ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Design of Sports Training Action Recognition and Feedback System Based on Deep Learning

Shan Song[1,*]

[1] Zhengzhou Academy of Fine Arts 451450, Henan, China

SUMMARY: *this paper introduces a deep learning-based sports training action recognition and feedback system to tackle issues such as motion deviation, unstable recognition and delayed correction in real-time training. The system uses a dual-channel perception layer to fuse RGB video and IMU data, extracts 33 skeletal keypoints and synchronized acceleration/gyroscope features, and forms a unified temporal tensor. A task decomposition module breaks down continuous movements into essential posture sequences by energy-based stage boundary reasoning and offers structured input for model learning. Based on the above, a hybrid Dilated Temporal Convolutional Network and Graph Attention Network (TCN+GAT) are used to address the problems of long-term temporal dependency and cross-joint spatial relationship. Node features contain pose coordinates and extremity IMU signals, and an attention-weighted loss gives more weight to joints with large angular excursions for faster convergence. A deviation-aware feedback mechanism maps the classification results and joint angle offsets to a dual feedback channel of voice prompts and vibration intensity, thereby forming a closed loop of "perception-reasoning-correction". Experiments on a combined dataset of a Kinetics-400 subset and 4,760 self-collected training clips (push-ups, squats, lunges and jumping jacks) evaluate Accuracy, Angle-Score and end-to-end Latency. The accuracy of the proposed model is 92.4±0.5%, the Angle-Score is 87.9±0.6%, and the latency is ≤1.8s; it outperforms TCN-Only and TCN+LSTM baselines in both recognition and response speed. Based on the ablation studies, the graph attention module, angle refinement strategy and dual feedback design have improved both the stability and feedback hit rate. Based on the above results, it can be seen that the TCN+GAT architecture with RGB-IMU fusion can offer practical real-time guidance for sports training and provide a deployable solution for intelligent coaching systems.*

KEYWORDS: *Deep learning, Sports training, Action recognition, Real-time feedback control*

1 Introduction

The Precision and Stability of movement execution in sports training directly affect training efficiency and injury risk control. The traditional way of manual demonstration and verbal correction has certain problems, such as high subjectivity, delayed feedback and lack of quantifiability, and is thus unable to meet the demands of high-frequency training and individualised guidance. Recently, training assistance systems based on video recognition and human pose estimation have gradually appeared. However, due to deficiencies in lighting interference, occlusion changes and various types of motion, their recognition accuracy is still inconsistent in actual use [1]. Some studies have applied keypoint detection combined with temporal analysis models for action classification [2], but they lack semantic interpretation

*7337373132@163.com
https://doi.org/10.65102/is2026879 Published: 30 April, 2026

capabilities in the judgment of rhythm fluctuations and fine-grained offsets. A second type of research is an acceleration pattern-matching model based on IMU sensors [3]. Although it has the advantage of a high sampling rate, its spatial correlation capability is relatively weak; thus, it cannot achieve the purpose of joint control for structural alignment and feedback generation.

Given the requirement of "continuity + structure" in a dynamic training environment for recognition, we need to build an all-encompassing model framework that supports multimodal perception, topological structure reasoning, and real-time feedback regulation. Deep learning technology has shown good generalisation ability in the field of action recognition, and the structure that integrates temporal convolution and graph attention can establish dual-channel associations in the temporal and joint dependencies [4]. RGB video can provide relatively coarse pose change information, and IMUs have small-angle drift. Together, they are expected to reduce the recognition fluctuations of single-mode under complex training conditions [5]. However, the current research has not yet connected the "identity-feedback" mechanism. Most models are still at the classification output level and lack the mapping logic for deviation measurement to feedback generation.

This study will focus on building a "deep learning-based sports training action recognition and feedback system", and a three-dimensional framework design will be used to organize the "perception mapping - structure recognition - feedback regulation" processes. The research questions are as follows: RQ1: Can the multimodal fusion model reliably distinguish among the different stages of action and output structural offset parameters? RQ2: Can the recognition results be converted into executable real-time error correction instructions? RQ3: Can the feedback mechanism form a high-frequency closed loop under the condition of controllable delay? This study intends to establish a high-stability, real-time-performance training-assistance scheme that provides algorithmic support for intelligent training systems.

2 Related Works

The current technical approaches in the research of sports training action recognition and feedback systems mainly fall into the following four categories: visual recognition based on key point estimation, sensor-driven time series modelling, fine-grained stage division and structured data modelling, and interactive deviation correction mechanisms. Cao and others (2017) proposed the Part Affinity Fields structure for visual keypoint estimation, and in order to achieve multi-person pose estimation, joint point heat maps and part correlation vectors were introduced for reasoning [7]. The above scheme has good stability in large-scale motions, such as swinging and jumping; however, there is skeleton drift during self-occlusion and turning actions, and thus cannot meet the demand for high-precision deviation correction. Wang et al. (2019) have studied how to build a sensor-driven time series model, and have also applied CNN-LSTM and time-frequency fusion networks to data recognition based on gyroscopes and accelerometers [6]. This kind of model is suitable for low-latency requirements of periodic movements, such as push-ups and squats, but it lacks the spatial topological representation and cannot indicate whether the amplitude of the movement has changed; thus, it has poor structural interpretability. Shao et al. (2020) proposed the FineGym dataset in the direction of fine-grained stage division, and divides gymnastic movements into stages such as take-off, somersaulting and landing [8] to promote the application of temporal convolution and converter models in stage discrimination. However, this framework has manual slicing. The new training project cannot set up stage divisions and therefore will be unable to perform continuous real-time assessment. Tharatipyakul et al. (2024) also noted that the current pose assessment systems are generally still at the stage of "recognition result display" and have not formed a closed loop of

"recognition - assessment - feedback" [9]. Liu and others (2022) have introduced a pose-embedding-based generation mechanism for running training advice in the direction of interactive deviation correction, and this mechanism has a delay of about 2.1 seconds [10]. This system has visualisation functions, but the feedback is only "tag-based reminders" and lacks executable correction instructions. Although the existing technical methods have built recognition, modelling and prompt mechanisms, there are still three main deficiencies: First, the robustness of the keypoint model to complex occlusions and non-standard postures is weak; Second, although the sensor model has low latency, it lacks semantic understanding of joint coordination. Thirdly, most of the feedback mechanisms are static displays without executable correction logic. Table 1 shows the differences among the performances of the above-mentioned main methods.

Table 1: Comparison and Summary of Training Action Recognition and Feedback Methods

Method Name	Input Modality	Model Type	Performance Metrics	Limitations
PAF-based Pose Estimation [7]	RGB Video	Keypoint Detection + Part Affinity Fields	COCO multi-person AP leading among peers, latency ~45 ms	Sensitive to occlusion, trajectory drift
IMU-CNN-LSTM [6]	IMU Sensors	CNN + LSTM	Accuracy 80%–90%, latency ≤50 ms	Lacks spatial topology representation
FineGym-based TCN [8]	Video Sequence	Temporal Convolution	~90% classification accuracy in small-sample stage	Requires manual clip annotation
PoseCoach [10]	RGB + Skeleton	Pose Embedding Contrastive	Latency ~2.1 s	Feedback not executable
Proposed Method	RGB + IMU	TCN + GAT Fusion	Accuracy 92.4% ± 0.5 / Feedback Delay ≤1.8 s / Angle-Score 87.9% ± 0.6	Requires synchronization between skeleton and sensors

3 Design of a Deep Learning-driven Sports Training Action Recognition and Feedback System

3.1 Task Decomposition of training movements and division of key posture stages

In order for it to be recognised by the recognition model, the continuous motion sequence of sports training needs to be divided into analysis-friendly stage units that can effectively help deep learning models learn the temporal dependency of posture change and other key nodes. In the task decomposition stage, the action flow is first divided into three parts: the beginning of an action, the contents of its execution, and the end of execution. A group of representative pose frames has been set up at each stage. The choice of a key posture should be based on the geometric stability and energy fluctuations of the skeletal points, and the extracted nodes should cover the main changes in the completion degree of the action. In the quantitative definition, it

is assumed that the position of the bone point of the training action at time t and its velocity vector are known. The total energy characterization can be expressed as: $t\ p_{t,j}\ v_{t,j} = p_{t,j} - p_{t-1,j}$

$$E_t = \sum_{j=1}^{J} \left\| v_{t,j} \right\|^2 + \beta \sum_{k=1}^{K} \left(\frac{d\theta_{t,k}}{dt} \right)^2 \tag{1}$$

$p_{t,j}\ j\ t\ v_{t,j}\ \theta_{t,k}\ k\ \beta$ Among them, represents the three-dimensional spatial position of the skeletal node at a certain time, and is used to describe the displacement amplitude of the attitude trajectory. It is obtained by taking the difference between adjacent frames to measure the intensity of local motion. Represents the angle change of the joint and shows the trend of joint opening and closing. Is the balance coefficient of velocity energy and angular energy. The formula comprehensively represents the energy state of a single frame by simultaneously taking into account linear displacement and changes in angular velocity, and is used to find the highly excited state in the time domain. To set the upper and lower bounds for the energy difference and acceleration, boundary scores were introduced.

$$S_t = \lambda \left| E_t - E_{t-1} \right| + (1-\lambda) \left\| a_{cm}(t) \right\| \tag{2}$$

$E_t\ t\ E_{t-1}\ a_{cm}(t)\ t\ \lambda$ Among them, represents the action energy value at moment t, which is composed of the speed term and the joint angle change term, and is used to describe the comprehensive movement intensity at this moment. Represents the energy value at the previous moment, and the difference between the two indicates the energy fluctuation between frames. Represents the vector acceleration of the whole center of mass at time t, and indicates whether the entire human body is in a state of posture transition. Is the weighting coefficient for sudden energy change and overall inertia change. This formula extracts the most probable stage transition points in the time series by detecting the joint peak of inter-frame energy fluctuations and inertial drift amplitudes, and provides a quantitative basis for the following piecewise reasoning. Keyposes in actual applications are not only based on the geometric features of skeleton points but also need local window pooling in conjunction with context information to improve the phased consistency expression of the model. Concatenate the feature vectors of all nodes with attitude angles, angular velocities and context-pooling vectors to maintain continuity and stability across stages.

As shown in Figure 1, the stages of the decomposition process are as follows: original signal analysis, bone topology extraction, energy mapping, boundary reasoning and pose screening. All steps output in a learnable format to ensure that the following recognition model can accept a unified structured sequence.

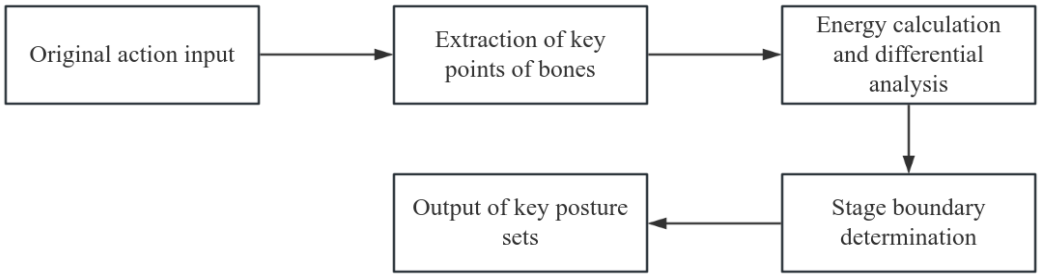

Figure 1: Breakdown of the process framework for the training action stage.

3.2 Perception Channel and input signal construction mechanism for Training actions

Deep recognition of sports training movements is based on the joint modelling of posture deformation and dynamic behaviour. Therefore, this system has built a dual-mode input mechanism consisting of skeletal key point channels and inertial sensing channels in the perception layer. Among them, a skeleton channel is used to get the spatial displacement structure, and an IMU sensor is used to acquire joint acceleration and rotational trends. Both are collected simultaneously in a frame-aligned way and encoded as a single input tensor. Let the joint point at time t in two or three dimensions be $\mathbf{r}_t$, and its velocity and linear acceleration are given by: $j\ t\ p_{t,j}$

$$f_{t,j} = \lfloor p_{t,j}, v_{t,j}, \alpha a_{t,j} \rfloor \tag{3}$$

$p_{t,j}\ j\ t\ v_{t,j} = p_{t,j} - p_{t-1,j}\ a_{t,j} = v_{t,j} - v_{t-1,j}\ \alpha$ Among them, represent the coordinate of the joint point at time t in two-dimensional or three-dimensional space; is the instantaneous velocity vector obtained from the difference between adjacent frames; is the linear acceleration vector obtained by taking the difference of the velocity; is the scaling coefficient of the acceleration term, which is used to suppress high-frequency jitter. This formula expresses the node dynamics in the form of the minimum splicing of "position + velocity + acceleration"; it is short in length, simple to implement, and convenient to align with the IMU channel on the time axis and directly input into the following model. The IMU sampling signal is the triaxial acceleration of the sensing module at time t and the triaxial angular velocity. Then the fused feature is as follows: $k\ t\ a_{t,k}\ g_{t,k}$

$$z_{t,k} = \gamma \cdot a_{t,k} + (1-\gamma) \cdot g_{t,k} \tag{4}$$

γ Among them, it is the weighting coefficient for linear acceleration and rotational velocity; it is used to balance the contributions of inertia-dominated motion and torsion-dominated motion in the feature space. The above formula reduces and fuses various sensor channels to help the model learn the general motion pattern in a low-dimensional space, thus avoiding training instability caused by a high number of feature dimensions.

$f_{t,j}\ z_{t,k}$ To achieve cross-modal alignment, the feature of the skeletal node and the feature of the sensor node are aligned to a unified sampling frequency in time via linear interpolation, and then concatenated along the channel dimension to form a combined input sequence. The above mechanism enables the model to learn both the structure of the pose and the trajectory of motion at a later time step simultaneously, providing a continuous and controllable signal foundation for the generation of feedback strategies.

3.3 Collaborative Design of Deep Learning Recognition Models and Feedback Strategies

The recognition of training movements needs to be precise in terms of joint movement trajectory modeling, and after discrimination, a feedback control mechanism should be activated immediately to prevent trainees from continuously performing in the wrong posture. A dual-channel fusion structure of "Dilated TCN + Graph Attention Network (GAT)" is proposed in this paper to address the rhythm variation of fast and slow movements over time and the connections among key nodes in space. The input signal is a sequence of skeletal key

points extracted from RGB images and the three-axis acceleration of an IMU sensor at the same time, denoted as and respectively, and synchronous mapping is performed in the fusion stage. The time-series modelling part is an extended convolution structure, and its basic form is: S_t A_t

$$H_t^{(l)} = \sigma\left(W_d^l * H_{t-d}^{(l-1)} + b^{(l)}\right) \tag{5}$$

$W_d^{(l)}$ d $H^{(l-1)}$ σ $b^{(l)}$ Among them, the convolution kernel has an expansion rate of 1, is the temporal feature of the previous layer, is the activation function, and is the bias term. The formula is used to expand the receptive field without increasing the number of parameters, and at the same time, both long-distance motion correlations and local fine-grained changes in motion can be captured by the model. A Graph Attention Network is employed in the spatial model to weigh the influence of various skeletal nodes differently. Update Form of nodeis: i

$$F_i = \sum_{j \in N(i)} \alpha_{ij} \cdot W_g H_j \tag{6}$$

W_g H_j α_{ij} j i Among them, is the linear transformation matrix; is the feature of neighboring nodes; and is the normalized attention weight that indicates the contribution intensity of node to node. Therefore, this method will give greater weight to the high-difference joints during motions such as knee lifting and arm extension and ignore the general motion trend.

After outputting the recognition result, it will be mapped to the dual-channel feedback module of voice and vibration. When the Angle deviation exceeds the threshold, a corrective voice command will be issued, and based on the amplitude of the error, the vibration intensity will be adjusted in real time for closed-loop training control.

3.4 System deployment Mode and Terminal Collaborative Operation Mechanism

The actual operating speed of the sports training action recognition system is based on the cooperation schedule for the acquisition device at the end and the reasoning module in the middle. The Camera and the IMU sensor will be responsible for acquiring the spatial position and acceleration data, respectively, in actual deployment. Synchronous input streams are formed by Bluetooth or Wi-Fi, aligned in time, and the initial feature encoding is carried out at the edge. To achieve millisecond-level closed-loop response for recognition and feedback, the server side and terminal devices use a hierarchical reasoning structure: lightweight TCN encoding and candidate-stage screening are performed at the end, and GAT path reasoning and feedback instruction generation are carried out on the server side. The Latency and Feedback accuracy of the various deployment modes differ considerably. Thus, an inference latency model is built for dynamic switching strategies:

$$D_t = \alpha \cdot D_e + (1-\alpha) \cdot D_c + \gamma \cdot R_s \tag{7}$$

D_e D_c R_s α γ Among them are the end-side inference delay, the server inference delay, the sensor synchronization cost, and an and a trade-off coefficient. This formula is employed to determine how much each deployment structure contributes to the response delay, and then a suitable path is selected at runtime.

To test the effect of the collaborative deployment mode, a comparison experiment as shown in Table 2 was conducted. Among them, local inference refers to all computations being carried out at the end, cloud inference refers to unified processing after transmission to the server, and

the collaborative mode has the terminal filter the candidate stage and completes fine recognition on the server.

Table 2: Comparison of Latency and Feedback Performance under Different Deployment Modes.

Deployment Mode	Average Response Latency (s)	Feedback Trigger Hit Rate (%)
Local Inference	1.1	84.7
Cloud-based Inference	2.6	91.3
Collaborative Hybrid Deployment	1.8	92.4

The comparison results show that the cooperative structure has a high recognition accuracy and controls for delay. It can still provide stable feedback in the case of complicated training, and the network will not be disturbed. The final deployment architecture will be a cloud-based inference for high-complexity action analysis and edge-side inference for the quick-prompt stage. Both have continuously produced output by instruction caching and timestamp calibration to establish a stable training feedback loop.

4 Results

4.1 Dataset

The construction of the dataset for sports training action recognition aims to cover different individual differences, changes in action amplitude and execution quality deviations, and uses a fusion strategy of "public set + self-built collected set". The public part selected clips from Kinetics 400 with labels similar to push-ups, squats, jumping jacks and dumbbell bends, and filtered out data with complex backgrounds and severe occlusion. Finally, about 2,960 video samples were kept. The self-built collection part had 60 subjects who performed the four types of target movements in three typical training environments: gym, dormitory and outdoors. Each subject has recorded 20 valid clips of that kind of movement. The collection equipment was a 120fps RGB camera and a six-axis IMU sensor (100Hz), and they were synchronized by a common time. MediaPipe Pose is used to extract a sequence of 33-dimensional keypoint nodes for skeletal tracking. At the same time, a bandpass filter and linear interpolation are applied to the IMU signal to build a unified RGB-IMU dual-mode pose matrix. To avoid sample imbalance among categories and thus bias in model training, a weight-based sampling strategy is used, and the sampling probabilities of each action category are defined as:

$$P_i = \frac{1/n_i}{\sum_{k=1}^{C}(1/n_i)} \tag{8}$$

n_i i C Among them, is the sample size of the category, and represents the total number of categories. The above formula increases the collection frequency of rare samples using the inverse-weight method and maintains a statistically balanced training batch.

In order to ensure that the constructed dataset can simultaneously show variations in individuals, different amplitudes of motion and deviations in execution quality, all four stages of data preparation were divided: multi-modal acquisition, keypoint extraction, temporal alignment and unified feature formatting. Synchronously acquire RGB video streams and IMU signals first, then perform skeletal keypoint estimation and sensor signal preprocessing. The resulting visual

and inertial features were then temporally aligned and fused into a single-module dual-modality pose representation for standard input to the following recognition and feedback model.

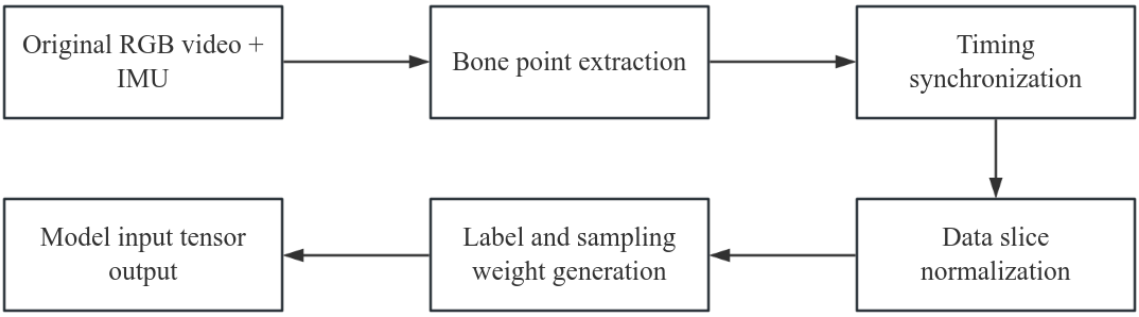

Figure 2: Framework for Dataset Construction and Alignment.

Figure 2 is the main process of dataset generation. Strictly standardised in terms of initial collection and construction of a unified modal format to ensure the consistency of input distribution for different individuals, devices and environments during subsequent model training, thereby improving the applicability of the recognition model in actual training scenarios.

4.2 Data Preprocessing

Multimodal collected signals of sports training movements have different characteristics in the time domain and amplitude. Synchronise the time and improve the amplitude stability using a single normalisation and noise-reduction method. The two channels of input in this study are a bone keypoint sequence and IMU sensor data. The frame rate of the bone is set to 120 fps, and the sampling rate of the IMU is 100 Hz. First, it is mapped to a unified time axis via linear interpolation, and the adjacent window average filling is carried out for the missing frames to avoid interference from pose jitter in model convergence. Process the three-dimensional coordinates of the skeletal nodes by zero-mean standardisation, and its formula is as follows:

$$\hat{P}_{t,j} = \frac{P_{t,j} - \mu_j}{\sigma_j} \tag{9}$$

$P_{t,j}$ j t μ_j σ_j Among them, is the original coordinate of the joint in the frame, and and are the mean and standard deviation of this node over the entire sequence, respectively. The formula is used for scale normalisation of posture data, and thus the difference in height or limb length among individuals does not affect the model's extraction of motion trends. Filter the sensor signal with first-order exponential smoothing, and the filtering formula is as follows:

$$\tilde{s}_t = \alpha \cdot s_t + (1-\alpha) \cdot \tilde{s}_{t-1} \tag{10}$$

s_t α Among them, the concatenation vector of the original acceleration and angular velocity is used, and the smoothing coefficient is set to 0.6 in this study. The purpose of this formula is to reduce high-frequency flutter noise, keep the main components that show the change trend of pose, and reduce the sensitivity of the model to slight jitter in feature learning.

To improve the stability of cross-channel alignment further, all windows are either truncated or repeatedly filled in units of 64 frames so that each sample has the same length. The final input format is a key-point matrix of (64×33×3) and a sensor matrix of (64×6), which are merged in a parallel-channel structure at the front end of the model. The above pre-processing steps will help the model learn muscle group force distribution and motion-change rhythm under a stable distribution for more stable input to the recognition and feedback control modules.

4.3 Performance Evaluation Indicators

Performance assessment of the sports training action recognition and feedback system does not rely on a single classification accuracy, but also needs to build a composite evaluation system with three dimensions: classification accuracy (Accuracy), joint angle deviation score (angle-score), and end-to-end delay (Latency). Frame-by-frame analysis was carried out on the three representative strength training exercises in the test set: push-ups, squats, and lunges, and both whole-segment and frame-level indicators were computed to ensure that the model performed stably at different time scales. The general recognition accuracy rate is given by the form of standard classification comparison as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

TP, TN, FP, FN Here, respectively represents the number of positive and negative samples identified as correct/incorrect. Using the formula above, we can find the total accuracy rate for action category detection; this indicates the proportion of all action categories correctly recognised by the model. According to the experiment, the accuracy is about 92.4% $\pm$ 0.5, and it has not changed much when changing the light and background. To evaluate the reliability of feedback in a posture-correction system, a joint angle deviation index is used, and its mean square error is defined as follows:

$$AS = 1 - \frac{1}{J}\sum_{j=1}^{J}\left|\theta_j^{pred} - \theta_j^{gt}\right| \tag{12}$$

J θ_j^{pred} θ_j^{gt} Among them, is the number of skeletal nodes, and the others are the predicted and true Angle values, respectively. The above formula is used to assess the quantitative performance of the model in pose offset. The Angle-Score is 87.9% $\pm$ 0.6, and thus the model shows a stable consistency in the deviation correction prompts for the elbow and knee joints.

To determine how long it has been since the last use by a person, this value is known as latency. The time from the collection of the input to the output of the feedback instruction is called the system response time. According to the above data, the delay has been regulated to 1.8 seconds. Based on the above indicators, it can be seen that this system balances the accuracy rate, deviation score and end-to-end delay, and thus provides performance support for the following real-time deviation correction and personalised evaluation.

4.4 Ablation Experiment

To verify the role of each functional module in the closed-loop process of training action recognition and feedback, a combined comparative experiment was designed based on dual-channel perception, graph attention reasoning and feedback strategy control, and key components were progressively removed to analyze the performance degradation. All the

scores in the performance are divided into three categories: Accuracy, Angle-Score and Latency; a general scoring function will be built simultaneously.

$$S = \frac{A_c \times AS}{L_a} \tag{13}$$

S S Among them, it is the corresponding feedback efficiency per unit time. If both the recognition accuracy and the angle assessment ability are enhanced at the same time and the response delay is shorter, then this value will be higher. The experiments removed the graph attention module (w/o GAT), turned off the attitude deviation correction mechanism (w/o Angle Refinement), and disabled the combined feedback of voice and vibration (w/o Dual Feedback) for comparison, respectively. The indices of all the strategy combinations are shown in Table 3.

Table 3: Response Performance for Various Strategy Combinations.

Strategy combination	Accuracy (%)	Angle-Score (%)	Latency (s)
w/o GAT	88.1 ± 0.6	82.7 ± 0.7	2.4
w/o Angle Refinement	90.3 ± 0.5	84.1 ± 0.6	2.1
w/o Dual Feedback	91.5 ± 0.4	86.3 ± 0.5	2.0
Complete Model	92.4 ± 0.5	87.9 ± 0.6	1.8

The results show that after removing graph attention, the dependency among pose nodes cannot be modelled properly, and recognition accuracy drops by about 4 per cent. After switching off the angle-correction mechanism, the Angle-Score dropped significantly; therefore, this module must be responsible for reducing the amplitude of the error. Disabling the dual-feedback mechanism will not affect the recognition result; however, it will be slower to respond. The full model has a better overall score, and thus the collaborative structure of compound perception reasoning and multimodal feedback strategy is necessary.

5 Model Training Process and Validation Analysis

5.1 Data Format construction and skeletal diagram structure conversion process

This study is based on the Kinetics-400 subset and a self-built dataset, with a total of 4,760 samples, all of which have been uniformly input with a length of 64 frames. MediaPipe Pose in the RGB video extracted 33 key points of bones. The self-built set synchronously collected the IMU signals at the extremities (both wrists and both ankles) and aligned them with a unified timestamp to form a dual-channel structure of "bone sequence + sparse IMU". The skeletal nodes are connected in a shoulder-elbow-wrist, hip-knee-ankle structure, and the main chain of the spine is also an adjacent node. A spatio-temporal graph sequence is built frame-by-frame in the time dimension to satisfy the parallel-input requirement of TCN+GAT. To reduce cross-individual differences, normalize the joint coordinates by resolution, and bandpass filter and linearly interpolate the IMU channels to fill in missing segments. The joint features at the node level are built according to the following formula:

$$f_{i,t} = \left(P_{i,t} \| \gamma \cdot u_{i,t}\right)$$

$P_{i,t} \in R^3$ i t $u_{i,t} \in R^6$ $\gamma \in [0,1]$Among them, represents the spatial coordinate of the skeletal node at time t, is the exclusive six-dimensional IMU signal for the wrist/ankle node, and the remaining nodes are automatically set to zero. is the modal balance coefficient, which is used to suppress the amplitude bias of the inertial channel. Although the 33 nodes are consistent, only the inertial information at the extremities of the limbs is fused in this formula to avoid noise amplification due to redundant sensors.

The format of the graph conversion is in the form of "node feature matrix + edge index", and the edges include topological edges of the human body and cross-frame time edges of the same node, forming a spatio-temporal skeletal graph. The three divisions of the data are training, validation and test, and they should be independent of each other.

5.2 Model Training Process and Description of Hyperparameter Settings

The training of the model has been carried out using the Kinetics-400 subset and the self-constructed training set. The total number of samples was 4760 segments, and of them, the training set was 70%, the validation set was 20%, and the test set was 10%. All sequences are uniformly converted into a bone + IMU time input. Extract the three-dimensional positions of 33 nodes from the bone part. Only the wrist and ankle nodes are linked to IMUs; all other nodes are set to zero for structural consistency and modal sparsity. The Construction Method of Node Features is as follows:

$$f_{i,t} = p_{i,t} + \sigma(W u_{i,t}) \tag{15}$$

$p_{i,t} \in R^3$ i t $u_{i,t} \in R^6$ W σ Among them, is the spatial coordinate of the skeletal node at time t, is the six-dimensional signal of the IMU (zero for nodes without IMU), is the trainable mapping matrix, and is the Sigmoid function, which is used to suppress the amplitude of the inertial mode and enable the IMU to participate in recognition as a skeletal error correction term. To increase the convergence speed of the training process for key nodes, a dynamic weighted loss based on attention weights is introduced:

$$L_{att} = \sum_{i=1}^{N} \alpha_i \cdot \| y_i - \hat{y}_i \|_2^2 \tag{16}$$

α_i y_i $\hat{y}_i$ Among them, the node importance score is output by the GAT module, and the angular deviation values of the real and predicted are respectively obtained. The formula above will increase the gradients of the high-weight nodes during training and help the model learn more rapidly from the key parts that have larger joint movement amplitudes.

The network structure is a single-layer TCN for capturing local temporal patterns, two layers of GAT for extracting bone topological dependencies, and a Dropout rate of 0.3 is added after the second layer to prevent overfitting. All of the above training was carried out on an RTX 4090 GPU, and it took about 4.6 hours to complete; the accuracy of the validation set reached 92.1%.

5.3 Structural Comparison and Applicability Analysis with other Recognition models

To verify the applicability of the model proposed in this paper to the scenarios of sports training action recognition and feedback, it is compared with and analyzed in relation to two typical structures: the TCN-only model that only uses dilated convolution, and the TCN+LSTM model

that adds a cyclic structure to improve long-term dependency modeling. To assess the overall performance of the three models on the three indices of recognition accuracy, angle-deviation control and response delay, a scoring function is as follows:

$$S = \alpha \cdot A_c + \beta \cdot AS - \gamma \cdot L_a \tag{17}$$

A_c AS L_a α, β, γ Among them, the first is the classification accuracy rate; the second is the joint angle deviation score; the third is the end-to-end response delay; and the fourth is an empirical weight set to 0.4, 0.4, and 0.2 respectively in this experiment to balance recognition ability and real-time feedback. This function can show the discriminative ability and interactive experience of the model at the same time. The comparison results of the three models are shown in Figure 3.

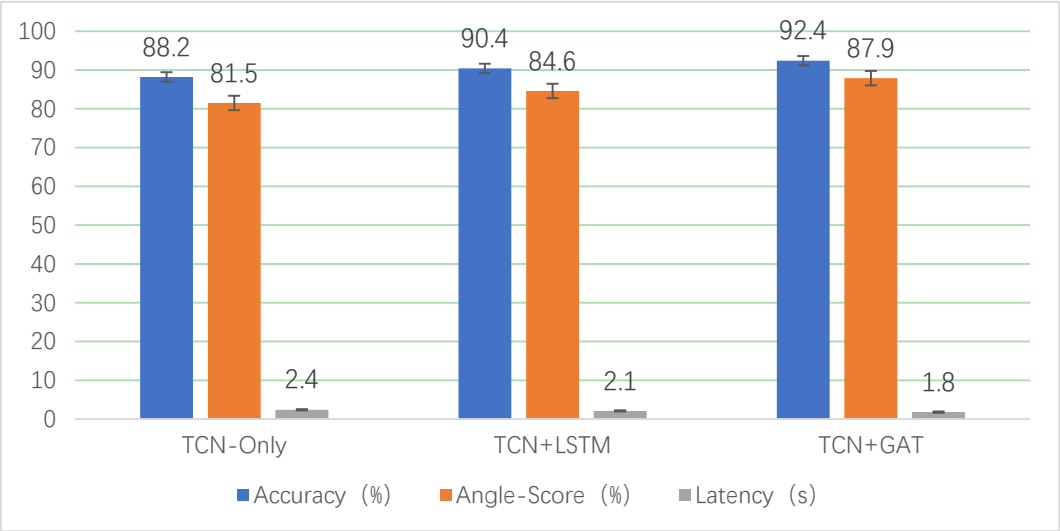

Figure 3: Bar chart of Model Structure Comparison.

Based on the above results, TCN-Only reached an accuracy of 88.2% ± 0.6, an angle-score of 81.5% ± 0.7, and a latency of about 2.4 seconds. TCN+LSTM increased the three indicators to 90.4±0.5, 84.6±0.6 and 2.1s. TCN+GAT in this paper also reached 92.4%±0.5, 87.9%±0.6 and 1.8s. A single layer of convolution is not generally sufficient to address the cross-joint dependency problem. A cycle has been added, but spatial topological models are still missing. A graph attention mechanism is added to the model to weigh the importance of different nodes more heavily, such as knees and elbows, and thus trigger deviation correction instructions more reliably. To further confirm that the results of the three independent experiments were statistically different, a two-sample t-test was performed on them and the results are shown in Table 4.

Table 4: shows the statistical significance test results of the performance comparison for the methods.

Indicator	TCN-Only vs TCN+LSTM	TCN+LSTM vs TCN+GAT	TCN-Only vs TCN+GAT
Accuracy	$p < 0.01$	$p < 0.05$	$p < 0.001$
Angle-Score	$p < 0.01$	$p < 0.05$	$p < 0.001$
Latency	$p < 0.05$	$p < 0.05$	$p < 0.01$

Experiments have shown that the model proposed in this paper performs significantly better than the other two structures in the three key indicators at the 0.05 level or lower, and thus the graph attention structure is required for the modelling of sports training actions and has promotional value. GAT can address the problem of skeleton drift in situations with high-frequency rhythm fluctuations and fine-grained deviation judgment more effectively, and improve the feedback trigger hit rate by 5% or more; it is also more stable and practical.

To further test the structural adaptability of the proposed TCN+GAT framework under different sensing conditions and runtime constraints, four complementary experiments were conducted: IMU binding strategies, temporal window length, multimodal robustness in the face of occlusion, and feedback threshold sensitivity.

As shown in Figure 4, the impact of different IMU mounting methods on recognition accuracy and joint angle consistency is displayed here. When inertial signals are removed and only RGB information is used, the system reaches an accuracy of 89.8% with an Angle-Score of 82.6, and its sensitivity to small changes in posture is relatively weak. Attach the IMU to the wrist or ankle to improve the accuracy of dynamic perception to 91.0% and 90.3%, respectively. The arrangement of the combined wrist-and-ankle sensor achieved the best results and had an accuracy of 92.6% and an angle score of 88.1%. Dense virtual binding at all joints introduces redundant inertial information and is thus slightly less efficient. Based on the above experiments, sparse extremity-based IMU fusion provides reasonable motion cues and does not amplify noise; thus, it is more suitable for continuous training.

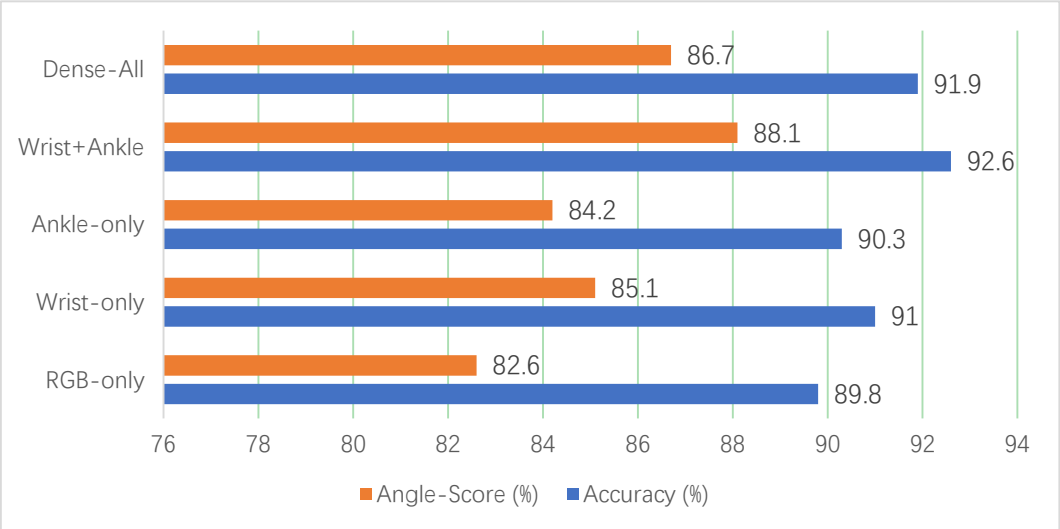

Figure 4: Performance Comparison of Different IMU Binding Strategies.

Figure 5 shows the results of different window lengths (from 32 to 96 frames) on the size of the temporal receptive field. At 32 frames, the recognition accuracy was 90.6%; after 48 frames, it had increased to 92.1%, then fallen slightly at 56 frames, and finally reached a high of 92.8% at 64 frames. With further expansion of the time window, the accuracy is 91.3% at 80 frames, shows a slight increase at 88 frames, and finally drops to 90.9% at 96 frames. At the same time, the end-to-end latency generally increases with an increase in window size from 1.52 s to 2.05 s for an 80-frame window, and then drops slightly at larger lengths. The above fluctuations indicate that an extended time window contains redundant motion patterns and increases inference overhead; thus, a suitable window size needs to be selected for temporal representation. A window size of 64 frames is thus a relatively small one that is both accurate in recognition and fast.

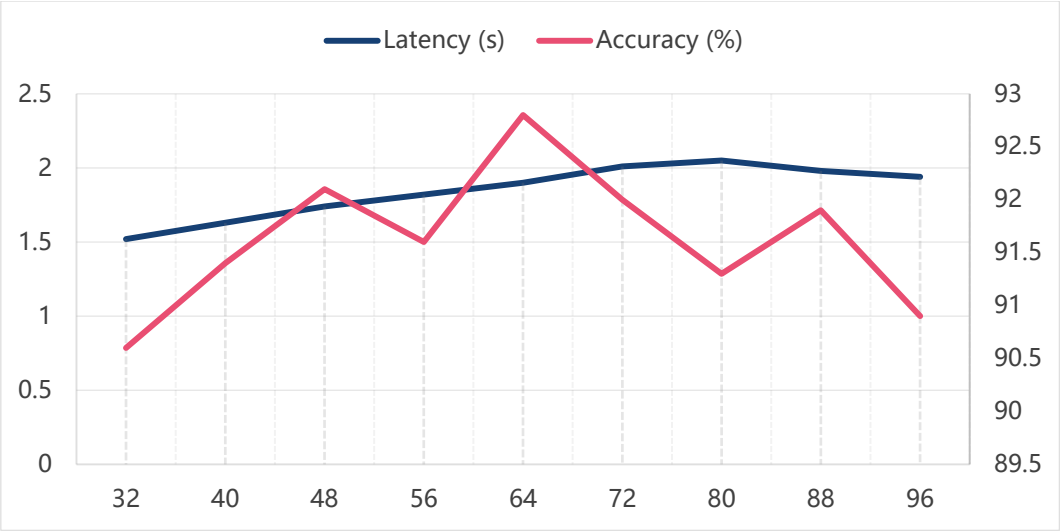

Figure 5: Effect of Temporal Window Length on Recognition Accuracy and Response Latency.

To assess the robustness to visual degradation, Figure 6 shows the results of RGB-only, IMU-only, and RGB+IMU modes at various degrees of occlusion. RGB-only performance drops significantly from 91.6% to 85.7% with an increase in occlusion to 45%, and IMU-only recognition is relatively stable but consistently lower. The multimodal fusion model has a high accuracy under all circumstances and reaches 92.9% with 30% occlusion; it is still above 91% when facing severe occlusion. A small rise in recovery under moderate occlusion can also be considered to show that inertial cues are sufficient to replace the missing visual information; thus, it can be used to address the challenge of difficult-to-train scenarios in cross-modal integration.

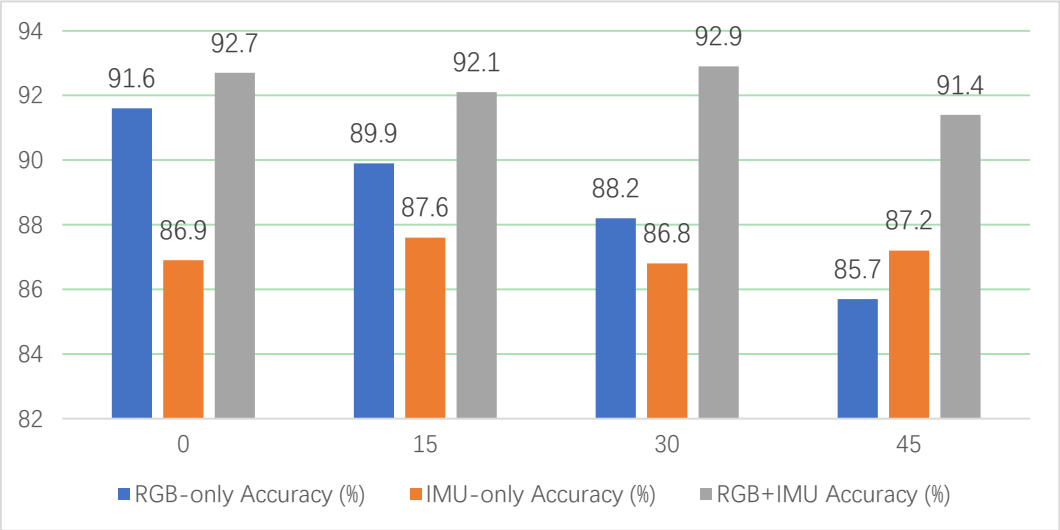

Figure 6: Robustness Comparison under Different Occlusion Levels for Single-Modal and Multi-Modal Inputs.

Figure 7 is the sensitivity of the closed-loop feedback mechanism to joint-angle-deviation thresholds. As the threshold increases from 3 to 7, the correction hit rate is 78.4% and 91.5% respectively, and the false trigger rate decreases to 18.9% and 11.8% respectively. At this time, the hit rate is relatively low and starts to oscillate again, reaching a maximum of 92.1% at a threshold of 11 before falling. At the same time, although the false trigger rate is still falling

generally, it shows a small increase near 10-11. Therefore, a low threshold will be prematurely triggered; otherwise, a high one will fail to respond in time. Based on the above fluctuations, it can be seen that both the optimal range for feedback accuracy and stability is between 7 and 11.

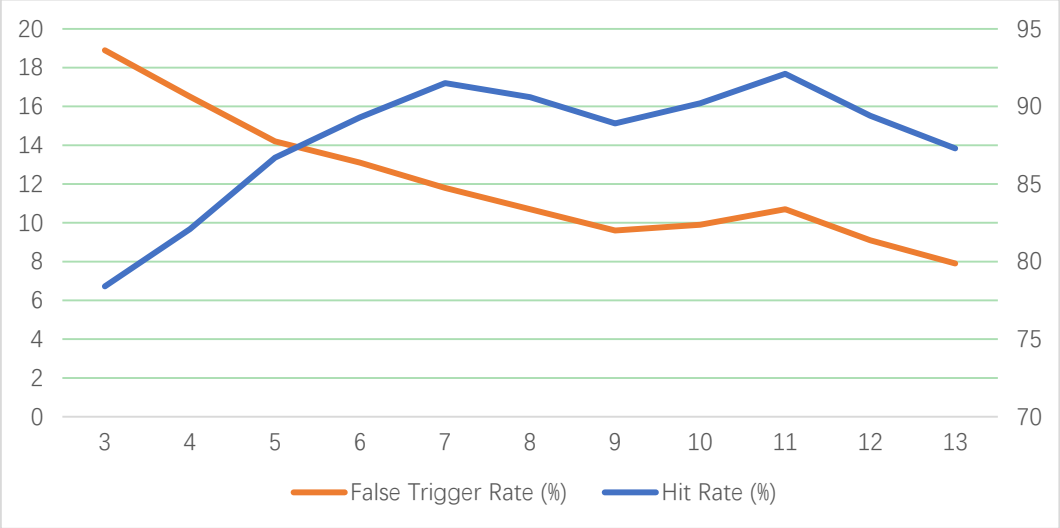

Figure 7: Effect of joint angle deviation threshold on feedback hit rate and false trigger rate.

Thus, we have introduced Graph Attention to improve the model for cross-joint dependency and reduce latency. With sparse IMU fusion and a moderate temporal window, the proposed framework maintains stable recognition and feedback performance in various sensing configurations to provide continuous support for continuous sports training.

5.4 Performance Index Evaluation and Error Visualization

To verify whether the proposed action recognition and feedback system can be used in a real-world training scenario, TCN-ONLY and TCN+LSTM were selected as reference models, representing the temporal convolutional classification structure and the temporal dependent enhancement structure, respectively. Training and testing were carried out on the same skeletal sequence dataset (a total of 4760 samples, with an average length of 64 frames). The main performance indicators for evaluating recognition accuracy, Angle-Score joint deviation score and end-to-end feedback delay comprehensively cover the cooperative operation of the system in terms of recognition stability, deviation correction ability and response real-time performance. The full-featured performance is as follows:

$$P = \frac{Accuracy \times AngleScore}{Latency} \tag{18}$$

Accuracy AngleScore Latency Among them, the first is the accuracy rate of the classification prediction; the second is a measure of the Angle matching degree between the predicted pose and the standard pose; and the third is the total time consumption for computational reasoning and prompt output. A relatively large P-value indicates that the entire system is relatively accurate and responds quickly. The three indicators of recognition accuracy, deviation correction ability and response delay are all within the same scale. A higher value indicates a shorter feedback delay that can be maintained at a high accuracy level and Angle score; thus, it is more suitable for comparing the all-around performance of different structures

in the "identification - feedback" closed-loop. The comparison results of the above models are as follows:

Table 5: Comparison Results of Model Structure and Performance.

Model structure	Accuracy (%)	Angle-Score (%)	Latency (s)
TCN-Only	88.2±0.6	81.5±0.7	2.4
TCN+LSTM	90.4±0.5	84.6±0.6	2.1
TCN+GAT	92.4±0.5	87.9±0.6	1.8

Based on the results, TCN+GAT has achieved the highest values for both Accuracy and Angle-Score, and simultaneously shows a shorter response time in the Latency dimension; thus, it can be seen that the attention mechanism significantly improves the multi-node fusion and keyframe weighting links. Error visualisation was also used to show the fluctuation curves of all the models at different action steps. Tcn-only showed significant displacement jitter in flexion and extension. TCN+LSTM can reduce short-term oscillations but still accumulates a lag in the face of a sudden change in rhythm. However, the angle offset curve of TCN+GAT was generally smoother and had a peak error of less than 5°. There is no persistent overshoot; thus, this structure can keep up with the speed requirements of the training and guidance environment more stably.

To further investigate the closed-loop behaviour of the proposed system during continuous training, three additional experiments were conducted to examine the convergence of the error across training rounds, the stability of latency under repeated interaction, and the consistency of feedback under different motion speeds.

Figure 8 shows the convergence behaviour of joint angle errors over five successive training rounds. As shown in Figure 8, all three models have a declining error with an increase in rounds. TCN-Only reduces the mean joint error to 5.3°, and TCN+LSTM shows a larger reduction from 6.8° to 4.2°. TCN+GAT is significantly faster in convergence, and after the first round, it has dropped from 6.9° to 3.6° in five rounds. Notably, the three curves are relatively well-separated after the third round; therefore, graph-based dependency modelling may be more suitable for accumulating correction information in an iterative training process.

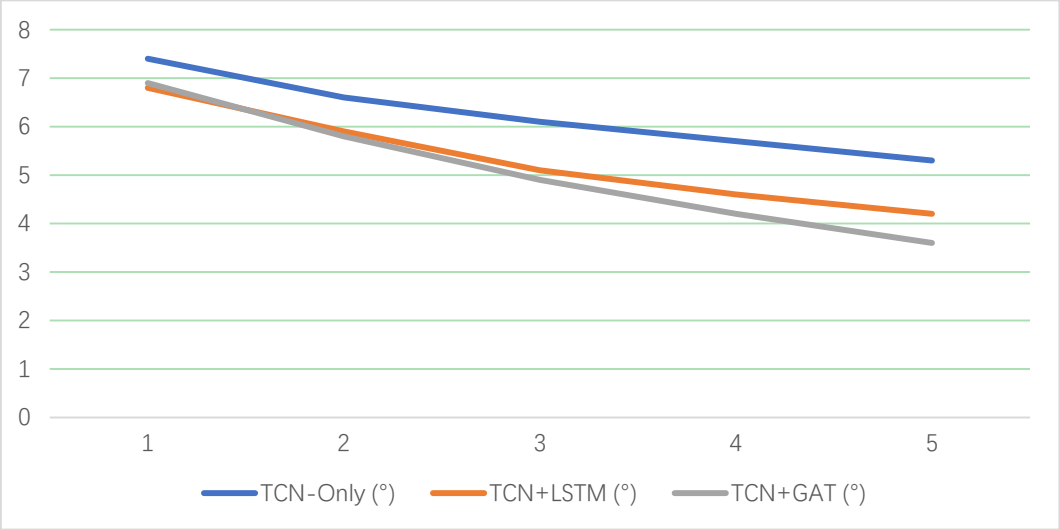

Figure 8: Comparison of mean joint angle error convergence after different numbers of training rounds.

Figure 9 is the variation of average feedback latency in the same five training rounds. TCN-Only has a relatively large fluctuation, decreasing from 2.45 s to 2.31 s in the second round, then rising back to 2.34 s in the third round, and finally fluctuating between 2.27 s and 2.29 s in the last two rounds. TCN+LSTM gradually decreased from 2.18s to 1.97s in the first three rounds, then increased to 2.06s in the fourth round, and fell again to 1.94s. TCN+GAT has a lower response delay than others; it decreased from 1.92s to 1.79s in the first few rounds, slightly increased to 1.85s at the fourth round, and then dropped further to 1.76s in the last round.

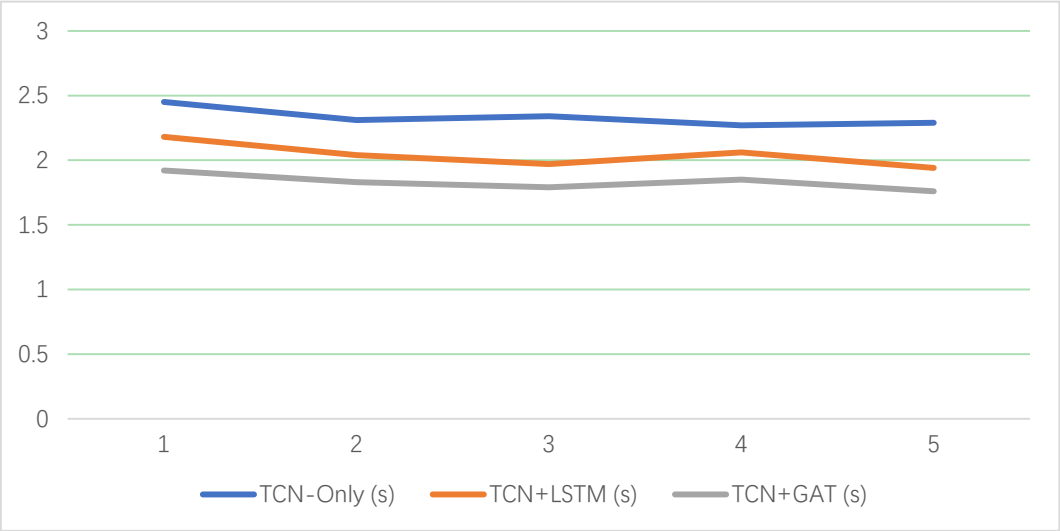

Figure 9: Mean feedback latency variation per training round.

Figure 10 is also divided according to various motion speeds. As the speed of movement increases from slow to fast execution, all models show a drop in performance. TCN-Only drops from 86.2% to 81.9%, and TCN+LSTM drops from 89.1% to 85.2%. TCN+GAT has better stability and, therefore, reaches 91.4% under slow speed, 90.6% at normal speed, and even 89.2% in fast execution.

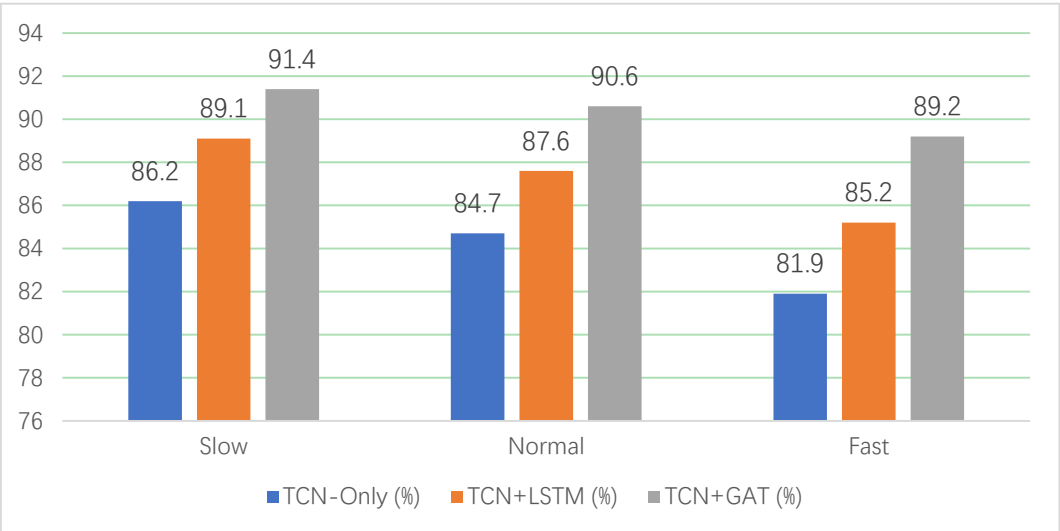

Figure 10: Feedback Hit Rate at Various Speeds of Motion.

The above results show that the proposed TCN+GAT framework can achieve a higher recognition accuracy and a lower response delay, and it is also stable in correction behaviour under continuous training. A faster error convergence rate, limited latency fluctuation, and stable feedback performance at all motion speeds indicate that a graph attention mechanism has improved inter-joint coordination and temporal response stability. Thus, the system can offer reliable deviation correction in repeated training scenarios and maintain the capacity for real-time interaction; otherwise, it would be impractical to apply in sports action guidance and feedback systems.

5.5 Discussion

As shown in the methods listed in Table 1, it can be observed that the TCN+GAT model has better recognition accuracy, reduced delay and correction of deviation compared to the previous pose estimation and time-series model structures. Although PAF-based Pose Estimation is good at multiple-person detection, it is easily affected by trajectory drift in waist-twisting and cross-occlusion actions. IMU-CNN-LSTM has good temporal stability but does not have a general skeleton topology model. FineGym-based TCN is good for classifying single-action segments but does not support continuous feedback. PoseCoach provides the semantic instruction but is not practical. TCN+GAT model is introduced in this paper to enhance multi-joint dependency by using graph attention and addresses both global rhythm and local angle deviations. It is also a feedback trigger strategy that can reliably transmit error correction instructions to the terminal device, and the accuracy is 92.4% $\pm$0.5. The Feedback Delay will be limited to 1.8s. However, the model still needs to synchronize the time of the skeleton data and the IMU sensor, and false triggers may still occur during very high-speed rotation. Event cameras or adaptive synchronisation modules can be added at this time to enhance cross-modal alignment and robustness for actual deployment.

6 Conclusion

The TCN+GAT model presented in this paper is used to address the multiple-input problem of action recognition and immediate feedback in sports training. It integrates the continuous modeling capability of temporal convolution with the key joint aggregation structure of graph attention, and combines angle deviation suppression and feedback triggering mechanisms to achieve a synergistic effect of dynamic recognition and deviation correction stability. Experimental data show that the three indices of accuracy, angle-score and feedback delay are all better than those of other methods. Among them, the recognition accuracy is 92.4% $\pm$ 0.5, the Angle Score is 87.9% $\pm$ 0.6, and the feedback delay is stable at 1.8s or less. Provide a practical, real-time guide output for continuous learning behavior.

However, there are still two shortcomings in the current research: firstly, the acquisition process of the action input is not synchronous with that of the skeleton sequence and the IMU sensor. Data consistency can be assured, but a problem of false triggering in an uncalibrated environment still exists due to time misalignment. Second, in cases of high-speed rotation and complex combinations of degrees of freedom for the model, sudden fluctuations can still occur and disrupt the continuity of the entire correction cycle. Next, a time-domain alignment mechanism for event cameras can be built to replace the traditional frame-based acquisition method and reduce the requirement for synchronisation. Self-supervised pre-training and cross-modal calibration modules can also be added to improve the model's generalization ability in extreme action and multi-person interaction scenarios, and scalable algorithmic support can be

provided for the engineering deployment of intelligent fitness equipment and virtual coaching systems.

Acknowledgements

This work was supported by the Teaching Reform and Research Project of Zhengzhou Academy of Fine Arts: “Research on the Reform and Innovation of Fitness Dance Teaching for Art College Students under the Background of National Fitness” (Grant No. ZMJGLX202418).

References

[1] Wang, J., Qiu, K., Peng, H., et al. (2019). Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. Proceedings of the 27th ACM International Conference on Multimedia, 374–382.

[2] Ingwersen, C. K., Xarles, A., Clapés, A., et al. (2023). Video-based skill assessment for golf: Estimating golf handicap. Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports, 31–39.

[3] Ju, C. Y., Kim, J. H., & Lee, D. H. (2023). GolfMate: Enhanced golf swing analysis tool through pose refinement network and explainable golf swing embedding for self-training. Applied Sciences, 13(20), 11227.

[4] Sideridou, M., Kouidi, E., Hatzitaki, V., et al. (2024). Towards automating personal exercise assessment and guidance with affordable mobile technology. Sensors, 24(7), 2037.

[5] Zhang, S., Li, Y., Zhang, S., et al. (2022). Deep learning in human activity recognition with wearable sensors: A review on advances. Sensors, 22(4), 1476.

[6] Wang, J., Chen, Y., Hao, S., et al. (2019). Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters, 119, 3–11.

[7] Cao, Z., Simon, T., Wei, S. E., et al. (2017). Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7291–7299.

[8] Shao, D., Zhao, Y., Dai, B., et al. (2020). Finegym: A hierarchical video dataset for fine-grained action understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2616–2625.

[9] Tharatipyakul, A., Srikaewsiew, T., & Pongnumkul, S. (2024). Deep learning-based human body pose estimation in providing feedback for physical movement: A review. Heliyon, 10(17), e36589.

[10] Liu, J., Saquib, N., Zhutian, C., et al. (2022). Posecoach: A customizable analysis and visualization system for video-based running coaching. IEEE Transactions on Visualization and Computer Graphics, 30(7), 3180–3195.

[11] Mennella, C., Maniscalco, U., De Pietro, G., et al. (2023). A deep learning system to

monitor and assess rehabilitation exercises in home-based remote and unsupervised conditions. Computers in Biology and Medicine, 166, 107485.

[12] Stoeve, M., Schuldhaus, D., Gamp, A., et al. (2021). From the laboratory to the field: IMU-based shot and pass detection in football training and game scenarios using deep learning. Sensors, 21(9), 3071.

[13] Liao, Y., Vakanski, A., & Xian, M. (2020). A deep learning framework for assessing physical rehabilitation exercises. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(2), 468–477.

[14] Liao, C. C., Hwang, D. H., & Koike, H. (2021). How can I swing like pro?: Golf swing analysis tool for self training. SIGGRAPH Asia 2021 Posters, 1–2.

[15] Pardos, A., Tziomaka, M., Menychtas, A., et al. (2022). Automated posture analysis for the assessment of sports exercises. Proceedings of the 12th Hellenic Conference on Artificial Intelligence, 1–9.

[16] Gupta, S. (2021). Deep learning based human activity recognition (HAR) using wearable sensor data. International Journal of Information Management Data Insights, 1(2), 100046.

[17] Shaikh, M. B., & Chai, D. (2021). RGB-D data-based action recognition: A review. Sensors, 21(12), 4246.

[18] Kim, S. E., Burket Koltsov, J. C., Richards, A. W., et al. (2023). Validation of inertial measurement units for analyzing golf swing rotational biomechanics. Sensors, 23(20), 8433.