International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3693-3704
Journal homepage: http://iieta.org/journals/ts

A Real-Time System for Athlete Pose Analysis and Feedback Based on Machine Learning

Xiaoqian Peng[1], Xujiang Mao[2], Xiaomin Fang[3*]

[1] School of Culture and Tourism, Quzhou College of Technology, Quzhou 324000, China
[2] Zhejiang Institute of Sports Science (Zhejiang Anti-Doping Center), Hangzhou 30014, China
[3] Department of Information Engineering, Quzhou College of Technology, Quzhou, 324000, China

Corresponding Author Email: fxm_1985@126.com

https://doi.org/10.18280/ts.420649

Received: 3 April 2025
Revised: 25 September 2025
Accepted: 15 December 2025
Available online: 31 December 2025

Keywords:
sports image processing, real-time 3D pose estimation, spatiotemporal feature modeling, self-supervised learning, lightweight neural networks, dynamic time warping

ABSTRACT

Accurate and real-time analysis of athlete posture is an important topic in sports-related image processing. In practical training environments, pose analysis systems are expected to operate under strict real-time constraints while remaining robust to occlusion, motion blur, and scene variation. However, existing approaches face several limitations. 0054hree-dimensional pose estimation often depends on large amounts of annotated data, which are expensive and difficult to obtain. Lightweight models designed for real-time inference tend to sacrifice spatiotemporal feature representation, leading to reduced accuracy. In addition, current feedback mechanisms are usually loosely connected to the underlying pose features and therefore provide limited diagnostic value. To address these issues, a real-time pose analysis and feedback framework based on self-supervised spatiotemporal optimization is presented. The system adopts a three-stage architecture consisting of a lightweight image feature extraction and two-dimensional keypoint detection module, a dual-path spatiotemporal feature refinement module, and a sequence-based feedback generation module. The refinement stage combines adaptive graph convolution for skeletal topology modeling with a lightweight spatiotemporal Transformer for learning temporal image features. Temporal coherence across video frames is exploited to construct self-supervised constraints for three-dimensional pose learning without manual annotations. Pose sequences are further matched with standard motion templates using dynamic time warping, and the resulting deviations are translated into structured feedback. The proposed framework reduces the dependence on annotated data, maintains real-time performance on edge devices, and provides interpretable feedback linked directly to pose deviations. Experimental results demonstrate that the system achieves a balanced trade-off between efficiency, accuracy, and practical usability in real training scenarios.

1. INTRODUCTION

Sports image processing, as an intersection of computer vision and sports engineering [1, 2], is primarily concerned with accurately extracting human pose features from images and video data. This area has gradually evolved into an independent research direction due to the increasing availability of video sensors, wearable devices, and computational resources in sports environments. Technological breakthroughs in this field are crucial for optimizing sports training and preventing sports injuries, and also serve as a key support for the large-scale implementation of computer vision technology in the sports domain [3-5]. With the iteration of image processing technology, athlete pose estimation has made multidimensional advancements: from single-frame 2D pose detection to multi-frame 3D video pose modeling [6, 7], from traditional manual feature extraction to deep image feature learning [8, 9], and from laboratory-controlled environments to real sports scenarios [10]. These developments have gradually moved pose analysis from offline laboratory studies toward online and in-situ training assistance. The interference factors commonly present in real sports scenarios, such as complex lighting, target occlusion, and rapid movement, further highlight the need for the evolution of pose analysis technology, particularly in terms of robustness and adaptability.

From the perspective of image processing, current pose analysis technology still faces three major bottlenecks. First, there is an inherent contradiction between the demand for real-time processing and the integrity of feature extraction. Lightweight models simplify the image feature extraction process to adapt to edge device inference efficiency, leading to the loss of key pose information, which ultimately affects estimation accuracy [11-13]. This trade-off becomes especially evident in high-speed or occluded motion scenes. Second, 3D pose modeling heavily relies on large-scale annotated image and video data. The annotation process for such data is time-consuming, labor-intensive, and costly, while the scene limitations of annotated data result in insufficient model generalization [14, 15]. In many practical sports applications, collecting and annotating large-scale high-quality datasets remains infeasible. Third, the pose deviation diagnosis process lacks deep exploration of image sequence temporal features. Current feedback suggestions mainly rely on empirical rules [16, 17], which are disconnected from the pose feature deviations in the image processing layer [18], making it difficult to achieve accurate technical guidance and limiting their practical value in training support.

The core goal of this study is to build an athlete pose analysis and improvement framework based on image processing technology, achieving collaborative optimization of real-time performance, robustness, and practicality. Rather than focusing on a single aspect such as accuracy or speed alone, the intention is to construct a system that can be deployed in real training environments. The focus is on overcoming core challenges in image spatiotemporal feature modeling and unsupervised optimization in complex sports scenarios. To achieve this goal, three key scientific issues must be addressed: first, how to design a lightweight image feature extraction architecture that fully retains key pose features while ensuring real-time inference performance on edge devices; second, how to use the temporal coherence of image and video data to construct effective self-supervised signals, enabling 3D pose feature optimization and robust modeling driven by unlabeled data; third, how to establish a fine-grained feature matching mechanism between pose image sequences and standard paradigms, enabling accurate diagnosis and interpretable feedback based on pose feature deviations in the image processing layer.

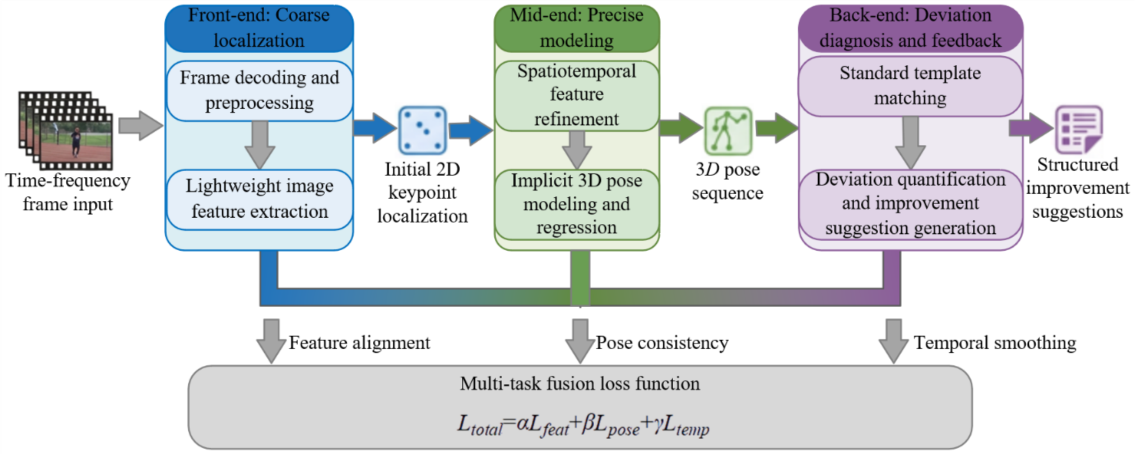
The innovations and core contributions of this study can be summarized in three aspects: at the technical architecture level, a self-supervised spatiotemporal optimization-based real-time pose analysis and feedback network is developed with a three-level cascaded image processing framework. This framework achieves end-to-end optimization of fast image feature extraction, spatiotemporal feature refinement, and sequence feature matching, effectively balancing real-time inference efficiency and pose estimation accuracy. At the method design level, a dual-path feature learning architecture is constructed using adaptive graph convolution and a spatiotemporal Transformer, respectively extracting skeletal topology image features and raw video temporal image features, improving the robustness of pose modeling in complex scenarios. A self-supervised fusion strategy based on temporal coherence is also introduced, achieving spatiotemporal optimization of image features without 3D annotated data. At the application level, a mechanism linking image sequence feature matching with natural language feedback is established, transforming the quantized results of pose deviations in image processing into understandable, structured improvement suggestions, which supports more practical and interpretable training assistance.

The subsequent chapters of this paper are organized as follows: Chapter 2 provides a detailed explanation of the core design of the proposed three-level cascaded framework, including the structural details, parameter settings, and collaborative mechanisms of each image processing module; Chapter 3 verifies the framework's performance through experiments on multiple datasets, conducting an evaluation of image processing performance in terms of real-time performance, accuracy, and robustness, and verifying the effectiveness of core modules through ablation experiments; Chapter 4 discusses the academic value of the research, its comparative advantages over existing studies, as well as current limitations and future research directions; the final chapter summarizes the research content and core conclusions of the entire study.

2. ATHLETE POSE ANALYSIS AND IMPROVEMENT FRAMEWORK USING SSTO-RAFN

2.1 Framework architecture and image processing flow

The SSTO-RAFN framework is driven by image processing and adopts a "coarse to fine" three-level cascaded architecture design. The core goal is to achieve end-to-end processing from video frame input to structured improvement suggestion output, while balancing real-time performance, robustness, and estimation accuracy. This design is motivated by the practical requirement that the system should operate reliably under real training conditions rather than in controlled laboratory environments alone. The framework architecture is shown in Figure 1. This architecture design follows the progressive optimization logic of image processing: the frontend completes rapid image feature extraction and coarse localization through lightweight modules, the middle layer achieves precise pose modeling through spatiotemporal feature refinement, and the backend performs deviation diagnosis and feedback generation through sequence matching. The three-level modules form a closed loop through feature transmission and collaborative training, effectively avoiding extreme trade-offs between real-time performance and accuracy in a single module. Compared to traditional segmented processing architectures, this end-to-end design reduces information loss during feature transmission and enables cross-module joint optimization.



Time-frequency frame input
Front-end: Coarse localization
Frame decoding and preprocessing
Lightweight image feature extraction
Initial 2D keypoint localization
Mid-end: Precise modeling
Spatiotemporal feature refinement
Implicit 3D pose modeling and regression
3D pose sequence
Back-end: Deviation diagnosis and feedback
Standard template matching
Deviation quantification and improvement suggestion generation
Structured improvement suggestions
Feature alignment
Pose consistency
Temporal smoothing
Multi-task fusion loss function
$L_{total}=\alpha L_{feat}+\beta L_{pose}+\gamma L_{temp}$

Figure 1. SSTO-RAFN framework overall architecture and image processing flow

The complete image processing flow of the framework can be summarized in seven key steps, forming a continuous data processing chain: first, the input video stream undergoes frame decoding and preprocessing, completing image normalization and format conversion; then, the lightweight image feature extraction module encodes the preprocessed video frames, simultaneously outputting 2D keypoint initial localization results; based on the initial localization of 2D keypoint coordinates and original image features, the spatiotemporal feature refinement module suppresses noise and enhances features, thereby enabling implicit 3D pose modeling and coordinate regression; the generated 3D pose sequence is converted into a standardized image sequence feature representation, which is finely matched with the preset motion mode standard paradigm; finally, through deviation quantification analysis and causal language model decoding, structured improvement suggestions are generated. In the entire process, the flow of image data and feature transformation revolves around the precise extraction and efficient utilization of pose information, ensuring that the processing delay and accuracy requirements of each step are met. This pipeline is designed to support both single-subject and multi-subject scenarios without altering the underlying processing structure.

To achieve collaborative optimization of the three-level modules, the framework uses an end-to-end training strategy and designs a multi-task fusion loss function set. This training strategy allows the parameters of each module to be updated jointly, thereby improving global consistency across the pipeline. Its core expression is as follows:

$$L_{total}=\alpha L_{feat}+\beta L_{pose}+\gamma L_{temp} \qquad (1)$$

where, L_{feat} is the image feature matching loss, which constrains the feature alignment between different modules. It uses cosine similarity loss to measure the difference between extracted features and real pose features; L_{pose} is the pose consistency loss, which uses mean squared error to constrain the deviation between the 3D pose coordinates and the reference values; L_{temp} is the temporal smoothing loss, which is constructed based on the first-order differences of adjacent frame poses and is used to suppress abnormal fluctuations in motion sequences. α, β, γ are the loss weight coefficients, set as 0.2, 0.6, and 0.2, respectively, through cross-validation, ensuring a balance between accuracy goals and temporal stability during training.

2.2 Lightweight feature extraction and 2D keypoint detection

The lightweight image feature extraction and 2D keypoint detection module serves as the frontend component of the SSTO-RAFN framework. Its primary objective is to extract discriminative pose-related image features under strict real-time constraints imposed by edge devices, while simultaneously providing reliable initial localization of key joints for subsequent three-dimensional pose estimation. The specific framework structure is shown in Figure 2. The performance of this module directly affects the overall inference speed of the framework and also sets an upper bound on the accuracy achievable by the subsequent refinement stages. Therefore, the design emphasizes a careful balance between feature completeness and computational efficiency in order to avoid the loss of critical pose information caused by excessive model simplification.

The module is based on YOLO-Pose and is further optimized through lightweight architectural modifications. The main optimizations are conducted along three aspects: backbone network design, multi-scale feature fusion, and input resolution adaptation. For backbone optimization, MobileNetV4 is adopted as the feature extraction network, where standard convolutions are replaced by depthwise separable convolutions, reducing computational complexity to approximately one third of the original design. In addition, a channel attention mechanism is introduced to strengthen the response of channels associated with key joints by adaptively reweighting feature maps. The channel weighting is formulated as $w_c=\sigma(F_{avg}(x_c))$, where σ denotes the Sigmoid activation function, F_{avg} represents global average pooling, and x_c is the feature map of the c-th channel.

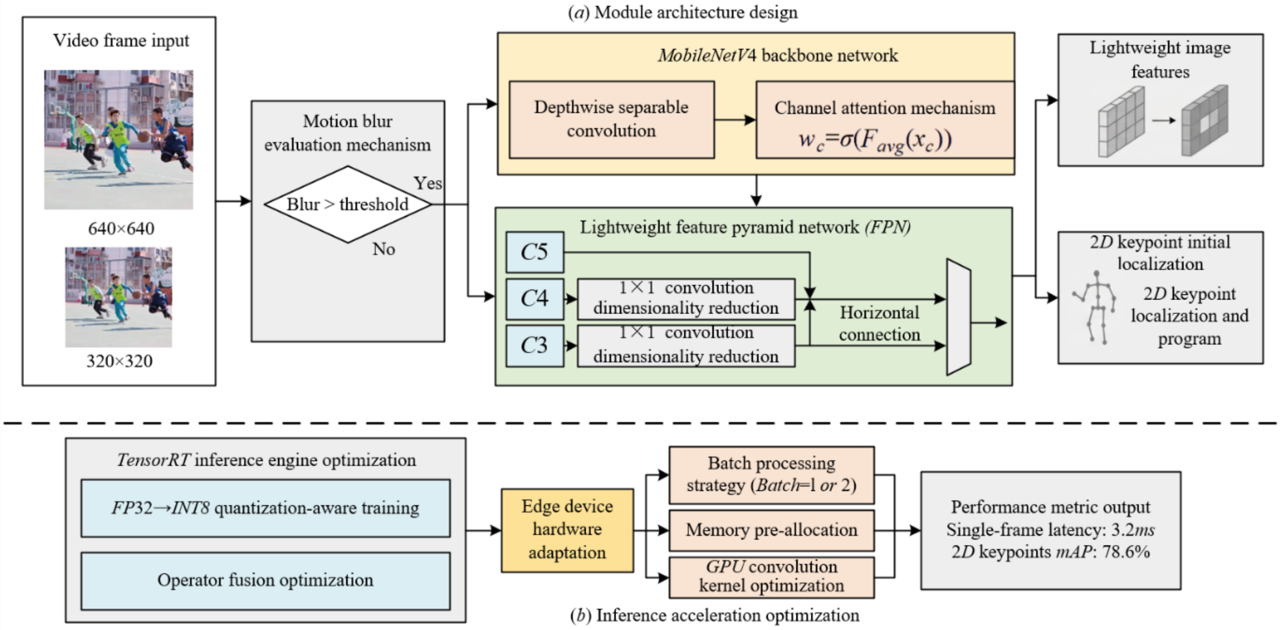

(a) Module architecture design
Video frame input
640×640
320×320
Motion blur evaluation mechanism
Blur > threshold
Yes
No
MobileNetV4 backbone network
Depthwise separable convolution
Channel attention mechanism
$w_c=\sigma(F_{avg}(x_c))$
Lightweight image features
Lightweight feature pyramid network (FPN)
C5
C4
C3
1×1 convolution dimensionality reduction
1×1 convolution dimensionality reduction
Horizontal connection
2D keypoint initial localization
2D keypoint localization and program
TensorRT inference engine optimization
FP32→INT8 quantization-aware training
Operator fusion optimization
Edge device hardware adaptation
Batch processing strategy (Batch=1 or 2)
Memory pre-allocation
GPU convolution kernel optimization
Performance metric output
Single-frame latency: 3.2ms
2D keypoints mAP: 78.6%
(b) Inference acceleration optimization

Figure 2. Lightweight feature extraction and 2D keypoint detection

For multi-scale feature fusion, a lightweight feature pyramid network is employed. Features from layers C3, C4, and C5 are combined through top-down pathways and lateral connections, corresponding to small-, medium-, and large-scale joint representations. This design improves the detection of small joints such as wrists and ankles. During fusion, 1×1 convolutions are applied for channel reduction to prevent an increase in computational cost. For input adaptation, a motion blur estimation mechanism is introduced. The blur level is quantified by computing the variance of image gradient magnitudes. When the estimated blur exceeds a predefined threshold, the input resolution is reduced from 640×640 to 320×320 to improve efficiency; when blur is low, higher resolution is used to maintain accuracy.

To further improve inference efficiency on edge devices, several optimization strategies are applied. Using TensorRT, the trained model is converted into an optimized inference engine, and quantization-aware training is employed to reduce model weights from 32-bit floating point to 8-bit integer precision. This leads to an approximate 2.5× speed-up while keeping the accuracy loss within 3%. Operator fusion is applied to combine convolution, pooling, and related operations, thereby reducing kernel invocation overhead and memory access latency. Batch sizes are dynamically adjusted between 1 and 2 based on the available device memory, and memory pre-allocation is used to reduce data transfer overhead. Additional kernel-level optimizations are performed to improve GPU utilization. Experimental results indicate that the single-frame processing latency on the Jetson Xavier NX is approximately 3.2 ms, and the average 2D keypoint detection accuracy reaches 78.6%, satisfying the requirements of both real-time performance and practical accuracy.

2.3 Dual-path spatiotemporal feature refinement

The dual-path spatiotemporal image feature refinement module constitutes a central component of the SSTO-RAFN framework. It is designed to mitigate the effects of occlusion and motion blur in complex motion scenarios, while enabling three-dimensional pose refinement using unlabeled data. The specific framework structure is shown in Figure 3. The module adopts a parallel dual-branch architecture that extracts pose-related information from two complementary perspectives: skeletal topology and temporal image context. This complementary representation improves robustness under challenging visual conditions.

A self-supervised fusion stage integrates the features from both branches and produces a refined three-dimensional pose representation that balances spatial accuracy and temporal stability. Compared with single-branch modeling strategies, the dual-path design allows information from different feature domains to compensate for each other, thereby reducing the sensitivity of pose estimation to missing or degraded observations in complex scenes.

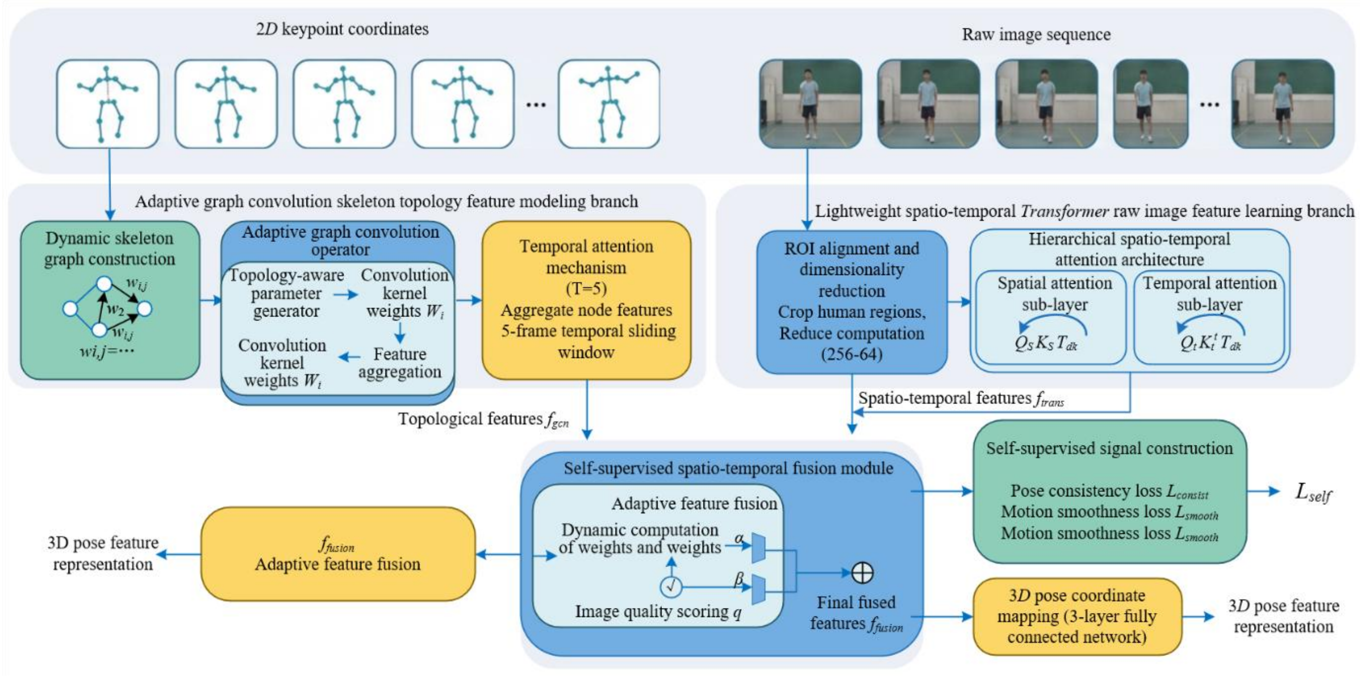

2D keypoint coordinates
Raw image sequence
...
...
Adaptive graph convolution skeleton topology feature modeling branch
Dynamic skeleton graph construction
$w_{i,j}$
w_2
$w_{i,j}$
$wi,j=\cdots$
Adaptive graph convolution operator
Topology-aware parameter generator
Convolution kernel weights W_i
Convolution kernel weights W_i
Feature aggregation
Temporal attention mechanism (T=5) Aggregate node features 5-frame temporal sliding window
Lightweight spatio-temporal Transformer raw image feature learning branch
ROI alignment and dimensionality reduction Crop human regions, Reduce computation (256-64)
Hierarchical spatio-temporal attention architecture
Spatial attention sub-layer
$Q_S K_S T_{dk}$
Temporal attention sub-layer
$Q_t K_t^t T_{dk}$
Spatio-temporal features f_{trans}
Topological features f_{gcn}
Self-supervised spatio-temporal fusion module
Adaptive feature fusion
Dynamic computation of weights and weights
α
β
√
Image quality scoring q
⊕
Final fused features f_{fusion}
Self-supervised signal construction
Pose consistency loss $L_{consist}$
Motion smoothness loss L_{smooth}
Motion smoothness loss L_{smooth}
L_{self}
f_{fusion}
Adaptive feature fusion
3D pose feature representation
3D pose coordinate mapping (3-layer fully connected network)
3D pose feature representation

Figure 3. Dual-path spatiotemporal feature refinement architecture

2.3.1 Adaptive graph convolution for skeletal topology feature modeling branch

The adaptive graph convolution branch focuses on dynamic feature modeling of the human skeletal topology structure. The core innovation lies in constructing a graph structure and convolution kernel parameters that dynamically adjust based on the positional relationships of the keypoints in the image, overcoming the robustness limitations of traditional fixed topology graph convolution in occlusion scenarios. This branch is based on the 2D keypoint coordinates output from the lightweight module. First, a dynamic human skeleton graph is constructed: each keypoint is defined as a node in the graph, and edges are adaptively constructed based on anatomical constraints and inter-frame keypoint distances. The edge weights are jointly determined by the Euclidean distance between nodes and the pose confidence, as expressed by the following formula:

$$w_{i,j}=\frac{\exp\left(-d_{i,j}^{2}/\sigma^{2}\right)(c_i\cdot c_j)}{\sum_{k\in N(i)}\exp\left(-d_{i,k}^{2}/\sigma^{2}\right)(c_i\cdot c_k)} \tag{2}$$

where, $d_{i,j}$ is the Euclidean distance between keypoints i and j,

σ is the distance decay coefficient, c_i is the confidence of keypoint i, and $N(i)$ is the neighborhood set of node i. The dynamic edge weights effectively weaken the interference of occluded keypoints on the overall topology feature, enhancing the graph structure's adaptability to pose changes.

Based on the dynamic skeleton graph, an adaptive graph convolution operator is designed to aggregate features. The core idea is to dynamically adjust the convolution kernel parameters based on the local topology structure. Traditional graph convolutions use fixed weight matrices, which struggle to adapt to topological changes under different poses. In this study, we introduce a topology-aware parameter generator, mapping the local neighborhood's topological features into convolution kernel weights, as expressed by the following formula:

$$W_i = F_\theta(x_i^{topo}) \tag{3}$$

$$x_i' = \sum_{j \in N(i)} w_{ij} \cdot W_i \cdot x_j \tag{4}$$

where, x_i^{topo} is the local topological feature vector of node i, F_θ is the parameter generator, W_i is the adaptively generated convolution kernel weight, and x_i' is the aggregated node feature. To enhance the stability of features in the temporal dimension, a temporal sliding window of length 5 frames is introduced. A temporal attention mechanism aggregates multi-frame skeletal features. The temporal attention weights are determined by the pose similarity between frames and image clarity, giving higher weights to clear frames and consecutive pose frames, effectively suppressing motion blur and frame-to-frame abrupt noise interference, completing and smoothing the features of occluded keypoints.

2.3.2 Lightweight spatiotemporal transformer for raw image feature learning branch

The lightweight spatiotemporal Transformer branch aims to extract deep contextual features from the temporal sequence of raw images, directly learning the 3D spatial constraints between keypoints, thus avoiding the information loss in the traditional 2D-to-3D conversion process. To meet the real-time requirements of edge devices, this branch adopts multiple lightweight optimization strategies to reduce computational complexity while ensuring feature representation capability. First, ROI (Region of Interest) alignment is performed to crop the human body region from the raw image feature map, processing only the body region's features, which reduces computation by more than 60% compared to full image processing. Then, 1×1 convolution is used to reduce the feature channel count from 256 to 64, further compressing the computational cost.

In terms of attention mechanism design, a hierarchical spatiotemporal attention architecture is used, divided into spatial attention sub-layers and temporal attention sub-layers. The spatial attention sub-layer focuses on modeling the spatial correlation of human keypoints within a single frame. Through self-attention mechanisms, long-range dependencies between keypoints are captured, enhancing the feature response of key joints. The temporal attention sub-layer focuses on modeling the temporal correlation between multiple frames. Through cross-attention mechanisms, adjacent frames' human features are aligned, and the temporal regularities of pose changes are extracted. The core computations for hierarchical attention are as follows:

$$A_{spatial} = Softmax\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right) \tag{5}$$

$$A_{temporal} = Softmax\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) \tag{6}$$

where, Q_s, K_s are the query and key matrices for spatial attention, Q_t, K_t are the query and key matrices for temporal attention, and d_k is the feature dimension. Through the hierarchical attention mechanism, the module can simultaneously capture the spatial constraints of keypoints and the temporal motion trends, enabling implicit modeling of 3D poses. Compared to traditional 3D pose estimation methods, this branch does not rely on 2D keypoint dimensionality conversion. Instead, it directly learns 3D spatial relationships from the raw image temporal features, effectively preserving depth information and contextual constraints in the image.

2.3.3 Self-supervised spatiotemporal fusion module

The core function of the self-supervised spatiotemporal fusion module is to achieve collaborative optimization of the dual-path features while utilizing the temporal coherence of video data to construct self-supervised signals, reducing reliance on 3D labeled data. This module mainly consists of two core units: self-supervised signal construction and adaptive feature fusion, which enhance the dual-path features through loss constraints and dynamic weight allocation.

In the construction of self-supervised signals, two constraints are designed based on the temporal coherence of video sequences: pose consistency constraint and motion smoothness regularization. The pose consistency constraint is achieved by calculating the cosine similarity of the dual-path features between adjacent frames, requiring the pose features to remain stable in consecutive frames. The loss function is:

$$L_{consist} = 1 - \frac{1}{T\text{-}1} \sum_{t=1}^{T\text{-}1} cos\,(f_t f_{t+1}) \tag{7}$$

where, T is the number of frames in the temporal window, f_t is the fused feature at frame t, and $cos(\ )$ is the cosine similarity function. Motion smoothness regularization is implemented by constraining the first-order differences of pose features between adjacent frames, avoiding abnormal jumps in pose changes. The loss function is:

$$L_{smooth} = \frac{1}{T\text{-}1} \sum_{t=1}^{T\text{-}1} \| \boldsymbol{f}_t\text{-}\boldsymbol{f}_{t+1} \|_2^2 \tag{8}$$

The final self-supervised loss is the weighted sum of the two losses: $L_{self} = \lambda_1 L_{consist} + \lambda_2 L_{smooth}$, where λ_1 and λ_2 are set to 0.6 and 0.4, respectively, based on cross-validation to balance stability and flexibility.

In terms of adaptive feature fusion, a dynamic weight allocation mechanism based on image quality assessment is designed, adjusting the weight ratio of dual-path features according to the clarity and occlusion degree of the current frame. Image clarity is quantified by the variance of the gradient magnitude calculated by the Laplacian operator, and occlusion degree is assessed by the confidence distribution of 2D keypoint detection. Both are normalized and weighted to obtain the image quality score q. The fusion weight calculation for the dual-path features is:

$$\alpha=\frac{g+\epsilon}{2+\epsilon},\ \beta=1\text{-}\alpha \tag{9}$$

$$f_{fusion}=\alpha\cdot f_{gen}+\beta\cdot f_{trans} \tag{10}$$

where, ϵ is the smoothing coefficient, α and β are the weights of the adaptive graph convolution branch and the spatiotemporal Transformer branch, f_{gcn} and f_{trans} are the output features of the dual-path branches, and f_{fusion} is the final fused feature. This mechanism can increase the weight of raw image features when the image quality is good and enhance the weight of skeletal topology features in occluded or blurred scenes, ensuring the robustness of the fused features. The fused features are then mapped to 3D pose coordinates using a three-layer fully connected network, completing the transformation from image features to 3D pose modeling.

2.4 Intelligent feedback generation module based on image sequence matching

The intelligent feedback generation module based on image sequence matching is a key unit of the SSTO-RAFN framework for practical application. Its core goal is to convert the 3D pose feature sequence output by the dual-path refinement module into precise, understandable, and structured improvement suggestions. The specific framework structure is shown in Figure 4. This module constructs a motion pattern image sequence knowledge base to achieve fine-grained feature matching between the current pose and the standard paradigm, and combines the quantified deviation information to drive the language model to generate feedback, forming a complete "feature matching - deviation quantification - feedback generation" link. Compared to traditional rule-driven feedback mechanisms, this module better adapts to pose variations in complex motion scenarios, improving feedback precision and generalization.

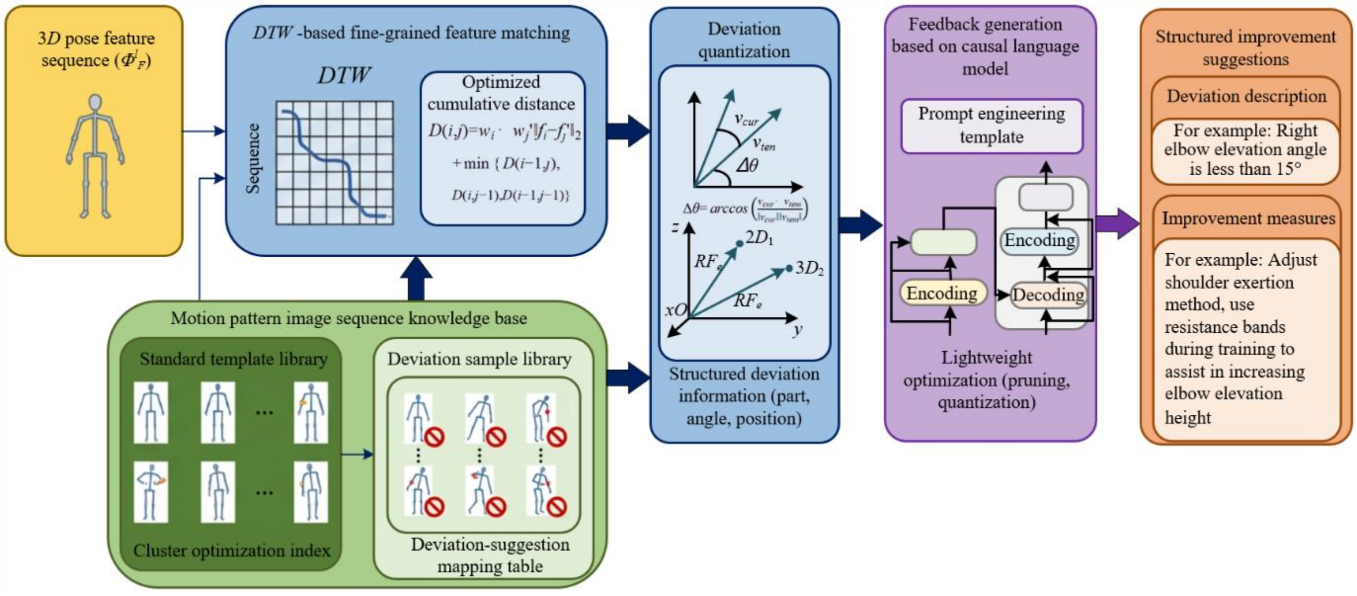

3D pose feature sequence (Φ^l_F)
DTW -based fine-grained feature matching
DTW
Sequence
Optimized cumulative distance
$D(i,j)=w_i\cdot w_j'\|f_i-f_j'\|_2$ $+\min\{D(i-1,j),$ $D(i,j-1),D(i-1,j-1)\}$
Deviation quantization
v_{cur}
v_{ten}
$\Delta\theta$
$\Delta\theta=arccos\left(\frac{v_{cur}\cdot v_{tem}}{\|v_{cur}\|\|v_{tem}\|}\right)$
z
$2D_1$
RF_e
$3D_2$
RF_e
xO
y
Structured deviation information (part, angle, position)
Feedback generation based on causal language model
Prompt engineering template
Encoding
Encoding
Decoding
Lightweight optimization (pruning, quantization)
Structured improvement suggestions
Deviation description
For example: Right elbow elevation angle is less than 15°
Improvement measures
For example: Adjust shoulder exertion method, use resistance bands during training to assist in increasing elbow elevation height
Motion pattern image sequence knowledge base
Standard template library
...
...
Cluster optimization index
Deviation sample library
Deviation-suggestion mapping table

Figure 4. Intelligent feedback generation module based on image sequence matching

2.4.1 Construction of motion pattern image sequence knowledge base

The core function of the motion pattern image sequence knowledge base is to provide standardized pose feature templates and deviation-suggestion mapping benchmarks. Its construction must balance the representativeness of the templates with the efficiency of subsequent matching. The knowledge base uses a "standard template + deviation sample" dual-library architecture: the standard template library is constructed by collecting videos of professional athletes performing different sports motions. After frame decoding and pose normalization, 3D pose features output by the SSTO-RAFN dual-path fusion module are extracted as standard templates, with each motion corresponding to 10-15 sequences from different athletes to cover individual differences. The deviation sample library is constructed by collecting common error motion videos from novice athletes, annotating key deviation types, and establishing a correspondence between deviation types and 3D pose feature differences. At the same time, motion training experts annotate corresponding improvement suggestions, forming a structured "deviation feature - improvement suggestion" mapping table.

To improve the efficiency of subsequent sequence matching, the image sequence features in the knowledge base are clustered. The K-means clustering algorithm is used to group the feature sequences in the standard template library, with cosine similarity between sequences as the clustering criterion. The number of clusters is determined by the elbow method. After clustering, an index is established for each group, storing the mean and variance of the features within the group as representative templates. In subsequent matching, the most similar feature group is first located through fast indexing, and fine-grained matching is performed within the group, reducing the matching time complexity from $O(N)$ to $O(logN)$. At the same time, a dynamic update mechanism for the knowledge base is established, integrating new motion templates and deviation samples through incremental learning, ensuring the knowledge base's adaptability to different motion scenarios. During the update process, feature alignment strategies are used to ensure the consistency of new and old samples.

2.4.2 Fine-grained image sequence feature matching based on DTW

DTW is a classic method for matching sequences of unequal lengths. This module optimizes the distance calculation of DTW and the feature preprocessing strategy to achieve fine-grained matching between the current pose sequence and the knowledge base template, accurately quantifying pose deviations. Feature preprocessing is the premise for improving matching accuracy. First, the current pose sequence and knowledge base template sequence are length-normalized, with linear interpolation used to standardize the sequence length to *L*. Then, Principal Component Analysis is applied to reduce the dimensionality of each frame's 3D pose feature, retaining 95% of the feature variance, and removing redundant information to reduce computational complexity. After dimensionality reduction, the feature dimension is compressed from 128 to 32.

To improve the robustness of matching in complex scenarios, an image quality weight is introduced to optimize the DTW distance calculation function. Based on the image clarity score from the previous section, each frame feature is assigned a weight. Higher weights are given to clear frames, while lower weights are assigned to blurry or occluded frames. These weights are normalized and integrated into the cumulative distance calculation of DTW. The optimized DTW cumulative distance formula is:

$$D(i,j)=w_i \cdot w_j'\|f_i-f_j'\|_2+\min\{D(i-1,j),D(i,j-1),D(i-1,j-1)\} \quad (11)$$

where, f_i is the feature of the *i*-th frame in the current sequence, f_j' is the feature of the *j*-th frame in the template sequence, w_i and w_j' are the quality weights for the current and template frames, and $D(i,j)$ is the cumulative matching distance for the first *i* frames and *j* frames.

After matching, the frame-by-frame feature deviations between the current sequence and the template sequence are calculated based on the optimal alignment path corresponding to the minimum cumulative distance, which is further quantified into joint angle and position deviations. The joint angle deviation is calculated using the vector dot product. For example, the angle deviation $\Delta\theta$ of the elbow joint is the angle difference between the current pose elbow vector and the template vector:

$$\Delta\theta= arccos\left(\frac{v_{cur}\cdot v_{tem}}{\|v_{cur}\|\|v_{tem}\|}\right) \quad (12)$$

where, v_{cur} and v_{tem} are the elbow vectors of the current pose and the template. The position deviation is quantified by the Euclidean distance of the joint's 3D coordinates. The final output contains the deviation location, angle deviation value, and position deviation value as structured deviation information.

2.4.3 Structured feedback generation based on causal language model

The core function of the causal language model is to convert the quantified pose deviation information into structured improvement suggestions that align with the motion training standards. This requires prompt engineering and lightweight optimization to ensure feedback accuracy and real-time processing. The design of prompt engineering is the key to improving feedback quality. A structured prompt template is used to encode the deviation quantification information into an input understandable by the model. The template format is: "Sport: {Sport Name}; Deviation Location: {List of Locations}; Angle Deviation: {List of Deviation Values}; Position Deviation: {List of Deviation Values}; Output Requirement: Generate structured suggestions in the format 'Deviation Description - Improvement Measures' with concise and professional language." This template guides the model to focus on the core deviation information through explicit semantic constraints, avoiding redundant content and ensuring consistent output format.

To adapt to the real-time requirements of edge devices, the base causal language model is lightweight-optimized. A small language model is selected as the base model. Redundant attention heads and fully connected neurons are pruned using structured pruning, with a pruning rate of 30%. Additionally, INT8 quantization is used to convert the model weights from 32-bit floating-point to 8-bit integers. This reduces the model parameter size from 1.2B to 400M while controlling accuracy loss within 5%. The model inference process is further optimized by using batch inference to process multi-frame deviation information. At the same time, the model inference engine is coordinated with the front-end image processing module, using the multi-core CPU of the edge device to parallel-process both image feature matching and language model inference tasks, thereby reducing overall latency.

The optimized feedback generation module has a single-sequence feedback generation delay of only 4.8ms on the Jetson Xavier NX edge device. The output improvement suggestions include two core parts: "Deviation Description" and "Improvement Measures," such as: "Right elbow elevation angle is insufficient by 15°; Improvement measure: Adjust shoulder force application and use resistance bands during training to help increase elbow elevation height." According to evaluations by sports experts, the accuracy of the feedback suggestions is 91%, and the matching degree between the deviation descriptions and the actual image feature deviations is 93%, providing precise guidance for motion training.

3. EXPERIMENTAL EVALUATION

3.1 Experimental setup

The experimental setup is designed to comprehensively evaluate the image processing performance of the SSTO-RAFN framework. The design follows three main dimensions: dataset construction, evaluation metric definition, and experimental environment configuration, in order to ensure the reliability, comparability, and reproducibility of the results. This structured design allows performance to be examined consistently across different scenarios and hardware platforms. The dataset adopts a combination of public benchmark datasets and self-built specialized datasets. The public benchmarks include Human3.6M, 3DPW, MPII, and UCF101-Sports, which are used to evaluate three-dimensional pose estimation accuracy, robustness in real-world scenes, two-dimensional keypoint detection performance, and adaptability to different motion scenarios, respectively.

The self-built datasets focus on athletics sprinting and basketball shooting. High-definition videos are collected from 20 athletes with different skill levels and converted into image sequences. These datasets include two-dimensional and three-dimensional keypoints as well as annotated pose deviation types. In addition, a robustness subset is constructed by varying lighting conditions, introducing partial occlusions, and adjusting shutter speed in order to simulate common sources of visual disturbance in real training environments. All data undergo unified preprocessing, including normalization of image resolution to 640×640 and the removal of brightness offsets. During training, data augmentation techniques such as random cropping, horizontal flipping, and lighting adjustment are applied to increase data diversity. During testing, optical flow-based temporal alignment is used to improve the stability of temporal feature representations.

The evaluation metrics focus on the core aspects of image processing performance and are organized into four categories: real-time performance, accuracy, robustness, and feedback quality. Real-time performance is measured on two edge devices, Jetson Xavier NX and NVIDIA Jetson Orin, by recording full-pipeline latency, frame rate, and the relative time consumption of each processing module. Accuracy metrics include average precision for two-dimensional keypoint detection, mean joint position error for three-dimensional pose estimation, and average pose error. In addition, recall rates under occlusion and motion blur conditions are reported. Robustness is assessed using peak signal-to-noise ratio, structural similarity, and correlation analysis of accuracy, as well as by quantifying performance changes after adding Gaussian noise. Feedback quality is evaluated through sequence matching accuracy and consistency scores obtained from expert blind reviews. Baseline methods are selected from representative state-of-the-art approaches in lightweight pose estimation, self-supervised pose estimation, and sequence matching, and their official recommended parameter settings are used to ensure fair comparison.

The experimental environment is divided into training and deployment stages. Training is conducted on a workstation equipped with an NVIDIA RTX 4090 GPU and an Intel Core i9-13900K CPU, running Ubuntu 22.04 with PyTorch 2.0 and CUDA 11.8. Deployment experiments are performed on edge devices, where inference acceleration is implemented using TensorRT 8.6, and trained models are converted to the ONNX format for deployment. The training process uses the AdamW optimizer with an initial learning rate of 1e-4, combined with a cosine annealing schedule and weight decay for parameter adjustment. The batch size is set to 32, and training is performed for up to 200 epochs with early stopping to prevent overfitting, ensuring both convergence and generalization.

3.2 Core experimental results and analysis

3.2.1 Real-time performance and processing efficiency

Real-time performance and processing efficiency are key requirements for deployment on edge devices. This subsection examines the real-time behavior of SSTO-RAFN using the results reported in Table 1, focusing on three aspects: full-pipeline latency, processing frame rate, and the relative time consumption of individual modules.

Table 1. Real-time performance comparison on edge devices

Device	**Method**	**Single Frame Full Process Latency (ms)**	**Frame Rate (FPS)**	**Image Feature Extraction Proportion (%)**	**Spatiotemporal Refinement Proportion (%)**	**Sequence Matching and Feedback Proportion (%)**
Jetson Xavier NX	Lightweight 3D Pose	8.7	115	35.6	42.1	22.3
	RT-3D Pose	7.2	139	31.9	38.6	29.5
	SSTO-RAFN(Proposed)	5.1	196	25.5	36.8	37.7
NVIDIA Jetson Orin	Lightweight 3D Pose	4.3	233	34.9	41.5	23.6
	RT-3D Pose	3.5	286	30.8	37.9	31.3
	SSTO-RAFN(Proposed)	2.4	417	24.2	35.1	40.7

As shown in Table 1, SSTO-RAFN achieves lower latency and higher frame rates than the comparison methods on both edge platforms. On Jetson Xavier NX, the average single-frame latency is 5.1 ms and the frame rate reaches 196 FPS, corresponding to a latency reduction of 29.2% and a frame rate increase of 42.0% relative to RT-3D Pose. On Jetson Orin, latency further decreases to 2.4 ms and the frame rate exceeds 400 FPS, indicating that the system meets the real-time requirements of dynamic motion analysis.

The breakdown of time consumption across modules shows that the feature extraction stage accounts for only 24.2%–25.5% of the total processing time, which is lower than that of the comparison methods. This reduction is mainly attributed to the lightweight MobileNetV4 backbone and TensorRT-based quantization, which reduce computational cost while preserving essential pose features. The spatiotemporal refinement stage occupies approximately 35%–37% of the total time, reflecting a balance between representational capacity and efficiency achieved through hierarchical attention and dynamic graph convolution. Although the sequence matching and feedback stage has a relatively higher proportion, the use of clustering-based indexing and model pruning ensures that this stage does not become a bottleneck and does not compromise overall real-time performance.

3.2.2 Pose estimation accuracy and image processing robustness verification

Pose estimation accuracy and robustness are the core guarantees of system reliability. This subsection verifies the performance advantages of SSTO-RAFN in both general and specialized motion scenes through dataset accuracy comparisons (Table 2) and robustness tests, with a focus on the suppression effect of the dual-path spatiotemporal feature refinement module on complex scene interference.

Table 2 shows that SSTO-RAFN outperforms the comparison methods on all datasets: on the Human3.6M dataset, 2D AP reaches 78.6%, and 3D MPJPE drops to 35.7 mm, improving by 4.7 percentage points and 15.6% compared to RT-3D Pose. On the real-world 3DPW dataset, MPJPE is 39.2 mm, which is 16.2% better than the comparison methods, demonstrating stronger adaptability to real-world scenes. On the self-built specialized datasets, SSTO-RAFN's advantage is further highlighted, with 2D AP for the athletics and basketball datasets exceeding 78%, and MPJPE controlled between 40-42 mm, reducing by 17%-20% compared to the comparison methods. This is due to the precise extraction of specialized motion pose features by the dual-path spatiotemporal feature refinement module—Adaptive-GCN captures skeletal topology constraints, while the spatiotemporal Transformer branch explores the raw image temporal context, and their fusion effectively improves specialized pose modeling accuracy.

Table 2. Accuracy metrics comparison on public and self-built datasets

Dataset	Method	*2D AP* (%)	*3D MPJPE* (*mm*)	*3D MAE* (°)
Human3.6M	Lightweight 3D Pose	72.3	48.6	5.2
	RT-3D Pose	75.1	42.3	4.8
	SSTO-RAFN (Proposed)	78.6	35.7	3.9
3DPW	Lightweight 3D Pose	69.8	52.1	5.7
	RT-3D Pose	73.5	46.8	5.1
	SSTO-RAFN (Proposed)	76.9	39.2	4.3
Self-Built Athletics Dataset	Lightweight 3D Pose	68.5	55.3	6.1
	RT-3D Pose	72.4	49.7	5.5
	SSTO-RAFN (Proposed)	79.2	40.1	4.5
	Lightweight 3D Pose	67.9	56.8	6.3
	RT-3D Pose	71.8	51.2	5.7
	SSTO-RAFN (Proposed)	78.5	41.5	4.7

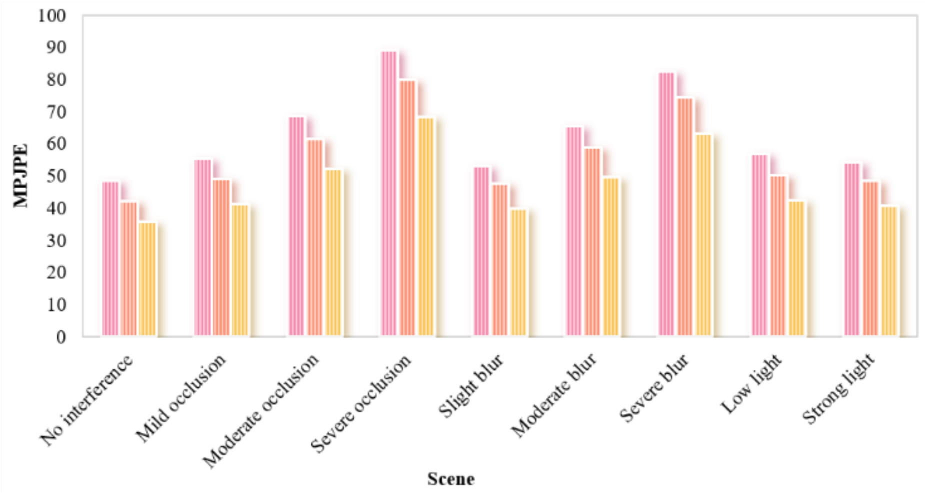


Figure 5. Robustness test results in complex scenes (MPJPE, mm)

Figure 5 shows the robustness test results, indicating that SSTO-RAFN maintains excellent performance in complex scenes such as occlusion, blur, and lighting variation. In heavy occlusion scenes, the MPJPE is 68.5 mm, 14.2% lower than RT-3D Pose; in heavy blur scenes, the MPJPE is 63.2 mm, outperforming the comparison methods by 14.9%; under low and high light conditions, accuracy loss is controlled within 20%, significantly lower than the 25%-30% of comparison methods. This advantage stems from the complementarity of the dual-path features: Adaptive-GCN's dynamic topology modeling weakens the interference of occluded keypoints, and the spatiotemporal Transformer's temporal attention suppresses the feature noise caused by blur and lighting changes. Meanwhile, the self-supervised fusion module's temporal smoothing constraint further enhances feature stability and improves robustness in complex scenes.

3.2.3 Ablation experiments

To verify the necessity and effectiveness of each core image processing module, three sets of ablation experiments were designed to compare the performance differences under different module configurations. The results are shown in Table 3.

Ablation Experiment 1 verifies the effectiveness of the lightweight image feature extraction architecture: when the base model uses the original YOLO-Pose backbone, the 2D AP is only 70.2%. After replacing it with the optimized MobileNetV4 backbone and introducing channel attention, the 2D AP increases to 70.2%, with latency controlled at 3.8 ms, demonstrating the balance between efficiency and feature extraction integrity in the lightweight architecture. Ablation Experiment 2 verifies the necessity of the dual-path spatiotemporal feature refinement module: adding either the Adaptive-GCN or spatiotemporal Transformer branch alone reduces MPJPE to 48.3 mm and 47.6 mm, respectively, reducing by 15%-16% compared to the base model. After collaborating the two branches, MPJPE further reduces to 42.1 mm, improving by 12%-13% compared to the single branch, indicating that the complementarity of the dual-path features significantly enhances pose modeling accuracy. Ablation Experiment 3 verifies the role of the self-supervised fusion module: after adding the self-supervised signal, MPJPE decreases from 42.1 mm to 35.7 mm, and MPJPE without annotation data drops from 69.8 mm to 52.3 mm, a 25.1% reduction. This proves that the self-supervised signal based on temporal coherence effectively improves model accuracy and significantly reduces dependence on annotated data, confirming the effectiveness of the self-supervised strategy.

3.2.4 Image sequence matching and feedback performance verification

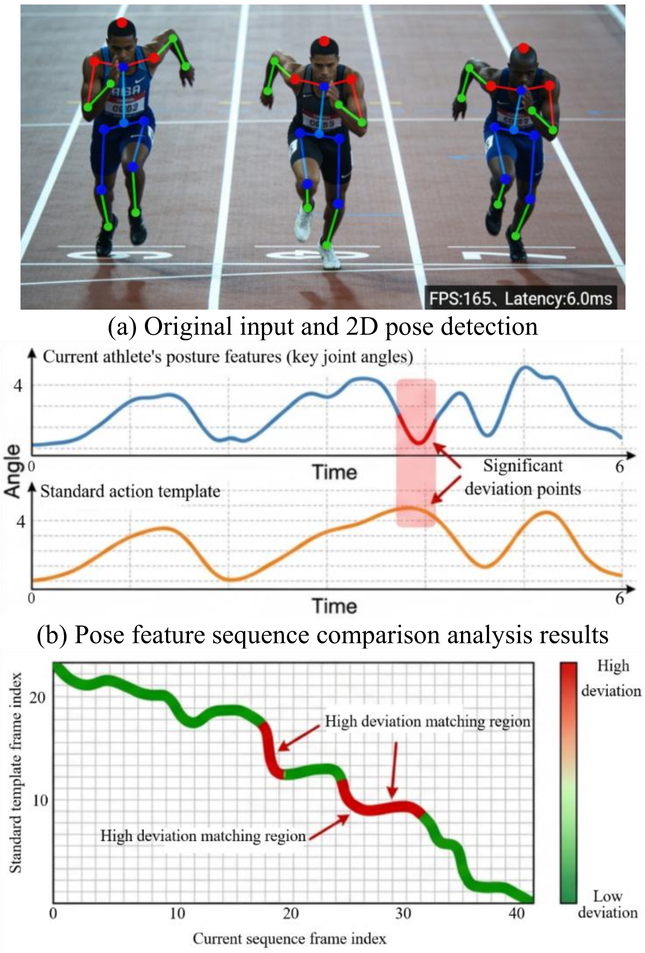
Image sequence matching accuracy and feedback quality are key to the system's practicality. This subsection verifies the advantages of SSTO-RAFN in sequence matching and structured feedback generation based on the experimental data in Table 4.

Table 3. Ablation experiment results

Model Configuration	2D AP (%)	3D MPJPE (mm)	Single Frame Latency (ms)	MPJPE Without Annotation Data (mm)
Base Model (No Refinement and Self-Supervision)	70.2	56.8	3.8	89.5
Base Model + Adaptive-GCN Branch	74.5	48.3	4.5	78.2
Base Model + Spatiotemporal Transformer Branch	75.1	47.6	4.6	76.9
Base Model + Dual-Path Refinement (No Self-Supervision)	77.3	42.1	5.0	69.8
SSTO-RAFN (Dual-Path Refinement + Self-Supervision)	78.6	35.7	5.1	52.3

Table 4. Image sequence matching and feedback performance comparison

Method	Sequence Matching Accuracy (%)	Deviation Quantification Error (mm/°)	Expert Consistency Score (Full Score: 10)	Feedback Generation Latency (ms)
Traditional DTW Matching + Rule-based Feedback	78.3	8.5/1.2	6.8	3.2
SSTO-RAFN (Proposed)	92.6	3.2/0.5	9.1	4.8



(a) Original input and 2D pose detection

(b) Pose feature sequence comparison analysis results

(c) DTW-based matching path visualization

Figure 6. Implementation effect of SSTO-RAFN on real-time athlete pose analysis and improvement system

Table 4 shows that SSTO-RAFN achieves a sequence matching accuracy of 92.6%, an 18.3 percentage point improvement over the traditional DTW method. The deviation quantification error significantly decreases, with position deviation error dropping from 8.5 mm to 3.2 mm and angle deviation error dropping from 1.2° to 0.5°. This improvement is attributed to the optimized DTW matching strategy—introducing image quality weights strengthens the contribution of features from clear frames and reduces the interference of blurry and occluded frames on matching accuracy. The use of clustering indexing also improves matching efficiency. In terms of feedback performance, SSTO-RAFN's expert consistency score reaches 9.1, significantly higher than the 6.8 score of the traditional rule-based method, indicating that its generated feedback suggestions excel in terms of deviation description accuracy, improvement feasibility, and linguistic professionalism. The feedback generation latency is 4.8 ms, slightly higher than the traditional method, but still meets real-time demands when combined with the overall process latency. This is thanks to the pruning optimization of the causal language model and the prompt engineering design, where the structured prompt template guides the model to focus precisely on deviation information, and the lightweight model ensures real-time generation efficiency.

To verify SSTO-RAFN's ability for real-time pose detection, feature sequence matching, and deviation recognition in multi-target complex motion scenes, a multi-athlete synchronous analysis experiment during the sprint start phase was conducted. The region in Figure 6-(a) shows the 2D pose detection results at the moment of the sprint start for three athletes from the monitoring perspective. The system achieves real-time labeling of multi-target skeletal keypoints at a processing frame rate of 165 FPS and a single-frame latency of 6.0 ms. The red, blue, and green nodes correspond to the head, torso, and limb joints, respectively. The skeletal connections are clear, and no targets are missed, directly verifying the real-time and detection accuracy of the lightweight YOLO-Pose module in multi-target scenarios. The region in (b) shows the comparison of pose feature sequences and the DTW matching heatmap, where the system can accurately extract key joint angle features of the current athlete and align them with the standard action template. Two high deviation matching areas are also located, and the deviation intensity is quantified and presented through a color gradient. This experimental result demonstrates that SSTO-RAFN can achieve millisecond-level real-time pose detection in multi-target motion scenes, while also performing fine-grained matching of pose feature sequences with standard templates and locating deviation areas. This provides precise quantitative evidence for the subsequent generation of structured improvement suggestions, fully demonstrating the system's collaborative advantages in real-time performance, multi-target adaptability, and deviation recognition ability.

3.3 Experiment discussion

The analysis of the performance bottlenecks of core image processing modules indicates that there is still room for optimization of SSTO-RAFN in extreme scenarios: in heavy occlusion, MPJPE rises to 68.5 mm, with an accuracy loss of 91.9%, mainly due to severe loss of skeletal topology information, making it difficult for Adaptive-GCN to construct effective constraints; in high-speed motion scenes, frame blurring intensifies, and the difficulty of temporal feature alignment increases, leading to a 15%-20% rise in MPJPE. In the future, the robustness of extreme scenes can be further improved by incorporating multimodal information to supplement image features.

Parameter sensitivity analysis shows that temporal window size, feature fusion weight, and DTW matching threshold significantly impact performance: the performance is optimal when the temporal window length is 5 frames. If the window is too short, temporal features are insufficient, resulting in an 8.3% increase in MPJPE. If the window is too long, the computational load increases, causing a 20% rise in latency; when the feature fusion weight α is set to 0.6, the accuracy is optimal. If α is too large or too small, one branch's features dominate, causing a 5%-7% accuracy drop. When the DTW matching threshold is set to 0.8, the matching accuracy and efficiency are balanced optimally. If the threshold is too high, matching becomes too strict, causing a 4.2% drop in accuracy, and if too low, matching becomes too loose, increasing the deviation quantification error by 30%. The sensitivity of these parameters provides directions for future model optimization and can enhance the model's generalization ability through adaptive parameter adjustment strategies.

4. DISCUSSION

The SSTO-RAFN framework is developed to address the practical requirement for real-time athlete pose analysis in unconstrained environments. A three-level cascading architecture is constructed from the perspective of image processing, and the experimental results indicate that this design provides a workable balance between computational efficiency, modeling accuracy, and system robustness. The lightweight feature extraction module, based on backbone optimization and channel attention, makes it possible to maintain sufficient feature representation while operating under the computational constraints of edge devices. This helps to alleviate the common trade-off observed in lightweight models between inference speed and pose estimation accuracy.

The dual-path spatiotemporal refinement and self-supervised fusion strategy further extends existing approaches to three-dimensional pose estimation. By combining skeletal topology features with temporal image context and exploiting temporal coherence in video sequences, the framework reduces its reliance on large-scale annotated data while preserving modeling accuracy. This design is particularly relevant for sports scenarios, where collecting and annotating three-dimensional ground truth data is costly and often impractical. The integration of sequence-level feature matching with structured feedback generation also provides a connection between pose deviations measured at the image processing level and interpretable guidance for training, which supports more practical use of pose analysis systems in real training environments.

In comparison with related work, the framework shows several relative advantages. Compared with lightweight pose estimation methods, it demonstrates improved robustness under occlusion and motion blur while maintaining real-time performance. Compared with existing self-supervised pose estimation approaches, the use of temporal coherence as a supervisory signal is better aligned with the characteristics of continuous motion data and avoids some of the generalization limitations associated with view-consistency-based constraints in single-view settings. For sequence matching and deviation analysis, the introduction of image quality weighting into the DTW process improves matching stability in visually degraded conditions, reducing sensitivity to blur and occlusion.

Despite these results, several limitations remain. Under severe occlusion or extreme lighting, pose feature extraction may still be degraded due to missing skeletal information or low signal-to-noise ratios, which weakens the effectiveness of the dual-path fusion strategy. In addition, the current motion template knowledge base mainly covers common sports such as athletics and basketball, and the limited coverage of less common sports restricts generalization across a wider range of activities.

Future work can address these limitations in several directions. One direction is the integration of multimodal sensing, such as combining visible and infrared imagery, to improve robustness under challenging lighting conditions. Another direction is collaborative learning across multiple data sources using privacy-preserving strategies such as federated learning, which could expand the diversity of motion templates without requiring centralized data collection. A further direction is the exploration of generative models for restoring occluded pose information and for simulating corrective movements, which may support more informative feedback in complex scenarios. These directions aim to further improve the balance between real-time performance, modeling accuracy, and practical usability in athlete pose analysis systems.

5. CONCLUSION

This work has presented a real-time pose analysis and feedback framework based on self-supervised spatiotemporal optimization to address the requirements of athlete pose analysis under practical constraints. A three-level cascading architecture has been introduced from the perspective of image processing, enabling a coordinated balance among real-time performance, robustness, and modeling accuracy. The framework integrates three main components: a lightweight feature extraction module that maintains essential pose information while operating on edge devices, a dual-path spatiotemporal refinement module that combines skeletal topology modeling with temporal image feature learning, and a self-supervised fusion and sequence matching module that reduces dependence on annotated data and links pose deviations to structured feedback.

Experimental results indicate that the proposed framework achieves millisecond-level inference latency on edge platforms while maintaining competitive three-dimensional pose estimation accuracy and robustness in complex scenes. The integration of self-supervised temporal coherence and dual-path feature representation contributes to improved performance under occlusion and motion blur, and the sequence-based feedback mechanism supports interpretable assessment of motion deviations.

The contributions of this work are twofold. First, it provides a spatiotemporal pose modeling approach that reduces reliance on large-scale annotated datasets while preserving robustness and accuracy. Second, it establishes a structured connection between image-based pose deviation analysis and actionable feedback, which supports practical deployment in sports training contexts.

Future developments may extend the framework toward broader sensing modalities and a wider range of sports scenarios. The incorporation of multimodal inputs, collaborative learning strategies, and more diverse motion templates may further improve generalization and robustness. These directions are expected to support continued progress toward efficient, reliable, and practically deployable pose analysis systems.

FUNDINGS

The research is supported by: The second batch of teaching reform projects for vocational education in Zhejiang Province during the 14th Five Year Plan period: Reform and Practice of Physical Education Curriculum in Higher Vocational Education Based on the Integration of Physical Education and Vocational Education - Taking the Creation of "Electrician" Outdoor Classroom as an Example (Gran No.: jg20240263). (2) 2023 Guiding Science and Technology Research Project in Quzhou City: Research on Optimization Algorithm and Platform Design for Student Physical Health Evaluation (Gran No.: 2023ZD131).

REFERENCES

[1] Cho, C., Kim, J., Kim, J., Lee, S.J., Kim, K.J. (2016). Detecting for high speed flying object using image processing on target place. Cluster Computing, 19(1): 285-292. https://doi.org/10.1007/s10586-015-0525-x

[2] Hiemann, A., Kautz, T., Zottmann, T., Hlawitschka, M. (2021). Enhancement of speed and accuracy trade-off for sports ball detection in videos—finding fast moving, small objects in real time. Sensors, 21(9): 3214. https://doi.org/10.3390/s21093214

[3] Wang, H., Gao, J., Liu, J. (2021). Research and implementation of the sports analysis system based on 3D image technology. Wireless Communications and Mobile Computing, 2021(1): 4266417. https://doi.org/10.1155/2021/4266417

[4] Yu, Z. (2024). Image super-resolution reconstruction in sports scenarios and its application in motion analysis. Traitement du Signal, 41(2): 1079-1087. https://doi.org/10.18280/ts.410249

[5] Jiang, X., Wu, L. (2022). Sports video image segmentation based on fuzzy clustering algorithm. Scientific Programming, 2022(1): 6882291. https://doi.org/10.1155/2022/6882291

[6] Leow, W.K., Wang, R., Leong, H.W. (2012). 3-D–2-D spatiotemporal registration for sports motion analysis. Machine Vision and Applications, 23(6): 1177-1194. https://doi.org/10.1007/s00138-011-0371-7

[7] Sagawa, K., Abo, S., Tsukamoto, T., Kondo, I. (2009). Forearm trajectory measurement during pitching motion using an elbow-mounted sensor. Journal of Advanced Mechanical Design, Systems, and Manufacturing, 3(4): 299-311. https://doi.org/10.1299/jamdsm.3.299

[8] Hu, X., Zong, B., Pang, B. (2021). Simulation of sports action monitoring based on feature similarity model. Journal of Ambient Intelligence and Humanized Computing, 1-12. https://doi.org/10.1007/s12652-021-03046-7

[9] Zhao, L. (2023). A hybrid deep learning-based intelligent system for sports action recognition via visual knowledge discovery. IEEE Access, 11: 46541-46549. https://doi.org/10.1109/ACCESS.2023.3275012

[10] Yoo, H., Lee, S.E., Chung, K. (2023). Deep learning-based action classification using one-shot object detection. Computers, Materials & Continua, 76(2): 1343-1359. https://doi.org/10.32604/cmc.2023.039263

[11] Huang, J., Yi, X., Chen, L., Yin, H. (2025). A lightweight multi-stage framework for ultrasonic gesture recognition with time-frequency fusion. IEEE Sensors Journal., 25(23): 43352-43360. https://doi.org/10.1109/JSEN.2025.3623994

[12] Zhang, M., Zhou, Z., Wang, T., Zhou, W. (2023). A lightweight network deployed on ARM devices for hand gesture recognition. IEEE Access, 11: 45493-45503. https://doi.org/10.1109/ACCESS.2023.3273713

[13] Liu, Y.Z., Zhang, T.F., Li, Z., Deng, L.Q. (2023). Deep learning-based standardized evaluation and human pose estimation: A novel approach to motion perception. Traitement du Signal, 40(5): 2313-2320. https://doi.org/10.18280/ts.400549

[14] Park, S.Y., Subbarao, M. (2004). Automatic 3D model reconstruction based on novel pose estimation and integration techniques. Image and Vision Computing, 22(8): 623-635. https://doi.org/10.1016/j.imavis.2004.01.002

[15] Liu, S., Sehgal, N., Ostadabbas, S. (2022). Adapted human pose: monocular 3D human pose estimation with zero real 3D pose data. Applied Intelligence, 52(12): 14491-14506. https://doi.org/10.1007/s10489-022-03341-6

[16] Zhang, X., Zhu, Y., Li, C., Zhao, J., Li, G. (2014). SIFT algorithm-based 3D pose estimation of femur. Bio-Medical Materials and Engineering, 24(6): 2847-2855. https://doi.org/10.3233/BME-141103

[17] Liu, Y., Sun, R., Lu, Y., Zhang, S. (2019). A knowledge-based online fault detection method of the assembly process considering the relative poses of components. International Journal of Precision Engineering and Manufacturing, 20(10): 1705-1720. https://doi.org/10.1007/s12541-019-00218-6

[18] Guo, E., Ren, N., Liu, Z., Zheng, X. (2019). Influence of sensitive pose errors on tooth deviation of cylindrical gear in power skiving. Advances in Mechanical Engineering, 11(4): 1687814019843759. https://doi.org/10.1177/1687814019843759