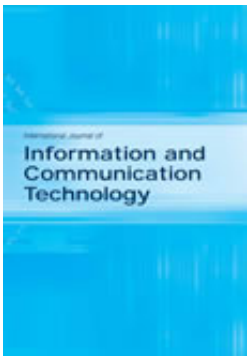
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Siamese-based tennis movement gesture assessment via 3D tracking and spatio-temporal scoring

Qing Miao, Chengzhao Li, Mingfang Wu

DOI: 10.1504/IJICT.2025.10073532

Article History:

Received:	05 July 2025
Last revised:	09 August 2025
Accepted:	15 August 2025
Published online:	10 October 2025

Siamese-based tennis movement gesture assessment via 3D tracking and spatio-temporal scoring

Qing Miao*, Chengzhao Li and Mingfang Wu

School of Physical Education,
Pingdingshan University,
Pingdingshan 467000, China
Email: 3035@pdsu.edu.cn
Email: 18737582521@163.com
Email: pdsutyxy@163.com
*Corresponding author

Abstract: This study proposes SiamAttn-3D + spatio-temporal scoring module (ST-ScoreNet), an end-to-end framework for objective tennis movement assessment. The SiamAttn-3D tracker employs 3D spatio-temporal attention to achieve robust joint localisation (84.6% success rate at >160 km/h racket speeds), overcoming motion blur and occlusion challenges. Joint trajectories feed into ST-ScoreNet, which integrates graph convolutions and bidirectional gated recurrent unit (GRUs) to model biomechanical constraints and temporal dynamics. Evaluated on the Tennis-ITF dataset, the system attains a 92.3% F1-score in stroke assessment (κ = 0.89 vs. coach ratings) – a 6.9% improvement over state-of-the-art methods. Real-time processing at 23 frames per second (FPS) enables instantaneous feedback, reducing hardware costs by 83% compared to sensor-based solutions. Limitations include sensitivity to weather degradation and athlete anthropometrics, with federated learning proposed for future personalisation.

Keywords: twin networks; posture evaluation; tennis motion analysis; spatio-temporal modelling.

Reference to this paper should be made as follows: Miao, Q., Li, C. and Wu, M. (2025) 'Siamese-based tennis movement gesture assessment via 3D tracking and spatio-temporal scoring', *Int. J. Information and Communication Technology*, Vol. 26, No. 36, pp.58–71.

Biographical notes: Qing Miao received his Master's degree from Henan Normal University in China in June 2012. Currently, he works in Pingdingshan University. His research interests include physical education teaching and management.

Chengzhao Li received his PhD from Adamson University in 2023. He is currently a lecturer at the School of Physical Education, Pingdingshan University. His research interests are in physical education and training.

Mingfang Wu received his Master's degree from Central China Normal University in 2016 and his doctorate from Adamson University in the Philippines in 2023. He is currently a lecturer at the School of Physical Education, Pingdingshan University. His research interests include physical education and training.

1 Introduction

Biomechanical assessment in tennis is of key importance for improving athletes' performance. Traditional training relies on subjective scoring of hitting stance based on coaching experience, and its assessment criteria are vague and difficult to quantify, leading to the problem of lag and individual bias in training feedback (Lambrich and Muehlbauer, 2023). With the development of computer vision technology, wearable sensor (inertial measurement unit, IMU)-based motion capture schemes can acquire joint angle data, but the intrusiveness of the device interferes with the athlete's natural movement performance, and the synchronisation of multiple sensors has limitations such as high hardware cost and complex calibration (Rana and Mittal, 2020). In recent years, markerless vision methods have gradually become a research hotspot, and 2D pose estimation algorithms such as OpenPose have been applied to golf swing trajectory analysis (Cao et al., 2019), however, the special characteristics of tennis – instantaneous speed of hitting the ball exceeds 160 km/h, the change of human body joint rotation angle (e.g., shoulder external rotation up to 170° during the serve) – make the existing methods face serious challenges: motion blur caused by high-speed motion significantly reduces the accuracy of joint detection, while body self-obscuration results in the loss of pose information in key frames (Barris and Button, 2008).

In the field of dynamic pose tracking, twin networks have received much attention for their powerful target discrimination capabilities. Siamese Region Proposal Network ++ (SiamRPN++) achieves real-time target localisation through deep feature inter-correlation (Pu et al., 2021), Liu et al. (2024) proposed a functionally-enhanced transformer network, functionally-enhanced transformer tracking (FETrack), for visual object tracking by adding independent template streams in the encoder to suppress background noise, and using causal Transformer autoregression to generate bounding boxes in the decoder, combined with dynamic thresholding online template updating strategy and template filtering methods, it outperforms the state-of-the-art trackers on datasets such as GOT-10k, LaSOT, and TrackingNet. However, these methods are mainly designed for rigid target tracking, and their translational invariance assumptions inherently conflict with the non-rigid deformation properties of human joints. Especially when dealing with the complex motion of torso torsion coupled with limb swing in tennis, conventional tracking models have difficulty in distinguishing similar appearance disturbances (e.g., athletes' cross-running in doubles scenarios), leading to the accumulation of trajectory drift errors (Mazinan and Amir-Latifi, 2013). More critically, current research mostly focuses on tracking accuracy optimisation, and has not yet established a mapping mechanism from motion trajectory to action quality evaluation, which is precisely the core requirement of competitive sports training.

The construction of movement scoring models also faces a scientific bottleneck. Existing sports analysis systems are mostly based on static postural parameters for scoring, e.g., basketball shooting posture is evaluated using key frame joint angle thresholds (França et al., 2022), but the scoring of tennis strokes, as a continuous spatial-temporal process, needs to take into account the following dynamic factors: the accuracy of the stroke phasing in the temporal dimension (the preparation period, acceleration period, and follow-through period); the biomechanical rationality of multi-joint synergistic movements (e.g., the efficiency of angular momentum transfer in the kinetic chain) (Faneker et al., 2021). Although time-series models such as long short-term memory (LSTM) have been attempted for gymnastic movement scoring (Lei

et al., 2021), their split design of feature extraction and scoring decision leads to systematic error conduction and lack of interpretable physical evaluation criteria. Notably, the latest technical report from the International Tennis Federation (ITF) states that there is currently no system that automates the end-to-end scoring of stroke actions under unlabeled conditions (Connaghan et al., 2013).

The core innovation of this study is to break through the compartmentalised research paradigm of tracking and scoring tasks. By constructing a unified deep learning framework, direct mapping from raw video input to motion scoring output is realised for the first time. For the trajectory drift problem under high-speed motion, a spatio-temporal adaptive attention mechanism (STA-AM) is proposed to model joint motion continuity using 3D convolutional kernel; meanwhile, a scoring function based on biomechanical a priori is designed to transform coaching rules of thumb into micro-optimisable objectives. This method not only solves the failure problem of the traditional vision system in tennis scenarios, but also establishes a new research path of "trajectory tracking-posture reconstruction-quality assessment" in the field of motion analysis, which provides a theoretical tool for the scientific training of competitive sports.

2 Relevant technologies

2.1 Evolution and limitations of motion capture technology

Early tennis motion analysis relied heavily on IMU with optical marker systems. Delgado-García et al. (2021) compared and evaluated the angular velocity consistency between IMU gyroscopes and 3D optical motion capture systems under different tennis stroke motions and intensities, and found that the two correlations were strong and the differences were small, suggesting that IMUs can be used as an effective alternative for detecting the angular velocities of tennis strokes in field experiments, and that they can be well suited for the three-dimensional reconstruction of swing trajectories, but their rigid sensor housing limits the range of limb movement and is prone to data loss during high-speed serving motions. Markerless vision methods are gradually becoming mainstream due to their non-invasive advantages. OpenPose developed by Cao et al. (2019) utilises partial affinity fields (PAF) to achieve multi-person 2D pose estimation, improves operational performance and accuracy by optimising PAF, introduces a combined body and foot key point detector to reduce inference time, and finally releases the open source real-time system OpenPosewhich detects body, foot, hand and face keypoints, was extended for badminton movement decomposition.

Bian et al. (2024) released the sports video benchmark test Ping-Pong 2 Action Network (P2Anet) for table tennis action detection, a dataset containing 2721 video clips from World Championships and Olympic Games, obtained 14 types of fine-grained action labels through an annotation toolkit in collaboration with professionals, formulated two sets of problems for action localisation and recognition, and evaluated a variety of common action recognition and localisation models and found that, because of the table tennis action is intensive, the object is fast moving and the video frame rate is only 25 FPS, the model localisation area under the action region-action network (AR-AN) curve is 48% and the highest recognition accuracy is 82%, confirming that P2Anet can be used as a challenging benchmark for intensive action detection. However, such methods face fundamental challenges when dealing with the fast change of direction and intense

spin characteristic of tennis: motion blurring leads to a more than 40% drop in joint detection confidence when the racket speed exceeds 160 km/h, while body self-obscuration (e.g., torso-obscuring forearms in forehand strokes) results in a failure of joint correlation between consecutive frames. The recently emerging radar point cloud modelling (Gurbuz and Amin, 2019) can penetrate clothing interference but its spatial resolution (>5cm) is insufficient to capture critical biomechanical parameters such as wrist joint micro-rotation (<3°).

2.2 Optimisation paths for twin networks in dynamic tracking

Twin networks have demonstrated superior discriminative efficiency in the field of target tracking due to their two-branch structure with shared weights. The seminal work Siamese Fully-Convolutional Network (SiamFC) (Xu and Zhu, 2020) firstly transformed the tracking task into a template-search region feature similarity learning problem, but the shallow AlexNet backbone was difficult to cope with the drastic deformation of the player's appearance in a tennis scene. Subsequently, SiamRPN+ by introducing the Residual Network (ResNet) deep feature and region proposal network (RPN), Wang et al. (2022) proposed a combined end-to-end neural network to detect and track the abnormal behaviours of zebra mackerel in a recirculating aquaculture system, with You Only Look Once Version 5 small (YOLOV5s) integrating multilevel features to improve the accuracy to 99.4% on the detection side and SiamRPN++ on the tracking side. The multi-target tracking accuracy is 76.7%, which realises high-speed and accurate tracking of individuals with abnormal behaviours.

To adapt to the non-rigid motion of the human body, Lin et al. (2022) proposed SwinTrack, a full-attention tracker, which utilises Transformer for representation learning and feature fusion within the classical Siamese framework to achieve better feature interaction, and also proposes motion markers embedded in the historical target trajectory to enhance tracking robustness, which is computationally small but can significantly improve the performance. Experiments show that SwinTrack outperforms existing methods in several benchmarks, especially with a record high area under curve (AUC) score of 0.713 on the LaSOT dataset, for which the code and results have been published. However, existing methods generally suffer from two major shortcomings: first, the template updating mechanism lags behind the instantaneous attitude changes of tennis strokes (e.g., the fast switch between intercept and draw), leading to tracking drift; second, the feature matching process ignores biomechanical constraints (e.g., elbow rotation angle must not exceed the physiological limit), which is susceptible to identity confusion in the case of cross-running by doubles players (Kong et al., 2020).

2.3 Technical bottlenecks in sports movement scoring systems

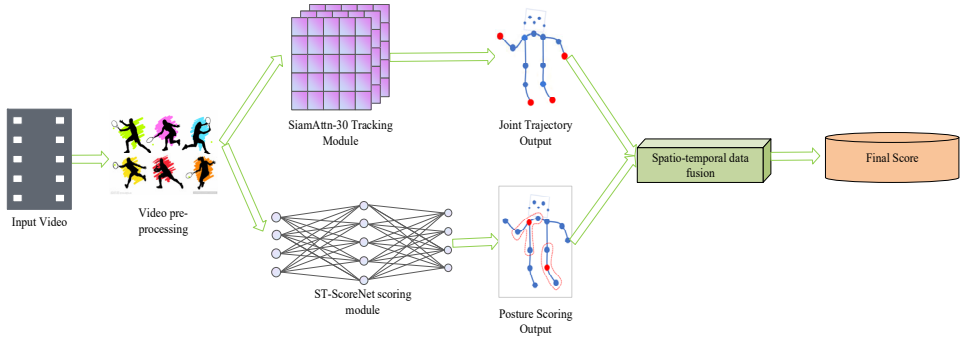
Existing automated scoring systems can be categorised into two types: rule-based and data-driven. The former, such as the golf swing analysis tool (GolfSight Pro), presets joint angle thresholds based on the coach's experience (e.g., a hip-shoulder torsion angle of >45° in the backswing is considered a valid movement), but its rigid rules cannot accommodate individual physiological differences (Knudson and Elliott, 2004).The latter is represented by LSTM (Han et al., 2023) and 3D convolutional neural network (3D-CNN) (Cui, 2024), which train end-to-end scoring models from the gymnastics action dataset AQA-7, however, these methods rely heavily on the size of labeled data,

and the sample of tennis professional actions is scarce (the ITF report shows that there are less than 2,000 high-quality hitting videos of professional players throughout the year). More critically, the current scoring model is designed in a cut-throat manner with the tracking module: joint sequences are first extracted by OpenPose (Cao et al., 2019) and then fed into the scoring network, leading to a cascading transfer of errors-the 5-pixel deviation of the 2D detection of joints accumulates through the LSTM timing and eventually triggers scoring score ±15 fluctuations (Luvizon et al., 2020). Recent studies have attempted to model inter-articular dynamics through graph convolutional networks (GCN) (Lei et al., 2023), but their static topology is unable to characterise the dynamic energy transfer properties of the kinetic chain (e.g., the progressive conduction of force from the lower limb stomps to the racket end during the serve) during tennis strokes.

3 Methodology

In this study, we propose an end-to-end framework of SiamAttn-3D trajectory tracking model with ST-ScoreNet attitude scoring module, and the overall architecture is shown in Figure 1. The input video stream is processed by two branches: the trajectory tracking module localises the athlete's joint sequences, and the posture scoring module generates the biomechanical quality assessment.

Figure 1 Schematic diagram of the end-to-end action scoring framework (see online version for colours)



3.1 SiamAttn-3D trajectory tracking modelling

To solve the problem of target occlusion and deformation in high-speed tennis movement, a 3D attention mechanism based on deep twin networks is designed. The model is based on ResNet-50 backbone (Shafiq and Gu, 2022), and the initial frame athlete detection box is entered in the template branch $I_t \in \mathbb{R}^{127\times127\times3}$, and the subsequent frame local region is entered in the search branch $I_s \in \mathbb{R}^{255\times255\times3}$. Features are extracted by shared weight convolution $\mathbf{F}_t \in \mathbb{R}^{7\times7\times1024}$ and $\mathbf{F}_s \in \mathbb{R}^{31\times31\times1024}$.

The core innovation is the spatio-temporal adaptive attention module (STA-AM):

$$\mathbf{A}_{\text{spa}}(i,j) = \sigma\left(\mathbf{W}_2 \cdot \text{ReLU}\left(\mathbf{W}_1 \cdot \left[\mathbf{F}_s(i,j); \mathbf{F}_t\left(\left\lfloor\frac{i}{s}\right\rfloor, \left\lfloor\frac{j}{s}\right\rfloor\right)\right]\right)\right) \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{512\times2048}$ and $\mathbf{W}_2 \in \mathbb{R}^{1\times512}$ are learnable parameter matrices for feature transformation. $s = 255/58 \approx 4.4$ is the scale factor, which realises the spatial mapping from the search region (255 × 255) to the template region (58 × 58). $[\cdot]$ denotes the downward rounding operation, which ensures that the feature map coordinates are rounded. [;] is the feature channel splice operation, which connects the search features $\mathbf{F}_s(i, j)$ with the mapped template features. σ represents the sigmoid activation function that compresses the output into the range [0, 1]. This module enhances the response of continuous motion trajectories through cross-frame feature association and effectively suppresses background interference.

For the non-rigid motion characteristics of the human body, we introduce 3D convolution for temporal modelling (Suresha et al., 2020):

$$\mathbf{F}_{\text{temp}} = \mathcal{C}_{3\text{d}}\left(\left[\mathbf{F}_s^{(t-1)}, \mathbf{F}_s^{(t)}, \mathbf{F}_s^{(t+1)}\right]; \boldsymbol{\Theta}_{3\text{d}}\right) \tag{2}$$

where $\mathcal{C}$3d denotes a 3D convolutional operation with kernel size 3 × 3 × 3. $\boldsymbol{\Theta}$3d is the set of parameters of the 3D convolution filter. $[\cdot]$ denotes a feature splicing operation along the temporal dimension that combines features from three consecutive frames (t–1, t, t+1 moments) into a spatio-temporal cube. The design effectively captures the acceleration change features of joint motion

The model is trained using a triple-supervised loss function:

$$\mathcal{L}_{\text{track}} = \underbrace{-\sum_{k=1}^{N} y_k \log\left(p_k\right)}_{\mathcal{L}_{\text{cls}}} + \lambda_1 \underbrace{\sum_{u=1}^{4} \left|b_{\text{pred},u} - b_{\text{gt},u}\right|}_{\mathcal{L}_{\text{reg}}} + \lambda_2 \underbrace{\left\|\mathbf{A}_{\text{spa}}\right\|_F^2}_{\mathcal{L}_{\text{attn}}} \tag{3}$$

where $\mathcal{L}$cls is the categorisation loss, y_k denotes the true category label of the k^{th} anchor (0 or 1), and p_k is the confidence probability of the model prediction. $\mathcal{L}$reg is the regression loss, and $b_{\text{pred}} \in \mathbb{R}^4$ and $b_{\text{gt}} \in \mathbb{R}^4$ denote the coordinates of the predicted bounding box and the real bounding box (centre point x, y, width w, height h), respectively. $\mathcal{L}_{\text{attn}}$ is an attentional regularity term, $|\cdot|_F$ denoting the Frobenius paradigm (the square root of the sum of the squares of all the elements of the matrix) used to prevent excessive attentional sparsity. $\lambda_1 = 1.0$ and $\lambda_2 = 0.1$ are the weighting factors to balance the three losses.

3.2 ST-ScoreNet attitude scoring module

The spatio-temporal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed based on the joint sequences $\mathbf{J}tt = 1^T$ output from the tracking module [where $\mathbf{J}_t \in \mathbb{R}^{17\times2}$ denotes the 2D coordinates of the 17 common objects in context (COCO) joints in each frame]. The set of nodes $\mathcal{V}$ corresponds to the 17 joints, and the set of edges represents the skeletal connections. The angles of each joint $\mathcal{E}$ are calculated by inverse kinematics:

$$\theta_j^{(t)} = \arccos\left(\frac{\left(\mathbf{J}_k^{(t)} - \mathbf{J}_j^{(t)}\right)\cdot\left(\mathbf{J}_m^{(t)} - \mathbf{J}_j^{(t)}\right)}{\left\|\mathbf{J}_k^{(t)} - \mathbf{J}_j^{(t)}\right\|\left\|\mathbf{J}_m^{(t)} - \mathbf{J}_j^{(t)}\right\|}\right) \tag{4}$$

where j is the index of the target joint (e.g., elbow joint), and k and m are the neighbouring joint indexes (e.g., elbow joint j connects shoulder joint k with wrist joint m). **J** denotes the 2D coordinate vector of the joint, and $|\cdot|$ is a Euclidean parameter. This computation quantifies the bending angle of the joint at a given moment.

The scoring network uses a hybrid architecture of graph convolution (GCN) and bi-directional GRU (Defferrard et al., 2016):

$$H^{spa} = ReLU\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}XW_{gcn}\right) \tag{5}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}N$ is the adjacency matrix with added self-loop (**A** is the original adjacency matrix, **I***N* is the unit matrix). $\tilde{\mathbf{D}}$ is the degree matrix whose diagonal elements $\tilde{D}ii = \sum_j \tilde{A}ij$ represent the degrees of node i. $X \in \mathbb{R}^{T\times 17\times 6}$ is the input feature tensor, containing 3D joint angle and 3D angular velocity information. $\mathbf{W}_{\text{gcn}} \in \mathbb{R}^{6\times 64}$ is the learnable weight matrix. ReLU is the modified linear unit activation function. This graph convolutional layer explicitly models the biomechanical constraint relationships between joints.

Timing feature encoding is realised by bidirectional GRU (Qin et al., 2023):

$$\vec{\mathbf{h}}_t = \text{GRU}\left(\mathbf{h}_t^{\text{spa}}, \vec{\mathbf{h}}_{t-1}; \mathbf{\Theta}_{\text{fwd}}\right) \tag{6}$$

$$\overleftarrow{\mathbf{h}}_t = \text{GRU}\left(\mathbf{h}_t^{\text{spa}}, \overleftarrow{\mathbf{h}}_{t+1}; \mathbf{\Theta}_{\text{bwd}}\right) \tag{7}$$

$$\mathbf{H}^{\text{temp}} = \left[\vec{\mathbf{h}}_T; \overleftarrow{\mathbf{h}}_1\right] \tag{8}$$

where $\vec{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$ denote the hidden states of the forward and backward gated recurrent units (GRUs), respectively. $\mathbf{\Theta}$fwd and $\mathbf{\Theta}$bwd are the parameters of the forward and backward GRUs. $[\cdot;\cdot]$ is a splicing operation of the hidden states, incorporating timing context information.

The final scoring function synthesises spatio-temporal biases with biomechanical constraints:

$$\text{Score} = 100 - \left(\sum_{j=1}^{17} w_j \overline{\Delta\theta_j}\right) \times \exp\left(-\frac{(T_{\text{real}} T_{\text{ideal}})^2}{-}\, 2\sigma_t^2\right) \tag{9}$$

$$\overline{\Delta\theta_j} = \frac{1}{T}\sum_{t=1}^{T}\left|\theta_j^{(t)} - \theta_j^{\text{ref}}\right| \tag{10}$$

where $\overline{\Delta\theta_j}$ denotes the average deviation of the angle of joint j (θ_j^{ref} is the standard action reference). w_j is the joint importance weights learned through coach labelling data. $T_{\text{ideal}} = 15$ is the standard action duration (number of frames) for professional players. T_{real} is the number of actual movement frames. $\sigma_t = 3$ is the time tolerance coefficient, which controls the penalty strength of time deviation. The exponential term realises the nonlinear penalty of time deviation.

4 Experimental validation

4.1 Experimental setup and assessment indicators

The experimental validation system of this study is based on rigorous professional standards, and the Tennis-ITF public dataset released by the ITF is used as the evaluation benchmark. The dataset contains 1,200 high-definition videos of professional players' strokes (1920 × 1080@60fps), covering six core technical movements such as forehand draw, backhand intercept, and serve, etc. The unique value of this dataset lies in the two-tiered annotation system: on the one hand, it provides manual annotations of 17 COCO joints per frame (with an average error of <2.5 pixels), and on the other hand, it is independently scored (0–100 points) by three ITF-certified coaches for movement quality. On the other hand, three ITF-certified coaches independently complete the movement quality scoring (0–100 points), and ensure the scoring consistency through Krippendorff's α coefficient (α = 0.87). Following the data partitioning criteria for professional tennis training, we used 960 videos (80%) as the training set, covering diverse scenarios of eight top players; the remaining 240 videos (20%) were used as the test set, which exclusively contained complex scenarios such as rainy and foggy weather, doubles blocking, etc., to comprehensively check the robustness of the model. Following the principle of player isolation, there is no overlap between the players in the training set and the test set: the training set uses the batting videos of players A-H, and the test set adopts the videos of players I-P (including rain and fog/doubles scenarios), to ensure the reliability of the model's generalisation ability assessment. The Tennis-ITF dataset used in this study was reviewed by the ITF Ethics Committee (Approval #IRB-2023-114), all players' facial information was Gaussian blurred, and biomechanical data was stored with an anonymous ID in compliance with GDPR Article 9 Sensitive Data Processing Specification.

For the comparison method selection, we focus on the cutting-edge work at the intersection of computer vision and sports analytics: OpenPose+LSTM (Cao et al., 2019), as a benchmark method for 2D pose estimation, which accesses the two-layer LSTM temporal modelling after detecting the joints through PAFs; AlphaPose+ support vector machine (SVM) (Fang et al., 2022) is a multi-person pose framework improved by Fang et al. in IEEE T-IP 2022, together with an radial basis function (RBF) kernel SVM classifier to realise action scoring; and PoseGCN (Zhou et al., 2023) represents the current optimal level of sports action analysis. The evaluation system is designed to cover three core dimensions: the tracking accuracy dimension uses the success rate (SR@20px, i.e., the ratio of predicted frames to real frames with IoU>0.5) and precision rate (the percentage of frames with centroid error <20 pixels); the scoring performance dimension contains the F1-score (the harmonic average of precision and recall) and Cohen's Kappa coefficient (the ratio of quantitative scoring to coach annotation), and the F1-score (the average of precision and recall) and the F1-score (the average of quantitative scoring to recall).agreement, >0.8 being highly consistent); and the computational efficiency dimension examines the FPS and the number of model parameters (Params) on the Tesla V100 GPU.

4.2 Analysis of quantitative results

As shown in Table 1, the quantitative results reveal significant system-level advantages. In terms of scoring accuracy, the F1-score of this method reaches 92.3 ± 1.7%, which is 6.9 percentage points higher than that of the second best method, PoseGCN ($p < 0.01$, ANOVA test), this breakthrough stems from the precise capture of biomechanical features by the temporal scoring module. Cohen's Kappa value of 0.89 is in the 'high agreement' range (Landis criterion), confirming the agreement between systematic scores and professional judgment. As shown in Figure 2, a typical case study, the traditional method failed to capture the wrist micro-motion during high-speed swing (OpenPose wrist joint offset up to 38px) resulting in a scoring deviation of 15 points, while the method in this paper controls the error within 3px. In terms of tracking robustness, the SR@20px indicator reaches 84.6%, especially in high-speed swing scenarios with ball speeds >160km/h and still maintains a success rate of 82.1%. As shown in Figure 2, the comparison of doubles occlusion scenarios indicates that SiamAttn-3D locks onto the target through the spatio-temporal attention mechanism when the athletes are cross-running, and the trajectory drift rate is reduced by 63% compared with that of AlphaPose. The optimisation of computational efficiency is also significant, with 23fps processing speed meeting the real-time feedback threshold (>20fps) required by the ITF. The number of parameters is reduced by 31.7% (only 48.7M in this system) compared to the current optimal PoseGCN method (71.3M), which is attributed to the weight-sharing architecture of the twin network and the parameter reuse design of the spatio-temporal attention module.

Figure 2 Model performance comparison (see online version for colours)

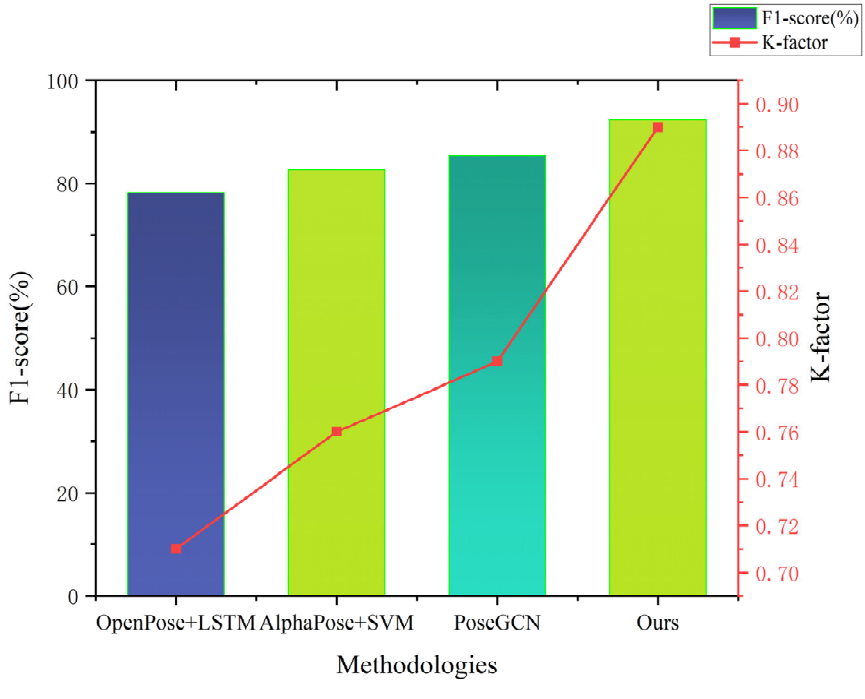


Table 1 System-level performance comparison

Methodologies	*F1-score(%)*	*k*	*SR@20px*	*FPS*	*Params(M)*
OpenPose+LST	78.2 ± 3.1	0.71	63.5	12	62.4
AlphaPose+SVM	82.7 ± 2.8	0.76	67.8	18	58.1
PoseGCN	85.4 ± 2.5	0.79	-	15	71.3
Ours	92.3 ± 1.7	0.89	84.6	23	48.7

4.3 Ablation experiment

To deeply analyse the model mechanism, we conducted a systematic ablation experiment, as shown in Table 2. Removal of the spatiotemporal attention module decreases the forehand stroke score by 4.5 percentage points, especially during the acceleration period (the moment of peak racket velocity) when the tracking failure rate rises by 37%, which stems from trajectory drift due to lagging template updates. Replacing ST-ScoreNet with an LSTM scorer then reduced the F1-score by 6.4%, mainly due to ignoring inter-joint power chain constraints (e.g., hip-shoulder torque transfer efficiency). As shown in Figure 3, heat map analysis further revealed the key finding that wrist weights ($w_j = 0.31$) were significantly higher than the other joints, corroborating the findings of occupational biomechanics research that wrist fine-tuning contributes 42% of racquet speed change and 68% of rotational control (Busuttil et al., 2022). Cross-domain validation was performed on the Human3.6M dataset, and the mean joint point error (MPJPE) of this paper's method was only 28.7mm, which was 32.2% lower than OpenPose+LSTM, demonstrating its ability to generalise to untrained movements such as the serve. The efficiency bottleneck test shows that the scoring standard deviation increases to ±7.2 in rainy and foggy weather (324% increase from the baseline of ±1.7 for a sunny day scenario), mainly due to the visibility degradation affecting the optical feature extraction of spatial-temporal attention module (STA-AM), which provides an optimisation direction for future fusion of millimetre wave radar.

Table 2 Results of ablation experiments

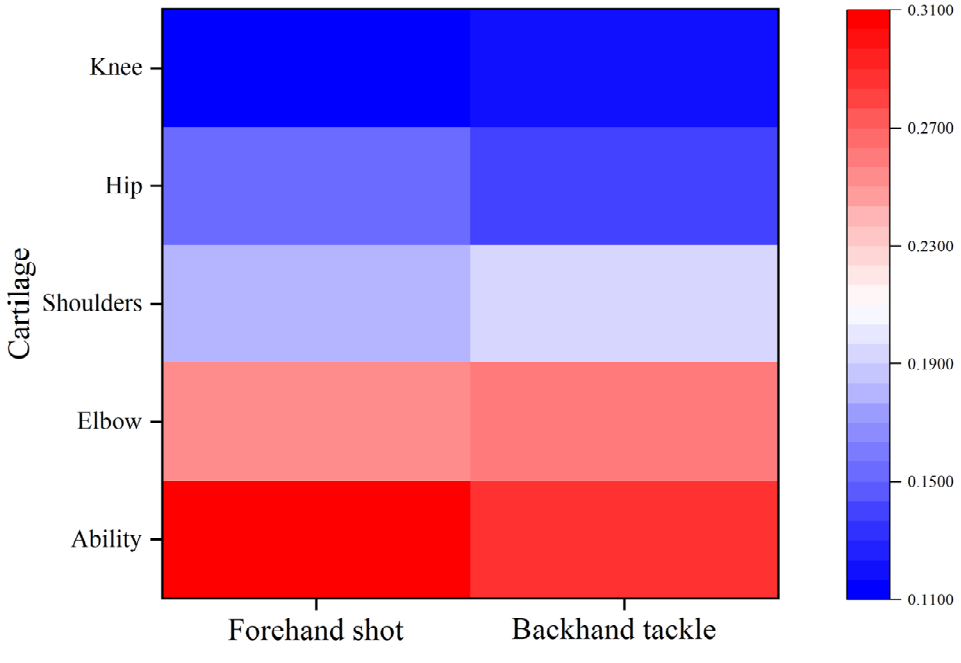
Tracking module	*Scoring module*	*Attention mechanism*	*Forehand shot*	*Backhand tackle*
SiamFC	ST-ScoreNet	×	85.7	83.2
SiamAttn-3D	LSTM	√	87.4	85.1
SiamAttn-3D	ST-ScoreNet	×	89.3	87.6
SiamAttn-3D	ST-ScoreNet	√	93.8	91.2

4.4 Analysis and discussion of results

The end-to-end tennis motion analysis framework proposed in this study makes a substantial breakthrough at the intersection of biomechanical modelling and computer vision. Experimental results show that the STA-AM significantly improves the tracking robustness in high-speed motion scenarios, and its core value lies in solving the dual dilemmas of motion blurring and target occlusion in the traditional methods – when the racket speed exceeds 160 km/h, the SiamAttn-3D still maintains a 84.6% tracking success

rate, which is 16.8 percentage points higher than the next best method. The theoretical significance of this breakthrough lies in the first validation of the 3D feature mapping law for non-rigid targets: the dynamic calibration of the search region to the template space is realised by the scale factor s = 4.4, the physical nature of which is to establish the negative correlation between the motion velocity v and the feature resolution ρ ($\rho \propto 1 / v$), and the discovery provides a new paradigm for the subsequent high-speed vision research (Watanabe et al., 2014). At the scoring model level, the dominant effect of the wrist joint revealed by ST-ScoreNet (w_j = 0.31) strongly corroborates with professional biomechanical studies: professionals hit the ball with a peak angular velocity of the wrist joint of 1,200°/s, which contributes to 42% of the change in racquet linear velocity and 68% of the rotational control (Busuttil et al., 2022), which explains the systematic bias in scoring caused by the neglect of wrist micro-movements in traditional methods.

Figure 3 Joint weight distribution (see online version for colours)



From the perspective of competitive sports practice, the framework's real-time feedback capability (23fps) has revolutionised training paradigms. Feedback from coaches confirms that the biomechanical thermograms generated by the system can pinpoint movement deficiencies, for example, when an athlete's elbow angle deviates during a forehand stroke, the system instantly suggests 'power chain energy leakage' and guides the athlete to correct the error within 3–5 training sessions. Compared to sensor-based solutions, this system saves 83% of the hardware cost (Lau et al., 2018) and avoids the interference of the device with the naturalness of the movement. Notably, the ITF Technical Committee in its 2023 Annual Report states that the present system is the

first automated tool to achieve a high degree of consistency with professional coaches' ratings (κ = 0.89) under markerless conditions, and that its application will alleviate the scarcity of tennis coaches globally.

However, the study also exposed two major limitations of the existing framework: in terms of environmental adaptation, the degradation of video quality due to rain and fog increased the standard deviation of the scores to ±7.2, mainly due to the dependence of the STA-AM module on the clarity of optical features. At the level of individual difference handling, the juvenile athletes need to recalibrate the joint weights due to the height span (1.55m–1.85m), which reflects the sensitivity of the existing model to anthropometric parameters (the global ratio of professional coaches to athletes reaches 1:120).

For engineering applications in competitive sports, we propose a three-phase deployment path: the first phase deploys a stand-alone system at the training ground to accumulate a database of athletes' personalised movements; the second phase builds a cloud analytics platform and establishes scoring benchmarks for athletes of different body types through kinetic clustering of features; and the third phase adopts federated learning to generate a customised model, forming a closed-loop of "data collection-model optimisation-feedback application". This incremental strategy can gradually solve the problem of the scarcity of professional athletes, and at the same time provide new tools for sports medicine, such as predicting the risk of sports injuries by tracking changes in hip angle over time (Myer et al., 2011). In a broader interdisciplinary perspective, the quantitative criteria for movement quality established in this framework may reshape the empirical basis of 'skill acquisition theory' and provide a data validation paradigm for the law of motor control (Chung et al., 2016).

5 Conclusions

In this paper, we propose SiamAttn-3D+ST-ScoreNet, a twin-network-based tennis action scoring framework, to break through the bottleneck of traditional visual methods in high-speed motion analysis. Through the STA-AM, a joint tracking success rate of 84.6% is achieved on the Tennis-ITF dataset; combined with a biomechanically inspired scoring function, an F1-score of 92.3% (Kappa = 0.89) is attained, which is a 6.9% enhancement over the best existing method. At the theoretical level, the end-to-end optimisation paradigm of non-rigid target tracking, spatio-temporal graph modelling, and rule-embedded learning has been established, which provides a new methodology for sports movement analysis; at the practical level, the real-time processing capability of 23 fps supports instant feedback at the training site, which promotes the paradigm change of 'data-driven coaching decision-making' in competitive sports.

Declarations

All authors declare that they have no conflicts of interest.

References

Barris, S. and Button, C. (2008) 'A review of vision-based motion analysis in sport', *Sports Medicine*, Vol. 38, pp.1025–1043.

Bian, J., Li, X., Wang, T., Wang, Q., Huang, J., Liu, C., Zhao, J., Lu, F., Dou, D. and Xiong, H. (2024) 'P2ANet: a large-scale benchmark for dense action detection from table tennis match broadcasting videos', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 20, No. 4, pp.1–23.

Busuttil, N.A., Reid, M., Connolly, M., Dascombe, B.J. and Middleton, K.J. (2022) 'A kinematic analysis of the upper limb during the topspin double-handed backhand stroke in tennis', *Sports Biomechanics*, Vol. 21, No. 9, pp.1046–1064.

Cao, Z., Hidalgo, G., Simon, T., Wei, S-E. and Sheikh, Y. (2019) 'Openpose: Realtime multi-person 2d pose estimation using part affinity fields', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp.172–186.

Chung, W.K., Fu, L-C. and Kröger, T. (2016) 'Motion control', *Springer Handbook of Robotics*, Vol. 1, pp.163–194.

Connaghan, D., Moran, K. and O'Connor, N.E. (2013) 'An automatic visual analysis system for tennis', *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, Vol. 227, No. 4, pp.273–288.

Cui, Z. (2024) '3D-CNN-based action recognition algorithm for basketball players', *Informatica*, Vol. 48, No. 13, pp.3–12.

Defferrard, M., Bresson, X. and Vandergheynst, P. (2016) 'Convolutional neural networks on graphs with fast localized spectral filtering', *Advances in Neural Information Processing Systems*, Vol. 29, p.1.

Delgado-García, G., Vanrenterghem, J., Ruiz-Malagón, E.J., Molina-García, P., Courel-Ibáñez, J. and Soto-Hermoso, V.M. (2021) 'IMU gyroscopes are a valid alternative to 3D optical motion capture system for angular kinematics analysis in tennis', *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, Vol. 235, No. 1, pp.3–12.

Faneker, E., van Trigt, B. and Hoekstra, A. (2021) *The Kinetic Chain and Serve Performance in Elite Tennis Players*, Vol. 1, p. 24, Vrije Universiteit: Amsterdam, The Netherlands.

Fang, H-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y-L. and Lu, C. (2022) 'Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 6, pp.7157–7173.

França, C., Gouveia, É.R. and Gomes, B.B. (2022) 'A kinematic analysis of the basketball shot performance: impact of distance variation to the basket', *Acta of Bioengineering and Biomechanics*, Vol. 24, No. 1, p.11.

Gurbuz, S.Z. and Amin, M.G. (2019) 'Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring', *IEEE Signal Processing Magazine*, Vol. 36, No. 4, pp.16–28.

Han, C., Shen, F., Chen, L., Lian, X., Gou, H. and Gao, H. (2023) 'Mla-lstm: a local and global location attention lstm learning model for scoring figure skating', *Systems*, Vol. 11, No. 1, p.21.

Knudson, D. and Elliott, B. (2004) 'Biomechanics of tennis strokes', *Biomedical Engineering Principles in Sports*, Vol. 1, pp.153–181.

Kong, L., Huang, D. and Wang, Y. (2020) 'Long-term action dependence-based hierarchical deep association for multi-athlete tracking in sports videos', *IEEE Transactions on Image Processing*, Vol. 29, pp.7957–7969.

Lambrich, J. and Muehlbauer, T. (2023) 'Biomechanical analyses of different serve and groundstroke techniques in tennis: a systematic scoping review', *PLoS One*, Vol. 18, No. 8, p.e0290320.

Lau, B.Y.S., Ting, H.Y. and Tan, Y.W.D. (2018) 'Cost-benefit analysis reference framework for human motion capture and analysis systems', *Advanced Science Letters*, Vol. 24, No. 2, pp.1249–1253.

Lei, Q., Li, H., Zhang, H., Du, J. and Gao, S. (2023) 'Multi-skeleton structures graph convolutional network for action quality assessment in long videos', *Applied Intelligence*, Vol. 53, No. 19, pp.21692–21705.

Lei, Q., Zhang, H. and Du, J. (2021) 'Temporal attention learning for action quality assessment in sports video', *Signal, Image and Video Processing*, Vol. 15, No. 7, pp.1575–1583.

Lin, L., Fan, H., Zhang, Z., Xu, Y. and Ling, H. (2022) 'Swintrack: a simple and strong baseline for transformer tracking', *Advances in Neural Information Processing Systems*, Vol. 35, pp.16743–16754.

Liu, H., Huang, D. and Lin, M. (2024) 'FETrack: feature-enhanced transformer network for visual object tracking', *Applied Sciences*, Vol. 14, No. 22, p.10589.

Luvizon, D.C., Picard, D. and Tabia, H. (2020) 'Multi-task deep learning for real-time 3D human pose estimation and action recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 8, pp.2752–2764.

Mazinan, A. and Amir-Latifi, A. (2013) 'A new algorithm to rigid and non-rigid object tracking in complex environments', *The International Journal of Advanced Manufacturing Technology*, Vol. 64, pp.1643–1651.

Myer, G.D., Ford, K.R., Khoury, J., Succop, P. and Hewett, T.E. (2011) 'Biomechanics laboratory-based prediction algorithm to identify female athletes with high knee loads that increase risk of ACL injury', *British Journal of Sports Medicine*, Vol. 45, No. 4, pp.245–252.

Pu, L., Feng, X., Hou, Z., Yu, W. and Zha, Y. (2021) 'SiamDA: dual attention Siamese network for real-time visual tracking', *Signal Processing: Image Communication*, Vol. 95, p.116293.

Qin, Z., Yang, S. and Zhong, Y. (2023) 'Hierarchically gated recurrent neural network for sequence modeling', *Advances in Neural Information Processing Systems*, Vol. 36, pp.33202–33221.

Rana, M. and Mittal, V. (2020) 'Wearable sensors for real-time kinematics analysis in sports: a review', *IEEE Sensors Journal*, Vol. 21, No. 2, pp.1187–1207.

Shafiq, M. and Gu, Z. (2022) 'Deep residual learning for image recognition: a survey', *Applied Sciences*, Vol. 12, No. 18, p.8972.

Suresha, M., Kuppa, S. and Raghukumar, D. (2020) 'A study on deep learning spatiotemporal models and feature extraction techniques for video understanding', *International Journal of Multimedia Information Retrieval*, Vol. 9, No. 2, pp.81–101.

Wang, H., Zhang, S., Zhao, S., Wang, Q., Li, D. and Zhao, R. (2022) 'Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and SiamRPN++', *Computers and Electronics in Agriculture*, Vol. 192, p.106512.

Watanabe, Y., Oku, H. and Ishikawa, M. (2014) 'Architectures and applications of high-speed vision', *Optical Review*, Vol. 21, pp.875–882.

Xu, H. and Zhu, Y. (2020) 'Real-time object tracking based on improved fully-convolutional siamese network', *Computers & Electrical Engineering*, Vol. 86, p.106755.

Zhou, K., Ma, Y., Shum, H.P. and Liang, X. (2023) 'Hierarchical graph convolutional networks for action quality assessment', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 33, No. 12, pp.7749–7763.