, which is defined as $\chi_t \in \mathbb{R}^{H_f\times W_f\times N_f}, \delta_t \in \mathbb{R}^{H_f\times W_f\times N_f}, \nu_t \in \mathbb{R}^{T\times N_j\times N_v}$. For pose estimation, prediction blocks take as input the multi-task features to predict body joint probability maps, expressed as

$$h_t = \Phi(W_h * \delta_t), h_t \in \mathbb{R}^{H_f\times W_f\times N_j} \tag{6}$$

The elastic net loss between the predicted human poses and ground-truth values is adopted for model training, which is defined as

$$L_p = \frac{1}{N_j}\sum_{j=1}^{N_j}(\left\|\hat{p}^j - p^j\right\|_1 + \left\|\hat{p}^j - p^j\right\|_2^2) \tag{7}$$

Furthermore, the multi-task features are multiplied by probability maps h_t at channel dimension to obtain the appearance features $V \in \mathbb{R}^{T\times N_j\times N_f}$ that describe the entire image sequences, thus recognizing sports actions by categorical cross-entropy loss on predicted actions.

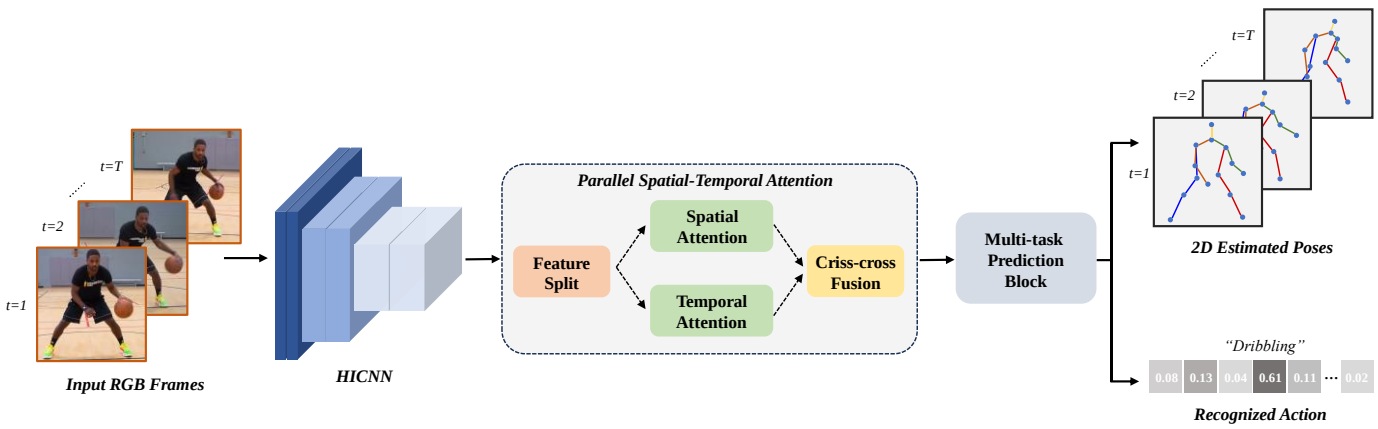

Fig. 1. Network structure diagram of sports action recognition and pose estimation based on HICNN-PSTA.

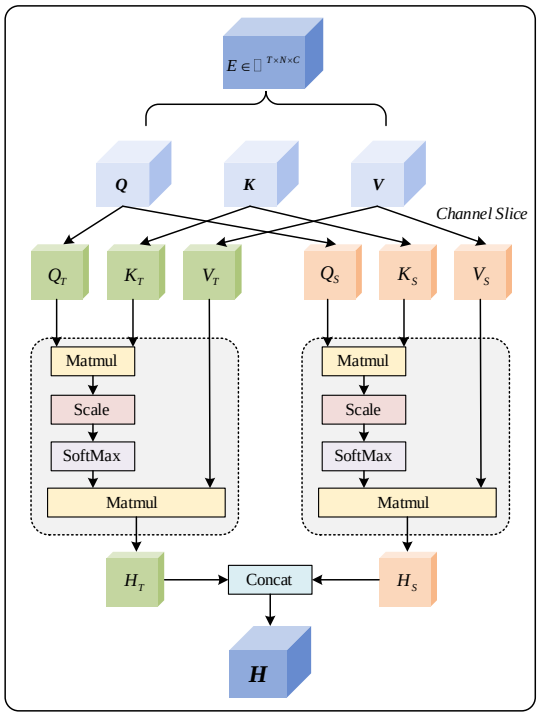

Fig. 2. An overview of the proposed PSTA.

C. MS-GCN for Sports Pose Prediction

Human pose prediction is of paramount importance in the field of sports skills training, as it assists participants in optimizing their skills and avoiding potential hazards. The key to effective prediction of human poses lies in a comprehensive understanding of the intrinsic correlations among sequences of human keypoints. To address these issues, this section introduces a Multi-Scale Graph Convolution Network (MS-GCN) that predicts future human poses based on the estimated sequences of human keypoints. MS-GCN extends conventional keypoints analysis by integrating single-scale graph and multi-scale graphs at various levels to connect body components. The single-scale graph provides a multi-granularity representation of the body skeleton, while the multi-scale graph, initialized by predefined physical connections, reflects the interconnections between different single-scale graphs and adjusts to poses sensitivity during training.

Suppose the estimated 2D skeleton-based poses are $\boldsymbol{P}_{-T_h:0}=[P_{-T_h},\ldots,P_0]\in\mathbb{R}^{M\times(T_h+1)\times 2}$ and the future poses are $\boldsymbol{P}_{1:T_f}=[P_1,\ldots,P_{T_f}]\in\mathbb{R}^{M\times T_f\times 2}$. The goal of pose prediction is to generate future poses by the past observed ones, which can be expressed as $\hat{\boldsymbol{P}}_{1:T_f}=M_{pred}(\boldsymbol{P}_{-T_H:0})$. To construct the MS-GCN, two body scales are first initialized, expressed as a trainable adjacency matris $A_s\in\mathbb{R}^{M_s\times M_s}$ at scale s. Based on the single-scale graph, the GCN block extract spatial features of body components as well as temporal features from poses sequences, defined as,

$$\boldsymbol{P}_{s,sp}=\mathrm{ReLU}(A_s\boldsymbol{P}_s\boldsymbol{W}_s+\boldsymbol{P}_s\boldsymbol{U}_s) \tag{8}$$

where, W_s, U_s are trainable parameters. To enable information exchange across scales, a cross-scake fusion block is adopted to convert features from one scale to another. The cross-scale graph is a bipartite graph that corresponds the nodes in one single-scale graph to the nodes in another graph. Assuming the cross-scale graph with adjacent matrix as $A_{s_1s_2}$, and the vectorized features of human joint and part are defined as $v_{s_1,i}=vec(conv_{s_1,\tau}((P_{s_1})_{:,i,:};\mu)), v_{s_2,k}=vec(conv_{s_2,\tau}((P_{s_2})_{:,k,:};\mu))$ to leverage temporal information, where τ,μ represent the temporal convolution kernel size and stride. Then, the edge weight between the joint and part can be inferred as

$$\begin{aligned}
r_{s_1,i}&=\sum_{j=1}^{M_{s_1}} f_{s_1}(v_{s_1,i},v_{s_1,j}-v_{s_1,i})\\
h_{s_1,i}&=g_{s_1}([v_{s_1,i},r_{s_1,i}])\\
r_{s_2,k}&=\sum_{j=1}^{M_{s_2}} f_{s_2}(v_{s_2,k},v_{s_2,j}-v_{s_2,k})\\
h_{s_2,k}&=g_{s_2}([v_{s_2,k},r_{s_2,k}])\\
(A_{s_1s_2})_{k,i}&=\mathrm{softmax}(h_{s_2,k}^T h_{s_1,i})\in[0,1]
\end{aligned} \tag{9}$$

where, the $f(\cdot), g(\cdot)$ denotes multi-layer perceptrons. Given the joint features at a certain time stamp, the part-scale feature can be updated by the edge weight.

To accurately predict the future human poses, the MS-GCN adopted encoder-decoder architecture, where a graph-based GRU is utilized to learn and update hidden states with the guide of a graph. Let $A_H\in\mathbb{R}^{M\times M}$ be the adjacent matrix of the inbuilt graph, which is initialized with the skeleton-graph, and $H^0\in\mathbb{R}^{M\times D_h}$ be the initial state of GRU. The processing procedures of GRU are defined as

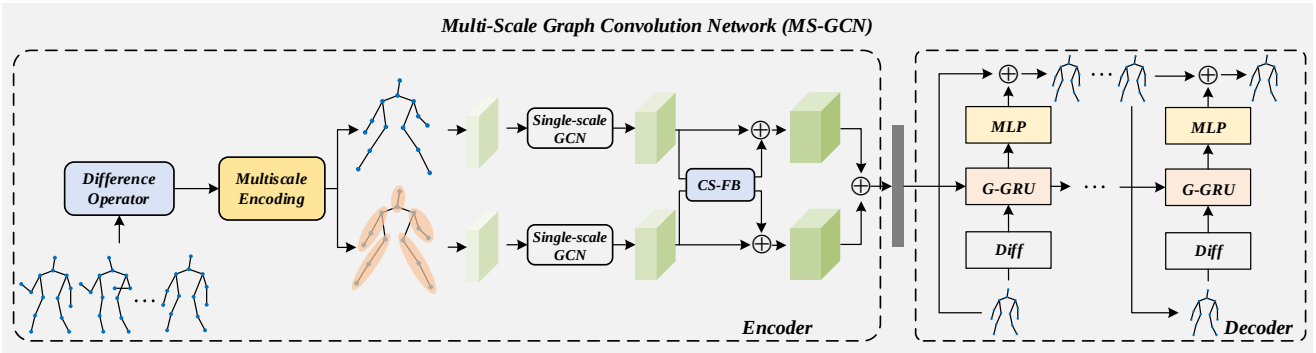

Fig. 3. Network structure diagram of pose prediction based on MS-GCN.

$$
\begin{aligned}
r^{(t)} &= \sigma(r_{in}(I^{(t)}) + r_{hid}(A_H H^{(t)} W_H)) \\
u^{(t)} &= \sigma(u_{in}(I^{(t)}) + u_{hid}(A_H H^{(t)} W_H)) \\
c^{(t)} &= \tanh(c_{in}(I^{(t)}) + r^{(t)} \odot c_{hid}(A_H H^{(t)} W_H)) \\
H^{(t+1)} &= u^{(t)} \odot H^{(t)} + (1 - u^{(t)}) \odot c^{(t)}
\end{aligned} \tag{10}
$$

At the end, the future pose is predicted by the decoder as

$$
\hat{P}^{(t+1)} = \hat{P}^{(t)} + f_{pred}(\mathrm{GRU}(diff(\hat{P}^{(t)}), H^{(t)})) \tag{11}
$$

where *diff* represents difference operator that calculate velocity and acceleration of human body keypoints. The entire structure diagram is shown in Fig. 3, which is trained end-to-end by the loss function as

$$
L_{pred} = \frac{1}{N} \sum_{n=1}^{N} \left\| (\boldsymbol{P}_{1:T_f})_n - (\hat{\boldsymbol{P}}_{1:T_f})_n \right\|_1 \tag{12}
$$

IV. Case Verification

This section will validate the effectiveness of the proposed method based on a self-made experimental dataset.

A. Dataset Preparation and Experimental Environment

To advance the application of action recognition and human pose understanding in sports skills training, this paper has developed a basketball motion dataset by collecting internet data and filming original content. This dataset comprises 400 RGB videos of various basketball actions such as dribbling, passing, shooting, and dunking, performed by different individuals in diverse settings. From this collection, approximately 6,000 images were meticulously selected and manually annotated with human keypoints for 2D pose estimation and prediction. The dataset is divided into training, validation, and testing sets in a ratio of [6:2:2]. Effective data augmentation techniques, including random flipping and the addition of random noise, have been applied. This dataset served as the basis for validating the proposed HICNN-PSTA and MS-GCN models. Detailed information on the hardware and software used in the experiments is provided in Table I.

TABLE I. Experimental Software and Hardware Environment Table

CPU	*Intel(R) Core(TM) i5-13400F*
GPU	*NVIDIA GeForce RTX 4070*
Operating System	*Ubuntu 18.04*
CUDA	*11.1*
Programming	*Pytorch1.10.0, Python 3.8*

To fully demonstrate the effectiveness of the proposed methods, this study conducted a series of comparative experiments to comprehensively evaluate the performance of the proposed algorithms in sports action recognition and human pose estimation and prediction. Given that the initial part of this research utilized a multi-task framework, we selected appropriate models for comparison based on the specific problems addressed. For sports action recognition, the AGC-LSTM [19] model was chosen as the benchmark, whereas the HPRNet [20] model was used as the comparative standard in the domain of 2D human pose estimation. The specific content of the experiments is as follows: the action recognition accuracy for different basketball skills, the results of 2D human keypoint estimation and pose prediction for various basketball skills, and ablation studies conducted on the proposed algorithm. During the training process, an Adam optimizer with an initial learning rate of 0.001 was employed, accompanied by a linear learning rate decay coefficient set at 0.95. The training batch size was configured to 64, and the number of epochs for iteration was set to 50.

B. Experimental Results

To fully demonstrate the effectiveness of the proposed algorithm, this section first explores the performance of the proposed HICNN-PSTA in basketball skill action recognition, building on the experimental setup described above. The model was trained using a supervised learning approach on the dataset constructed for this study, and the classification cross-entropy error variation curve during the training process is shown in Fig. 4. It is evident that the error rapidly decreased and stabilized shortly after training commenced, reflecting the model's capability to effectively adjust model weights using error gradients and ultimately achieve convergence. Additionally, the proposed HICNN-PSTA was also subjected to quantitative experiments, as shown in Table II, which includes the recognition accuracies of various models for different basketball skill actions. The proposed algorithm achieved the highest recognition accuracy across all skills, likely benefiting from the PSTA module's superior ability to capture temporal-spatial information, particularly crucial for the rapid and complex movements characteristic of basketball and other sports skills.

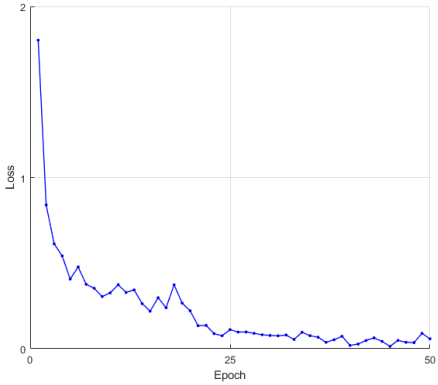

Fig. 4. Training loss rate curve.

TABLE II. Quantitative Comparison of Sports Action Recognition

Method	**Accuracy**			
	Dribbling	*Shooting*	*Passing*	*Dunk*
AGC-LSTM	94.53	90.24	92.50	85.02
Proposed	**96.11**	**91.15**	**93.87**	**91.52**

Furthermore, to delve deeper into the performance of the proposed method in sports action recognition, we conducted several ablation experiments, the results of which are presented in Table III. This experiment compared the proposed model with two variants: one substituting the backbone network with ResNet, referred to as ResNet-PSTA, and another omitting the PSTA module, referred to simply as HICNN. The assessment

criterion was the F1-score on the test set, and performance across four types of sports skills was evaluated. The results indicate that the proposed HICNN-PSTA model achieved the best performance in recognizing different sports skills. This demonstrates the superior capability of the proposed modules in capturing latent movement information and modeling spatio-temporal relationships, which are crucial for sports action recognition. The effectiveness of the HICNN-PSTA model underscores its significant role in accurately recognizing complex sports actions.

TABLE III. ABLATION STUDY OF SPORTS ACTION RECOGNITION

Method	F1-score			
	Dribbling	*Shooting*	*Passing*	*Dunk*
ResNet-PSTA	0.81	0.79	0.70	0.68
HICNN	0.76	0.73	0.61	0.71
Proposed	**0.88**	**0.81**	**0.76**	**0.80**

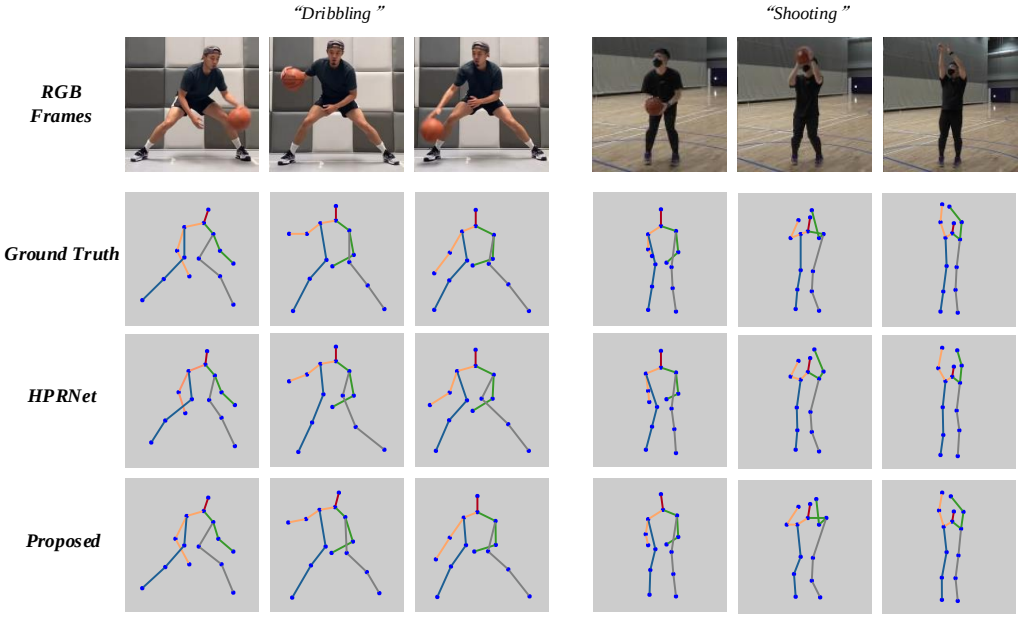

Fig. 5. Qualitative comparison of sports pose estimation based on the HICNN-PSTA.

Additionally, we further explored the performance of HICNN-PSTA in sports pose estimation, as shown in Table IV and Fig. 5. Table IV employs the Percentage of Correct Keypoint Percentage (PCK) as the evaluation metric under thresholds of [0.2, 0.1, 0.05], demonstrating the model's performance in dribbling pose estimation. It is observable that HICNN-PSTA achieved the best recognition results across all thresholds, reflecting the model's capability to utilize effective information from action recognition to enhance the accuracy of human pose estimation in a multi-task framework. The results of human pose estimation for dribbling and shooting are illustrated in Fig. 5.

In addition, we conducted multiple experiments to evaluate the performance of the proposed MS-GCN in human pose prediction and compared it with two widely-used models, TP-RNN [21] and Traj-GCN [22], as shown in Table V. This table employs the Mean Angle Error (MAE) as the evaluation metric, detailing the prediction results for different sports skills over various time intervals. It is evident that, compared to the other two algorithms, MS-GCN achieved the best prediction results in the pose prediction for shooting and dunking across different time intervals. Although Traj-GCN outperforms MS-GCN in predicting these two skills, overall, MS-GCN still demonstrates substantial potential and practicality in predicting sports skills.

TABLE IV. QUANTITATIVE COMPARISON OF SPORTS POSE ESTIMATION

Method	PCK@0.2	PCK@0.1	PCK@0.05
HPRNet	82.11	75.8	70.01
Proposed	**90.08**	**81.47**	**79.34**

TABLE V. QUANTITATIVE COMPARISON OF SPORTS POSE PREDICTION

Sports Skills	ms	TP-RNN	Traj-GCN	Proposed
Dribbling	80	0.34	**0.32**	0.33
	160	0.61	0.50	**0.42**
	320	1.25	1.19	**0.88**
Shooting	80	0.56	0.45	**0.41**
	160	1.48	0.86	**0.78**
	320	1.97	1.28	**1.01**
Passing	80	0.66	0.59	**0.47**
	160	1.01	1.13	**0.92**
	320	1.68	**1.45**	1.47
Dunk	80	0.30	0.38	**0.28**
	160	0.75	0.49	**0.50**
	320	1.32	1.06	**0.92**

V. CONCLUSION

This study aims to explore how to better utilize neural network models to optimize sports skill training, with a focus on achieving sports action recognition and the estimation and prediction of athletes' poses, thereby advancing the application of neural networks and other artificial intelligence algorithms in the field of sports training. To address these challenges, we first propose a multi-task framework-based HICNN-PSTA model. This model enhances the feature extraction capabilities of the conventional CNN by integrating the Involution operator into the backbone network. Additionally, this study constructs a PSPA module based on the attention mechanism to fully capture the latent spatio-temporal information of sports actions, thereby improving the efficiency of the algorithm with the help of the multi-task framework. Furthermore, to accurately predict future poses of athletes and provide training recommendations, this paper introduces an MS-GCN model based on a multi-scale graph. This algorithm considers the constraints between human body keypoints and segments, significantly enhancing the capability to model the complex sports skills. Detailed experiments validate that the proposed algorithms can effectively recognize sports actions and also demonstrate excellent performance in human pose estimation and prediction. In the future, we plan to integrate more advanced neural network algorithms to address the generalization deficiencies across different sports, thereby further optimizing sports skill training.

REFERENCES

[1] P. Wang, "Research on Sports Training Action Recognition Based on Deep Learning," Scientific Programming, vol. 2021, pp. 1–8, Jun. 2021, doi: 10.1155/2021/3396878.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[3] K. Host and M. Ivašić-Kos, "An overview of Human Action Recognition in sports based on Computer Vision," Heliyon, vol. 8, no. 6, p. e09633, Jun. 2022, doi: 10.1016/j.heliyon.2022.e09633.

[4] K. Soomro and A. R. Zamir, "Action Recognition in Realistic Sports Videos," in Computer Vision in Sports, T. B. Moeslund, G. Thomas, and A. Hilton, Eds., in Advances in Computer Vision and Pattern Recognition. , Cham: Springer International Publishing, 2014, pp. 181–208. doi: 10.1007/978-3-319-09396-3_9.

[5] N. FarajiDavar, T. De Campos, J. Kittler, and F. Yan, "Transductive transfer learning for action recognition in tennis games," in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain: IEEE, Nov. 2011, pp. 1548–1553. doi: 10.1109/ICCVW.2011.6130434.

[6] J. Calandre, R. Péteri, and L. Mascarilla, "Optical Flow Singularities forSports Video Annotation: Detection of Strokes in Table Tennis".

[7] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos".

[8] B. X. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA: IEEE, Jun. 2015, pp. 1293–1301. doi: 10.1109/CVPR.2015.7298734.

[9] J. Lin, C. Gan, and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 7082–7092. doi: 10.1109/ICCV.2019.00718.

[10] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Body Joint Guided 3-D Deep Convolutional Descriptors for Action Recognition," IEEE Trans. Cybern., vol. 48, no. 3, pp. 1095–1108, Mar. 2018, doi: 10.1109/TCYB.2017.2756840.

[11] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 469–478. doi: 10.1109/CVPR.2018.00056.

[12] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA: IEEE, Jun. 2014, pp. 1653–1660. doi: 10.1109/CVPR.2014.214.

[13] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in Computer Vision – ECCV 2016, vol. 9912, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9912. , Cham: Springer International Publishing, 2016, pp. 483–499. doi: 10.1007/978-3-319-46484-8_29.

[14] L. Pishchulin et al., "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 4929–4937. doi: 10.1109/CVPR.2016.533.

[15] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, Jun. 2020, pp. 5385–5394. doi: 10.1109/CVPR42600.2020.00543.

[16] C.-H. Chen and D. Ramanan, "3D Human Pose Estimation = 2D Pose Estimation + Matching," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 5759–5767. doi: 10.1109/CVPR.2017.610.

[17] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 7745–7754. doi: 10.1109/CVPR.2019.00794.

[18] D. Li et al., "Involution: Inverting the Inherence of Convolution for Visual Recognition," Apr. 11, 2021, arXiv: arXiv:2103.06255. Accessed: Aug. 21, 2024. [Online]. Available: http://arxiv.org/abs/2103.06255

[19] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA: IEEE, Jun. 2019, pp. 1227–1236. doi: 10.1109/CVPR.2019.00132.

[20] N. Samet and E. Akbas, "HPRNet: Hierarchical point regression for whole-body human pose estimation," Image and Vision Computing, vol. 115, p. 104285, Nov. 2021, doi: 10.1016/j.imavis.2021.104285.

[21] H.-K. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, "Action-Agnostic Human Pose Forecasting," in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA: IEEE, Jan. 2019, pp. 1423–1432. doi: 10.1109/WACV.2019.00156.

[22] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning Trajectory Dependencies for Human Motion Prediction," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 9488–9496. doi: 10.1109/ICCV.2019.00958.