Research on Real-Time Recognition and Posture Assessment of Tai Chi Cloud Hands Movement Based on Improved YOLOv7

Yazhou Song[1]
[1]Department of Physical Education and Teaching, Hebei Finance University, Baoding 071051, Hebei, China
Corresponding author: Yazhou Song (e-mail: syzwushu@163.com)

ABSTRACT Traditional Tai Chi Cloud Hands teaching is facing the challenges of frequent arm-crossing and large rhythmic variations , which make it hard for traditional methods to stably track joints and assess posture in real time. In this paper, we design an improved YOLOv7 dual task framework with spatiotemporal attention and human prior knowledge. Firstly, a lightweight coordinate attention mechanism is introduced to improve the perception of occlusion on joint. Secondly, a temporally aligned neck design is implemented. Optical flow is used to fuse inter-frame features and suppress jitter. Then, template geometric constraints are introduced. Sag and arc deviation angles are introduced into loss function for quantitative scoring. Finally, reparameterization and simplification of detection head guarantee real-time application. The experimental results show that our model can achieve an average recall rate of 91.7% for occluded joints, the improvement of 19.6% over the baseline in fully occluded situation, parameter amount of 38.3M, running time of 14. 2ms and correlation coefficient between score and judge of 0.882. This model effectively balances computational efficiency, joint localization accuracy, and professional biomechanical evaluation, providing a robust and feasible framework for the digitized instruction of traditional martial arts. The real-time posture framework can be combined with wearable motion sensors in traditional martial arts training.

INDEX TERMS human pose estimation,tai chi cloud hands,YOLOv7,posture evaluation,motion sensing

I. INTRODUCTION

TAI Chi Chuan, as an important traditional Chinese martial art and national fitness activity, directly affects the practice effect and sports injury prevention due to the standardization of its movements and the accuracy of its force transmission. Cloud Hands, as one of the core movements of Tai Chi Chuan, is characterized by continuous and smooth movements. It requires practitioners to accurately control the spatial position and trajectory of the wrist, elbow, and shoulder during the alternating arc of the arms, and to smoothly transition between the four forces of “Peng, Lu, Ji, and An”. However, in the actual teaching and assessment of Cloud Hands, the evaluation of the quality of the movements has long relied on the coach’s visual observation and subjective experience. Due to the characteristics of Cloud Hands practice, such as frequent limb crossing and occlusion and significant differences in individual practice rhythm, manual evaluation is not only inefficient, but also difficult to objectively quantify fine posture indicators such as elbow joint drop and wrist joint arc deviation angle. Traditional computer vision methods often suffer from problems such as missed joint detec tion, false detection, and failure to judge posture compliance when faced with the above continuous occlusion and rapid dynamic changes [1-3]. Therefore, developing a Tai Chi Cloud Hands movement recognition and posture evaluation algorithm that can balance real-time performance and accuracy is of great theoretical and practical significance for promoting the intelligentization and standardization of traditional martial arts teaching.

In terms of human posture estimation and movement recognition, most of the current research focus on two aspects: improving joint detection accuracy and spatiotemporal feature modeling. For example, Yang et al. [4] used HRNet to design a multi-scale feature refinement network to solve the problems of missing keypoint detection and positioning offset in occluded scenes. Jiang et al. [5] used importance weighting to design a multi-type feature fusion network. Fu et al. [6] combined attention mechanism with single stage detection framework and applied to the posture estimation and recognition task of fitness movement. Niu et al. [7] introduced attention mechanism into cascaded

pyramid network structure, and greatly improved the joint positioning accuracy under the condition of complex background and keypoint overlap.In terms of the intelligent recognition of traditional martial arts movement, most of the current research attempt to apply deep learning method to Tai Chi movement classification, and most of the works only focus on discrimination of discrete movement label and lack of in-depth research on the evaluation of continuous posture norm in the process of movement.

To solve the above problems, embedding attention mechanism and single-stage detection framework bring new ideas to the pose estimation in occluded scene in recent years. For instance, Hou and Chen [8] enhanced the idea of embedding attention mechanism in YOLOv7-Pose, and Li et al. [9] published their experience in lightweight multi-scale coordinate attention in human pose estimation. Second, as shown in the spatiotem poral learning Transformer framework proposed by Gai et al. [10], its rationality can reduce the computational overhead without sacrificing the real-time performance. Some scholars have added coordinate attention modules in the network to enhance the network's sensitivity to key directional information or used optical flow guided temporal feature alignment method to suppress the single-frame jitter. Inspired by this, this paper constructs an integrated algorithm framework for real-time recognition and pose evaluation of Tai Chi cloud hand movement. While specialized backbones like HRNet provide high-resolution refinement, they often carry a heavy computational burden that hinders real-time feedback. YOLOv7 is selected as the baseline because its single-stage detection paradigm offers an optimal trade-off between joint localization precision and inference latency, which is critical for the fluid, continuous monitoring required in Tai Chi teaching. Based on YOLOv7 as the baseline model, this paper firstly embeds lightweight adaptive coordinate attention module into the backbone network to enhance the spatial position awareness of occluded joint. Second, it designs temporal features to adapt neck structure, and uses the motion information of adjacent frames to guide feature fusion and suppress the pose misjudgment caused by the difference of rhythm. Third, it introduces geometric constraints from standard Tai Chi movement template into loss function to transfer joint regression results into quantifiable pose specification indicators. Finally, it uses reparameterization method to simplify inference structure to guarantee the real-time deployment of model.

II. OCCLUSION JOINT FEATURE ENHANCEMENT BASED ON ADAPTIVE COORDINATE ATTENTION

When doing Tai Chi Cloud Hands exercise, the arms will cross and arc in front of the chest repeatedly and the wrists will often cover the elbows and the elbows will often cover the shoulders and the shoulders will often cover the wrists.Although the original YOLOv7-Pose network has the ability to extract multi-scale features well, there is no explicit modeling of spatial positional relationship in the feature transfer between neck and detection head, which causes the response region of joint heatmap to shift or diffuse in case of occlusion, leading to missed detections and false detections.Therefore, this paper embeds a lightweight adaptive coor dinate attention unit into the end of the ELAN module of YOLOv7 backbone network to improve the sensitivity of network to the directionality of key joint spatial positions.

Firstly, the unit conducts global average pooling on the input feature map in both horizontal and vertical directions of the plane to obtain a pair of one-dimensional feature descriptors with directionality awareness. Then, one-dimensional convolution and nonlinear activation are used in each path to model long-range feature dependencies in row and column directions, and spatial attention weights are generated through gating mechanism finally. At last, the weight map is multiplied by the original input feature element by element to realize the adaptive enhancement of information at different spatial locations. Compared with traditional channel attention or two-dimensional spatial attention, coordinate attention can accurately reserve the positional prior of joint position with less computational overhead and is more suitable for joint localization with clear direction activities such as Tai Chi Cloud Hands [12].

In terms of module deployment strategy, this paper embeds the coordinate attention unit in parallel after the last three ELAN convolutional groups of the YOLOv7 backbone network using residual connections, ensuring that feature maps at high, medium, and low levels can all obtain enhancement of spatial structure information. During the training phase, this module is optimized together with the backbone network; during the inference phase, no additional computational branches are required, maintaining the forward efficiency of the overall network structure.

III. A TEMPORAL FEATURE ALIGNMENT MODULE FOR TAI CHI CLOUD HANDS THAT ADAPTS TO PERFORMANCE RHYTHM

Tai Chi Cloud Hands emphasizes "continuous and uniform movements," but in actual practice, there are significant individual differences in the speed and rhythm of movements among different practitioners. This paper introduces a temporal feature alignment module adapted to performance rhythm into the YOLOv7 neck feature pyramid structure, using motion cues between adjacent frames to guide the weighted fusion of feature maps. This module is located in the feature fusion stage between the backbone network output and the detector head, specifically consisting of a motion optical flow extraction layer and a deformable feature alignment layer. First, using the current frame as a reference, it backtracks one frame and calculates the sparse optical flow field between the two frames to obtain the motion vector distribution of each local region on the two-dimensional plane. Given that the Tai Chi Cloud Hands movements are generally smooth and mainly consist of circular motions, to reduce the computational overhead of

optical flow, this paper adopts a downsampling version of dense optical flow estimation based on the Farneback algorithm. To prevent motion aliasing during slow movements, the video sampling rate is maintained at 30 FPS, and the Farneback parameters are tuned with a larger window size to capture subtle displacements. A sensitivity analysis determined that a lower-bound displacement threshold is necessary to distinguish intentional slow motion from sensor noise, ensuring the temporal alignment remains active even during the most deliberate phases of the Cloud Hands cycle. Optical flow extraction is only performed on the low-resolution feature map at the top layer of the feature pyramid, and the motion vector is passed to the lower high-resolution branch as guiding information. To eliminate potential latency from inter-frame calculations, the system employs an asynchronous pre-fetching strategy where the optical flow estimation is performed in parallel with the back bone feature extraction. This optimization ensures that the temporal alignment module adds only minimal overhead (approximately 1.9ms), maintaining a total single-frame inference time of 14.2ms on standard GPU hardware, thus fulfilling the "real-time" requirement without perceptible lag.

After the inter-frame motion field is extracted, deformable feature alignment layer takes the feature map of corresponding level of last frame as input, and offsets the spatial position of feature map pixel by pixel by optical flow vector. Specifically, for any position on the feature map of current frame, based on the motion direction and amplitude predicted by optical flow field, bilinear sampling is implemented in the neighborhood of corresponding offset on the feature map of previous frame, and the finally sampled feature is concatenated with the feature of current frame in channel dimension. Finally, the concatenated features are adaptively weighted and summed together by the lightweight 1×1 convolutional layer to output the aligned temporally enhanced feature map. The whole process can be formalized as

$$F_{\text{align}}(t) = \text{Conv}_{1\times1}\big(\text{Concat}(F^{(t)}, \text{Warp}(F^{(t-1)}, \text{Flow}(t-1, t))\big) \quad (1)$$

where $F(t)$ represents the feature map of the current frame, $F(t-1)$ is the feature map of the previous frame, and $\text{Warp}(\cdot)$ represents the spatial sampling operation performed based on the optical flow field [13]. Considering that the first frame of a video does not contain a previous frame, and that optical flow estimation errors may occur during some scene transitions, this paper implements an adaptive gating switch in the temporal alignment module. When the average optical flow amplitude of adjacent frames is below a set threshold, the action is determined to be relatively static or extremely slow, and the module directly outputs the original features of the current frame to reduce invalid computation. When the optical flow amplitude is within a reasonable range, the module performs feature alignment and fusion normally. This design allows the network to dynamically adjust the utilization of temporal information according to the actual training pace, avoiding the introduction of redundant feature interference during slow training. To prevent potential feature "flicker" or temporal artifacts caused by frequent gating toggles, a temporal smoothing buffer is implemented. The gating switch only triggers when the optical flow amplitude consistently crosses the threshold for five consecutive frames, ensuring a stable transition between original and temporally enhanced features during the continuous Cloud Hands movement.

In terms of module deployment, the temporal feature alignment unit is embedded in parallel before the three output levels (P3, P4, P5) of the feature pyramid, ensuring that feature maps at different scales can obtain compensation for inter-frame motion information. During training, optical flow extraction and feature align ment operations are performed in conjunction with forward propagation, requiring no additional annotation; during the inference phase, only the cached features of the current and previous frames are retained, keeping the additional GPU memory overhead within an acceptable range.

IV. CONSTRUCTION OF A POSE EVALUATION BRANCH INTEGRATING STANDARD TEMPLATE GEOMETRIC CONSTRAINTS

Accurate detection of human joint coordinates is fundamental to action recognition. However, for Tai Chi Cloud Hands instruction, simply obtaining joint position information is insufficient to judge the standardization of the movement. The quality evaluation of Cloud Hands involves many professional details, such as whether the elbow joint remains droopy, whether the wrist joint's arc is smooth and fluid, and whether the distance between the arms remains relatively constant. This paper decouples and adds an independent pose evaluation branch based on the YOLOv7 detection head. By embedding geometric constraint terms of the standard Tai Chi movement template into the loss function, the network learns the standardized spatial structure features of the Cloud Hands movement simultaneously while regressing joint coordinates.

The input of the pose evaluation branch is the multi-scale fusion feature map output by the neck feature pyramid. After the channel dimensionality reduction and the feature refinement of the subsequent two con volutional layers, it is input into joint heatmap regression head and pose compliance scoring head respectively. The heatmap regression head follows the original structure of YOLOv7-Pose, outputting a two-dimensional Gaussian heatmap distribution of 17 joints for the subsequent coordinate decoding. While posture compliance scoring head is designed in this paper. Its structure is a lightweight fully convolutional sub-network containing three 3×3 convolutional layers and one global average pooling layer. It finally outputs a scalar to represent the comprehensive posture compliance score of the cloud

hands movement in the current input frame, and the value range of this score is [0,1].

To align the evaluation outputs with professional judging criteria, this paper establishes a comprehensive reference database of standardized Cloud Hands kinematics and designs a loss function incorporating three distinct geometric constraints. The specific construction process of template library is as follows: This paper invites three national level-two or above Tai Chi referees to annotate the joint frame by frame in the standard cloud hands performance video, and then extracts the normalized motion trajectory of the wrist, elbow and shoulder joint in the complete movement process. After temporal alignment and mean normalization, the spatial reference template of the standard cloud hands movement is formed. This paper defines three kinds of geometric constraint loss terms based on this template.

As shown in formula (2), the first term is the constraint of elbow joint drop. In Tai Chi cloud hands, the elbow can not be lifted or bent outward, so it must be kept in a dropped state. This paper takes the line connecting the shoulder joint and wrist joint as the reference axis, and calculates the vertical distance from the elbow joint to this line. It is compared with the standard drop distance in the template library, and then L2 norm of the deviation between the two is used as the drop loss term. To account for variations in body proportions among different practitioners, the elbow drop distance d is normalized by the individual's shoulder-to-wrist length. This transforms the absolute spatial coordinate into a relative posture ratio, ensuring that the scoring remains objective regardless of the practitioner's height or limb length.

$$\mathcal{L}_{\text{elbow}} = \frac{1}{N}\sum_{i=1}^{N}\left\| d\left(p_{\text{elbow}}^{(i)}, \overline{p_{\text{sho}}^{(i)} p_{\text{wrnst}}^{(i)}}\right) - d_{\text{ref}}^{(i)} \right\|_2 \quad (2)$$

Where $p_{\text{elbow}}^{(i)}$ is the predicted elbow joint coordinate of the i-th sample, $\overline{p_{\text{shoulder}}^{(i)} p_{\text{wrist}}^{(i)}}$ represents the straight line connecting the shoulder joint and wrist joint, $d(\cdot)$ is the Euclidean distance from the point to the straight line, $d_{\text{ref}}^{(i)}$ is the reference drop distance of the corresponding frame in the standard template, and N is the batch sample number.

The second term is the wrist joint arc deviation angle constraint. In the cloud hand movement, the wrist joint should run smoothly along an approximately circular arc trajectory. Deviating from this trajectory means that the movement has sharp angles or pauses. This paper calculates the cosine similarity between the local motion direction vector formed by the wrist joint coordinates of three consecutive frames and the reference motion direction vector of the corresponding frame in the template library. The mean of the angle between the two is used as the arc deviation loss term, as shown in formula (3):

$$\mathcal{L}_{\text{wrist}} = \frac{1}{N}\sum_{i=1}^{N}\arccos\left(\frac{\vec{v}_{\text{pred}}^{(i)} \cdot \vec{v}_{\text{ref}}^{(i)}}{\left\|\vec{v}_{\text{pred}}^{(i)}\right\| \left\|\vec{v}_{\text{ref}}^{(i)}\right\|}\right) \quad (3)$$

Where $\vec{v}_{\text{pred}}^{(i)}$ is the predicted motion direction vector of the wrist joint between the current frame and the previous two frames, and $\vec{v}_{\text{ref}}^{(i)}$ is the corresponding reference direction vector in the standard template. Third term is the arm spacing stability term. Since the cloud hand prohibits the arms to be apart or squeezed together during the alternating arc drawing process, the cloud hand sets the arms with a relatively stable spacing, such as a distance between the left and right wrist joints. The paper calculates the Euclidean distance between the left and right wrist joints, and then compares it with the reference spacing in the template library. L1-norm of the deviation is used as the spacing stability loss term as follows in equation (4):

$$\mathcal{L}_{\text{dist}} = \frac{1}{N}\sum_{i=1}^{N}\left| \left\| p_{\text{wrist\_L}}^{(i)} - p_{\text{wrist\_R}}^{(i)} \right\|_2 - d_{\text{ref\_dist}}^{(i)} \right| \quad (4)$$

Finally, the total loss function of the posture evaluation branch is composed of the weighted sum of the heatmap regression loss and the above three geometric constraint losses, as shown in formula (5):

$$\mathcal{L}_{\text{pose}} = \lambda_{\text{hm}}\mathcal{L}_{\text{heatmap}} + \lambda_e\mathcal{L}_{\text{elbow}} + \lambda_w\mathcal{L}_{\text{wrist}} + \lambda_d\mathcal{L}_{\text{dist}} \quad (5)$$

After experimental optimization, the weight coefficients are set to $\lambda_{\text{hm}} = 1.0$, $\lambda_e = 0.5$, $\lambda_w = 0.3$, $\lambda_d = 0.2$, respectively. During the training phase, the geometric constraint loss term also acts on the regression process of the joint coordinates, forcing the network to optimize the accuracy of the heatmap while taking into account the spatial consistency between the output joint configuration and the Tai Chi standard template. During the inference phase, the posture compliance scoring head directly outputs the comprehensive stan dard score, eliminating the need to recalculate geometric constraint terms and ensuring the real-time nature of the evaluation process. Table 1 summarizes the symbol definitions and corresponding weight configurations of the three geometric constraint loss terms and regression loss mentioned above.

TABLE 1. Geometric Constraint Loss Terms and Weights for Posture Evaluation Branch

Constraint Term	Symbol	Weight
Elbow droop constraint	$\mathcal{L}_{\text{elbow}}$	0.5
Wrist arc deviation angle constraint	$\mathcal{L}_{\text{wrist}}$	0.3
Arm spacing stability constraint	$\mathcal{L}_{\text{dist}}$	0.2
Heatmap regression loss	$\mathcal{L}_{\text{heatmap}}$	1

V. SIMPLIFIED REPARAMETERIZED INFERENCE STRUCTURE FOR REAL-TIME DEPLOYMENT

The essence of reparameterization technology is that training and inference are structurally decoupled. In training, the network uses multiple branches, which not only can make feature extraction more diverse but also can improve the stability of optimization; in inference, the multi-branch structure is effectively combined into one path directly connected convolutional layer, which can lower the times of memory access and computational cost. The updated detection head in this paper contains a joint heatmap regression branch and a pose compliance scoring branch. Both of two branches use multiple branches in training, and convolutional layers and identity mappings are used in the branch. Particularly, for the input feature map, this multi-branch structure has three paths, and the response of multi-branch is obtained by element-wise sum operation along the channel dimension of the output of three paths..

After training, this paper performs a reparameterization transformation on the multi-branch detection head. The transformation process consists of three steps. The first step involves fusing the kernels of the 1×1 convo lutional layer with those of the adjacent 3×3 convolutional layer. Since a 1×1 convolution can be considered a convolution operation with a spatial size of 1×1 and a stride of 1, its combination with the subsequent 3×3 convolution is equivalent to first linearly recombining the 3×3 convolution kernel with the 1×1 convolution kernel along the channel dimension. Specifically, the weight matrix of the 1×1 convolution kernel is weighted and superimposed channel-wise with the two-dimensional spatial kernel of the corresponding input channel of the 3×3 convolution kernel to obtain a new equivalent 3×3 convolution kernel. The second step is to convert the identity mapping branch into a 3×3 convolutional form. The identity mapping is equivalent to a 3×3 depthwise convolutional kernel with a center weight of 1 and weights of 0 at the other positions, with a one-to-one correspondence between input and output channels. The third step involves adding the parameters of all the 3×3 convolutional kernels obtained from the three path transformations according to channel alignment rules, while simultaneously accumulating the bias terms of each branch to a unified bias, ultimately fusing them into a standard 3×3 convolutional layer. After this transformation, the detection head structure in the inference stage degenerates from the original three-branch parallel topology to a single-layer 3×3 convolution, significantly reducing both the computation graph depth and the number of parameters. The reparameterized model only includes a single forward propagation path during inference, eliminating the need for element-wise addition of multi-branch results, effectively reducing memory read/write bandwidth pressure and computational latency.

VI. CLOUD HAND RECOGNITION EVALUATION EXPERIMENT SETUP AND INDICATOR DEFINITION

A. EXPERIMENTAL DATASET PREPARATION AND PLATFORM PARAMETER CONFIGURATION

To clarify the source of experimental data and the basic conditions for model training, the dataset size, annotation specifications, and hardware/software environment configuration details are shown in Table 2.

TABLE 2. Experimental Data and Platform Configuration

Category	Item	Specification
Dataset	Subjects	20 participants (12 novices, 8 professionals)
	Viewpoints	Front, left 45°, right 45°
	Sample size	15,680 frames, 17 keypoints annotated
	Data split	7:2:1 (training : validation : testing)
Annotation method	Labelme + cross-validation by 2 referees	
Hardware	CPU/GPU	i7-13700K / RTX 4090 (24GB)
	Operating system	Ubuntu 20.04
Software	Framework	PyTorch 2.0.1 + CUDA 11.8
Hyperparameters	Optimizer	AdamW, initial learning rate 0.001
Training setup	Batch size	16, 300 epochs
	Input size	640×640

B. ABLATION EXPERIMENTS

To further validate the effectiveness of each improved module in our method for the accuracy of cloud hand gesture recognition, we design the ablation experiments. We take the original YOLOv7-Pose as the basic model, and sequentially superimpose adaptive coordinate attention module, temporal feature-aligned neck structure loss branch and geometric constraint loss branch fusing Tai Chi standard template respectively, and form four kinds of variant models respectively. The same dataset and hyperparameters configuration are used to train all the models, and the average recall rate of occluded joints on test set is used as the evaluation index.

TABLE 3. Comparison of Ablation Experiment Results

Model Configuration	Average Recall (%)	Parameters (M)	Inference Time (ms)
Baseline (YOLOv7-Pose)	83.4	37.2	11.8
Baseline + Coordinate Attention (CA)	87.6	37.5	12.1
Baseline + CA + Temporal Alignment (TA)	90.2	38.1	13.7
Full Model (+ Geometric Constraint GC)	91.7	38.3	14.2

Table 3 shows that the baseline model achieved an average recall of 83.4% for occluded joints. After intro ducing the coordinate attention module, the network's spatial perception of the arm intersection area was enhanced, increasing the recall to 87.6% with only a 0.3M increase

in parameters. Further addition of a tem poral feature alignment structure resulted in a recall of 90.2%, indicating that inter-frame motion information effectively suppressed missed detections caused by single-frame pose jitter. Finally, the complete model, after adding a geometric constraint branch, achieved a recall of 91.7%, an improvement of 8.3 percentage points compared to the baseline. These results demonstrate that each improved module positively contributes to the cloud hand gesture recognition performance, and the increase in model parameters and inference time is within an acceptable range..

C. OCCLUSION ROBUSTNESS COMPARISON EXPERIMENT

To evaluate the reliability of our method under commonly occurring occlusion conditions in Tai Chi Cloud Hands, we randomly choose 300 frames of Cloud Hands pose (unoccluded, occluded, and fully occluded) from the test set. The average recall rate of wrist and elbow keypoints were calculated under different levels of occlusion for joint in both the original YOLOv7-Pose and our improved model, to evaluate the spatial inference ability and occlusion resistance of the model under the condition of limb crossing..

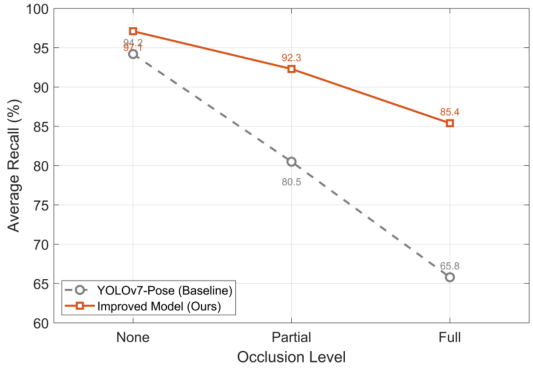

FIGURE 1. Occlusion Robustness Comparison Evaluation

As shown in Figure 1, the baseline model had an average recall rate of 94.2% under unoccluded conditions, which dropped to 80.5% under partial occlusion and further declined to 65.8% under full occlusion, indicating that the original YOLOv7 was more sensitive to the Cloud Hands double-arm crossing scenario. The improved model achieved a recall of 97.1% under unoccluded conditions, a 2.9 percentage point improvement over the baseline; a recall of 92.3% under partial occlusion, an 11.8 percentage point improvement over the baseline; and a recall of 85.4% under full occlusion, a 19.6 percentage point improvement over the baseline. These data demonstrate that by introducing coordinate attention and temporal alignment mechanisms, the model can effectively utilize spatial context and motion information to infer the position of occluded joints, significantly enhancing detection robustness in extreme occlusion scenarios.

D. HORIZONTAL COMPARISON EXPERIMENTS WITH MAINSTREAM ALGORITHMS

In order to objectively analyze the comprehensive performance of the proposed enhanced model in this paper, the horizontal comparison with current mainstream human pose estimation algorithms and object detection structure is performed, such as OpenPose, HRNet, original YOLOv7-Pose, YOLOv8-Pose. All comparison methods are retrained to converge on the self-built Yunshou dataset with the same training configuration. The input resolution is uniformly adjusted to 640×640 during the comparison method test. The average recall, model parameter quantity, and single frame inference time are selected as the evaluation index to evaluate the balance of recognition accuracy and application efficiency of each algorithm..

TABLE 4. Horizontal Comparison and Evaluation of Mainstream Algorithms

Algorithm Model	Average Recall (%)	Parameters (M)	Inference Time (ms)
OpenPose	85.6	51.4	46.2
HRNet	90.3	63.8	35.7
YOLOv7-Pose (Original)	83.4	37.2	11.8
YOLOv8-Pose	87.9	42.5	13.2
Improved Model (Ours)	91.7	38.3	14.2

As shown in Table 4, OpenPose, as a classic bottom-up method, achieves a recall rate of 85.6%, but its inference time reaches 46.2 ms, which is difficult to meet real-time requirements. HRNet achieves a recall rate of 90.3% due to its high-resolution representation, but its number of parameters is as high as 63.8 M, which is not conducive to lightweight deployment. The original YOLOv7-Pose inference takes only 11.8 ms, but its recall rate for occluded joints is only 83.4%. Our improved model achieves a recall rate of 91.7%, an improvement of

8.3 percentage points compared to the original YOLOv7-Pose and 3.8 percentage points higher than YOLOv8- Pose, with only 38.3 M parameters and an inference time of 14.2 ms, still meeting real-time processing requirements. These results demonstrate that our method achieves a better balance between accuracy and speed, possessing significant engineering application value.

E. SUBJECTIVE SCORING CONSISTENCY EXPERIMENT

In order to check the professional authenticity of the output scores of the posture assessment branch in this paper, we randomly select three national level-two or above Tai Chi referees to perform independent blind evaluation on 200 cloud hand movement samples in the test set. The scoring range is 0-10 points, and the average score of three referees is used as the referee reference score of that sample. At the same time, the improved model output is used to obtain a comprehensive quantitative score of the

elbow joint drop and wrist joint arc deviation angle for the same sample . Calculate the Pearson correlation coefficient between the two scores to evaluate the consistency between the evaluation results output by the model and the subjective judgment of human experts.

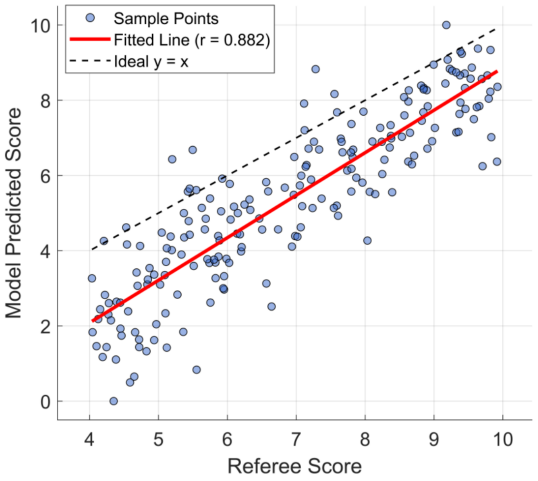

FIGURE 2. Subjective Scoring Consistency Assessment

As shown in scatter plot 2, the distribution trend of the model output score and the mean of the blind evalu ation by the three professional judges is highly consistent. Statistically, the Pearson correlation coefficient r between the two sets of scores is 0.882, indicating that the comprehensive index of elbow joint drop and wrist joint arc deviation angle quantified by the posture assessment branch in this paper has a strong positive correlation with the subjective experience judgment of human experts. The results confirm that by embedding the geometric constraints of the standard Tai Chi movement template into the loss function, the model can effectively learn and reproduce the judge's evaluation logic for the quality of the cloud hands movement, providing reliable algorithmic support for the subsequent construction of an objective and standardized Tai Chi teaching evaluation system.

VII. CONCLUSION

To solve the problems of being highly continuous, frequently occluded and subject to a high degree of subjectivity in evaluation of the Tai Chi cloud hands movement, this paper makes an improved YOLOv7 dualtask framework with spatiotemporal attention prior human constraints. Firstly, this paper adopts coordinate attention to improve joint perception in the case of occlusion, constructs temporal alignment to suppress the single-frame misjudgment and embeds geometric constraints to realize quantitative evaluation, and finished the design of lightweight model structure to guarantee the real-time. The experimental results show that this method can effectively improve the joint recall rate in the occluded case, and the quantitative score is highly consistent with the expert judgment. At present, this method is only applied in the occlusion of a single movement type and a fixed view. Although the high correlation (r=0.882) achieved in Cloud Hands demonstrates the model's accuracy for specific circular trajectories, the "professionalism" of the framework lies in its ability to translate traditional martial arts principles into quantifiable geometric constraints. This modular evaluation logic provides a validated methodology that can be extended to other core Tai Chi movements, representing a significant step toward an objective and standardized intelligent teaching system. Subsequently, we will extend this method to the evaluation of the whole Tai Chi routine and multi-view fusion, and attempt to deploy on mobile device.

AUTHOR CONTRIBUTIONS

Yazhou Song designed, collected and analyzed the data, and drafted the manuscript. Yazhou Song conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Yazhou Song participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Huang Qiong, "Application Research of Visualization Based on Attitude Estimation in 24 Style Tai Chi Teaching," Journal of Guangzhou Sport University, vol. 44, no. 1, pp. 1-9, 2024, doi: 10.3969/j.issn.1007-323X.2024.01.001.

[2] X. Feng, X. Lu, and L. Si X, "Taijiquan auxiliary training and scoring based on motion capture technology and DTW algorithm," International Journal of Ambient Computing and Intelligence, vol. 14, no. 1, pp. 1-15, 2023, doi: 10.4018/IJACI.330539.

[3] Z. Fan and F. Huang, "Tai Chi movement segmentation and recognition on the grounds of multi-sensor data fusion and the DBSCAN algorithm," Nonlinear Engineering, vol. 14, no. 1, Art. no. 20250151, 2025, doi: 10.1515/NLENG-2025-0151.

[4] N. Yang Q, D. Ji X, and H. Yang X, "Multi-scale feature refined network for human pose estimation," International Journal of Pattern Recognition and Artificial Intelligence, vol. 37, no. 13, Art. no. 2356022, 2023, doi: 10.1142/S0218001423560220.

[5] H. Jiang J, N. Xia, and Y. Zhou S, "A multi-type feature fusion network based on importance weighting for occluded human pose estimation," IEEE/CAA Journal of Automatica Sinica, vol. 12, no. 4, pp. 789-805, 2025, doi: 10.1109/JAS.2024.124953.

[6] C. Fu H, W. Gao J, and B. Liu H, "Human pose estimation and action recognition for fitness movements," Computers & Graphics, vol. 116, pp. 418-426, 2023, doi: 10.1016/j.cag.2023.09.008.

[7] Niu Yue, Wang Annan, and Wu Shengxi, "Attitude estimation based on attention mechanism and cascaded pyramid network," Journal of East China University of Science and Technology (Natural Science Edition), vol. 49, no. 5, pp. 724-734, 2023, doi: 10.14135/j.cnki.1006-30800.20220715003.
[8] Hou Xiangjun and Chen Yajun, "Human pose estimation algorithm based on improved YOLOv7 Pose," Journal of Chengdu University of Information Science and Technology, vol. 40, no. 4, pp. 441-445, 2025, doi: 10.16836/j.cnki.jcuit.2022.04.005.
[9] G. Xu and C. Wu X, "Human pose estimation method based on LSA-HRnet and its application in Taijiquan," Journal of South-Central Minzu University (Natural Science Edition), vol. 42, no. 6, pp. 839-845, 2023, doi: 10.12130/znmdzk.20230608.
[10] Y. Gai D, Y. Feng R, D. Min W, X. Yang, P. Su, Q. Wang, and Q. Han, "Spatiotemporal learning transformer for video-based human pose estimation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 9, pp. 4564-4576, 2023, doi: 10.1109/TCSVT.2023.3269666.
[11] J. Ding, W. Niu S, G. Nie Z, and Y. Zhu W, "Research on human posture estimation algorithm based on YOLO-Pose," Sensors, vol. 24, no. 10, pp. 3036, 2024, doi: 10.3390/s24103036.
[12] H. Ni Y, H. Wang, X. Mao J, et al., "Quantitative detection of typical bridge surface damages based on global attention mechanism and YOLOv7 network," Structural Health Monitoring, vol. 24, no. 2, pp. 941-962, 2025, doi: 10.1177/14759217241246953.
[13] F. Xu J, S. Komorita, and K. Kawamura, "FSPose: A heterogeneous framework with fast and slow networks for human pose estimation in videos," IEICE Transactions on Information and Systems, vol. E106-D, no. 6, pp. 1165-1174, 2023, doi: 10.1587/transinf.2022EDP7182.