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

Dynamic posture evaluation system for athletes based on computer vision and sensor fusion

Yue Wang[1], *Meiwei Zhou*[2] *and Shujin Liu*[3,*]

[1] Emilio Aguinaldo College, 1123, Philippines
[2] Harbin Sport University,Harbin 150006, Heilongjiang , China
[3] Guangxi College for Preschool Education, Nanning, 530022,Guangxi ,China

SUMMARY: *With the development of computer vision and multi-modal perception, the dynamic posture assessment of athletes has become the research content in the field of intelligent sports analysis. In order to build a posture assessment system, this paper designs a deep vision detection module to locate athletes and extract skeleton key points in competition videos. Then, the inertial measurement, plantar pressure and joint Angle signals were synchronized by the time alignment strategy, and fused by the gated attention fusion network. The system is trained with 38640 synchronized samples from 86 athletes in sprint, basketball, football and combat training. Experimental results show that the visual branch of FusionNet is stable in complex scenes, the AP of clear scenes is 97.2%, and the AP of crowded scenes is 86.4%. Compared with the single-modal recognition method, the fusion model achieves a better classification effect, with a recall rate of 94.87%, a precision rate of 95.31%, and a F1 value of 95.09%. The overall posture evaluation accuracy reaches 96.18%, and the average inference delay remains at 31.6 ms, which supports real-time athlete evaluation in the deployment phase.*

KEYWORDS: *Computer vision; Sensor fusion; Dynamic pose estimation; Multimodal time series modeling*

1 Introduction

Dynamic posture assessment of athletes is a typical task in intelligent sports analysis and human-computer interaction perception, which involves the calculation process of video target localization, skeleton key point extraction, inertial signal synchronization, pressure data analysis and time state discrimination. With the development of deep vision networks and wearable sensors, motor performance no longer relies only on manual observation records, but can be jointly described by image sequences, 3D pose, acceleration, angular velocity, and plantar pressure. Uhlrich et al. constructed a framework for analyzing human motion dynamics based on smartphone video and proved that low-cost video acquisition can also form a computable kinematic data link [1]. Cronin et al. verified the feasibility of OpenPose markerless motion analysis in real track and field competition scenes, indicating that visual pose estimation can enter the motion measurement process in open environments [2]. Fukushima et al. conducted verification research on sports motion capture and pointed out that human pose estimation has application potential in kinematic parameter reconstruction and motion process analysis [3]. These studies provide a visual computing basis for dynamic

*yue.wang.mnl@eac.edu.ph
https://doi.org/10.65102/is2026966
Published: 30 April, 2026

posture assessment of athletes, and also show that non-contact posture recognition in competition scenes has experimental support.

Only relying on the human skeleton in the video frame cannot completely express the body control state in high-speed motion. Jumping, stopping, turning, sprinting and confrontation will produce obvious inertial fluctuations. Image occlusion, illumination change and perspective shift will also affect the stability of key points. Aleksic et al. compared markerless pose estimation and 3D marker-based motion capture systems in automated reverse jump analysis, and the results showed that the visual method had good measurement consistency in vertical jump analysis [4]. Milone et al. proposed a neural network motion capture method enhanced by DeepLabCut to improve the stability in markerless pose analysis, which reflects the ability of deep networks to correct keypoint jitter and pose drift [5]. In the direction of sensor Fusion, Kim and Lee proposed the Fusion Poser method, which uses sparse IMU and head tracker to realize real-time 3D human pose estimation, providing a technical path for dynamic pose recovery with a small number of sensor nodes [6]. Amadi and Agam proposed PosturePose, which uses semi-supervised monocular 3D pose estimation to complete pose analysis, further demonstrating that vision models can be combined with pose constraints, sample augmentation, and spatial inference [7].

The dynamic posture evaluation system of athletes needs to process the visual model and sensor data in the same time series framework. Rajendran and Sethuraman reviewed yoga pose recognition research and emphasized the influence of pose categories, key point structure and action stage division on recognition results [8]. Avogaro et al. summarized the method system of mark-free human pose estimation in biomedical applications, indicating that deep learning pose models have formed a relatively complete technical route in complex human motion analysis [9]. Roggio et al. analyzed machine learning pose estimation models for human motion and pose analysis and pointed out that different models had obvious differences in accuracy, stability and applicable scenarios [10]. These results show that dynamic pose evaluation cannot stay at the level of single frame pose recognition, but should unify video detection, skeleton sequence, inertial variation, pressure distribution and classification scoring into a reviewable system process.

Based on this, this paper constructs a dynamic posture evaluation system for athletes based on computer vision and sensor fusion. Firstly, the system uses the deep vision network to complete the athlete target detection and motion trajectory tracking, and then extracts the key points of the shoulder, elbow, hip, knee, ankle and other skeletons to form a continuous posture sequence. Then, the IMU, plantar pressure and joint Angle sensor data were connected, and the multi-source feature alignment was completed by timestamp calibration, sliding window segmentation and gated attention fusion. Finally, the time series classification network was used to output the posture category, action stability, joint offset and comprehensive evaluation results. At the data level, the system uses video frames, inertial measurement, pressure sampling and manual labeling to form sample units together to avoid mismatching of action stages caused by a single modality. At the algorithm level, the visual branch is responsible for spatial structure localization, the sensing branch is responsible for capturing fast acceleration/deceleration and support transitions, and the fusion layer adjusts the contribution of different features according to temporal consistency and confidence weights. At the evaluation level, the system not only outputs the action category, but also gives the joint stability, posture offset amplitude and abnormal segment index, so that the evaluation results can correspond to specific frame segments and sensing waveforms. This structure enables posture assessment to have a clear data source and calculation basis, and can maintain a consistent processing logic in the training field, laboratory and mobile device acquisition environment. At the same time, cross-item action samples are introduced into

model training, which is convenient to test the generalization ability and real-time inference stability under different motion types. And retain the manual review data interface.

The content of this paper is divided into five parts: introduction explains the research background and technical basis; Related work combs visual pose estimation and sensor fusion methods. The system constructs part of the design object detection, key point extraction and multi-source fusion model. In the performance evaluation part, the detection accuracy, trajectory error, recognition accuracy, inference delay and robustness tests are used to verify the performance of the system. The conclusion section summarizes the experimental results and the application boundaries of the system.

2 Related work

The dynamic posture assessment of athletes belongs to the directions of computer vision, wearable sensing, time-series signal analysis and multi-modal fusion. The visual end is responsible for human detection, skeleton key point extraction and trajectory tracking. The sensor end mainly records acceleration, angular velocity, plantar pressure and joint Angle changes, and the two need to complete feature alignment in a unified time axis. Tharatipyakul et al. studied the application of human pose estimation based on deep learning in body motion feedback, and summarized the processes of 2D skeleton extraction, 3D pose recovery, and action deviation suggestion, indicating that the pose estimation model has been able to shift from pure recognition to a computational framework for action quality feedback [11]. Ghosh et al. conducted a review from the perspective of artificial intelligence applications, cutting-edge technologies and algorithms in sports analysis, and pointed out that visual recognition, predictive modeling, data mining and real-time computing formed a technical link in sports performance analysis [12]. Naik et al. summarized the research progress of computer vision in the field of sports, involving target detection, action recognition, motion trajectory analysis and game scene understanding, which provided a visual algorithm basis for dynamic posture evaluation of athletes [13]. Host and Ivašić-Kos studied the method of sports action recognition based on computer vision, analyzed the relationship between video input, feature representation, action classification and deep models, and showed that the human body region, skeleton structure and time series changes should be considered for action recognition in sports scenes [14].

In multimodal activity recognition, a single visual model is easily affected by occlusion, illumination and viewpoint changes, and a single sensor model is difficult to express the complete spatial posture structure. Islam et al. proposed a multi-level feature fusion method to integrate local features, temporal features and high-level semantics hierarchically-to improve feature complementarity in complex activity recognition [15]. Akter et al. studied the human training method based on the combination of whole body wearable Internet of things sensors and machine learning, and used multi-node sensor data to describe the body movement process, which reflected the data value of sensor networks in action monitoring and training feedback [16]. Khan et al. proposed a hybrid deep learning model for human activity recognition, which combines the convolutional structure with the time series network to extract local changes and continuous state features from sensing data, and is suitable for continuously sampled signals such as acceleration and gyroscope [17]. Tanigaki et al. studied the method of predicting the performance improvement of activity recognition model through supplementary data collection, and showed that sample size, action coverage and data distribution would affect the stability of the model in new scenarios [18].

In terms of spatial localization and motion scene measurement, the joint use of visual pose

estimation and sensor signals is more conducive to identity tracking and pose recovery in complex environments. De Marchi et al. proposed to combine 3D human pose estimation with IMU sensor for human recognition and tracking in multi-person environment. This method shows that visual skeleton and inertial measurement can complement each other in occlusion and interactive scenes [19]. Merker et al. compared the position measurement accuracy of HTC VIVE Tracker 3.0 and Vicon system, which provided a measurement basis for the effectiveness of position feedback in virtual reality environment, and also provided a reference for the error calibration of positioning equipment in dynamic attitude assessment system [20]. Russomanno et al. studied the sports position detection method based on UAV and completed the verification, indicating that the mobile visual acquisition platform can expand the spatial scope of athlete position recognition [21].

To further sort out the correspondence between the existing research and the proposed system design, Table 1 summarizes three aspects: research direction, main methods, and implications for dynamic posture assessment systems. This table not only presents the technical connection of visual pose estimation, sports action recognition and sensor activity recognition, but also illustrates the supporting role of related research on the vision branch, sensing branch and fusion evaluation module of this paper.

Table 1: Correspondence between related studies and the proposed system design

Literature	Research Direction	Main Method	Implication for Dynamic Posture Assessment System
[11]	Human posture feedback	Deep learning-based pose estimation	Supports action quality feedback modeling
[12]	Intelligent sports analytics	AI and data mining	Supports the computational framework for athletic performance data
[13]	Computer vision in sports	Object detection and action recognition	Supports video-based posture analysis
[14]	Sports action recognition	Video features and classification models	Supports temporal action discrimination
[15]	Multimodal activity recognition	Multi-level feature fusion	Supports visual and sensor feature fusion
[16]	Wearable training monitoring	IoT sensors and machine learning	Supports full-body motion data acquisition
[17]	Sensor-based activity recognition	Hybrid deep learning model	Supports inertial signal temporal modeling
[18]	Data acquisition evaluation	Incremental data prediction	Supports sample expansion and generalization analysis
[19]	Pose and IMU fusion	3D pose estimation and IMU	Supports tracking in occluded scenarios
[20]	Position feedback measurement	VR positioning device accuracy validation	Supports positioning error calibration
[21]	Sports position detection	Drone-based visual positioning	Supports open-space trajectory acquisition

In summary, existing researches cover visual pose estimation, sports action recognition, sensor activity recognition and position measurement, but most of them focus on the performance verification of a single task or single modality, and there is still a lack of unified

modeling for multi-source time alignment, skeleton trajectory correction, inertial signal fusion and systematic output in athlete dynamic pose evaluation. In the vision branch, the target detection, skeleton key point extraction and motion trajectory tracking are completed. In the sensing branch, the IMU, plantar pressure and joint Angle signals are analyzed. The system integrates the spatial structure of the image and the change of the sensing time sequence into the same evaluation process, so that the posture recognition results have a clear data source and review path.

3 The construction of dynamic posture evaluation system for athletes based on the fusion of visual perception and multi-source sensing

3.1 Athlete target detection and skeleton key point extraction method based on deep vision Network

In the dynamic posture evaluation system of athletes, the vision branch mainly assumes the tasks of athlete target location, skeleton key point extraction and trajectory sequence generation. The original video frames are entered into the deep vision network after scale normalization, noise suppression and frame sequence calibration. The model needs to maintain stable detection results in fast displacement, limb occlusion and multi-person overlapping scenes. In order to ensure a reliable visual basis for subsequent pose assessment, we incorporate object detection, keypoint heatmap generation, skeleton connection decoding and timing correction into the same calculation process, and form traceable and reviewable visual pose feature output.

To illustrate the feature flow relationship after the video frame enters the visual network, Fig. 1 shows the processing link from the original motion picture to the skeleton sequence output. In this flow, image preprocessing, athlete detection, key point heat map generation and skeleton timing cache are put into the same calculation path, which is convenient for subsequent time alignment with inertial sensing and pressure sensing data.

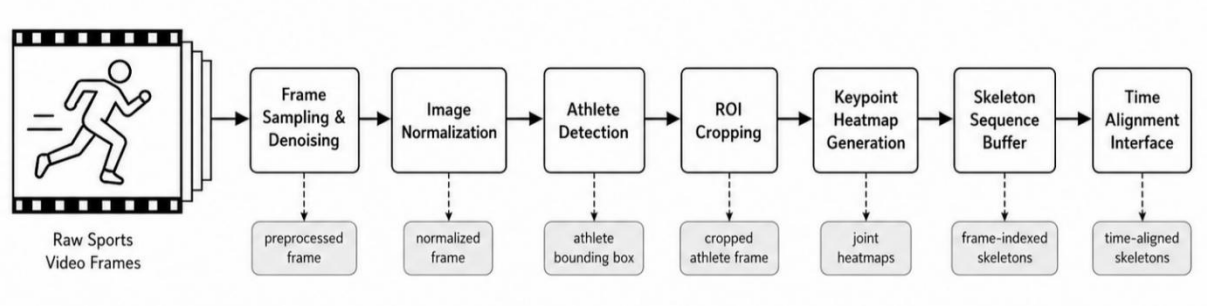

Figure 1: Process of athlete visual object detection and skeleton keypoint extraction

In order to describe the response relationship of video frames of different scales in the convolutional coding layer, local texture, limb edge and central region of human body are uniformly mapped into multi-layer feature representation. The calculation process is shown as follows:

$$F_t^l = \delta\left(BN(W_l * D_l(I_t) + b_l)\right), \quad l = 1,2,3,4 \tag{1}$$

Here, I_t represents the t frame motion image, $D_l(\cdot)$ represents the l layer

downsampling transformation, $W_l$and b_l represent the convolution kernel weight and bias term, $BN(\cdot)$ represents the batch normalization operation, $\delta(\cdot)$ represents the nonlinear activation function, and F_t^l represents the visual feature map output by this layer. The function of this formula is to compress the athlete edge, clothing texture, limb contour and background separation information at different resolutions into a learnable feature space, avoiding the direct dependence of object detection on a single scale image response.

The deep vision network adopts a hierarchical coding method, which retains the boundaries of athletes and joint edges at the low level, and expresses the semantics of the overall human body area and action posture at the high level. Fig. 2 shows the internal structure of the detection network. The input frame is gradually compressed in the four-level coding layer, and the shallow edge information is returned in the feature fusion layer. Finally, the target center heat map, scale regression map and human region confidence map are output.

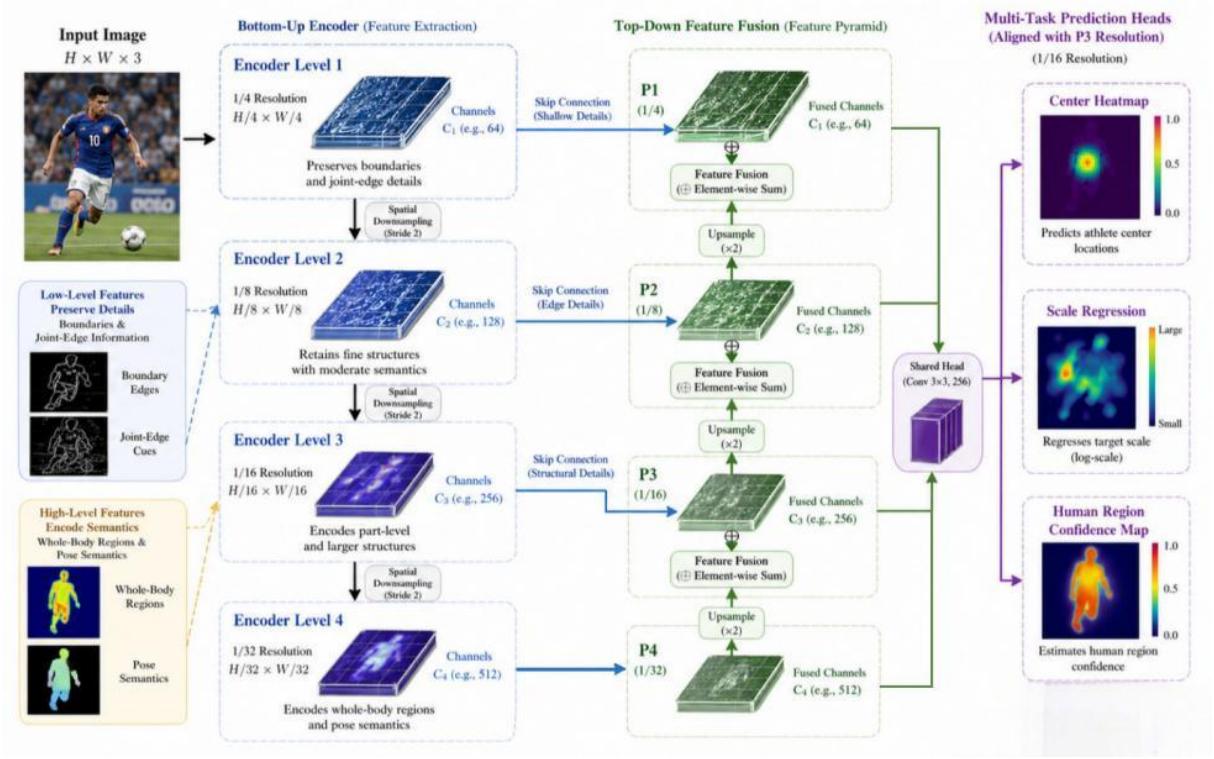

Figure 2: Network structure for multi-scale athlete object detection

In order to obtain the center position of the athlete target and reduce the dependence on anchor box parameters, the network represents the human body center response in the form of heat map. The closer the center point is to the real human body area, the higher the response value is. The calculation process is as follows:

$$H_t(x,y) = \max_{n \in \Omega_t} \exp\left[-\frac{(x-x_n)^2 + (y-y_n)^2}{2\sigma_n^2}\right] \tag{2}$$

Here, Ω_t represents the set of athlete targets in frame t,(x_n, y_n) represents the true center coordinates of the n athlete, σ_n represents the Gaussian spread coefficient related to the scale of the human frame, and $H_t(x,y)$ represents the center response value at pixel position (x,y). This formula replaces discrete candidate boxes by continuous heat maps, so that the detection branch can identify the central region in multi-player sports scenes, and

provide a coordinate basis for cross-frame trajectory association of the same athlete.

After the target region is determined, the skeleton key point extraction branch generates multiple joint heatmaps according to the human body region features. In order to ensure the distinguibility of shoulder, elbow, hip, knee, ankle and other nodes in rapid motion, this paper superposes the target region features and context features, and the key point response is calculated as follows:

$$P_{k,t}(u,v) = \mathrm{sigmoid}(W_k * \Psi(F_t^1, F_t^2, F_t^3, F_t^4))_{u,v} \tag{3}$$

Here, $P_{k,t}(u,v)$ represents the response probability of the k skeleton keypoint at position (u,v), $\Psi(\cdot)$ represents the cross-layer feature aggregation function, W_k represents the prediction convolution kernel of the corresponding joint point, and $\mathrm{sigmoid}(\cdot)$ is used to map the response value to the probability interval. This formula combines the overall semantics of the human body with the local joint texture, so that the adjacent limbs can still maintain a clear node response when they overlap.

Fig. 3 is used to show how the keypoint heatmap is decoded to the skeleton structure. Heatmaps of each joint are shown on the left, coordinate regression and connection screening are shown in the middle, and athlete skeleton sequence is shown on the right. The structure not only outputs the coordinates of joint points, but also records the joint confidence, limb connection relationship and frame-level skeleton number, which provides fine-grained visual evidence for posture stability calculation.

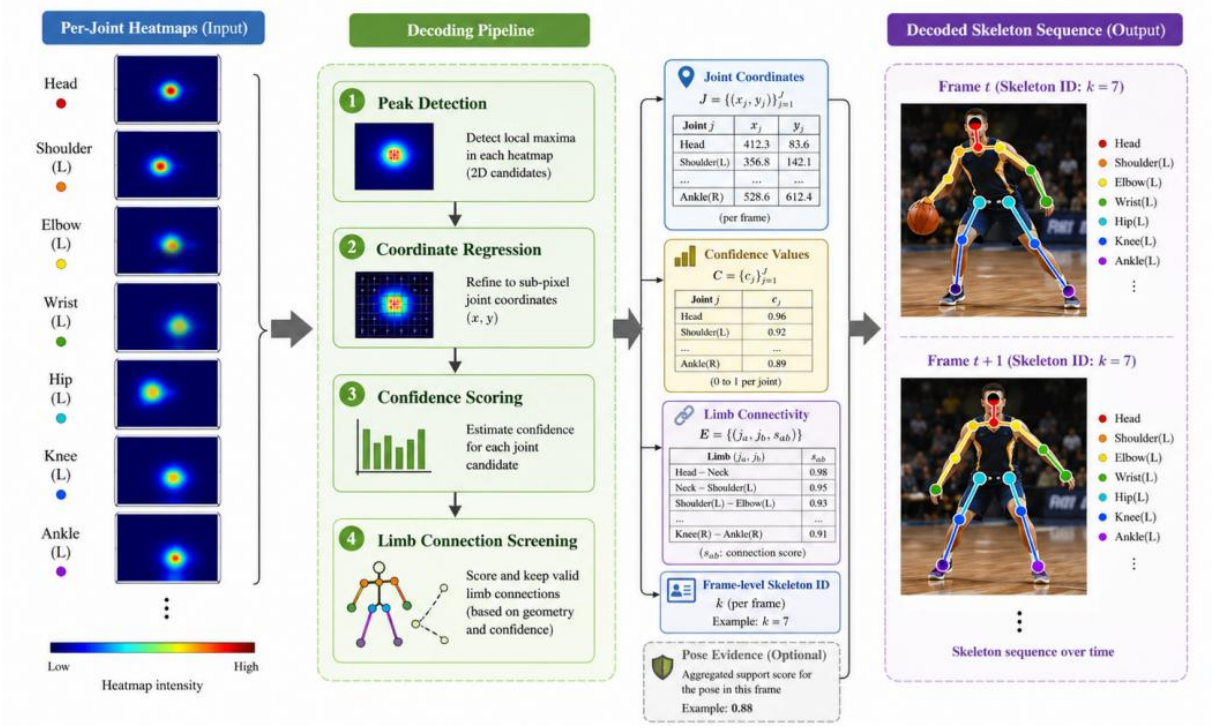

Figure 3: Keypoint heatmap generation with skeleton connection decoding structure

In order to reduce the coordinate offset caused by the discretization of the heat map, the coordinates of the key points are determined by the response weighted coordinates and the local offset. The network calculates the sub-pixel position near each candidate peak, and the calculation process is shown in the following equation:

$$\hat{q}_{k,t} = \frac{\sum_{(u,v)\in R_k} P_{k,t}(u,v)[u,v]}{\sum_{(u,v)\in R_k} P_{k,t}(u,v)} + O_{k,t} \tag{4}$$

Here, $\hat{q}_{k,t}$ represents the predicted coordinate of the k joint point in the t frame, R_k represents the keypoint candidate region, $P_{k,t}(u,v)$ represents the response probability in the candidate region, and $O_{k,t}$ represents the local offset regressors. This formula combines the probability distribution of the heat map with the offset correction, which can alleviate the error caused by the low-resolution heat map for the localization of small-scale joints such as elbow, wrist and ankle.

In order to convert discrete joint points into skeleton graphs that can be used for dynamic posture analysis, the limb connection features are established according to the human motion topology. The spatial distance, direction Angle and confidence weight between the nodes together constitute the edge features, which are defined as follows:

$$E_{i,j,t} = \left[\hat{q}_{i,t}, \hat{q}_{j,t}, \left\|\hat{q}_{i,t} - \hat{q}_{j,t}\right\|_2, \arctan\frac{y_{j,t} - y_{i,t}}{x_{j,t} - x_{i,t}}, c_{i,t}c_{j,t}\right] \tag{5}$$

Here, $E_{i,j,t}$ represents the skeleton edge feature between the i and j joint point in the t frame; $\hat{q}_{i,t}$ and $\hat{q}_{j,t}$ represent the two joint point coordinates; $c_{i,t}$ and $c_{j,t}$ represent the node confidence. This formula extends the human skeleton from a point set to a graph structure with direction and weight, so that the upper limb swing, lower limb support and torso tilt can enter the same pose representation.

Due to the existence of partial occlusion and short-term missed detection in motion video, the keypoint results of a single frame need to be corrected for temporal consistency. Fig. 4 shows the process of cross-frame skeleton tracking and node correction. The system establishes a matching relationship according to the human body center trajectory, joint confidence and the displacement of adjacent frames, and smooths and compensates the low confidence nodes.

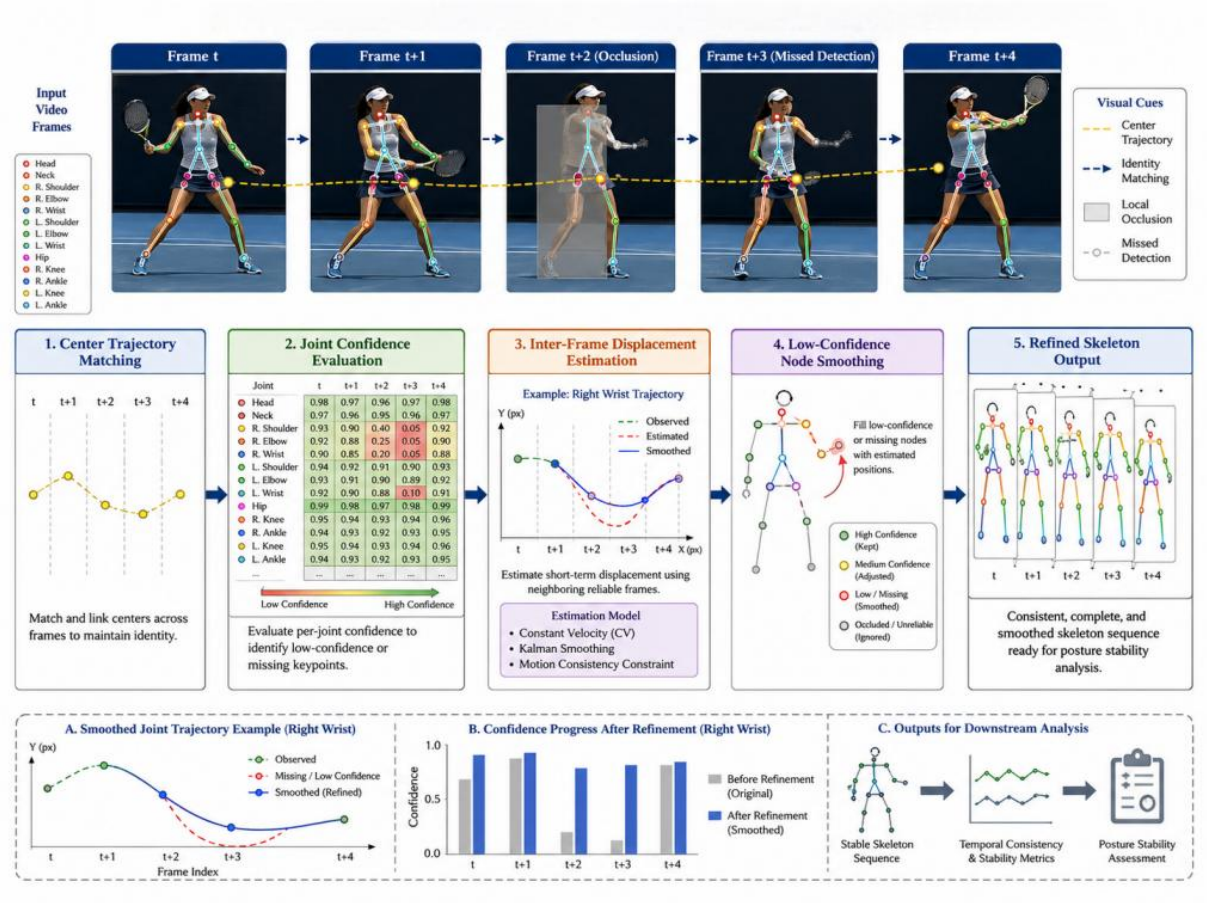

Figure 4: Cross-frame skeleton tracking with keypoint timing correction process

In order to maintain the continuity of the trajectory of the joint points during the motion, this paper introduces a time series smoothing mechanism with confidence constraints to fuse the predicted coordinates of the current frame and the historical state. The calculation process is shown as follows:

$$\tilde{q}_{k,t} = \lambda c_{k,t} \hat{q}_{k,t} + (1 - \lambda c_{k,t}) \tilde{q}_{k,t-1} \tag{6}$$

Here, $\tilde{q}_{k,t}$ represents the corrected joint point coordinates, $\hat{q}_{k,t}$ represents the predicted coordinates of the current frame, $\tilde{q}_{k,t-1}$ represents the corrected coordinates of the previous frame, $c_{k,t}$ represents the confidence of the current node, and λ represents the smoothing coefficient. This formula makes the high confidence nodes retain the current visual observation more, and makes the low confidence nodes inherit the historical trajectory more, so as to reduce the impact of occlusion and transient blur on the skeleton sequence.

In order to output reliable visual evidence to the subsequent multi-source sensor fusion module, the system calculates visual credibility for each frame of skeleton result, which integrates node confidence, trajectory continuity and structure deviation amplitude. The definition process is as follows:

$$R_t = \frac{1}{K} \sum_{k=1}^{K} c_{k,t} \exp\left(- \frac{\left\| \hat{q}_{k,t} - \tilde{q}_{k,t} \right\|_2^2}{\eta} \right) \tag{7}$$

Here, R_t represents the visual skeleton credibility of frame t, K represents the number of skeleton keypoints, and η represents the trajectory offset scaling coefficient. This equation is used to screen the visual frame segments that can enter the fusion model. When the confidence of the skeleton point is low or the trajectory deviation is too large, the system reduces the visual weight of this frame, and the inertial and pressure sensing features are supplemented in the subsequent attitude evaluation.

After multi-scale feature extraction, center point detection, key point regression, skeleton connection and timing correction, the visual branch can form a more complete representation of the athlete's posture. The detection frame is used to limit the area of the human body, the center trajectory is used to describe the motion displacement, the skeleton node is used to represent the limb structure, the edge feature is used to describe the spatial relationship between the joints, and the frame-level credibility is used to judge the stability of the visual result. This output form avoids the jitter error caused by the single frame detection results directly participating in the pose evaluation, and also enables the visual information to participate in the subsequent calculation as structured data. For dynamic actions such as sprinting, jumping, turning and landing, the skeleton sequence can record the changes of body center of gravity, limb swing direction and joint Angle, and the target trajectory can reflect the continuous movement state of the athlete in space. The visual branch is no longer just to complete the picture recognition, but to convert the video content into computable pose data, which provides a stable data basis for dynamic pose assessment.

3.2 Dynamic posture assessment model construction and system design based on multi-source sensor fusion

After the visual skeleton sequence is formed, the system continues to access inertial measurements, plantar pressure, and joint Angle sensing data. The acquisition frequency of the sensing end is different from the video frame rate, and the original waveform must undergo time calibration, noise filtering and segment segmentation before entering the same pose evaluation process with the visual features. In this paper, multi-source sensing data are represented as continuous time Windows, each corresponding to a segment of motion action, and bound with skeleton coordinates, center trajectory and visual credibility in the same time range to form a sample unit for pose evaluation.

To illustrate the computational relationship of multi-source data after entering the evaluation model, Fig. 5 shows the fusion links of visual skeleton, IMU, pressure array, and joint Angle data. All kinds of signals first generate timestamps at the acquisition end, and then enter the synchronization layer, encoding layer and fusion layer, and finally output the pose category, stability score and abnormal segment index.

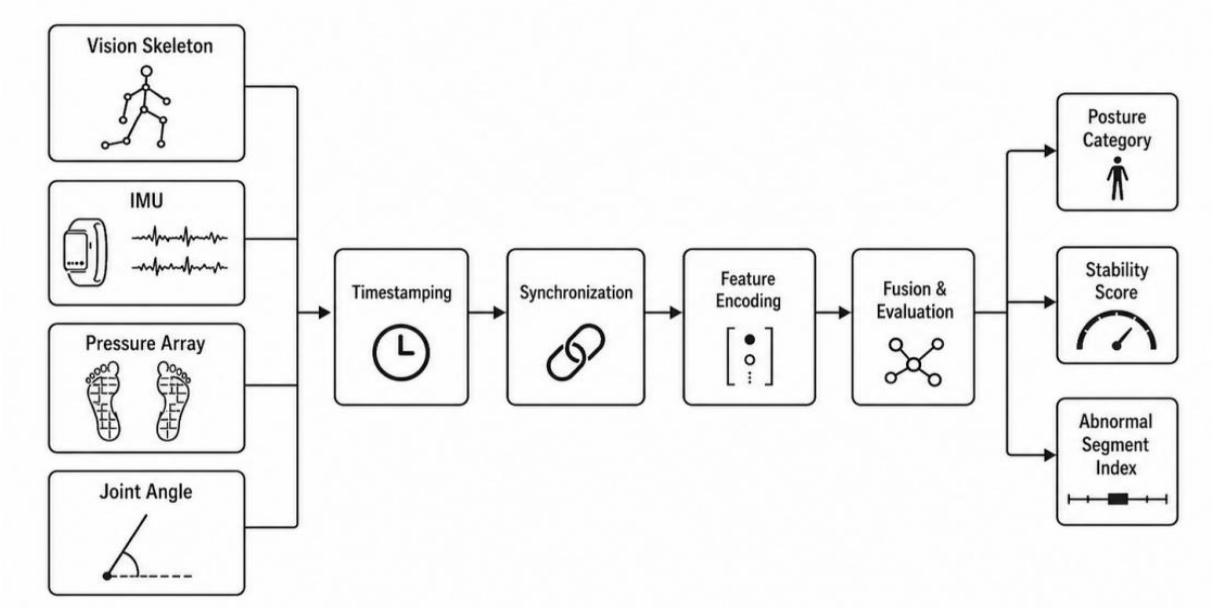

Figure 5: Process of multi-source sensing data fusion and pose assessment

In order to ensure that the sampling sequences of different devices can be mapped to the same motion window, the system establishes the sensor resampling index based on the video time axis, and uses the interpolation function to correct the offset caused by asynchronous acquisition. The time alignment process is shown in the following equation:

$$\bar{s}_{m,t} = \sum_{r=1}^{R_m} s_{m,r} \cdot \max\left(0,1 - \frac{|\tau_t - \tau_{m,r}|}{\Delta_m}\right) \tag{8}$$

Here, $\bar{s}_{m,t}$ represents the alignment signal of the m sensor at video frame t, $s_{m,r}$ represents the original sensing sampling value, τ_t and $\tau_{m,r}$ represent the video frame time and sensor sampling time, respectively, and Δ_m represents the interpolation window allowed by this sensor. This equation maps the discrete sensing waveform into a sequence of video frames, so that the visual skeleton and the sensing signal have consistent timing coordinates.

In order to weaken the influence of different sensor dimension differences on the fusion results, the system performs robust standardization on the acceleration, angular velocity, pressure and Angle signals, and retains the fluctuation amplitude within the window. The normalization calculation process is shown in the following equation:

$$z_{m,t} = \frac{\bar{s}_{m,t} - \text{median}(\bar{s}_m)}{\text{IQR}(\bar{s}_m) + \varepsilon} \tag{9}$$

Here, $z_{m,t}$ represents the normalized sensing value, $\text{median}(\bar{s}_m)$ represents the median of the m class sensor sequence,$\text{IQR}(\bar{s}_m)$ represents the interquartile range, and ε is the stability term that prevents the denominator from being zero. This formula is suitable for processing motion data with large differences in motion amplitude, and can reduce the interference caused by individual body shape, sensor wearing position and instantaneous impact peak.

Fig. 6 shows the internal structure of the dynamic pose evaluation model. The visual encoder extracts the skeleton graph features, the sensor encoder extracts the multi-channel temporal features, the gated fusion layer assigns the modal weights according to the

confidence, and the temporal discrimination layer outputs the action stages and pose scores.

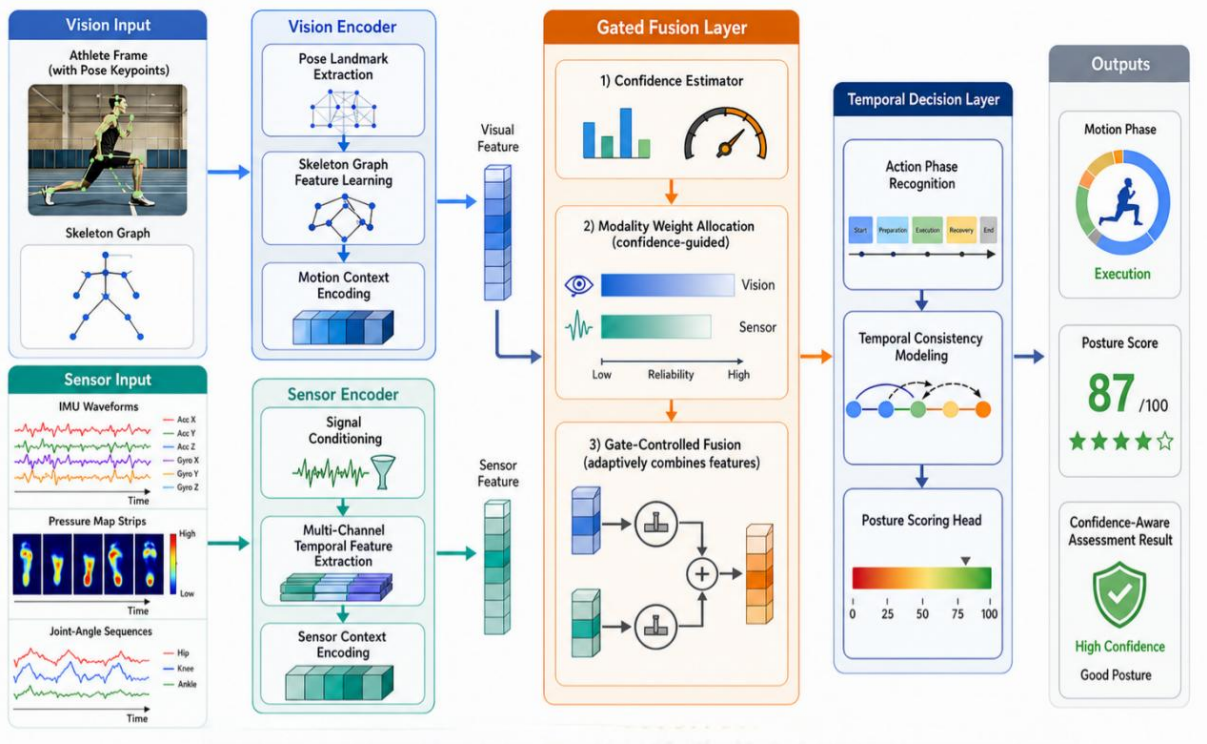

Figure 6: Network structure for vision-sensor gated fusion evaluation

In order to extract local motion changes from the aligned sensing sequence, in this paper, one-dimensional dilated convolution is used to encode acceleration and pressure fluctuations at different time scales, which is calculated as shown in the following equation：

$$g_{m,t}^{d} = \sigma\left(\sum_{i=0}^{L-1} K_{m,i}^{d}\, z_{m,t-d\cdot i} + b_{m}^{d}\right) \tag{10}$$

Where $g_{m,t}^{d}$ represent the timing characteristics of the m sensor under the expansion rate d, $K_{m,i}^{d}$ represent the convolution kernel parameters,L represents the convolution length, b_{m}^{d} represents the bias term, and $\sigma(\cdot)$ represents the activation function. This formula can capture the short-term changes in the transition of takeoff, landing, steering and support with fewer parameters, so that the sensing branch has the ability of continuous action analysis.

In order to fuse the visual skeleton and sensing features according to reliability, the model introduces cross-modal attention weight, and calculates the modal contribution by combining visual reliability and sensing waveform stability. The calculation process is shown as follows:

$$\alpha_{v,t}, \alpha_{s,t} = \mathrm{softmax}\left([W_v h_{v,t} + \rho R_t,\ W_s h_{s,t} + \mu Q_t]\right) \tag{11}$$

Here, $h_{v,t}$ and $h_{s,t}$ represent visual coding features and sensing coding features respectively, R_t represents visual credibility, Q_t represents sensing stability, W_v, W_s, ρ and μ are learning parameters, $\alpha_{v,t}$ and $\alpha_{s,t}$ represent two types of modal weights. This formula ensures that the sensing mode obtains higher weight when the occlusion is serious, and the visual skeleton results can bear more judgment basis when the sensing waveform is abnormal.

In order to form a single attitude representation vector, the system inputs the weighted visual features, sensing features and historical states into the gated update unit to complete the dynamic attitude feature fusion. The calculation process is as follows:

$$u_t = \Gamma_t \odot (\alpha_{v,t} h_{v,t}) + (1 - \Gamma_t) \odot (\alpha_{s,t} h_{s,t}), \quad \Gamma_t = \mathrm{sigmoid}(W_g [h_{v,t}; h_{s,t}; p_{t-1}]) \tag{12}$$

Here, u_t represents the fused attitude feature, Γ_t represents the gating matrix, p_{t-1} represents the attitude hidden state at the last time, $[h_{v,t}; h_{s,t}; p_{t-1}]$ represents feature concatenation, and $\odot$ represents element-wise multiplication. This formula compressed the visual-spatial structure and the sensing time waveform into a unified vector, so that the subsequent classification and scoring modules could share the same motion state expression.

In order to depict the posture changes in continuous actions, the fusion features are entered into the temporal state propagation layer, and the posture memory is updated together with the hidden state at the last moment. The calculation process is shown as follows:

$$p_t = \tanh(W_p u_t + U_p b_{t-1} + B_p a_t) \tag{13}$$

Here, p_t represents the dynamic attitude state at the current time, b_{t-1} represents the historical state, a_t represents the action phase prior vector, and W_p, U_p, and B_p represent the state propagation parameters. This formula enables the model to preserve the continuous information of action initiation, force generation, air clearance, support and recovery stages, avoiding the local fluctuations of a single window directly changing the evaluation results.

Dynamic pose evaluation requires not only to give the category, but also to quantify the action stability and the joint offset amplitude. The system establishes a comprehensive score based on posture state, joint Angle and sensing impact strength, and the calculation process is shown in the following equation:

$$A_t = \omega_1 \mathrm{CE}(y_t, \hat{y}_t) + \omega_2 \frac{1}{J} \sum_{j=1}^{J} |\theta_{j,t} - \theta_{j,t}^{\mathrm{ref}}| + \omega_3 \|z_{p,t} - z_{p,t}^{\mathrm{ref}}\|_2 \tag{14}$$

Here, A_t represents the posture deviation value in the t window, $\mathrm{CE}(\cdot)$ represents the category error, $\theta_{j,t}$ and $\theta_{j,t}^{\mathrm{ref}}$ represents the j joint Angle and its reference value, $z_{p,t}$ and $z_{p,t}^{\mathrm{ref}}$ represents the pressure characteristics and reference pressure distribution, ω_1 to ω_3 represent the weight coefficients. This equation unifies the classification error, joint offset and support pressure variation into the same evaluation value.

Fig. 7 shows the system deployment structure. The acquisition end accesses the camera and sensor, the computing end completes the preprocessing, model reasoning and result caching, the application end displays the pose score, trajectory curve and abnormal segment, and the database saves the sample index, model version and evaluation record.

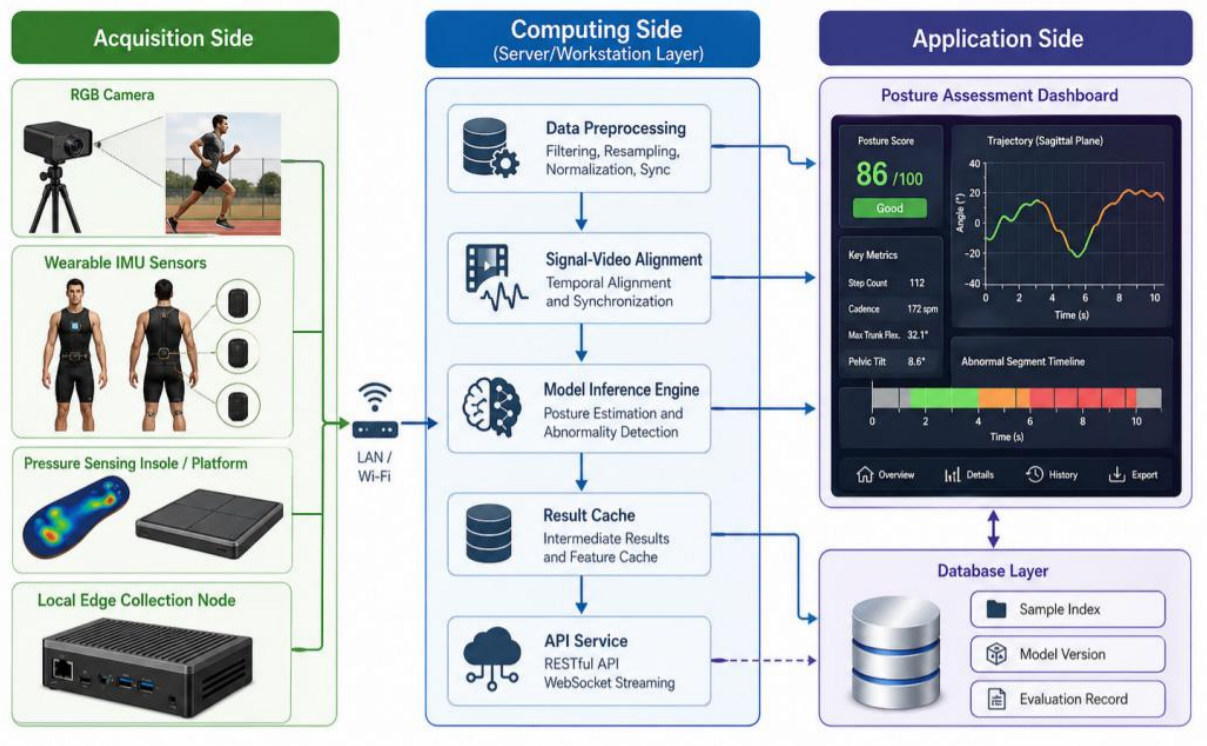

Figure 7: Dynamic pose assessment system deployment architecture

In order to output interpretable comprehensive evaluation results, the system maps window-level deviations into stability scores and combines them with consecutive segments to form the dynamic posture level of the athlete, which is calculated as follows:

$$S = 100 \times \exp\left(-\frac{1}{T}\sum_{t=1}^{T} A_t\right), \quad \hat{c} = \arg\max_{c} \sum_{t=1}^{T} \pi_{c,t} S_t \tag{15}$$

Here, S represents the overall pose stability score, T represents the number of evaluation Windows, $\pi_{c,t}$ represents the probability that the t window belongs to class c, S_t represents the window stability weight, and $\hat{c}$ represents the final pose level. This formula synthesizes the segmented evaluation results into a global output, which enables the system to give numerical scores, category judgments, and corresponding segment positions simultaneously.

At the system design level, the dynamic attitude assessment platform adopts the structure of acquisition layer, algorithm layer, business layer and data layer. The acquisition layer is responsible for the data access of the camera, IMU, pressure pad and joint Angle module. The algorithm layer completed the frame sequence calibration, sensor standardization, modal fusion and attitude score. The business layer generates athlete posture reports, action clip indexes and abnormal tips. The data layer holds the original segments, fusion features, model parameters, and review records. The system output includes posture category, stability score, joint offset and abnormal time period, and the evaluation results can correspond to specific video frames and sensing waveforms, which ensures that the dynamic posture evaluation process has a clear data source and review basis.

4 Performance evaluation and discussion of dynamic posture evaluation system for athletes

4.1 Performance test of visual object detection and motion trajectory tracking

In order to verify the recognition ability of the visual object detection and trajectory tracking module in complex motion scenes, this paper completes comparative experiments on the self-built dynamic posture dataset of athletes. The experimental platform uses Python 3.10, PyTorch 2.1 and CUDA 12.1, and the RTX 4090 graphics card is used at the training end. The video resolution is unified to 1280×720, and the sampling frequency is set to 60 fps. The data covers five types of environments: clear field of view, local reduced visibility, fast movement, backlight interference and multi-person crowding, and contains a total of 128 training videos and 36 testing videos. The average accuracy AP is used as the detection evaluation index, and the center tracking error is used as the trajectory evaluation index. SSD, Faster R-CNN, YOLOv8 and the FusionNet visual detection branch constructed in this paper are selected as comparison methods to observe the recognition stability under different visibility conditions.

As shown in Fig. 8, different methods all show AP decrease after scene complexity increase, but the decrease magnitude is obviously different. FusionNet has an AP of 97.2% in clear scenes, and maintains 93.6%, 91.1%, 89.2% and 86.4% in local visibility loss, fast motion, backlight and crowded scenes, respectively, which is overall higher than YOLOv8, Faster R-CNN and SSD. The AP of YOLOv8 is 94.5% in clear scenes, but drops to 81.3% in crowded scenes. The attenuation of Faster R-CNN and SSD is more obvious in complex background. The results show that the multi-scale visual coding and motion boundary enhancement mechanism of FusionNet can maintain a relatively stable detection ability in complex motion images.

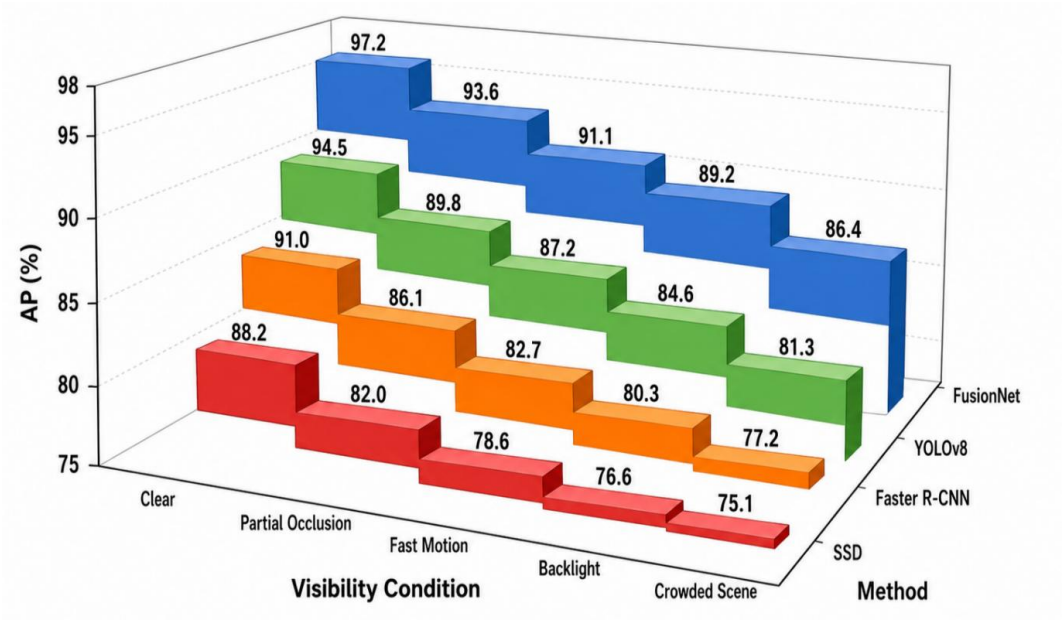

Figure 8:3D waterfall plot of detection accuracy in visibility versus motion condition

As shown in Fig. 9, the center tracking errors of the four types of motion actions show different changes in consecutive frames. Sprint motion has the highest error, mainly distributed between 4.6 and 5.3 pixels, which indicates that high-speed linear motion will

enlarge the inter-frame displacement. The Basketball COD error is concentrated between 4.0 and 4.8 pixels, and the direction switching affects the tracking continuity. Soccer Sprint error is about 3.2 to 4.2 pixels; Combat Footwork has the lowest error, mainly between 2.7 and 3.4 pixels. This figure illustrates that the target trajectory has differential error distributions under different motion rhythms, and the visual branch needs to maintain both detection accuracy and trajectory stability.

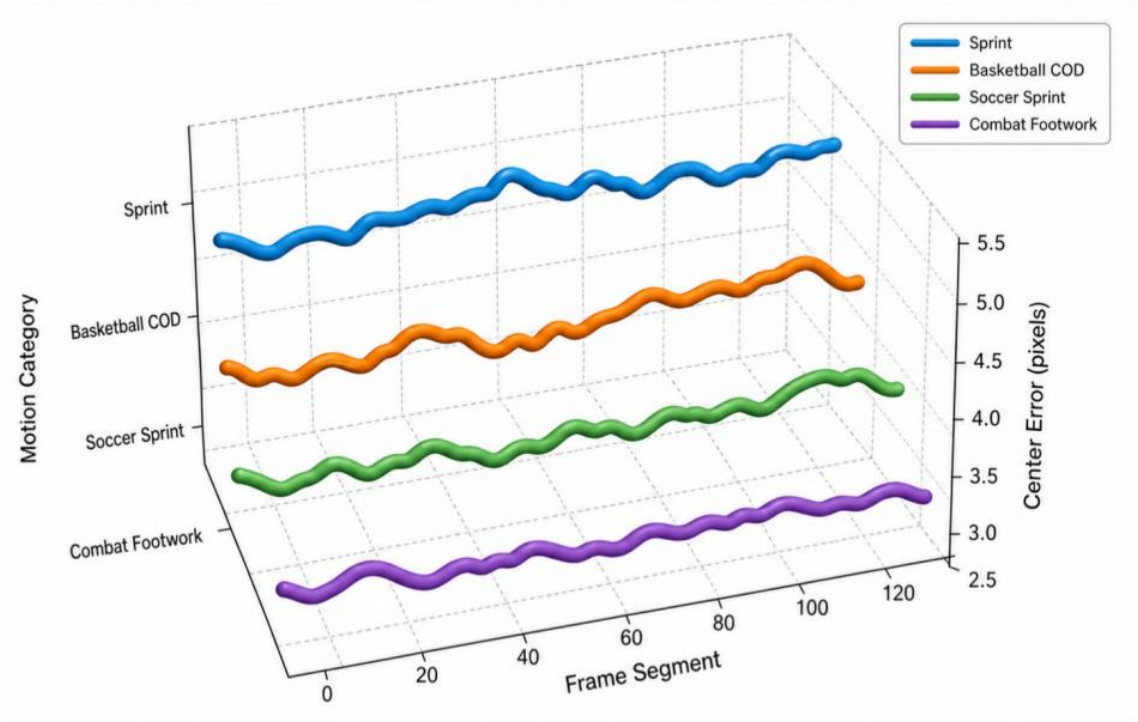

Figure 9:3D trajectory plot of the center tracking error in continuous motion

Table 2 compares the results from four aspects: detection time consuming, ID switching times, tracking success rate, and trajectory smoothness.

Table 2: Comparison of comprehensive performance of visual object detection and trajectory tracking

Method	Single-frame Detection Latency/ms	ID Switches	Tracking Success Rate/%	Trajectory Smoothness
SSD	22.4	46	84.7	0.812
Faster R-CNN	48.6	31	88.5	0.846
YOLOv8	31.2	24	91.6	0.889
FusionNet	26.8	15	95.3	0.934

Table 2 shows that FusionNet achieves 95.3% tracking success rate and 0.934 trajectory smoothness while maintaining a low inference delay with the least number of ID switches. This result corresponds to the previous one, indicating that the visual branch in this paper can not only improve the target detection AP, but also reduce the target number jump in continuous motion scenes.

In summary, the visual object detection and trajectory tracking module can provide stable video input for skeleton key point extraction, trajectory sequence modeling and dynamic pose assessment, and achieve a good balance between accuracy, speed and trajectory continuity.

4.2 Performance test of multi-source sensor fusion attitude recognition algorithm

The core of multi-source sensor fusion posture recognition is to map the visual skeleton, IMU signals, plantar pressure and joint Angle changes into the same representation space, and complete the category discrimination in the action window. In this paper, samples are extracted from six typical actions: jump, landing, stop, turn, swing arm sprint and support, and the fusion feature distribution and modal weight changes are jointly analyzed. The visual skeleton is used to describe the human body structure, the IMU is used to record the inertial change, the plantar pressure is used to represent the support migration, and the joint Angle is used to characterize the local posture synergy. Four types of data enter the fusion network after timestamp calibration and window segmentation.

As shown in Fig. 10, the six types of actions form a clear cluster distribution in the 3D feature space. The Jump cluster is located in the high angular velocity region, indicating that the jump stage has a strong joint angular velocity response. Landing is concentrated in the medium IMU amplitude and positive angular velocity range, reflecting the inertial absorption characteristics of the landing buffer stage. The Stop samples are in the region of low angular velocity and small pressure transfer. Turn is located in the middle region, indicating that the turn motion is accompanied by both a support transition and a moderate angular velocity change. Sprint Arm Swing was more active in plantar pressure change and joint angular velocity. The Support samples are in the region of high IMU amplitude but low angular velocity. The distribution indicates that the multi-source sensing features have better class separability after joint coding.

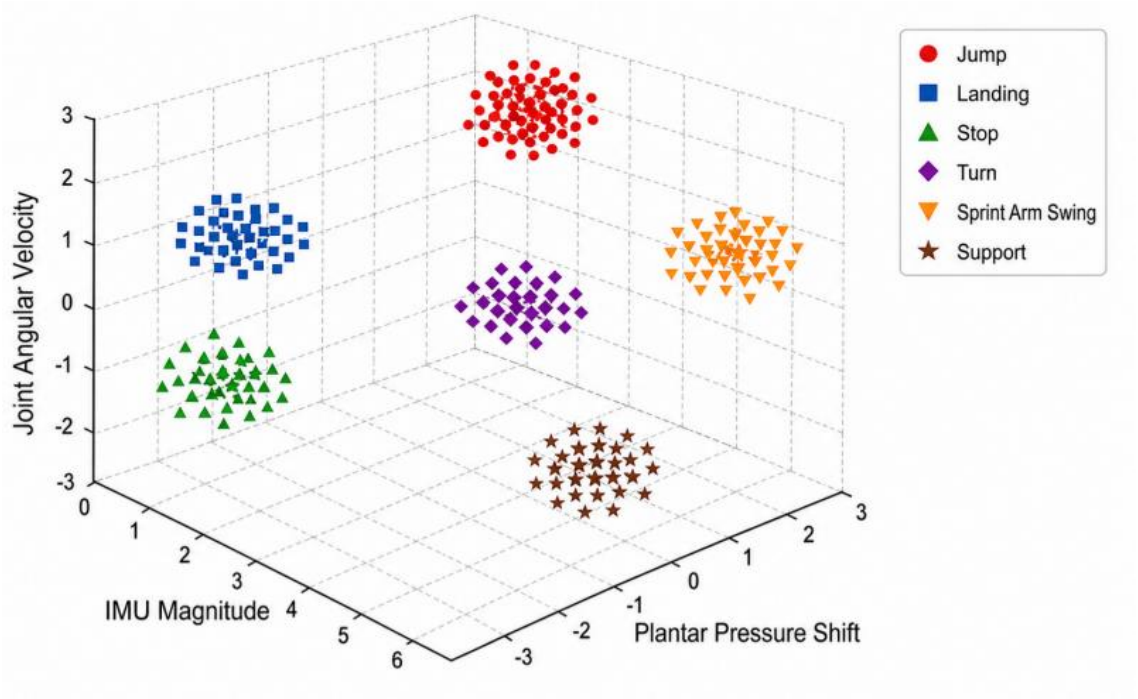

Figure 10:3D plot of feature manifold for multi-sensor pose recognition

As shown in Fig. 11, different sensing modalities do not have the same weight in the action phase. The Vision Skeleton has the highest weight in the Take-off stage, reaching 0.42, indicating that the torso expansion and limb extension at the takeoff moment mainly depend on the visual skeleton judgment. The weight of IMU increased to 0.32 in the Mid-swing stage, reflecting the inertia change during the swing. Plantar Pressure reached 0.42 in the Support

Transfer stage, indicating that support transfer was more dependent on plantar pressure transfer. Joint Angle reaches 0.38 and 0.34 in the Support Transfer and Landing Buffer phases, respectively, indicating that joint Angle has a strong explanatory power for action buffer and support change.

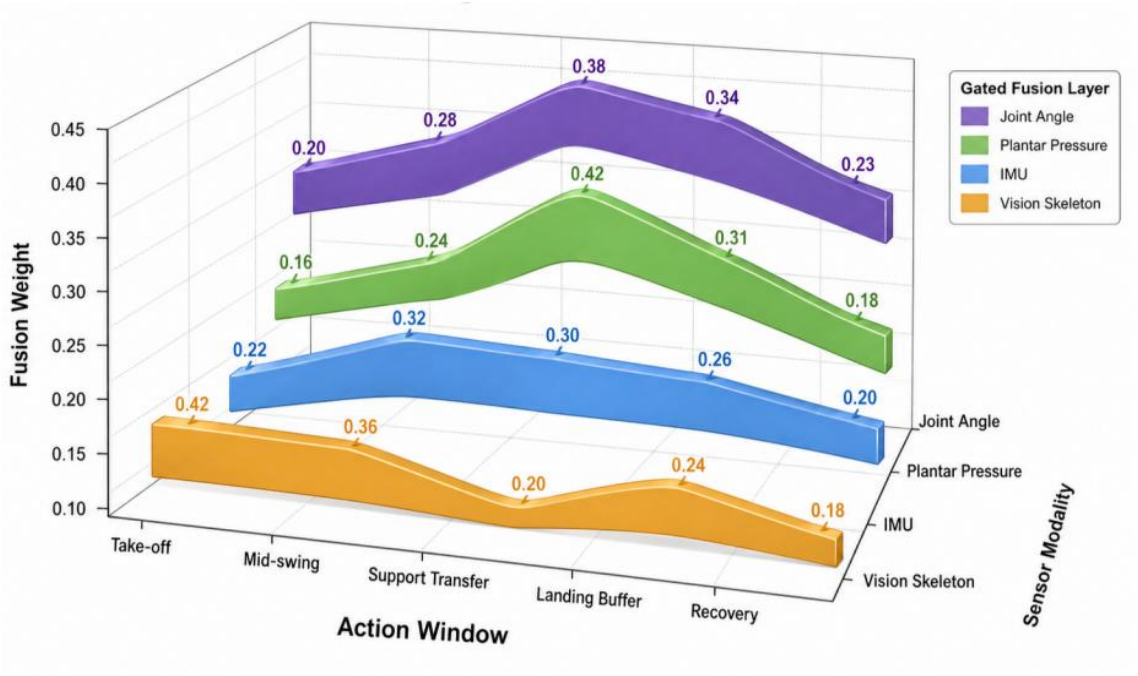

Figure 11:3D hierarchical strip diagram of fusion weights of action window and sensing modality

To further verify the influence of multi-source fusion on posture recognition performance, Table 3 lists the recognition results under different modal inputs.

Table 3: Comparison of posture recognition performance with different modal inputs

Modality Combination	Accuracy/%	Recall/%	Precision/%	F1 Score
Vision Skeleton	91.8	90.6	91.3	90.9
IMU	90.6	89.8	90.3	90.0
Plantar Pressure	88.9	87.5	88.1	87.8
Joint Angle	91.4	90.7	91.0	90.8
Full Fusion	95.28	94.87	95.31	95.09

Table 3 shows that the full fusion model has higher Accuracy, Recall, Precision, and F1 values than the single modality. The visual skeleton provides the action structure, the IMU provides the inertial change, the plantar pressure reflects the support migration, and the joint angles characterize the limb synergy.

It can be seen from the above that multi-source sensor fusion can make up for the problem of insufficient data expression of single vision or single sensor. The fusion model automatically adjusts the modal weights in different action stages, so that the posture recognition results are more in line with the real data changes during the movement, and also provides a more complete feature basis for dynamic posture scoring.

4.3 Dynamic posture evaluation model performance test

The dynamic posture assessment model not only needs to complete the discrimination of action categories, but also needs to score the quality of action completion, spatial coordination and temporal coherence. Therefore, this paper further analyzes the performance of the model in posture score and abnormal segment localization, focusing on the relationship between dynamic posture score, action stage and localization confidence. The model inputs include skeleton node sequence, motion trajectory, IMU timing features, plantar pressure distribution and joint Angle variation, and the output results include posture category, evaluation score, stability and abnormal segment index.

As shown in Fig. 12, the dynamic posture score forms a continuous surface under the joint synergy index and temporal consistency. When the joint synergy index and time series consistency were close to the medium-high range, the Assessment Score was close to 1.00. When the two types of indicators decrease, the score gradually decreases to near 0.75. The wireframe surface in the figure does not appear obvious fracture, which indicates that the pose score output by the model has good continuity. The bottom contour further shows that the areas with high scores are concentrated in the regions with stable action structure and smooth timing change, which is consistent with the logic of action quality judgment in the athlete posture assessment task.

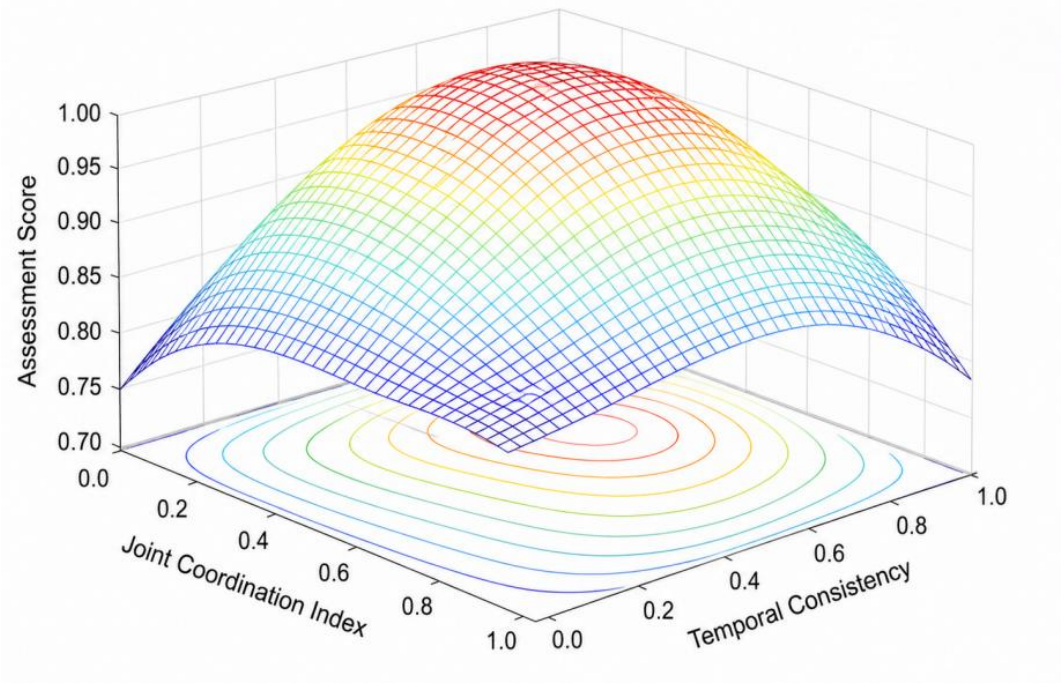

Figure 12: Dynamic pose evaluation score 3D wireframe response plot

As shown in Fig. 13, the abnormal segment localization confidence changes with Motion Phase and Offset Magnitude. In Takeoff phase, the confidence is 0.97 under low offset condition, 0.96 in Mid-swing phase, 0.95 in Support Transfer phase, 0.94 in Landing Buffer phase, and 0.92 in Recovery phase. With the Offset Magnitude increasing from 5° to 20°, the confidence of each stage decreases, and the lowest value is 0.76 in the Recovery stage. This result shows that the greater the joint offset, the higher the uncertainty of the action segment, but the model can still maintain a clear positioning trend in the critical stage.

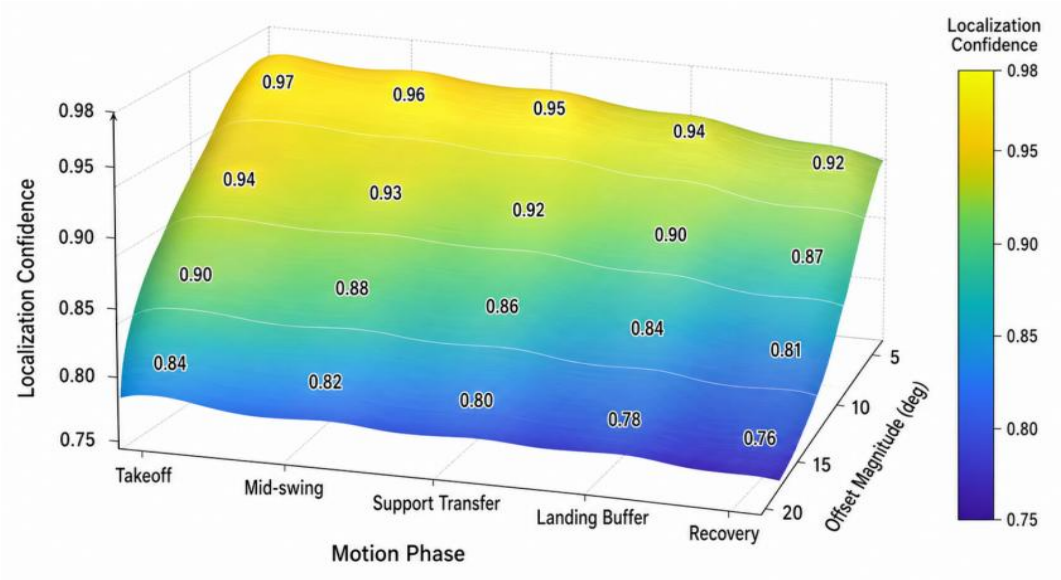

Figure 13:3D topographic map of location confidence of abnormal motion segments

To compare the differences between the proposed model and the commonly used temporal model and skeleton model in dynamic posture assessment tasks, Table 4 lists the comprehensive accuracy, average scoring error, stability correlation coefficient, and anomaly localization rate.

Table 4: Comparison of comprehensive performance of dynamic pose assessment models

Model	Accuracy/%	MAE/Score	Stability Correlation Coefficient	Abnormal Localization Rate/%
ST-GCN	91.8	4.36	0.873	88.2
TCN	92.4	4.08	0.881	89.5
BiLSTM	90.7	4.91	0.856	86.9
Transformer	94.3	3.47	0.904	91.6
Proposed Model	96.18	2.84	0.932	94.8

Table 4 shows that the comprehensive accuracy of the proposed model reaches 96.18%, the MAE score is 2.84, the correlation coefficient of stability reaches 0.932, and the abnormal location rate reaches 94.8%. Compared with ST-GCN, the proposed model not only relies on the skeleton topology, but also uses sensor data to supplement action shocks and support changes. Compared with TCN and BiLSTM, the proposed model has stronger expressive ability for action phase transitions. Compared with Transformer, the proposed model is more stable in anomaly localization rate and scoring error.

The comprehensive results show that the dynamic posture evaluation model can complete posture quality scoring and abnormal segment location at the same time, and the scoring surface is continuous and the positioning trend is clear, which is suitable for supporting the fine-grained evaluation of athletes 'motion quality.

4.4 Practical application and robustness analysis of the system

When the system enters the real application scenario, noise, data packet loss and end-side load will affect the stability of the evaluation results. This paper analyzes the practical application performance of the system from three aspects: anti-interference ability, end-side computing performance and module contribution, and verifies the role of each component through

ablation experiments. In the test phase, the system is deployed on the edge computing terminal, and the lateral camera, IMU node, plantar pressure module and joint Angle sampling unit are synchronously connected to the acquisition terminal. The output terminal records the posture evaluation results, abnormal segment index and inference delay.

As shown in Fig. 14, inference latency increases as CPU and GPU utilization goes up. The Normal operating point corresponds to CPU utilization of about 52%, GPU utilization of about 62.4%, and inference delay of 31.6ms. The Full-load point corresponds to about 88% CPU utilization, 79% GPU utilization, and 38.9 ms inference latency. The overall change of the resource response plane is smooth, which indicates that the system delay increase is controllable after the end-side computing load rises, and there is no burst-type delay peak. This result shows that the proposed system can complete the real-time pose assessment task at the edge end.

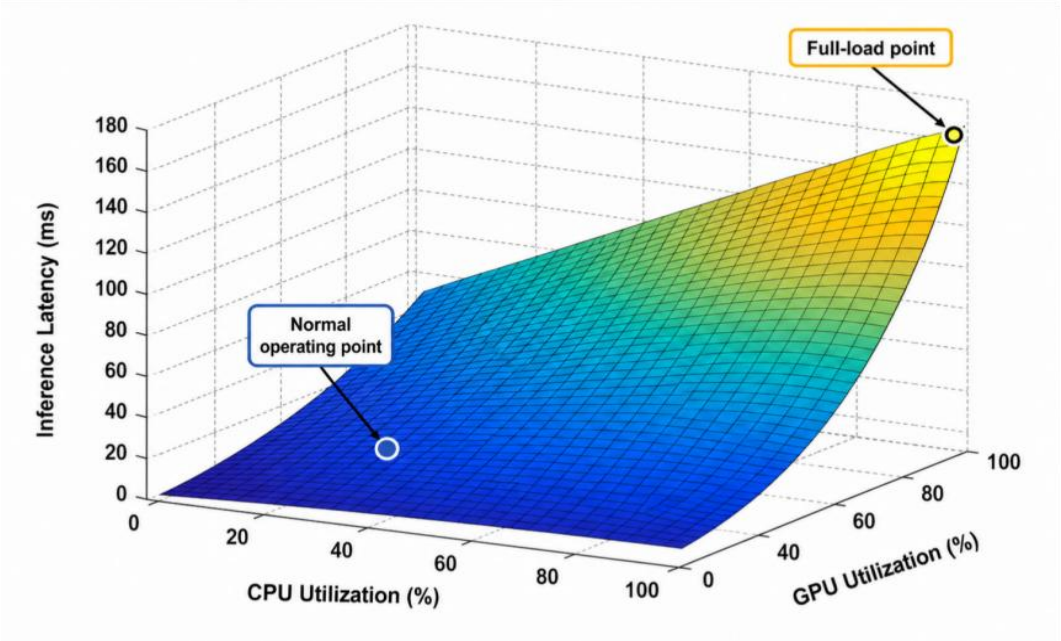

Figure 14:3D response floor plan of resource occupancy versus inference latency in edge-side deployment

As shown in Fig. 15, different slices of data packet loss rate are set in the Normal Run scenario, so three blue trend lines are presented in the figure. The F1 Score of the three curves is close to 96%-98% under low Noise and low Packet Loss conditions. With the increase of Noise Level and packet loss, the curves gradually decrease, but still maintain at about 92% under high interference conditions. It shows that the fusion results of vision and sensing are stable under normal running conditions. The scene of Fast Sprint corresponds to the green trend line, and the F1 Score decreases from about 95% to about 85%, which is significantly larger than that of Normal Run, indicating that high-speed sprint will amplify the disturbance of sensing waveform and the offset between video frames, and form a stronger interference to the attitude recognition results. The Sensor Dropout scenario corresponds to the red trend line, and the F1 Score decreases from about 91% to about 82%, which is the most obviously affected among the three types of scenarios. This result shows that the system has some tolerance to general noise and mild data packet loss, but the stability of the multi-source fusion evaluation will decrease significantly when the motion speed increases or the sensor data is missing.

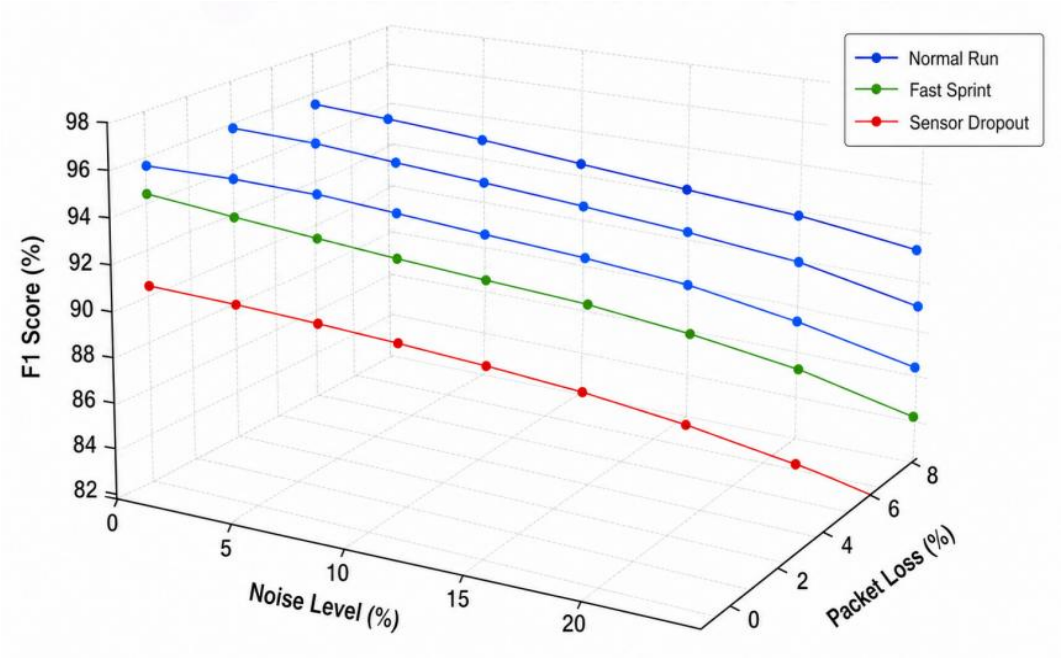

Figure 15: Plot of system robustness under noise and data packet loss conditions

To further examine the contribution of each module of the system to the overall performance, Table 5 presents the ablation experiment results. Different from Fig.s 14 and 15, this table focuses on analyzing the influence of visual branch, IMU branch, pressure branch, joint Angle branch and gated fusion mechanism on system accuracy, F1 value and abnormal localization rate.

Table 5: Ablation experiment results of dynamic pose assessment system

Model Setting	Accuracy/%	F1 Score	Average Latency/ms	Abnormal Localization Rate/%
Without Vision Branch	91.2	0.902	25.8	86.7
Without IMU Branch	92.5	0.918	28.4	88.3
Without Pressure Branch	93.1	0.924	29.1	89.6
Without Joint Angle Branch	93.8	0.931	29.5	90.2
Without Gated Fusion	92.0	0.912	30.8	87.5
Complete System	96.18	0.951	31.6	94.8

Table 5 shows that removing either modality causes performance degradation. After removing the visual branch, the Accuracy is reduced to 91.2%, and the abnormal location rate is reduced to 86.7%, which indicates that the visual skeleton plays a fundamental role in the judgment of action structure. After removing the IMU branch, the dynamic response ability of the fast movement stage decreases. After removing the pressure branch, the recognition stability is weakened in the stage of support conversion and landing buffer. After removing the gated fusion, the F1 value decreases to 0.912, indicating that the dynamic weight allocation mechanism has an obvious effect on multi-source feature integration.

Based on the above, it can be seen that the system still maintains a relatively stable output ability under the conditions of enhanced noise, data packet loss, sensor loss and increased end-side load, which has the application basis for sports training, attitude monitoring and action technology analysis.

4.5 Discussion

The experimental results of the proposed system show that the dynamic pose assessment needs to be completed by visual perception, sensing timing and fusion discrimination. In the visual detection stage, the AP of FusionNet reaches 97.2% in clear scenes, and maintains 93.6%, 91.1%, 89.2% and 86.4% in reduced visibility, fast motion, reverse light and crowded conditions, which indicates that the multi-scale visual coding retains the characteristics of human boundaries and motion regions. The center tracking error shows differences in different actions, with Sprint error ranging from 4.6 to 5.3 pixels, Combat Footwork remaining at 2.7 to 3.4 pixels, and high-speed displacement amplifying the center drift between frames. In the sensing fusion stage, the F1 value of the complete fusion model reaches 95.09%, which is higher than the single modal results of visual skeleton, IMU, plantar pressure and joint Angle. The difference distribution of fusion weights in the Take-off, Support Transfer and Landing Buffer Windows indicates that the gating structure can adjust the data source according to the action stage. In the dynamic evaluation stage, the accuracy of the model in this paper reaches 96.18%, the MAE score is 2.84, and the anomaly location rate is 94.8%, which is better than ST-GCN, TCN, BiLSTM and Transformer. Posture quality should not be explained by the skeleton topology alone, but should also be synergetic in combination with support pressure, inertial variation, and joint angles. In the robustness test, Normal Run maintains about 92% F1 value under high interference, and Sensor Dropout drops to about 82%, which can weaken the fusion stability due to the lack of sensing. The conventional delay of the edge side is 31.6ms, and the full load state is 38.9ms. On the whole, the system converts video structure information and sensor timing information into repeatable posture assessment results, which is suitable for continuous motion monitoring and technical state analysis.

5 Conclusion

In this paper, a system framework based on computer vision and sensor fusion is constructed around the task of dynamic posture assessment of athletes. The system uses the deep vision network to complete the athlete target detection, skeleton key point extraction and trajectory tracking, and uses IMU, plantar pressure and joint Angle sensing signals to supplement the inertial change, support migration and local posture coordination that are difficult to express directly in the video. Then a unified dynamic posture representation is formed through the gated fusion mechanism. Experimental verification shows that the vision branch, the sensing branch and the fusion evaluation module can form a continuous data processing link, so that the system output not only contains the action category, but also gives the attitude stability, abnormal segment and end-side running state. The design enables the athletes' actions to be transformed from video images into computable and traceable multimodal features, which provides technical support for sports training analysis, action quality review and on-site monitoring. There are still some limitations in this paper. The sample collection mainly focuses on a number of typical motion actions, and complex confrontation, multi-player high-speed interaction and open field changes have not been fully covered. The sensor wearing position has an impact on the signal stability, and some actions still need to rely on manual review and confirmation. Although the edge-side deployment has real-time performance, the model compression, cache management and cross-device synchronization on low-power devices still need further verification. The follow-up research can be carried out from three directions. First, the cross-project, cross-field and cross-individual samples should be expanded to enhance the adaptability of the model to different body types and motion

styles. Second, self-supervised temporal pre-training and cross-modal alignment mechanisms are introduced to reduce the dependency on manual labeling. Thirdly, it combines lightweight network, model pruning and end-cloud collaborative reasoning to reduce deployment cost and maintain evaluation accuracy. In the future, the posture assessment results can be combined with training load, fatigue recognition and injury risk warning to form a more complete intelligent motion analysis system, and support long-term motion data management and individualized technical file construction, which is convenient for subsequent training review and model iteration.

References

[1] Uhlrich S D, Falisse A, Kidziński Ł, et al. OpenCap: Human movement dynamics from smartphone videos[J]. PLoS computational biology, 2023, 19(10): e1011462.

[2] Cronin N J, Walker J, Tucker C B, et al. Feasibility of OpenPose markerless motion analysis in a real athletics competition[J]. Frontiers in Sports and Active Living, 2024, 5: 1298003.

[3] Fukushima T, Blauberger P, Guedes Russomanno T, et al. The potential of human pose estimation for motion capture in sports: a validation study[J]. Sports Engineering, 2024, 27(1): 19.

[4] Aleksic J, Kanevsky D, Mesaroš D, et al. Validation of automated countermovement vertical jump analysis: Markerless pose estimation vs. 3D marker-based motion capture system[J]. Sensors, 2024, 24(20): 6624.

[5] Milone D, Longo F, Merlino G, et al. MocapMe: DeepLabCut-enhanced neural network for enhanced markerless stability in sit-to-stand motion capture[J]. Sensors, 2024, 24(10): 3022.

[6] Kim M, Lee S. Fusion poser: 3d human pose estimation using sparse imus and head trackers in real time[J]. Sensors, 2022, 22(13): 4846.

[7] Amadi L, Agam G. Posturepose: Optimized posture analysis for semi-supervised monocular 3D human pose estimation[J]. Sensors, 2023, 23(24): 9749.

[8] Rajendran A K, Sethuraman S C. A survey on yogic posture recognition[J]. IEEE Access, 2023, 11: 11183-11223.

[9] Avogaro A, Cunico F, Rosenhahn B, et al. Markerless human pose estimation for biomedical applications: a survey[J]. Frontiers in Computer Science, 2023, 5: 1153160.

[10] Roggio F, Trovato B, Sortino M, et al. A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review[J]. Heliyon, 2024, 10(21).

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

[12] Ghosh I, Ramasamy Ramamurthy S, Chakma A, et al. Sports analytics review: Artificial intelligence applications, emerging technologies, and algorithmic perspective[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2023, 13(5): e1496.

[13] Naik B T, Hashmi M F, Bokde N D. A comprehensive review of computer vision in sports: Open issues, future trends and research directions[J]. Applied Sciences, 2022, 12(9): 4429.

[14] Host K, Ivašić-Kos M. An overview of human action recognition in sports based on computer vision[J]. Heliyon, 2022, 8(6).

[15] Islam M M, Nooruddin S, Karray F, et al. Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things[J]. Information Fusion, 2023, 94: 17-31.

[16] Akter N, Molnar A, Georgakopoulos D. Toward Improving Human Training by Combining Wearable Full-Body IoT Sensors and Machine Learning[J]. Sensors, 2024, 24(22): 7351.

[17] Khan I U, Afzal S, Lee J W. Human activity recognition via hybrid deep learning based model[J]. Sensors, 2022, 22(1): 323.

[18] Tanigaki K, Teoh T C, Yoshimura N, et al. Predicting performance improvement of human activity recognition model by additional data collection[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(3): 1-33.

[19] De Marchi M, Turetta C, Pravadelli G, et al. Combining 3-D Human Pose Estimation and IMU Sensors for Human Identification and Tracking in Multiperson Environments[J]. IEEE Sensors Letters, 2024, 8(6): 1-4.

[20] Merker S, Pastel S, Bürger D, et al. Measurement accuracy of the HTC VIVE Tracker 3.0 compared to vicon system for generating valid positional feedback in virtual reality[J]. Sensors, 2023, 23(17): 7371.

[21] Russomanno T G, Blauberger P, Kolbinger O, et al. Drone-based position detection in sports—validation and applications[J]. Frontiers in physiology, 2022, 13: 850512.