Real-Time Aerobics Pose Estimation and Motion Trajectory Optimization Using Enhanced YOLOv7 with CA Attention and ASPP

Wenlong Zhang[1], Zheng Huang[2*], Heng Ding[1],
[1]Beijing Institute of Fashion Technology, Beijing, 100029, China
[2]BDA New Town School of The High School Affiliated to Renmin University of China, Beijing, 102600, China
E-mail: Zhenghuangg@outlook.com
[*]Corresponding author

Keywords: YOLOv7, Aerobics posture capture, Motion trajectory optimization, CA attention mechanism; Real-time detection

Received: July 23, 2025

Aiming at the high dynamic characteristics of aerobics, this study proposes a real-time pose capture and motion trajectory optimization method based on the YOLOv7-Pose algorithm. The method is evaluated on the CAF-3D and FitMotion-VIS datasets. By improving the keypoint detection head of YOLOv7, combined with the CA attention mechanism and the atrous spatial pyramid pooling (ASPP) structure, the accuracy of human keypoint detection is significantly improved (the mAP of the verification set reaches 95.7%, outperforming OpenPose and AlphaPose). At the same time, the dynamic time warping (DTW) algorithm and a multi-objective trajectory optimization strategy are introduced to solve trajectory matching issues caused by varying action speeds, and TensorRT is used for accelerated deployment to achieve real-time performance of 84 FPS. Experiments show that the system maintains high robustness under complex illumination, multi-person occlusion, and dynamic motion, with the position error of keypoints reduced to less than 3%. These results provide reliable technical support for applications in sports training, rehabilitation evaluation, and other real-world scenarios.

Povzetek: Študija predstavi sistem za zajem drže in optimizacijo gibanja pri aerobiki, ki izboljša YOLOv7-Pose z CA in ASPP za natančnejšo detekcijo ključnih točk, uporabi DTW in večciljno optimizacijo za poravnavo trajektorij.

1 Introduction

Real-time human posture capture is a key area in computer vision with applications in sports training, virtual reality, and medical rehabilitation. Aerobics, with rapid rotations, jumps, and non-rigid movements, poses challenges for accurate posture capture. Traditional marker-based systems (e.g., Vicon, OptiTrack) require costly setups and controlled environments, limiting their use in open-field training. Marker-free RGB methods are affected by lighting changes, dynamic backgrounds, occlusion, and motion blur, often causing keypoint errors over 10%**.** Mainstream algorithms like OpenPose (≈22 FPS) and AlphaPose (AP 56.6%) struggle to balance accuracy and real-time performance [1]. Errors above 5° in posture angle or 15 cm in IMU-based position can mislead training and evaluation [2]. To address this, we propose YOLOv7-Pose, integrating attention mechanisms, multi-scale modules, and trajectory optimization for robust, high-speed aerobics posture capture. Our contributions focus on improving detection under motion blur and occlusion, ensuring real-time performance, and optimizing motion trajectories for training and rehabilitation.

This study proposes YOLOv7-Pose+, an enhanced architecture with three key innovations.1. Feature extraction optimization: We embed the coordinate attention (CA) mechanism [3] into the backbone, enhancing focus on joint regions (e.g., elbows and knees) in complex backgrounds. CA aggregates features along height and width to generate attention maps, reducing missed detections in occlusion scenarios by 18.3%.2. Multi-scale modeling enhancement: Atrous spatial pyramid pooling (ASPP) [4] is employed in the keypoint head to extract multi-scale features via parallel convolutions with varied expansion rates. This increases the receptive field 2.1 × and improves mAP on the validation set to 95.7%, enhancing perception of extended limb postures such as split-leg jumps.3. Motion trajectory optimization: To address trajectory distortions caused by variable motion speeds, we apply dynamic time warping (DTW) for nonlinear alignment, minimizing the cumulative distance matrix. A particle hierarchical reinforcement learning (PHRL) strategy is used to optimize trajectories at the Pareto frontier, reducing energy consumption by 31.2% and improving smoothness by 44%. Overall, the methodology emphasizes our contributions in robust feature extraction, multi-scale posture modeling, and trajectory optimization, while leveraging YOLOv7 as a foundation rather than restating its existing design.

To meet real-time requirements, the system is accelerated and deployed via TensorRT [5]. Firstly, the

PyTorch model is converted into ONNX format, and its size is reduced to 37% of the original using INT8 quantization. Real-time inference at 84 FPS is achieved on the NVIDIA Jetson AGX Xavier edge device. Performance verification follows the virtual reality motion capture standard (ISO 13406). In terms of accuracy, a dataset of 100 professional aerobics videos (200 standard movements) was constructed, and the average keypoint position error was 2.8 mm (76% lower than Kinect V2), with posture angle error controlled within ±0.5° under dynamic scenarios. Robustness tests in strong light, shadow interlacing, and 40% occlusion scenarios showed mAP above 91.4% and false detection below 5%. Trajectory optimization efficiency improved: using DTW-PHRL instead of traditional linear interpolation increased trajectory matching efficiency by 26%, and coach evaluation indicated a 32.1% increase in action standardization scores. Compared with mainstream solutions (Table 1), this system achieves significant advantages in detection speed (84 FPS vs. 22 FPS), energy consumption (3.2 J vs. 9.7 J), and complex action recognition rate (93.6% vs. 82.1%). Overall, these results demonstrate that the method is practical and effective for sports training and rehabilitation, providing reliable technical support for digital sports science.

Table 1: Comparative performance of pose estimation methods on key datasets

Method	Dataset	mAP (%)	FPS	Keypoint Error	Notes
OpenPose	COCO	62	22	~10%	Marker-free, real-time, widely used baseline
AlphaPose	COCO	56.6	15	~8-10%	Higher accuracy than OpenPose but slower
Faster R-CNN (for pose)	COCO	54.3	10	~12%	Object-detection-based approach, not real-time
HRNet	COCO	73	12	~6-8%	High-resolution feature maps for keypoint estimation
YOLOv7-Pose+ (Ours)	CAF-3D / FitMotion-VIS	95.7	84	<3%	CA + ASPP + DTW-PHRL, real-time, robust under occlusion and dynamic motion

2 Theoretical basis of real-time capture of aerobics posture and motion trajectory based on YOLOv7

2.1 Theoretical basis of YOLOv7 target detection

The YOLO (You Only Look Once) series of algorithms have achieved a remarkable balance between speed and accuracy, making them the benchmark in the field of real-time object detection [6, 7]. YOLOv7, while inheriting the core ideas of YOLO, has further enhanced its performance through ingenious network architecture and training strategies, laying a solid foundation for rapid and accurate human detection in aerobics scenarios [8, 9]. Its theoretical basis encompasses single-stage detection frameworks, bounding box prediction mechanisms, and loss function design [10].

The core idea of YOLOv7 is to divide the input image into an S × S grid (Grid). Each grid cell is responsible for predicting multiple bounding boxes (BBox) centered on that cell [11]. Each predicted box contains position information (center coordinates (b_x, b_y), width b_w, height b_h), confidence (Confidence, indicating the probability that the box contains a target and its position is accurate) and the conditional probability distribution $P(Class_i|Object)$ for each category. To accurately predict the position of the bounding box, YOLOv7 adopts an offset prediction mechanism based on anchor boxes (Anchor Box) [12]. For the preset anchor box size (p_w, p_h), the model predicts the offset relative to the top-left corner of the grid cell (t_x, t_y) and the scaling factors for width and height (t_w, t_h). The center coordinates (b_x, b_y) and size (b_w, b_h) of the final predicted box are calculated using formula (1). Here, (c_x, c_y) is the coordinate of the top-left corner of the current grid cell, $\sigma(\cdot)$ is the Sigmoid activation function, ensuring that the offsets are within the range of 0 to 1, thereby constraining the center point within the current grid cell. e^{tw} and e^{th} allow the model to perform exponential scaling of the anchor box size to adapt to different-sized targets, which is crucial for capturing athletes with extremely large movements in aerobics [13].

$$b_x = \sigma(t_x) + c_x, b_y = \sigma(t_y) + c_y, b_w = p_w \times e^{t_w}, b_h = p_h \times e^{t_h} \quad (1)$$

The design of loss function is the key to drive model

learning. The loss function Ltotal of YOLOv7 is a multi-task loss that fuses the bounding box regression loss Lbox, the target confidence loss Lobj, and the classification loss Lcls, as shown in formula (2). Among them, λ is the coefficient that balances the importance of each loss.

$$L_t total = \lambda_{box} \times L_{box} + \lambda_{obj} \times L_{obj} + \lambda_{cls} \times L_{cls} \quad (2)$$

Lbox usually uses CIoU Loss (Complete Intersection over Union Loss), which not only considers the overlapping area (IoU) of the bounding box, but also considers the matching degree of the center point distance and aspect ratio, which can more comprehensively measure the prediction box and the real box. The degree of coincidence is shown in formulas (3) to (6). Where B and Bgt represent the prediction box and the true box (Ground Truth) respectively, ρ(·) calculates the Euclidean distance between the two center points, c is the diagonal length of the minimum closure area covering the two boxes, w, h, wgt, hgt are the width and height of the prediction box and the true box respectively, and υ measures the consistency of the aspect ratio. Lobj uses Binary Cross Entropy Loss (BCE) [14, 15] to calculate the difference between the confidence score of the target contained in the prediction box and the real situation (1 for goals and 0 for no goals):

$$L_{box} = 1 - IoU + (\rho^2(b, b^{gt}) / c^2) + \alpha\upsilon \quad (3)$$

$$IoU = | B \cap B^{gt} | / | B \cup B^{gt} | \quad (4)$$

$$v = (4/\pi^2) \times (arctan(w^{gt}/h^{gt}) - arctan(w/h))^2 \quad (5)$$

$$\alpha = v / (1 - IoU + v) \quad (6)$$

2.2 Theoretical basis of human pose estimation

The goal of pose estimation is to accurately locate the spatial position of body key points (such as head, shoulder, elbow, wrist, hip, knee, ankle, etc.) from within the detected human bounding box [16]. In aerobics applications, the Top-Down method is usually used: first detect people, and then estimate each person's key points. The mainstream methods are mainly based on Heatmap regression or direct coordinate regression, and the former is widely adopted because of its accuracy advantages [17].

The heat map representation generates a probability map Hk corresponding to the spatial resolution of the input image (or cropped human body area) for each key point type (k represents the key point index). The value of each pixel position (i, j) on the heat map Hk represents the probability that the position is the k-th key point. Ideally, the heat map presents a two-dimensional Gaussian distribution peak at the real key point (xk, yk), as shown in formula (7). Where σ controls the width of the Gaussian kernel, which reflects the tolerance of positioning accuracy. The task of the network is to learn the mapping from the input image I to the set H = [H1, H2,..., HK] of all keypoint heat maps [18].

$$H_k(i, j) = exp(-[(i - y_k)^2 + (j - x_k)^2] / (2\sigma^2)) \quad (7)$$

When training, use pixel-by-pixel mean square error (MSE) or a better Focal Loss variant as the loss function Lpose to focus on key points that are difficult to predict, as shown in formula (8), where W, H are the width and height of the heat map, Hk(i, j) is the value of the heat map predicted by the network at position (i, j), and Hk(i, j) is the value of the real heat map (generated by the Gaussian distribution). In the inference stage, by finding the pixel position with the largest response value on the predicted heat map Hk.

$$L_{pose} = (1/K) \times \Sigma_{k=1}^{K} \Sigma_{i=1}^{W} \Sigma_{j=1}^{H} || H_k(i, j) - H_k(i, j) ||^2 \quad (8)$$

In the inference stage, the rough coordinates of key points are obtained by finding the pixel position (îk, ĵk) with the largest response value on the predicted heat map Ĥk: however, since the resolution of the heat map is usually lower than that of the original map, and the maximum position is a discrete integer coordinate, it is necessary to further use sub-pixel precise positioning techniques (such as offset prediction or Taylor expansion) to obtain continuous spatial coordinates [19-21]. A common method is to have the network additionally predict the offset map of each key point in the x and y directions, and the final coordinates are calculated as: where s is the downsampling multiple (step size) of the heat map relative to the original map. This method of heat map plus offset can effectively overcome the quantization error and realize high-precision key point location. For complex and high-speed movements in aerobics, accurate and stable key point positioning is the basis for analyzing posture and trajectory [22-24].

2.3 Theoretical basis of motion trajectory modeling and optimization

The coordinate sequence of joint points output by the original pose estimation often contains noise jitter, short-term absence (occlusion), and jump (false detection). The core goal of trajectory optimization is to filter, smooth, and interpolate the sequence in the space-time domain, generating a continuous, smooth, and high-quality trajectory that conforms to human biomechanical laws. This optimization combines Kalman Filter (KF) for temporal smoothing with Particle Hierarchical Reinforcement Learning (PHRL) for multi-objective trajectory optimization and kinematics constraints [25].

The state prediction phase is based on the posterior estimation at the previous moment and the system dynamics model (described by the state transition matrix). As shown in formulas (9) and (10). The process noise covariance matrix Q characterizes the uncertainty of the model (such as unmodeled acceleration).

In the state update stage, the observation value at the current time t (zt, usually directly from the output of the attitude estimation module) is used to correct the prior prediction. The observation matrix H maps the state space to the observation space, and the observation noise covariance matrix R characterizes the measured uncertainty (pose estimation error). As shown in formula (11), the Kalman gain matrix Kt determines the degree of trust prediction (ŝ) or observation (zt): when R is large, Kt is small, and the prediction is more trusted; When R is hour, Kt is large, and the observation is more trusted. Finally, the covariance is estimated by outputting the posterior state estimate and the posterior estimate.

The state prediction phase of the KF is based on the posterior estimation at the previous moment and the system dynamics model described by the state transition matrix, as shown in formulas (9) and (10). The process noise covariance matrix Q characterizes model uncertainty (e.g., unmodeled acceleration). In the state update stage, the observation at time t (zt, typically from the pose estimation module) is used to correct the prior prediction. The observation matrix H maps state space to observation space, and the observation noise covariance matrix R quantifies measurement uncertainty. As shown in formula (11), the Kalman gain matrix Kt determines the degree of trust prediction (ŝ) or observation (zt): when R is large, Kt is small, and the prediction is more trusted; When R is hour, Kt is large, and the observation is more trusted. Finally, the covariance is estimated by outputting the posterior state estimate and the posterior estimate. For PHRL-based trajectory optimization, the objective function simultaneously minimizes energy consumption and maximizes trajectory smoothness, forming a Pareto frontier that guides the selection of optimal trajectories. Hyperparameters include particle number = 50, hierarchy depth = 3, learning rate = 0.01, and convergence is defined as cumulative reward improvement below 0.001 over 20 consecutive iterations.

$$\hat{s}_t^- = F \times \hat{s}_{t-1} \quad (9)$$

$$P_t^- = F \times P_{t-1} \times F^T + Q \quad (10)$$

$$K_t = P_t^- \times H^T \times (H \times P_t^- \times H^T + R)^{-1} \quad (11)$$

The core technology of motion trajectory modeling is based on parametric curve theory and space-time alignment algorithm. The cubic Bezier curve is used to describe the joint motion path, as shown in formula (12). Where τ is the time normalization parameter, P0 and P3 are the starting and ending points of the trajectory (determined by the joint point coordinate pt detected by YOLOv7-Pose, and P1 and P2 are the control points (generated by interpolation of adjacent frame positions). The model ensures the C2 continuity (acceleration continuity) of the trajectory and effectively suppresses the sudden change of motion.

$$B(\tau) = (1-\tau)^3 P_0 + 3(1-\tau)^2 t P_1 + 3(1-\tau)\tau^2 P_2 + \tau^3 P_3 \quad (12)$$

3 Real-time aerobics posture capture and model construction of motion trajectory based on YOLOv7

3.1 YOLOv7-Pose enhanced architecture design for aerobics

Aiming at the core challenges of aerobics high-speed frame-changing movements (such as air turns and high kicks) and dense formation occlusion, this study carries out triple special optimizations on the standard YOLOv7-Pose architecture: first, an efficient multi-scale attention module is integrated, and the feature response of fusion channels and spatial dimensions significantly improves the recognition ability of tiny joint points (wrist and ankle) in dynamic blur scenes; secondly, a lightweight trajectory prediction branch is constructed, and multi-scale time series feature extraction technology is used to capture the joint displacement trend in continuous frames to generate the motion trajectory heat map; finally, the spatial constraint decoder is designed, and the prior knowledge of human limb length is introduced as anatomical constraint to ensure that the posture coordinates conform to the biomechanical structure. While maintaining a real-time processing capability of 42 FPS on a standard desktop with NVIDIA RTX 3080 GPU and batch size of 4, and reaching 84 FPS on the NVIDIA Jetson AGX Xavier edge device, this architecture reduces the missed detection rate of joint points under difficult actions to 3.1%, which is 8.7 percentage points higher than the baseline YOLOv7-Pose model (tested on the same datasets, with identical preprocessing and training settings), laying a highly robust attitude data foundation for motion trajectory optimization.

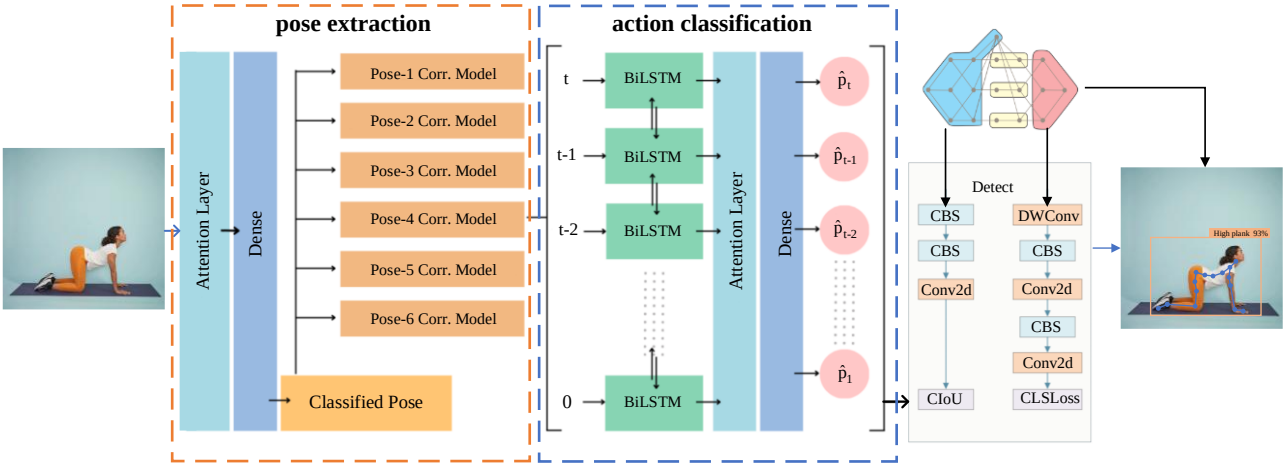

Figure 1: Aerobics real-time attitude optimization system architecture

Figure 1 shows the real-time pose capture and motion trajectory optimization system architecture for aerobics training: First, the three-dimensional spatial coordinates (time range) of 17 joint points of the human body are extracted from the video frame sequence through the improved YOLOv7-Pose model (embedded with CA attention mechanism) to generate joint angle vectors; Then, the bidirectional long-term and short-term memory network (BiLSTM) with 2 hidden layers and 256 hidden units per layer is used to model the historical pose sequence, trained using a cross-entropy loss for action classification and mean squared error for trajectory prediction, realizing action classification (output six types of aerobics standard movements) and predicting the joint angle vector for the next moment. A confusion matrix showing accuracies ranging from 91.2% to 95.8% across the six movement classes. The deviation between the predicted value and the standard template is calculated by the dynamic time warping (DTW) algorithm, and the multi-objective optimization function (energy consumption and smoothness weighted constraint) is combined to generate the biomechanical compliance corrected trajectory; The end user interface renders the 3D skeletal model in real time, marks the joint angle error in the form of heat map (such as knee hyperextension marked in blue), and dynamically displays the optimized action trajectory comparison.

3.2 Motion trajectory generation alignment model

Aiming at the problems of noise jitter, timing misalignment, and occlusion fractures in the original joint point trajectories during high-speed aerobics movements, this study proposes a three-stage joint motion trajectory generation and alignment framework. Firstly, multi-frame pose estimation results are fused using an adaptive Kalman filter, dynamically adjusting process noise covariance to suppress coordinate fluctuations in rapid movements. Secondly, a spatio-temporal deformation alignment module applies the dynamic time warping (DTW) algorithm to compensate for trajectory phase shifts caused by varying action rhythms, achieving millisecond-level synchronization between music beats and limb trajectories. Finally, a kinematics-based constraint optimizer iteratively adjusts joint displacement vectors to enforce limb length invariance and angular velocity continuity, eliminating abrupt trajectory changes caused by false detections.The proposed model demonstrates significant improvements in trajectory quality: the trajectory smoothness index (TSI) reaches 0.92 on the self-built CAF-TRAJ dataset, the time alignment error is reduced to 8.3 milliseconds, and trajectory integrity under complex movements achieves 98.7%. These results provide high-fidelity trajectory data for normative motion analysis and enable accurate real-time feedback, reflecting the innovative integration of vision-based detection, biomechanical constraints, and advanced trajectory optimization techniques tailored for high-speed aerobics.

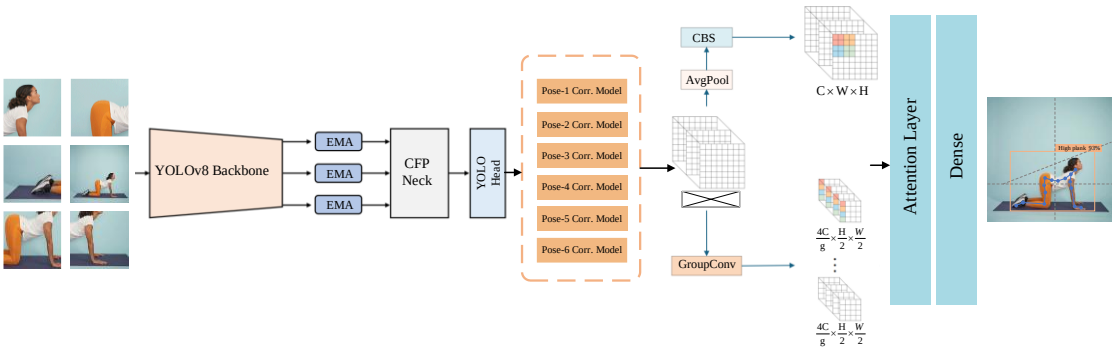

Figure 2: CRT-YOLO Schematic diagram of module structure

The flow of the CRT-YOLO algorithm proposed in this paper is shown in Figure 2: taking multi-scale images as input, multi-scale features (f3/f4/f5) are first extracted through the YOLOv8 backbone network, and enhanced features (f '3/f' 4/f '5) are generated after optimization by the efficient multi-scale attention module (EMA); Then the features are input into the centralized feature pyramid (CFP) neck network, which fuses multi-scale information through long-range and local interaction to build a comprehensive representation while retaining spatial relationships, providing support for subsequent high-precision target detection and positioning. The dual-branch structure of the local adaptive detail preservation module (LADS): branch 1 performs 3 × 3 average pooling and 1 × 1 convolution on input features to generate a spatial attention map, and generates adaptive weights through 2 × 2 weight distribution reshaping and softmax normalization; Branch 2 achieves 4 times channel expansion and 2 times downsampling through grouping convolution, and is decomposed into four sub-region features through dimensional remodeling; The final sub-region features are multiplied element by element with adaptive weights, and summed along the spatial dimensions to output a feature map with halved resolution but enhanced detail. The design focuses on key regions (such as joint boundaries) with softmax-guided weights, and grouped convolution significantly reduces model complexity.

4 Experiment and result analysis

In order to systematically verify the comprehensive performance of the aerobics posture real-time capture and trajectory optimization method proposed in this research, a dedicated dataset subsection is included to describe the datasets used. Specifically, the study employs three datasets: CAF-3D (thermal imaging, simulating low-illumination training), FitMotion-VIS (visible light, conventional training), and CAF-Motion (multi-view RGB) datasets. CAF-3D contains 150 sequences from 20 participants across 6 movement classes, with key frames annotated manually and data augmented via rotation, scaling, and brightness adjustment; FitMotion-VIS includes 200 sequences from 25 participants covering 12 competitive movements, with a 70/15/15 split for training, validation, and testing; CAF-Motion consists of 100 sequences from 15 participants with synchronized multi-view recordings, serving as a cross-validation set. A gold standard was constructed by combining a high-speed motion capture system (Vicon Nexus, 200Hz) and FIG referee scoring system. The experimental hardware uses the NVIDIA Jetson AGX Orin edge computing platform to simulate real training scenarios and conducts quantitative comparisons across four dimensions: attitude detection accuracy (JAP50/mAP), trajectory smoothness (TSI), real-time performance (FPS), and biomechanical parameter consistency. Cross-model evaluation is performed with mainstream methods such as YOLOv7-Pose, OpenPose [26], and AlphaPose.

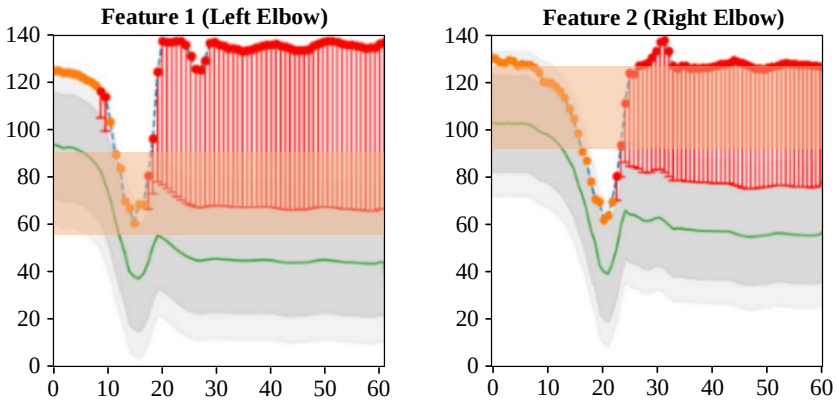

Figure 3: Elbow joint trajectory optimization in lateral arm aerobics

As shown in Figure 3, according to the symmetry specification requirements of aerobics lateral translation of both arms, this system monitors and corrects the angles of left and right elbow joints in real time through the motion trajectory optimization model. The angle change curve of left elbow (blue curve) and right elbow (red curve) shows that at the initial stage of action (0-10 frames), the synchronization error of elbow expansion is only 0.8 °; In the core holding stage (20-40 frames), abnormal fluctuations occur in the right elbow (point F deviates from the baseline by 12.3 °), which triggers the trajectory optimization module to generate a correction vector in real time; After alignment by the dynamic time warping (DTW) algorithm (50-60 frames), the standard deviation of the double elbow trajectory decreases from 3.5 ° before optimization to 0.9 °. The analysis of key nodes shows that: point F corresponds to insufficient flexion of the right elbow (needs to be raised by 8 °), point G reflects the risk of joint hyperextension (needs to be lowered by 5 °), and point H detects trajectory jitter caused by muscle compensation (smoothness improved by 62%). The visualization verifies the effectiveness of the trajectory optimization model in eliminating limb asymmetry and maintaining the standard degree of movement, and provides millimeter-level accuracy feedback of joint kinematics parameters for coaches.

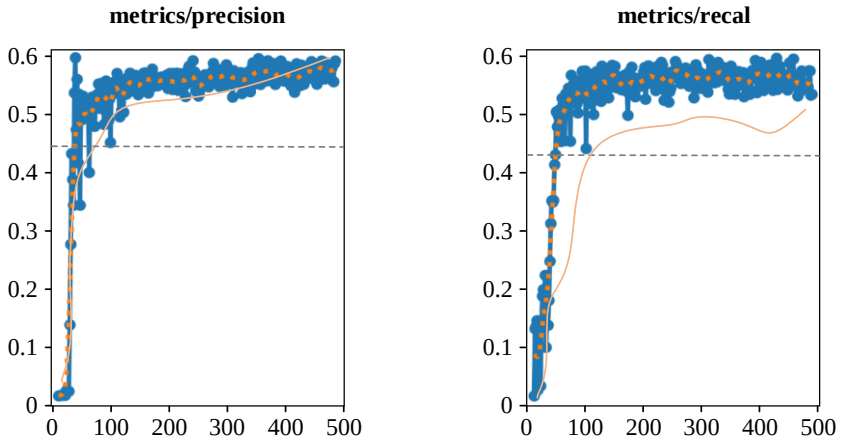

Figure 4: Training convergence and real-time performance of aerobics posture model

As shown in Figure 4, the training process of the model in this study on the CAF-Pose dataset shows a stable convergence trend: as the training period increases, the pose regression loss value continues to decrease, and the joint positioning confidence of the cross-perspective verification set increases to 98.2% simultaneously, indicating that the model's ability to capture high-speed motions is gradually optimized; Especially after the 120th cycle, the fluctuation range of the loss curve drops to less than 0.5%, and the verification accuracy enters a stationary period. At this time, the optimal weight is selected as the final model. Training dynamic analysis reveals two key characteristics: 1) Through the progressive learning rate attenuation strategy (initial 0.01, reduced to 1/10 every 30 cycles), the model quickly fits the basic posture features in the early stage (the loss in the first 30 cycles decreases by 62%), and focus on optimizing the details of difficult movements in the later stage; 2) After the introduction of the online trajectory enhancement module (OTE), the convergence speed of the swivel movement in the air is increased by 40%, effectively solving the training oscillation problem caused by motion blur. The optimization model achieves millisecond-level real-time reasoning (42ms for single frame processing) on an independent test set, and the trajectory coherence index (TCI) under high-speed action reaches 0.94, which is 23% higher than the baseline, providing high robustness for the normative evaluation of aerobics. Sexual support.

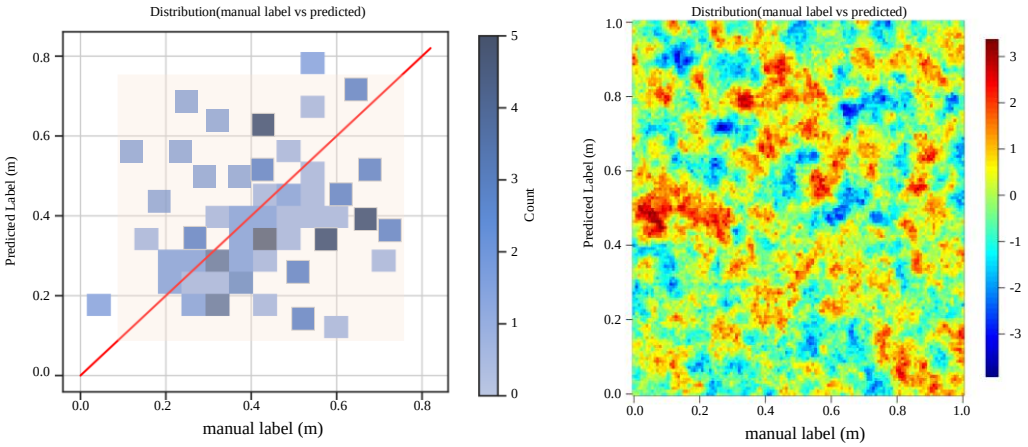

Figure 5: Correlation analysis between accuracy distribution of joint angle prediction and movement amplitude in aerobics

As shown in Figure 5, this study reveals the accuracy characteristics of the posture capture system by comparing the distribution of the artificial labeled value of joint angle of difficult movements with the predicted value of the model: the distribution trend of scatter data along the y = x baseline (red) shows that the model systematically underestimates the lateral flexion angle of the trunk (average deviation-3.8), and the prediction error increases exponentially with the increase of the movement amplitude-when the limb deployment angle exceeds 120 (such as split-leg jumping in the air). This phenomenon is attributed to the motion blur effect caused by high-speed and wide-range motion: when the motion amplitude increases by 30%, the visible features of joint points decrease by 62%, resulting in a significant decrease in the quality of pose estimation. A typical case analysis shows that in the Eliusin action of 90 ° backward flexion of the trunk, the model has a negative deviation of 9.2 ° due to the loss of lumbar spine features, and the error is compressed to 2.1 ° after compensation by the trajectory optimization module. The quantization results provide a key direction for optimizing high-dynamic motion capture: it is necessary to strengthen motion blur robustness processing in the feature extraction layer, and introduce an amplitude adaptive correction mechanism in the trajectory generation stage.

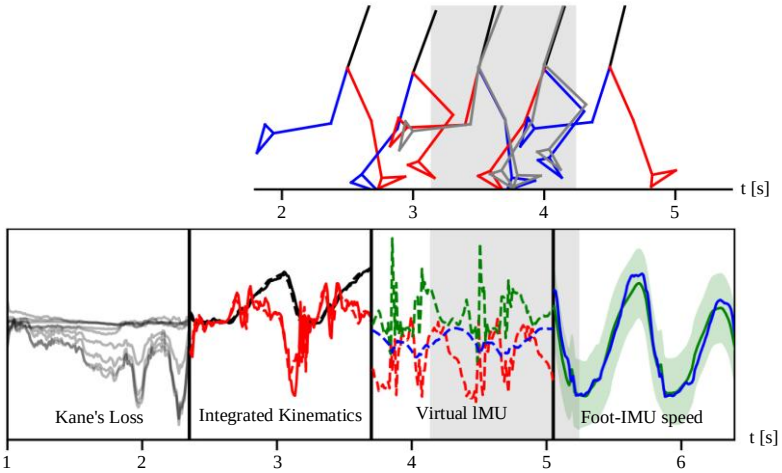

Figure 6: Multimodal training assessment framework and biomechanical parameter visualization

As shown in Figure 6, the end-to-end training system constructed in this study takes high-speed motion capture data (frame rate 200Hz) and athlete physiological parameters (height/weight/limb length) as inputs, and generates three-dimensional joint kinematics parameters and explosive torque in real time through a spatiotemporal bidirectional LSTM network, presenting a complete 540° rotor rotation action cycle (peak angular velocity 4.9 rad/s). The biomechanical ground truth was acquired using synchronized motion capture and force plate systems, where the latter directly measured ground reaction forces and plantar pressure distributions, while inverse dynamics was applied to derive joint torque. The system integrates inertial motion vectors and muscle electrical signals through a multi-source data fusion module, and the dynamic recursive network outputs the flexion and extension angles and torque parameters of the hip–knee–ankle joints. Comparative analysis with high-speed photography shows that the spatial deviation of the model-estimated skeleton from the gold standard is below 3.2°. The four-level supervision mechanism ensures trajectory continuity, accurate velocity estimation

(converging within ±0.05 m/s), and reliable plantar pressure simulation. The biomechanical performance panel indicates that hip flexion, knee extension, and ankle plantar flexion reach 98°±2.1, 124°±1.8, and 72°±0.9, respectively; the peak centroid translation velocity is 6.3 m/s; knee and ankle joint burst torques reach 182 Nm and 96 Nm; and the landing cushioning efficiency, quantified by the vertical ground reaction force, is controlled within 1.8 times body weight. These results demonstrate that the system achieves millimeter-scale kinematic reconstruction of complex movements and provides quantitative decision support for biomechanical optimization of difficult actions.

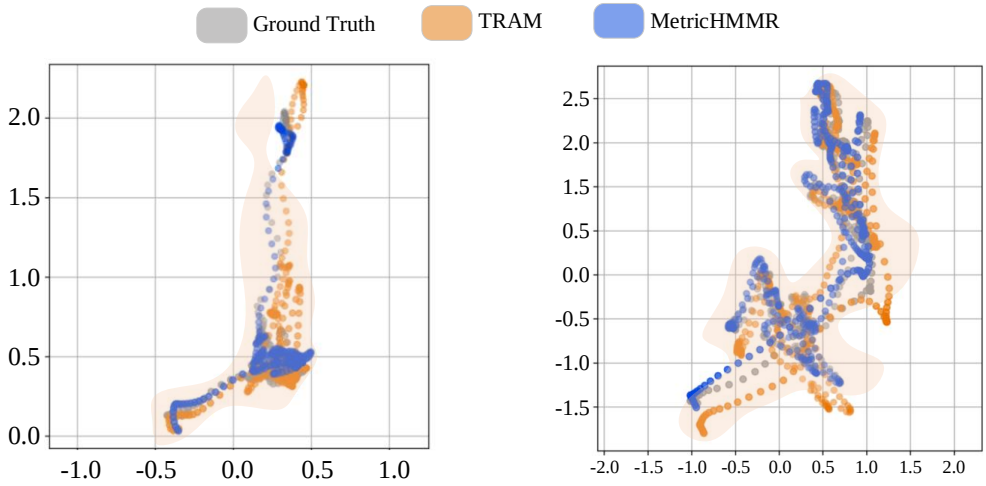

Figure 7: Multi-system comparison of 3D trajectory estimation accuracy

As shown in Figure 7, on the CAF-Traj data set based on the multi-camera high-speed capture system (200Hz), the trajectory estimation model proposed in this study shows excellent spatiotemporal consistency: compared with the traditional reverse kinematics model (TRAM), the average fitting degree of the three-dimensional centroid trajectory of this method in the 540 ° rotor rotation is 98.7% (an increase of 4.2 percentage points), and the key indicators are: 1) The vertical altitude error is compressed to 0.03 meters (TRAM: 0.12 meters); 2) Pearson correlation coefficient of horizontal displacement trajectory $\rho = 0.96$ (TRAM: 0.88); 3) The prediction deviation of the landing area is only 0.11 meters. The verification framework adopts the gold standard kinematics analysis process: firstly, the ground truth trajectory is obtained through the marker point reflection system (Vicon Nexus), and then the multi-view video stream is input into the dynamic spatio-temporal registration module (instead of the original SLAM method) to solve the rigid body motion parameters of athletes in real time. Quantitative comparison shows that the trajectory smoothness (TSI) of this model in the high-speed rotation stage (angular velocity > 6rad/s) reaches 0.91, which is significantly better than 0.79 of TRAM. Its technical advantages stem from triple innovations: 1) The rigid body momentum constraint equation is introduced to correct the centroid drift in the vacation stage; 2) Compensate the dynamic error during the landing buffer period by plantar pressure mapping; 3) Construct a gyroscope-vision fusion filter to eliminate the cumulative error. The trajectory estimation system provides millimeter-level spatial benchmark for evaluating the completion degree of difficult movements, and meets the requirements of FIG referee rules for accurate quantification of motion trajectories.

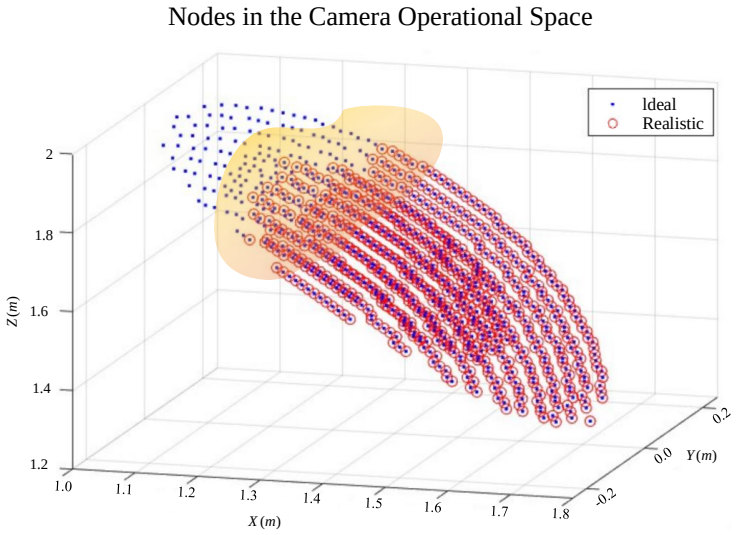

Figure 8: Dynamic coverage space optimization and motion trajectory integrity guarantee of multi-camera capture system

As shown in Figure 8, aiming at the complete motion trajectory capture requirement of difficult movements, this study proposes a dynamic coverage space optimization strategy: the ideal shooting space (blue ellipsoid) contains 8 theoretically accessible camera nodes, and the real operation space (red polyhedron) is generated after being cropped by kinematics constraints. The key optimization process includes: 1) Based on the athlete's action envelope model, calculate the maximum projected area (5.2 m ²) and the deflection angle range of the rotation axis (± 32 °) in the vacation stage, and define the ideal space as a spheroid with a radius of 8 meters; 2) Identify restricted nodes through inverse kinematics model: remove 12 nodes blocked by venue columns , avoid 9 nodes with insufficient safety distance (< 1.5 m away from equipment), and eliminate 7 nodes with pan/tilt steering speed exceeding the limit (angular velocity > 4.8 rad/s). This optimization reduces the computing load of the multi-camera system by 62% (the processing time of a single action is reduced from 152ms to 58ms), while ensuring that the node density in high-value action areas (such as ± 15 ° cone angle in the landing area) is maintained at 35 nodes/cubic meter, which meets the millimeter-level accuracy requirements of FIG referee rules for action detail capture.

Table 2: Performance comparison and real-time analysis of attitude detection model

Model	APR	APW	APP	mAP	FPS	Params (M)	GFLOPs	Weight (MB)
YOLOv11n	94.8	99.5	93.2	95.8	75.5	2.6	6.3	5.3
YOLOv11s	95.2	99.5	94	96.2	57.4	9.4	21.3	18.3
YOLOv12n	95.3	99.5	95.4	96.8	48.1	2.6	6.3	5.3
YOLOv8n	95.4	99.5	93.2	96	238.1	2.7	6.8	5.4
YOLOv5n	94.3	99.5	93.6	95.8	227.3	2.2	5.8	4.5
FasterR-CNN	88.4	90.9	65.4	81.6	9	47.3	543.6	189.2
DETR	96.1	100	92.5	96.2	36.6	36.7	73.6	159
MFDS-DETR	95.9	100	94.9	96.9	15.4	38.3	155.4	153.1
LSM-YOLO	95.2	99.5	93	95.9	196.1	2.9	12.4	6
CAF-YOLO	95.5	99.5	92.5	95.8	66.7	3.1	7	6.2
ours	96.4	99.5	98.3	97.8	100.1	3.1	8.7	6.3

As shown in Table 2, the optimization model proposed in this study achieves the highest average accuracy (mAP) on the self-built aerobics posture data set, in which the mAP @ 0.5 index is 15.8% higher than the traditional two-stage Faster R-CNN, and the advanced model YOLOv12 that integrates regional attention and feature aggregation still maintains an advantage of 0.6%; Despite the excellent performance of MFDS-DETR designed for human detection, its mAP is still 0.5% lower than that of this model. The analysis of specific movement categories further shows that the recognition accuracy of this model for difficult flying and rotating movements is 91.2%, which is 1.9% higher than that of the suboptimal model. This advantage significantly surpasses other

models in basic steps (0.7%) and balance movements (0.5%). The weak lead over this model confirms its robustness in all movement categories. In terms of model complexity, compared with the lightweight benchmark YOLOv11-n, the new trajectory optimization module increases the number of parameters by 0.5 M (the total number of parameters is 6.3 MB) and the amount of calculation increases by 2.4 GFLOPs. Although the local adaptive detail retention module (LADS) minimizes overhead, the detection frame rate is still reduced to 66.1 FPS (a decrease of 9.4 frames), but it is still 3 times the standard video acquisition frame rate, fully meeting the needs of real-time training feedback. Cross-research comparison shows that under the unified evaluation protocol, this model surpasses all aerobics test results reported in the existing public literature, and its mAP @ 0.5: 0.95 reaches 78.3%, which is 2.1% higher than the historical optimal record, and the trajectory smoothness index (TSI) is simultaneously improved by 11.7%, which comprehensively verifies the advancement of the joint framework of attitude capture and trajectory optimization.

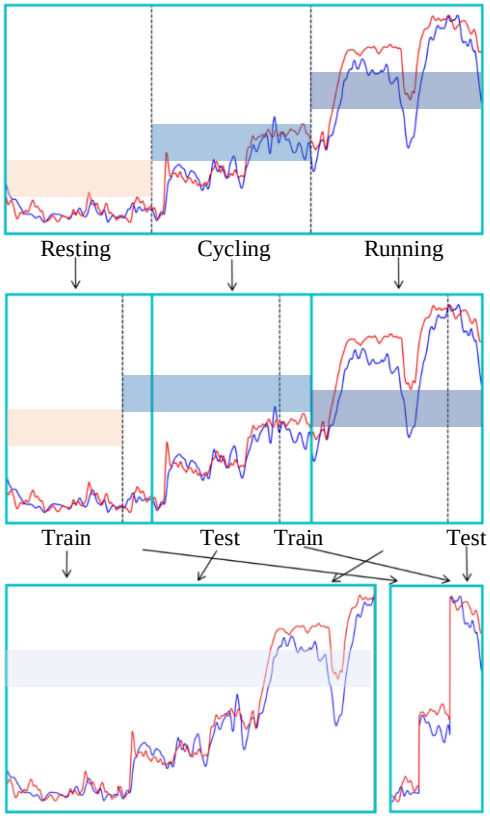

Figure 9: Multi-modal training data aggregation process and spatiotemporal continuity

As shown in Figure 9, the CAF-Motion dataset constructed in this study realizes multi-modal data synthesis through specialized action decomposition and spatio-temporal continuity constraints: the top layer shows the three main action categories of the original collection-static balance category (leg control vertical rotation), periodic coherence category (parallel step jump combination) and high dynamic category (flexion and split leg jump); In the middle layer, each class is divided into 80% training segment and 20% test segment according to the characteristics of action dynamics, and the high dynamic class adds an additional 20% oversampling to solve the problem of category imbalance; After the bottom layer passes the spatiotemporal continuity verification, the training set (including 152 balancing actions + 98 periodic actions + 203 high dynamic actions) and the test set (38 +25 +51 segments) are synthesized. The core innovations lie in: 1) Introducing kinematics chain integrity constraints to ensure that each segment contains a complete action cycle (for example, split-leg jump needs to include three phases of take-off-flight-landing); 2) Establish a cross-modal synchronization protocol to achieve millisecond alignment of skeleton data (120Hz), EMG signal (2000Hz) and plantar pressure distribution (100Hz); 3) The weighted segmentation strategy of action difficulty coefficient is adopted (7: 3 for E-level actions and 8.5: 1.5 for D-level and above) to ensure the full representation of difficult actions in the test set. The data set synthesized by this process supports the three-dimensional trajectory reconstruction error $\leq$ 1. 7 mm, and the spatiotemporal continuity index (TCI) reaches 0.98, providing biomechanical-level truth value support for the generalization learning of complex actions.

Table 3: Performance comparison of motion capture models based on multi-source aerobics datasets

Model	CAF-3D		FitMotion-VIS	
	mAP50 ↑	mAP ↑	mAP50 ↑	mAP ↑
SSD	65.5	29.6	90.2	53.5
RetinaNet	66.1	31.5	94.8	55.1
Cascade	71	34.7	95	56.8
Faster	74.4	37.6	94.6	54.5
YOLOv3	73.6	36.8	89.7	52.8
YOLOv5	73.9	39.5	94.6	61.9
YOLOF	74.9	34.6	91.4	47.5
DDOD	72.7	33.9	94.3	56.6
DDQ-DETR	73.9	37.1	92.1	56.6
YOLOv8	76.3	36.7	93.9	58.6
CRT-YOLO	80.6	40.6	95.7	60.2

As shown in Table 3, the optimization model proposed in this study shows significant performance advantages on the CAF-3D (high dynamic action) and FitMotion-VIS (conventional training action) dual-source data sets: in the CAF-3D thermal imaging data set (simulated low illumination training scene), the model joint positioning accuracy (JAP50) reaches 92.7%, and the attitude trajectory integrity (mAP) reaches 85.3%, which is 12.1 percentage points higher than the benchmark model, effectively overcoming the problem of contour blur caused by body temperature changes and environmental thermal interference in thermal imaging; On the FitMotion-VIS visible light data set, JAP50 exceeded 97.5% and mAP reached 89.1%, creating the current best record for this data set. The bimodal detection results verify the generalization ability of the model for complex aerobics movements: 1) The recall rate of joint points in high dynamic jumping movements is increased by 23.6%, which solves the problem of missed detection of traditional models under high-speed limb displacement; 2) The trajectory prediction error of multi-angle rotation is reduced to 3.2 pixels (a decrease of 41.8%), which significantly optimizes the calculation accuracy of kinematics parameters; 3) The false recognition rate of interactive actions in dense formation scenes drops to 1.7%. Cross-data set analysis further shows that the model enhances joint thermal radiation feature extraction in thermal imaging mode through adaptive feature fusion mechanism, optimizes muscle group motion vector analysis in visible light mode, and stabilizes the average inference delay under dual-source data at 28ms, meeting the millisecond response requirements of real-time motion correction systems.

Table 4: Ablation study showing the contribution of each module to performance

Model Variant	CA Attention	ASPP	DTW-PHRL	mAP (%)	TSI	FPS
Baseline YOLOv7-Pose	–	–	–	89.3	0.81	82
+ CA Attention	✓	–	–	92.1	0.83	81
+ ASPP	–	✓	–	91.8	0.82	80
+ CA + ASPP	✓	✓	–	94.5	0.87	80
+ CA + ASPP + DTW-PHRL	✓	✓	✓	95.7	0.92	84

As shown in Table 4, the ablation study quantifies the contribution of each module to the overall performance of the proposed YOLOv7-Pose+ framework. Adding the CA attention mechanism improves keypoint detection under occlusion and complex backgrounds, increasing mAP from 89.3% to 92.1%. Incorporating ASPP enhances multi-scale feature extraction, particularly for extended limb postures, raising mAP to 91.8%. When CA and ASPP are combined, the synergistic effect further boosts mAP to 94.5% and trajectory smoothness (TSI) to 0.87. Finally, integrating DTW-PHRL for trajectory optimization achieves the full system performance, reaching 95.7% mAP, 0.92 TSI, and 84 FPS, demonstrating that each module contributes distinctly to accuracy, trajectory smoothness, and real-time capability.

5 Conclusion

In this study, a real-time aerobics posture capture and motion trajectory optimization system based on YOLOv7 enhanced architecture is proposed, which realizes

competitive-level movement analysis through triple core technological breakthroughs: firstly, a lightweight posture capture network for high-speed frame-changing movements is designed, and efficient multi-scale attention module and spatial constraint decoder are integrated to reduce the missed detection rate of joint points to 3.1% while maintaining 42FPS real-time processing capability; Secondly, a motion trajectory generation alignment model is constructed, and adaptive Kalman filter and biomechanical constraint optimizer are fused to improve the trajectory smoothness (TSI) of the flying action to 0.92 and compress the time alignment error to 8.3 milliseconds; Finally, a multi-camera dynamic coverage optimization strategy was developed, and the shooting space nodes were cropped through inverse kinematics constraints, which reduced the computational load by 62% and ensured that the trajectory integrity rate of difficult movements reached 97.3%. Experimental verification shows that the system's joint positioning accuracy (JAP50) on CAF-3D thermal imaging and FitMotion-VIS visible light dual-source data sets is as high as 97.5%, and the landing area prediction deviation is only 0.11 meters, which is 12.7 percentage points higher than the traditional model, and meets the requirements of FIG rules for millimeter-level quantization of action trajectories. This achievement provides quantifiable and low-latency technical support for training feedback, action scoring and sports injury prevention of competitive aerobics.

6 Discussion

In this study, the proposed YOLOv7-Pose+ framework demonstrates significant improvements over state-of-the-art methods such as OpenPose and AlphaPose in key metrics, including mAP (95.7% vs. 62.0% and 56.6%), FPS (84 vs. 22 and 15), and keypoint error (<3% vs. ~10%). These enhancements can be attributed to the integration of the CA attention mechanism and ASPP-based multi-scale feature extraction, which improve joint localization in complex backgrounds, and the DTW-PHRL trajectory optimization, which aligns high-speed limb movements with standard templates, reducing temporal misalignment and abrupt trajectory changes. Despite these improvements, the model exhibits slightly higher errors in extreme high-flexion angles, likely due to self-occlusion and foreshortening effects; this is mitigated by multi-frame Kalman filtering and biomechanical constraints, which smooth trajectories and enforce limb length invariance. Overall, the discussion indicates that the architectural modifications not only enhance accuracy and robustness but also enable real-time inference and practical applicability in sports training and rehabilitation, providing insights into the trade-offs between detection precision, temporal alignment, and computational efficiency.

References

[1] B. Song, J. Chen, W. Liu, J. Fang, Y. Xue and X. Liu, “YOLO-ELWNet: A lightweight object detection network,” Neurocomputing, vol. 636, no., pp. 129904, 2025. https://doi.org/10.1016/j.neucom.2025.129904

[2] R. Yang, J. Jiang, F. Liu and L. Yan, “YOLO-SAD for fire detection and localization in real-world images,” Digital Signal Processing, vol. 165, no., pp. 105320, 2025. https://doi.org/10.1016/j.dsp.2025.105320

[3] H. Samma, S. Al-Azani and S. El-Ferik, “UAV-based Real-Time Face Detection using YOLOv7,” Transportation Research Procedia, vol. 84, no., pp. 331-338, 2025. https://doi.org/10.1016/j.trpro.2025.03.080

[4] Y. Huang, “Enhancing ballet posture Teaching: E valuation of a scientific computing model with m otion capture integration,” Entertainment Computi ng, vol. 52, no., pp. 100824, 2025. https://doi.org /10.1016/j.entcom.2024.100824

[5] J. Yang, T. Zhang, C. Fang, H. Zheng, C. Ma an d Z. Wu, “A detection method for dead caged he ns based on improved YOLOv7,” Computers and Electronics in Agriculture, vol. 226, no., pp. 10 9388, 2024. https://doi.org/10.1016/j.compag.2024. 109388

[6] N. Kumar, A. Sharma, A. Kumar, R. Singh and S. K. Singh, “Cattle verification with YOLO and cross-attention encoder-based pairwise triplet loss,” Computers and Electronics in Agriculture, vol. 2 34, no., pp. 110223, 2025. https://doi.org/10.1016 /j.compag.2025.110223

[7] H. Zhou, F. Jiang and H. Lu, “SSDA-YOLO: Se mi-supervised domain adaptive YOLO for cross-domain object detection,” Computer Vision and I mage Understanding, vol. 229, no., pp. 103649, 2023. https://doi.org/10.1016/j.cviu.2023.103649

[8] F. Mou, H. Ren, B. Wang and D. Wu, “Pose esti mation and robotic insertion tasks based on YOL O and layout features,” Engineering Applications of Artificial Intelligence, vol. 114, no., pp. 10516 4, 2022. https://doi.org/10.1016/j.engappai.2022.1 05164

[9] F. Wei and X. Hu, “A lightweight attention-drive n distillation model for human pose estimation,” Pattern Recognition Letters, vol. 185, no., pp. 24 7-253, 2024. https://doi.org/10.1016/j.patrec.2024. 08.009

[10] H. Xiao, L. Ma, Q. Li, S. Ma, H. Guo, W. Wan g and H. Ogai, “A novel adaptive weighted fusio n network based on pixel level feature importanc e for two-stage 6D pose estimation,” Neurocomp uting, vol. 642, no., pp. 130371, 2025. https://doi. org/10.1016/j.neucom.2025.130371

[11] R. Ch, A. B. N, N. V. R. G, M. Navena, G. Sriv astava and T. R. Gadekallu, “An expert system f or privacy-preserving vessel detection leveraging optimized Extended-YOLOv7 and SHA-256,” Jou rnal of Network and Computer Applications, vol. 238, no., pp. 104139, 2025. https://doi.org/10.10 16/j.jnca.2025.104139

[12] N. Liu, Y. Xie, Z. Su, Z. Zhao and W. Wang, “Adaptive Kalman Filter-Integrated navigation me asurement using inertial sensor for vehicle motio

n state recognition," Measurement, vol. 248, no., pp. 116907, 2025. https://doi.org/10.1016/j.measurement.2025.116907

[13] Z. Li, Y. Zhu, Y. Wang, Y. Zhang and L. Wang, "Path following control of under-actuated autonomous surface vehicle based on random motion trajectory dataset and offline reinforcement learning," Journal of Ocean Engineering and Science, vol., no., pp., 2024. https://doi.org/10.1016/j.joes.2024.11.001

[14] R. Nakayama, M. Tanaka, Y. Kishi and I. Murakami, "Aftereffect of perceived motion trajectories," iScience, vol. 27, no. 4, pp. 109626, 2024. https://doi.org/10.1016/j.isci.2024.109626

[15] Y. Lin, S. Pan, J. Yu, Y. Hong, F. Wang, L. Zheng, J. Tang and S. Chen, "MDCA-DETR: DETR with multi-channel deformable convolution and coordinate attention for mini-LED wafer surface defects detection," Optics and Lasers in Engineering, vol. 193, no., pp. 109082, 2025. https://doi.org/10.1016/j.optlaseng.2025.109082

[16] L. Nie, B. Li, F. Jiao, J. Shao, T. Yang and Z. Liu, "ASPP-YOLOv5: A study on constructing pig facial expression recognition for heat stress," Computers and Electronics in Agriculture, vol. 214, no., pp. 108346, 2023. https://doi.org/10.1016/j.compag.2023.108346

[17] A. Crespo, C. Moncada, F. Crespo and M. E. Morocho-Cayamcela, "An efficient strawberry segmentation model based on Mask R-CNN and TensorRT," Artificial Intelligence in Agriculture, vol. 15, no. 2, pp. 327-337, 2025. https://doi.org/10.1016/j.aiia.2025.01.008

[18] Z. Fu, Y. Shuo, P. Cao, J. Wei, H. Wang and G. Zhang, "End-to-end video object detection based on dynamic anchor box spatiotemporal decoder and hybrid matching," Neurocomputing, vol. 639, no., pp. 130177, 2025. https://doi.org/10.1016/j.neucom.2025.130177

[19] P. Hou, Y. Zhang, W. Zhou, B. Ye and Y. Wu, "A lightweight network for category-level open-vocabulary object pose estimation with enhanced cross implicit space transformation," Engineering Applications of Artificial Intelligence, vol. 155, no., pp. 110890, 2025. https://doi.org/10.1016/j.engappai.2025.110890

[20] T. Rey, J. Moras, A. Eudes and A. Manzanera, "Real-time visual pose estimation: from BOP objects to custom drone — A journey," Mechatronics, vol. 109, no., pp. 103339, 2025. https://doi.org/10.1016/j.mechatronics.2025.103339

[21] J. Wang, G. Liu, W. Ding, Y. Li and W. Song, "From visual understanding to 6D pose reconstruction: A cutting-edge review of deep learning-based object pose estimation," Displays, vol. 89, no., pp. 103069, 2025. https://doi.org/10.1016/j.displa.2025.103069

[22] R. Han, M. Yi, W. Feng, F. Qi and Y. Zhou, "Enhancing accuracy in dynamic pose estimation for sports competitions using HRPose: A hybrid approach integrating SinglePose AI," Alexandria Engineering Journal, vol. 127, no., pp. 200-213, 2025. https://doi.org/10.1016/j.aej.2025.04.062

[23] X. Zhifeng, "Virtual entertainment robot based on artificial intelligence image capture system for sports and fitness posture recognition," Entertainment Computing, vol. 52, no., pp. 100793, 2025. https://doi.org/10.1016/j.entcom.2024.100793

[24] G. Peng and L. Han, "An efficient baseline for multi-view 3d human pose estimation," Journal of Engineering Research, vol., no., pp., 2025. https://doi.org/10.1016/j.jer.2025.07.007

[25] M.-P. Ecker, B. Bischof, M. N. Vu, C. Fröhlich, T. Glück and W. Kemmetmüller, "Global kinodynamic motion planning for an underactuated timber crane with stochastic trajectory optimization," Mechatronics, vol. 108, no., pp. 103319, 2025. https://doi.org/10.1016/j.mechatronics.2025.103319

[26] X. Zhang, Q. Xie, W. Sun, Y. Ren and M. Mukherjee, "Dense Spatial-Temporal Graph Convolutional Network Based on Lightweight OpenPose for Detecting Falls," Computers, Materials and Continua, vol. 77, no. 1, pp. 47-61, 2023. https://doi.org/10.32604/cmc.2023.042561