JJMIE

Volume 20, Number 2, Special issue: Instrumentation & Measurements in Mechanical Engineering, April 2026
ISSN 1995-6665
Pages 217– 227
https://doi.org/10.59038/jjmie/200208
Editor-in-Chief: Prof. Faris M. AL-Oqla

Jordan Journal of Mechanical and Industrial Engineering

HRNet-Driven Key Point Detection and Action Synchronization in Dance Robot Choreography

Tingyu He*

Luoyang Vocational College of Science and Technology, Luoyang, 471000, China

Received 15 Sep 2025 *Accepted 12 Feb 2026*

Abstract

To solve the key point detection and motion synchronization in dance robot choreography, and meet the needs of real-time dance choreography, a technique for detecting key points and synchronizing actions based on high-resolution network design is designed. First, a multi-dimensional weighting and attention mechanism was introduced to improve the accuracy of high-resolution network pose estimation. Parallel multi-scale feature extraction and fusion mechanism enhances the detection of key points in orchestration. Second, considering the dynamism and coherence of continuous actions, an end-to-end object detection model and trajectory tracking technology are designed under cross-scale encoding to achieve motion adaptive adjustment. The results show that the improved high-resolution network outperforms other algorithms in keypoint detection accuracy, and the computing resource consumption is only 6.3% of that of high-resolution networks. The detection accuracy in both styles exceeds 90%. The tracking accuracy of the research method in jump, spin, and Thomas total spin exceeds 93%, which is superior to other comparative algorithms and performs well in action pose extraction and matching tasks. The proposed method has good applicability and robustness in key point detection accuracy and action synchronization effect, providing technical support and reference for dance robot choreography and personalized interaction.

Keywords: HRNet; Dance; Robot; Choreography; Attention mechanism; Key point detection; Object detection model; Cross-scale coding; Action tracking.

1. Introduction

Driven by artificial intelligence and robotics technology, dance robots, as an important branch of the integration of technology and art, are gradually moving from laboratories to more practical application fields, becoming a bridge connecting technology and humanities, and exerting enormous potential for application [1]. Dance robots mainly use high-precision sensors to capture the details of dance actions. They can also achieve dance choreography and present body language by imitating muscle action trajectories and programming content. This technology embodies a new paradigm of human-machine collaborative creation. The key to implementing this technology lies in accurately detecting the key points of choreographed actions and synchronously imitating actions [2, 3]. Dance robots are usually composed of mechanical structures, sensors, control systems, and software algorithms. They use sensors and control systems to obtain pose information and drive mechanical structures. Key point detection refers to identifying and locating the position of dance actions with representative nodes [4]. The choreography for dance robots requires high professional knowledge. Traditional key point detection methods for dance robots mainly rely on pre-programming and manual marking, which have limitations such as low efficiency, poor flexibility, and difficulty in adapting to complex actions, making it difficult to meet the detailed requirements of fine-grained dance actions (such as finger joint extension, wavy motion of the spine, etc.) [5].

Xu et al. utilized deep neural networks with support capabilities and key points for joint training to achieve robot key point detection. A new dataset was designed. The results indicated that the method had good accuracy in detecting key points, and the operation position could effectively determine the object's pose [6]. Murali et al. used scene representation methods and quaternion filters to achieve visual tactile interactive perception in robot pose estimation research. The results indicated that the method improved pose accuracy by over 35% in random scenes [7]. Matsuyama et al. analyzed the dance recognition robot. The long short-term memory network and time step trajectory of the robot could effectively improve its accuracy in character classification, with values much higher than the baseline results [8]. In response to the low efficiency and inability to process video data in traditional Laban dance score generation, Cai et al. extracted keyframes from dance videos, using multi-scale high-resolution network fusion to detect two-dimensional joint points, and then mapping them into a three-dimensional pose sequence through pose projection to generate an adversarial network. The results

* Corresponding author e-mail: hty1993425@163.com.

showed that this method could effectively digitize dance videos into dance scores, with efficiency far exceeding manual recording [9]. To address the challenges of dance pose estimation in complex scenes, multi-target tracking, and real-time performance, Zhao et al. proposed a DanceFormer model that integrated visual and time series transformers to fuse multi-modal features and real-time feedback. The results showed that the DanceFormer surpassed the existing models in pose estimation accuracy and multi-target tracking accuracy, and reduced the design delay of edge computing to 35.2 ms [10]. In response to the untimely guidance and subjective evaluation in traditional dance teaching, Liu et al. used OpenPose to detect body joints and then extracted posture information through deep neural networks. This method could quantify and visually display the differences between learners and standard actions, improving the accuracy of motion perception and quantitative evaluation [11]. The experiment showed that this improved method significantly enhanced the model's ability to handle scale changes. To address the challenges of pose recognition caused by occlusion and deformation in multi-person dance, Kao utilized cross-progressive multi-resolution representation integration technology and hierarchical pose recognition to achieve feature scale fusion and joint point matching. The results showed that this method significantly improved the robustness and accuracy of pose recognition, especially when dealing with complex and varied dance motions [12]. To achieve real-time and effective capture of dancer postures, Miao proposed a motion capture system based on heterogeneous sensors. The system utilized a single node heterogeneous sensor to collect real-time raw posture data of dancers, and filters and fuses it through data processing algorithms to solve the interference of magnetic fields and acceleration on posture calculation. The results showed that the deviation between the sensor output angle and the video extraction angle of the system was controlled within 1° [13]. To cope with the synchronized arrangement of music and dance, Fan and An extracted pose features by modifying visual transformers, captured spatiotemporal relationships between joints using graph convolutional networks, and quantified poses using methods such as K-means. The beat alignment loss function was optimized. The results indicated that the model could generate dance sequences that were highly synchronized with the rhythm of music, and had good performance in music motion correlation. It has strong application potential in AI choreography, virtual teaching, and interactive entertainment fields [14].

The current motion synchronization technology of dance robots is mostly based on rule-based temporal alignment or offline optimization, which is difficult to adapt to the rhythm changes and improvisational interactions in real-time performances. The research focuses on building a dance key point detection model adapted to High-Resolution Network (HRNet) to address the robustness in dynamic occlusion, fast motion blur, and multi-person collaborative scenes. Meanwhile, a motion feature encoding strategy is designed to convert the detected key point sequence into motion instructions that can be analyzed by the robot. This study adopts the HRNet to design key point detection for dance robot choreography, while recognizing dance robot actions based on multi-modal information. Compared with previous research, the innovation lies in two aspects. The HRNet is improved by introducing multi-dimensional weighting and Attention Mechanism (AM) to enhance pose estimation accuracy and address the information loss in resolution feature maps. Through parallel multi-scale feature extraction and fusion mechanisms, human key points can be accurately located while preserving spatial details, thereby detecting key points. In addition, to capture the dynamic changes of sequential actions, an end-to-end object detection model is designed, and cross-scale encoders and trajectory tracking techniques are introduced for target action recognition and adaptive adjustment.

2. METHOD

2.1. Key point detection of dance robot choreography based on HRNet

The key points of dance robot choreography are mainly used to recognize the accuracy and coherence of robot joints and body parts. It can use the results of human pose recognition to transform dance action execution sequences, and map joint movements by extracting key data [15]. Pose estimation can use image or video data to recognize joint positions and 3D poses, and its accuracy directly affects its visual presentation effect. Dance robots have high requirements for flexibility and accuracy. Strengthening key point detection is the technical challenge they face. High-resolution networks such as HRNet can effectively improve the extraction effect of local joint points. They can use a parallel structure to achieve fusion processing of feature information at different scales, and their image resolution after stage progressive processing is good, which can effectively ensure the accuracy of information extraction [16]. However, the HRNet structure has a high computational cost, and its weighting processing inevitably leads to fine-grained loss in pose estimation [17]. Therefore, the study improves the HRNet by introducing multi-dimensional weighting and AM to enhance pose estimation accuracy and address the information loss in resolution feature maps, resulting in an improved HRNet (Multi-dimensional Convolution High-Resolution Network Attention Mechanism, MC-HRNet-AM). Figure 1 is a schematic diagram of the MC-HRNet-AM.

In Figure 1, the MC-HRNet-AM consists of four stages, and each stage's parallel branch performs feature fusion processing. The sub-networks of the network gradually undergo resolution halving and width doubling processing as the stages progress. The first stage extracts features and expands the network width through 3×3 convolution and shuffle blocks. The subsequent stages repeat feature extraction and fusion, with each module performing Multi-dimensional Convolution (MC) operations. The AM in Figure 1 includes obtainable channel information and position information. The MC part consists of two parts: global information modeling and dynamic convolution. Global information modeling can model the context of space and channels. Spatial context modeling uses 3*3 convolution kernels to establish remote dependencies and achieve information exchange in high-resolution branches. Channel context modeling and dynamic convolution fuse receptive field features in the first and second branches to obtain key information related to joint points and achieve channel, spatial, and convolution weighting. Equation (1) is

the mathematical expression for modeling contextual information [18].

$$Z' = Z \otimes w(re(W_c \otimes W_1)) \quad (1)$$

In Equation (1), $w(re(W_c \otimes W_1))$ represents the long-range dependency weight matrix. $\otimes$ represents the tensor multiplication. Z and Z' represent the feature tensors of input and output. W_c is the tensor dimension of $N \times C \times HW$ (batch size × number of channels × feature height and width). W_1 is the tensor dimension of $N \times HW \times 1$ (batch size × feature height width × 1). w represents information fusion operation. re is the dimension for recovering weights after compression. Long distance dependency refers to the relationship between nodes in an image that are spatially distant but semantically strongly correlated. Traditional convolutional neural networks are difficult to directly capture this dependency due to their inherent local receptive fields. Introducing a mechanism similar to self-attention in spatial context modeling can calculate the correlation weight between any two positions on the feature map, thereby directly establishing a global dependency model. The weight matrix compresses the features in a predetermined dimension and performs normalization on the minimum channel features to obtain long-distance information. Afterwards, information integration is achieved through 1*1 convolution operation. Figure 2 shows the channel weighting structure.

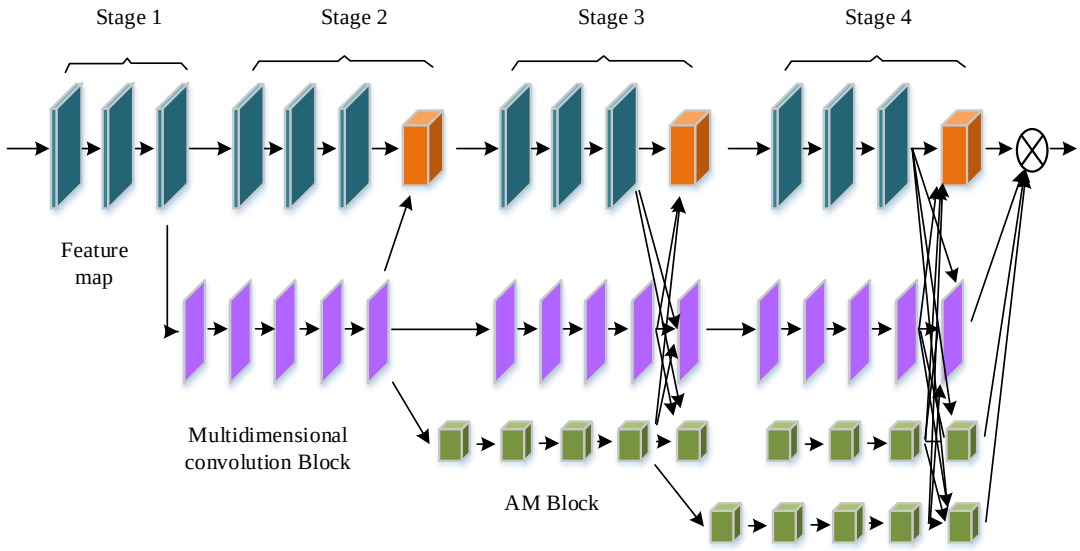


Figure 1. Schematic diagram of the network structure of MC-HRNet-AM

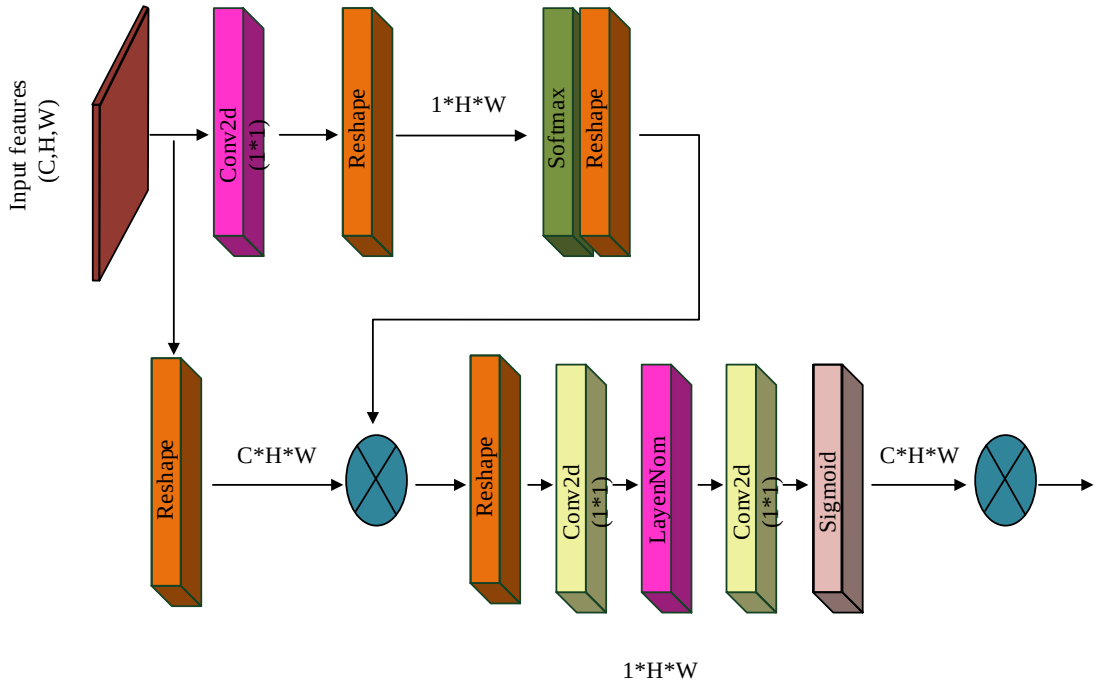


Figure 2. Channel weighting structure

Channel weighting has two compression matrices W_c and W_1. The non-linear information is added through normalization operation. Channel branching reduces global average pooling to perform information fusion. The parallel branches in the HRNet stage represent different network widths and resolutions, while spatial weighting in MC can compress channel spatial information and restore resolution after information fusion [19]. The study uses channel-by-channel weighting operation to achieve spatial weighting. Equation (2) is the mathematical expression of spatial weighting.

$$Z_s = Zus(Sig(conv(nor(conv(Z_b^{'} \in R^{c\times hm\times wm}))))) \quad (2)$$

In Equation (2), $conv$ is the convolution operation. nor is the normalization operation. us is the upsampling operation. Sig is a nonlinear function. b is the input number. $h_m \times w_m$ is the minimum resolution size. $Z_i^{'}$ is the compressed heatmap. In the dynamic convolution section, the study introduces channel information interaction and achieves dynamic information aggregation through channel information and parallel convolution kernels. Figure 3 shows the dynamic convolution processing.

Parallel convolutional kernels discard paranoid terms when inputting relevant attention and embed local channel information interaction content during information extraction, enriching aggregated information content with cross-channel attention and controlling computational costs. Equation (3) is the mathematical expression of dynamic convolution operation [20].

$$Y = \{\sum_{i}^{n}[Sig(conv2d(3\times3)(conv1d(3\times3)(Gap(z))))]W_i\} * Z \quad (3)$$

In Equation (3), information exchange can be achieved by performing convolution operations on the attention weights through dimensional convolution $conv$ and global pooling Gap and feature tensor Z. W_i is the weight matrix of the i-th convolution kernel. n is the network stage. $*$ is the convolution operation. To reduce memory consumption, the MC-HRNet-AM performs bilinear interpolation on the upsampling when generating new branches, while the downsampling part utilizes 3*3 convolution for operation. To avoid the loss of position and channel information in the feature map during the sampling process, this study utilizes the AM mechanism to implement attention weight processing for different position information to better manage feature information. Specifically, a coordinate attention module is introduced in the third stage of the HRNe to extract channel and position information at the fourth stage resolution. The score features obtained through fusion processing can be decoded to obtain the final key point coordinates, thereby achieving position prediction. The coordinate attention module encodes spatial coordinate information into the generated attention map by performing one-dimensional pooling along both horizontal and vertical directions, while capturing channel relationships and direction aware position information, enabling the network to more accurately locate the spatial positions of joint points. The SE module compresses spatial information into a single value through global average pooling, which results in the loss of positional information. The coordinate position block can be regarded as a computing unit that enhances the representation of the transformation tensor. It first embeds coordinate information and then performs pooling operations along the horizontal and vertical directions of the input feature map to obtain one-dimensional features $[T^h, T^w]$ in both horizontal and vertical directions. The obtained features are concatenated to obtain an attention map, whose mathematical expression is shown in equation (4).

$$f = \delta(F_1([T^h, T^w]) \quad (4)$$

In Equation (4), δ is the nonlinear activation function. F_1 is the intermediate feature mapping. Feature mapping can achieve spatial decomposition and obtain tensors with the same number of channels through convolutional transformation. By reducing the number of channels in feature mapping, the computational complexity can be reduced. Afterwards, the channel information and position information are connected using the self-attention module. Equation (5) is the mathematical expression of the self-attention operation [21].

$$c_h = W_c y_h + x_s \quad (5)$$

In Equation (5), c_h is the output feature. W_c is a learnable weight matrix. y_h is the output of the self-attention mechanism. x_s is a branch jump connection structure. The self-attention module can overcome the shortcomings of ordinary convolution and capture the internal correlation between channels and space without damaging the network structure [22].

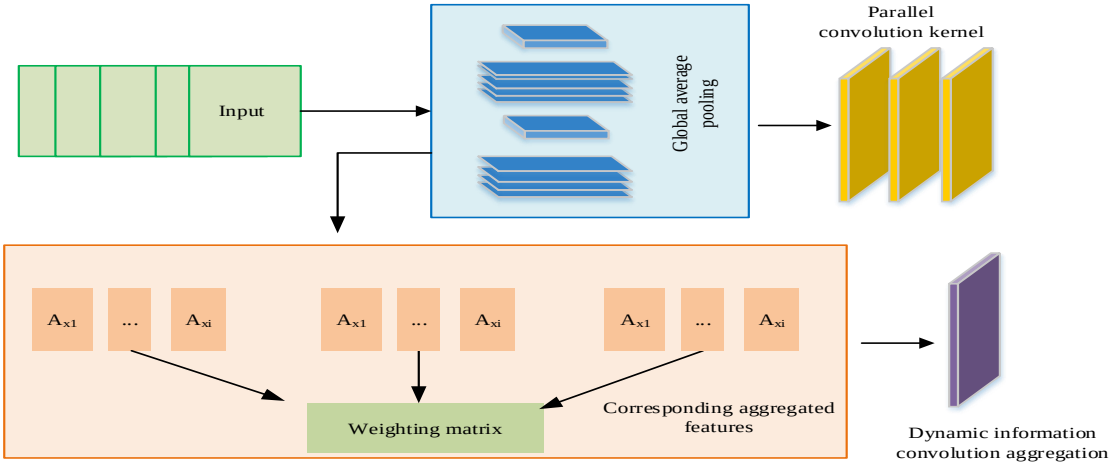


Figure 3. Content of dynamic convolution processing

2.2. Action recognition of dance robots based on multi-modal information

When compiling actions, dance robots should not only consider their accuracy, but also pay attention to their music emotions and rhythm. Simple pattern matching may reduce the flexibility and fun of the robot [23]. Music information is an important part of robot motion trajectory, and strengthening its motion pose recognition and adaptive adjustment can effectively achieve robot motion synchronization. To capture the dynamic changes of sequence actions, an end-to-end object detection model (Detection Transformer, DETR) is designed, and a cross-scale encoder is introduced to locate target actions. Figure 4 is a schematic diagram of the improved DETR structural framework.

In Figure 4, in the improved DETR structure, the input values are pre-convolved to generate an initial feature map. Then, multiple Hourglass modules are used to gradually extract multi-scale features. The output layer of each Hourglass module adds intermediate supervision and guides the network to learn more effective feature representations through a loss function. Input values can achieve fusion processing of different semantic features under cross-scale processing, where self-attention with position encoding can capture long-distance information. The Fusion module performs feature fusion processing. In the decoding feature query section, the study introduces the intersection over union ratio score into the loss function and defines the classification loss criterion function, as shown in Equation (6).

$$Loss = -\sum_{i}(\alpha \cdot (1-S_i)^{Y} \cdot IoU(P_i, G_i) \cdot CE(S_i, y_i)) \quad (6)$$

In Equation (6), P_i is the boundary box to be processed. S_i is the classification reliability. y_i is the actual label. CE signifies the cross entropy loss function. G_i signifies the real frame. IoU is the intersection over union ratio. α is a hyperparameter. The improved DETR can perform position encoding on the selected queries and feed the denoised samples into the encoder. The Across Attention module in the decoder can increase query selection, while the Deformable Attention module can process multi-scale features of convolutional structures while accelerating convergence speed [24]. Equation (7) is the mathematical expression of attention weights for Deformable Attention processing.

$$A_{mlqk} = soft\max(AttWe(z_q)) \quad (7)$$

In Equation (7), m represents the attention head. l represents the level. q is the query location. k is the sampling location. The attention weight $AttWe$ can be used to query the tensor z_q and calculate the sampling position using normalized $soft\max$ and sampling offset, achieving linear combination and output of different position values. Considering that the action information in different backgrounds may vary under changes in lighting and occlusion, it can lead to certain pose errors in action recognition. Therefore, to better integrate key point pose, based on the improved DETR model, this study combines RGB modal information to complete action recognition. The action recognition technology under dual modality is designed. Specifically, this action recognition technology includes two parts: modal information path and pose information path, with feature fusion achieved through bidirectional lateral connections between the two paths. Each path undergoes independent loss, with modal information reflecting pixel details and pose information reflecting semantic information of key points. The connection part uses time-step convolution operation to achieve fusion, that is, aligning the time dimension with convolution kernels, and concatenating RGB information and key point information in the channel dimension. Finally, the concatenated features are restored to the original number of channels to complete feature fusion. Afterwards, to achieve the accuracy of robot motion control, trajectory tracking technology (Cerebellar Model Articulation Controller, CMAC) is introduced to achieve adaptive control of local actions. CMAC, as a local approximation neural network, achieves nonlinear mapping of variable data by dividing the input space into multiple local regions such as associative space, receptive space, and weight space [25, 26]. Figure 5 is a schematic diagram of robot tracking control.

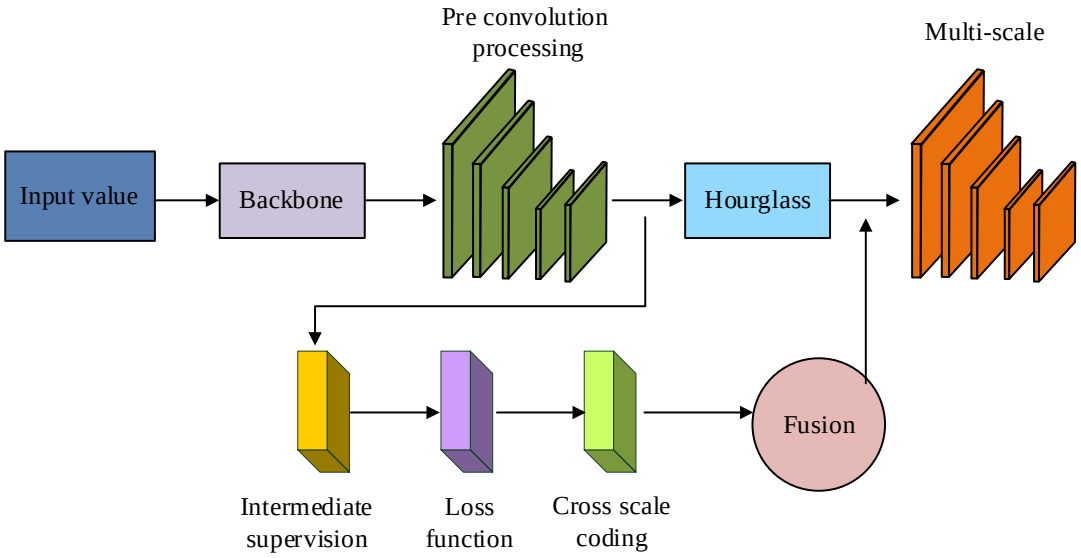


Figure 4. Schematic diagram of the improved DETR structural framework

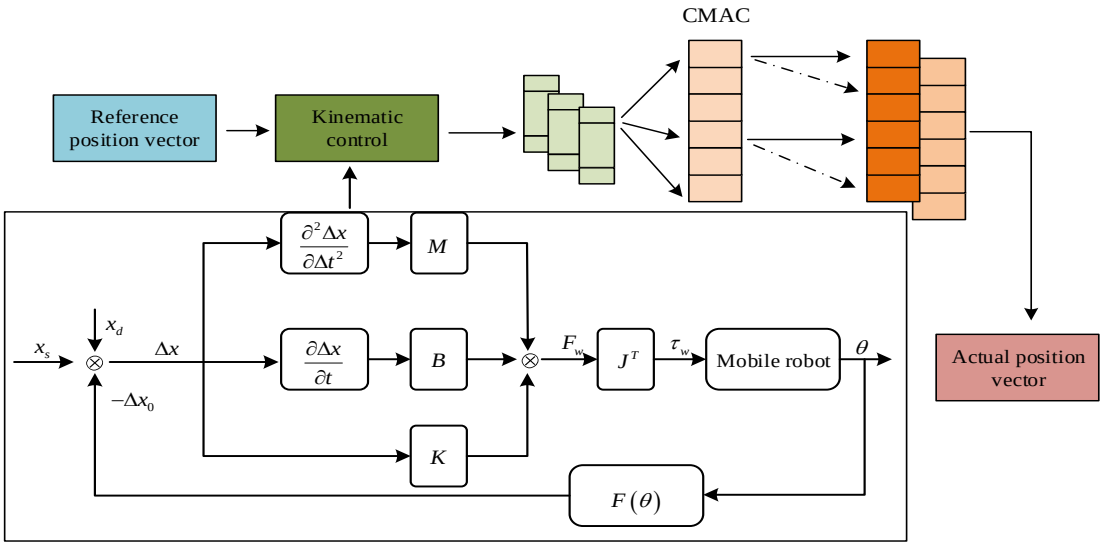


Figure 5. Trajectory tracking content

The key point information executed by MC-HRNet-AM and improved DETR model is saved as a coordinate set. The extracted two-dimensional pose is expressed as a heatmap. Stacking the heatmaps along the time dimension yields a sequence of key point heatmaps. The redundant parts of the heatmap sequence are removed. First, the rectangular box containing all key coordinate information is located and preliminarily cropped. The frames under each segment are uniformly sampled. Meanwhile, this study implements the robot dance system from both hardware and software aspects to better achieve robot dance action recognition and control. Dance actions are dynamically debugged and edited, and joint coordinate information can be recorded to generate action data tables and control robot actions based on position data to achieve synchronized tracking of actions. The key parts of the robot are equipped with specific motors to complete dance actions. As the control center, the research adopts ATmega8515L micro-controller (responsible for multi-channel servo control and path planning) and CPLD (detecting speed and providing motion feedback) for collaborative control [27].

3. RESULTS

To verify the effectiveness of HRNet-based dance robot choreography key point detection and action synchronization technology, a hardware environment (CPU: Intel i9-10900K, GPU: NVIDIA RTX 3090, and memory: 64 GB) is designed for model training and testing. The robot hardware platform adopts a multi-joint humanoid robot equipped with an IMU inertial measurement unit. The operating system is Ubuntu 20.04, the deep learning framework is PyTorch 1.9, and the programming language is Python 3.8. The upper computer development environment is Visual C++6.0, supplemented by Matlab to implement upper and lower computer control. The study pre-trained on a standard COCO dataset and fine-tuned and ultimately evaluated on a self-built dance motion dataset. This self-built dataset aims to compensate for the lack of diversity in dance motions in existing public datasets. This dataset contains approximately 2,000 video clips, each lasting 5-10 seconds, recorded by 10 amateur dancers of different body types under uniform background and lighting conditions. The content covers various styles, such as street dance, modern dance, and basic ballet actions. A semi-automatic annotation process is adopted. First, OpenPose is used to generate 17 preliminary COCO format keypoints, and then three annotators are manually refined frame-by-frame to ensure annotation quality. All data are strictly divided into training and testing sets in a 6:4 ratio, ensuring that video footage of the same dancer does not appear in both sets to avoid data leakage. To train the MC-HRNet-AM and improved DETR models, the Adam optimizer is uniformly used, with parameters set to β1=0.9, β2=0.999, and weight decay of 1e-4. The initial learning rate is set to 1e-4 and dynamically adjusted using cosine annealing strategy. The training batch is set to 32, with a total of 120 iterations. To improve the generalization ability of the model, different data augmentation techniques have been applied, including random horizontal flipping, random rotation within ± 15°, and random scaling from 0.75 to 1.25 times. In terms of loss function, the balance weight between classification loss and regression loss is set to 0.25 based on experience. To measure the accuracy of keypoint detection, the research adopts Percentage of Correct Key points (PCK) and mean Average Precision (mAP) for detection. The PCK is defined as the proportion of keypoints whose normalized distance between predicted keypoints and true keypoints is less than 0.5. The tracking performance of action sequences is measured using Multiple Object Tracking Accuracy (MOTA) and analyzed using joint angle sequence errors such as Mean Absolute Error (MAE). The HRNet model adopts the HRNet-W32 structure, with an input resolution of 256×256. Figure 6 shows the application results of different algorithms in key action detection. Firstly, the loss results of the MC-HRNet-AM are analyzed, as displayed in Figure 6.

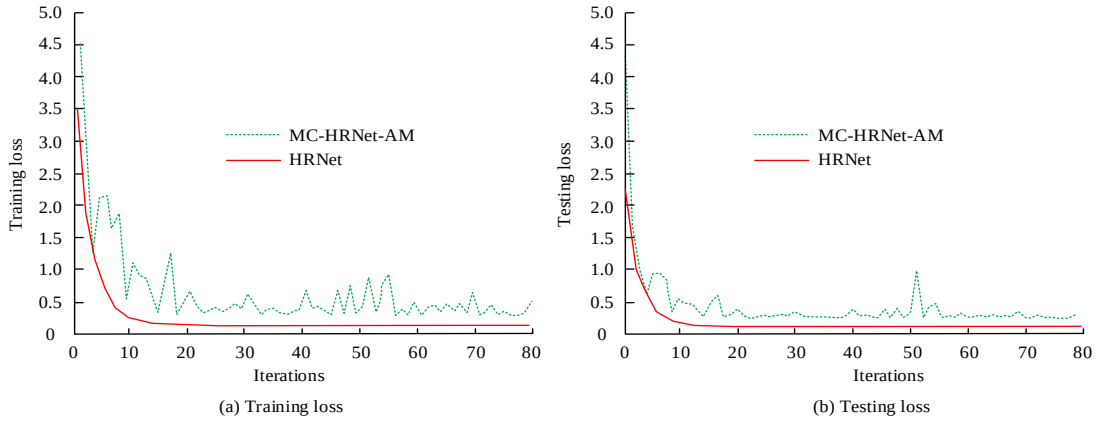
In Figure 6, the training and testing losses of the MC-HRNet-AM showed a significant downward trend with increasing iteration times, with average training and testing losses of 0.28 and 0.26, respectively. The loss value of the HRNet was also relatively small, but the loss curve exhibited significant fluctuations. The study introduces baseline models such as MobileNetV3 Large and ViTPose, and further compares them with the MC-HRNet-AM model to test their performance. Taking mAP, PCKh@0.5,

Robustness, Giga Floating-point Operations Per Second (GFLOPs), etc. as indicators, the results are shown in Table 1.

As shown in Table 1, compared with the original HRNet-W32, the mAP and PCKh@0.5 increased by 3.4 and 1.1%, respectively, while the computational and parameter requirements were relatively small. Although ViTPose-S performs well on the COCO dataset, the computational cost of MC-HRNet-AM is much lower on dance datasets that require higher dynamics and details. Especially in robustness testing for fast motion and occlusion, MC-HRNet-AM showed the smallest decrease in mAP (only 4.2%), which fully demonstrates the effectiveness of multi-dimensional weighting and AM in capturing key features in complex dynamic scenes. Subsequently, the accuracy and speed results of the proposed MC-HRNet-AM model are compared, as displayed in Table 2.

In Table 2, the test results of MC-HRNet-AM on the dataset achieved an Average Precision (AP) score of 72.15. The AP 50 and AP 75 values reached 92.55 and 78.67, respectively. The accuracy and efficiency were significantly better than those of other comparison networks. In terms of computational complexity, the GFLOPs value of MC-HRNet-AM was the smallest, reaching 0.65, which performed better than the accuracy under the same conditions. The AP values of Lite-HRNet and Dite-HRNet did not exceed 71, and the AP score of MC-HRNet-AM improved by more than 1% compared with other methods. The Average Recall (AR) value reached 77.06, indicating good performance in testing and application speed. Afterwards, ablation experiments are conducted to analyze the accuracy of action key point detection using the above methods, and two types of slow-paced and fast-paced are selected. Table 3 displays the comparison results.

The results in Table 3 indicated that MC-HRNet-AM model outperformed other algorithms in both music genres. In slow-paced music, HRNet-W32 algorithm had a maximum accuracy feature extraction difference of over 10%, while Lite-HRNet algorithm had a difference range of less than 5%. The accuracy ranking of algorithms for detecting action key points in fast-paced music is as follows: MC-HRNet-AM (90.52) > Lite-HRNet (88.75) > Dite-HRNet (84.33) > ShuffleNetV2 (83.48) > MobileNetV2 (80.11) > Resnet-50 (74.31) > HRNet-W32 (70.44). The above results indicate that the MC-HRNet-AM model can achieve good key point action detection and has good adaptability to music with different rhythms. Afterwards, the motion synchronization technology proposed in the study is subjected to motion pose analysis, as displayed in Figure 7.



(a) Training loss (b) Testing loss

Figure 6. Training loss and testing loss results of the fusion algorithm

Table 1. Comparison of performance of datasets with different pose estimation algorithms

Model	Dataset	mAP	PCKh@0.5	Robustness testing	GFLOPs	Params (M)
HRNet-W32	COCO	74.4	92.5	/	15.30	28.52
	Self-built dataset	85.1	95.3	7.8%		
MobileNetV3-Large	COCO	66.1	89.5	/	0.54	3.26
	Self-built dataset	79.2	91.8	11.7%		
ViTPose	COCO	75.8	93.1	/	10.28	20.17
	Self-built dataset	86.2	95.9	9.5%		
MC-HRNet-AM	COCO	72.8	91.9	/	0.64	1.54
	Self-built dataset	88.5	96.4	4.2%		

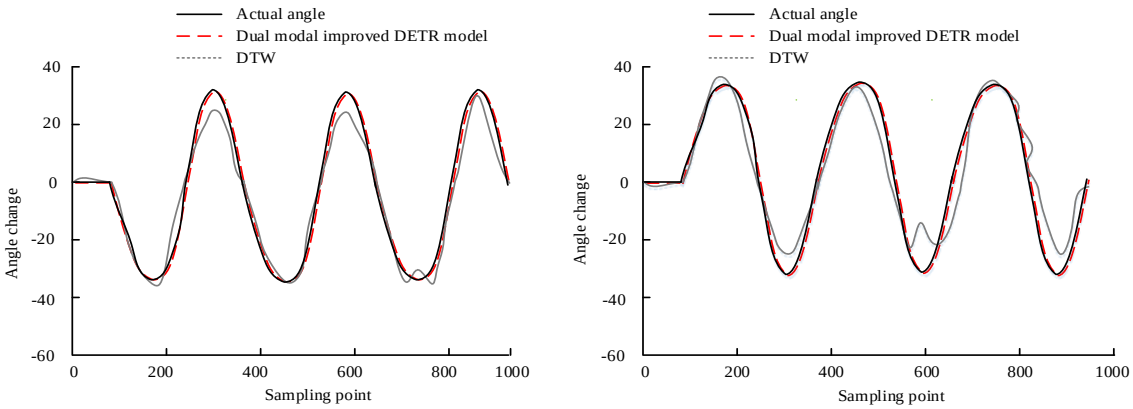
Table 2. Comparison of training experiment results of different models

Method	GFLOPs	AP	AP 50	AP 75	AR
Resnet-50	9.15	70.37	90.45	77.58	75.73
HRNet-W32	5.15	75.15	86.45	69.37	77.01
MobileNetV2	3.55	67.05	88.35	74.26	72.39
ShuffleNetV2	0.75	63.15	84.45	70.36	68.51
Lite-HRNet	1.35	69.95	90.45	77.16	75.33
Dite-HRNet	0.95	70.85	90.75	78.06	76.21
MC-HRNet-AM	0.65	72.15	92.55	78.67	77.06

In Figure 7, under a single action training, the recognition fit rate between the pose curve and the actual under the improved DETR model was over 94%, while the traditional Dynamic Time Warping (DTW) algorithm showed significant node fluctuations, with a maximum deviation of 3.17% from the improved DETR model. Under repeated action testing, the deviation amplitude of the pose curve between the two algorithms was expanded, reaching a maximum of 6.38%. Subsequently, three actions of jump, spin, and Thomas total spin are selected to analyze the synchronized tracking results of the improved DETR model. YOLOv7-Pose algorithm and hierarchical recurrent neural network (HQP-SWET-RNN) are selected as comparative literature, as displayed in Figure 8.

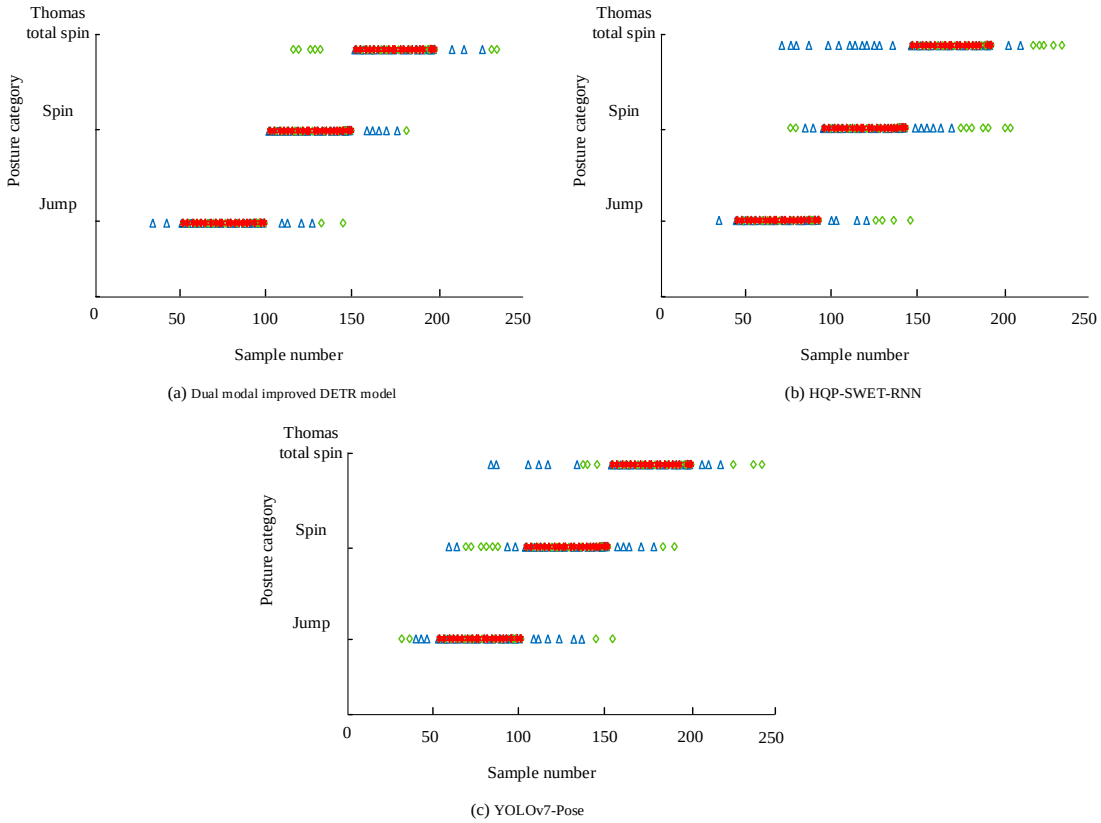
Table 3. Accuracy of action key point detection

Method	Slow-paced			Fast-paced		
	Accuracy	Recall	F value	Accuracy	Recall	F value
Resnet-50	74.35	78.23	75.48	72.34	75.36	74.31
HRNet-W32	77.47	80.52	79.21	70.32	70.23	70.44
MobileNetV2	82.25	81.49	81.71	80.07	79.32	80.11
ShuffleNetV2	84.36	85.32	84.38	83.25	82.43	83.48
Lite-HRNet	86.43	87.25	88.12	87.12	89.24	88.75
Dite-HRNet	85.31	84.36	87.12	83.24	81.16	84.33
MC-HRNet-AM	91.22	90.32	92.23	90.47	91.16	90.52



(a) Posture recognition results under a single action (b) Posture recognition results under repeated actions

Figure 7. Action pose curves of two algorithms



(a) Dual modal improved DETR model (b) HQP-SWET-RNN

(c) YOLOv7-Pose

Figure 8. Action synchronization tracking results

From Figure 8, the improved DETR model showed significant differences in tracking and matching accuracy compared with YOLOv7-Pose algorithm and HQP-SWET-RNN algorithm in jump, spin, and Thomas total spin actions. The tracking and recognition accuracy of the improved DETR model on three actions reached 94.8%, 93.4%, and 96.12%, respectively. The HQP-SWET-RNN algorithm had a higher number of motion pose matching errors than the research algorithm, while YOLOv7-Pose algorithm had poor matching performance. The improved DETR model can effectively reduce pose matching errors, judge action sequence similarity, and improve action synchronization accuracy. Subsequently, the recall performance of the three algorithms is analyzed, as displayed in Figure 9.

In Figure 9, during the training process on two datasets, the recall rate of the improved DETR model remained consistently above 0.60 and 0.65, and the recall rate curve showed minimal fluctuations regardless of the number of nodes. The recall rate of HQP-SWET-RNN algorithm was significantly affected by the sample size, while YOLOv7-Pose algorithm had the highest recall rate, approaching 0.70, indicating poor performance in extracting pose actions. The above results indicate that the improved DETR model can better adapt to pose actions of different difficulty levels, and the recognition effect is significant. However, YOLOv7-Pose lacks cross-frame correlation mechanism and HQP-SWET-RNN has poor timing dependency, which results in poor action tracking performance, and their robustness still needs further improvements.

4. DISCUSSION

A key point detection algorithm and action synchronization algorithm for dance robot choreography were designed based on HRNet. The proposed MC-HRNet-AM performed well in key point detection tasks, with an AP score of 72.15, significantly better than that of other small networks. The computational resource consumption was only 5.37% (parameter count) and 4.25% (FLOPs) of HRNet. High efficiency performance allows the MC-HRNet-AM to be deployed on embedded computing units within robots (such as NVIDIA Jetson series), achieving real-time local processing. This not only reduces dependence on expensive cloud GPU servers, but also eliminates network latency, providing possibilities for improvisational human-robot interaction and live performances, and greatly enhancing the system's robustness and cost-effectiveness. The improved MC-HRNet-AM model significantly improves the pose estimation accuracy and effectively addresses the information loss in resolution feature maps by introducing multi-dimensional weighting and AM. Moreover, the key point detection accuracy of MC-HRNet-AM was superior to that of the comparison algorithm in both slow-paced and fast-paced music, especially in fast-paced music, with an accuracy of 90.52%, which was 20.08% higher than that of HRNet-W32. MC-HRNet-AM still achieved 90.52% accuracy at fast paces, which means that the robot can accurately capture fleeting action details. For example, in Thomas total spin, higher detection accuracy allows the robot to reproduce leg trajectories more stably and avoid motion deformation. This indicates that MC-HRNet-AM can better adapt to complex dynamic scenes and has strong robustness. In terms of motion synchronization, the dual-mode improved DETR model designed in this study significantly enhances the dynamic capture capability of motion sequences by introducing cross-scale encoders and trajectory tracking techniques. Its pose curve fitting rate under single action training reached over 94%, showing higher accuracy than traditional time warping algorithms (the maximum deviation of 4.17%). In the repeated action test, the deviation amplitude of the pose curve of the improved DETR model was significantly smaller than that of the comparison algorithm, with a maximum deviation rate of only 6.38%. In addition, in the tracking experiments of jump, spin, and Thomas total spin, the accuracy of the improved DETR model reached 94.8%, 93.4%, and 96.12%, respectively, significantly higher than that of YOLOv7-Pose algorithm and HQP-SWET-RNN algorithm. The improved DETR model achieved a tracking accuracy of up to 96.12% and smaller pose curve deviation, which can effectively enhance the perfect synchronization of robot movements with music rhythm, thereby improving the watchability of dance robot movements and the immersive experience of human-robot collaborative creation.

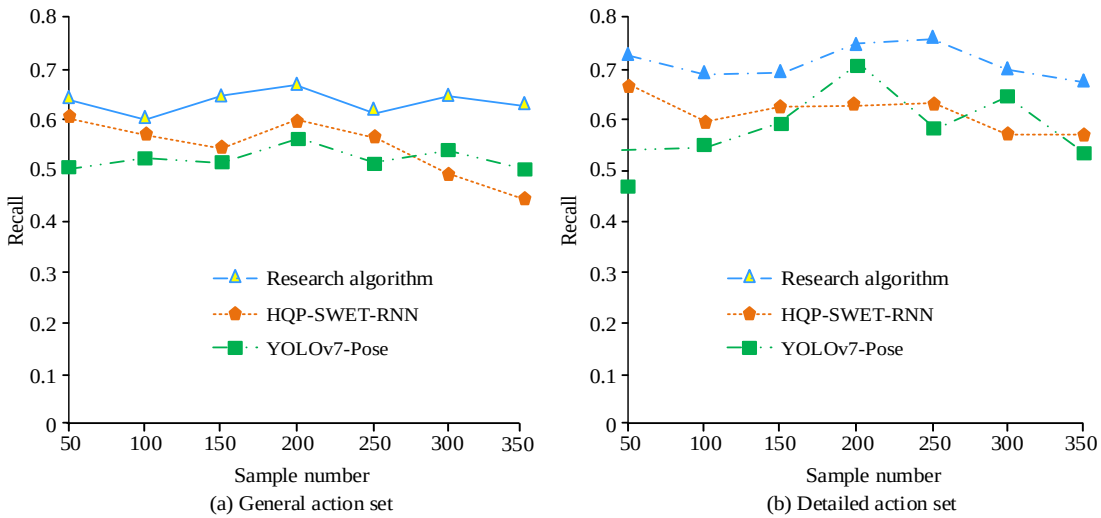


Figure 9. Comparison of recall under different models

YOLOv7-Pose, as a single-stage object detection algorithm, although has advantages in speed, it lacks the ability to model global contextual information in pose estimation tasks, which leads to key point missed or false detections in complex actions such as Thomas total spin. Moreover, YOLOv7-Pose relies on post-processing steps such as non-maximum suppression, which may lead to the loss of key point information and further reduce the accuracy of action tracking [28]. Although HQP-SWET-RNN can capture time series information, RNN is prone to gradient vanishing or exploding problems when processing long sequences, resulting in a decrease in the action tracking accuracy. Moreover, it is difficult to capture both global and local details of actions simultaneously, resulting in poor performance in complex actions such as spin. The improved DETR model shows a stable trend in recall rate on both datasets, while HQP-SWET-RNN algorithm and YOLOv7-Pose algorithm have significant fluctuations in recall rate and poor performance. The improved DETR model has higher stability and reliability in action pose extraction and matching tasks. The MC-HRNet-AM compensates for the shortcomings of standard convolution in capturing long-range dependencies between joint points through multi-dimensional and AM, reducing computational complexity compared to HRNet-W32. Lite/Lite HRNet mainly achieves lightweighting through ShuffleNet module or structural search, while the MC-HRNet-AM allocates computing resources more finely through composite weighting, and can optimize computing resources with attention guidance. The MOTR model aims to continuously track the IDs and bounding boxes of multiple common targets in the scene, while the MC-HRNet-AM is to recognize and understand dance action sequences of specific targets. It inputs bimodal information into the framework, using pose information as a strong prior to help the model focus on the action itself and achieve intelligent structural control.

5. CONCLUSION

Both the MC-HRNet-AM model and the improved DETR model demonstrate significant advantages in key point detection and action synchronization tasks. Multi-dimensional weighting and AM effectively improves feature extraction capabilities, while cross-scale encoders and trajectory tracking techniques enhance the dynamic capture ability of action sequences. The MC-HRNet-AM can effectively improve the accuracy of dance robot choreography, and can also adapt to different styles and rhythms of dance actions, with strong universality and robustness. The MC-HRNet-AM has good performance in dance robot choreography and synchronization tasks, but it still has certain limitations and challenges in practical deployment. Although the self-built dataset has been optimized for dance motions, it is still limited to specific dance styles and scenes. Faced with new dance styles that have not appeared in the training set, complex stage lighting effects, or unconventional clothing such as long dresses and props, the detection accuracy of the model may be challenged. The CMAC does not fully consider the dynamic constraints of the robot, such as joint torque limitations, self-collision, and interaction with the environment. When performing large and high-speed actions, the actual physical performance of the robot may deviate from the simulation results, and may even lead to hardware damage. The control model proposed in the study may experience cumulative delays in extreme situations, which can affect the perfect synchronization of robot actions with music beats, especially in extremely fast-paced music. Future research can further optimize model technology, design lightweight structures, and strengthen multi-modal data fusion and human-computer interaction technology processing to better explore its potential applications in more complex scenarios and broaden the applicability in different fields. Specifically, the aim is to introduce more diverse public datasets and utilize generative AI technology for data augmentation to enhance the model's generalization ability. It is possible to explore a robot controller based on reinforcement learning to integrate the dynamic model of the robot directly into the training process and conducts end-to-end performance analysis and optimization of the entire system to further compress the total system delay and achieve higher-level real-time synchronization.

Funding

This study was supported by Educational and Powerful Province Research of Henan Provincial Philosophy and Social Science (Project No. 2025JYQS1264).

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of data and material

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Competing interests

The authors have declared that no competing interests exist.

Authors' contributions

Tingyu He wrote the main manuscript text, prepared figures, tables and equations. Tingyu He reviewed the manuscript.

References

[1] L. Roda-Sanchez, C. Garrido-Hidalgo, A. S. García, T. Olivares, A. Fernández-Caballero, "Comparison of RGB-D and IMU-based gesture recognition for human-robot interaction in remanufacturing", The International Journal of Advanced Manufacturing Technology, Vol. 124, No. 9, 2023, pp. 3099-3111. https://doi.org/10.1007/s00170-021-08125-9

[2] S. Tsuchida, "Dance information processing: computational approaches for assisting dance composition", New Generation Computing, Vol. 42, No. 5, 2024, pp. 1049-1064. https://doi.org/10.1007/s00354-024-00273-2

[3] Y. Jang, I. Jeong, M. Younesi Heravi, S. Sarkar, H. Shin, Y. Ahn, "Multi-camera-based human activity recognition for human–robot collaboration in construction", Sensors, Vol. 23, No. 15, 2023, p. 6997. https://doi.org/10.3390/s23156997

[4] A. Tharatipyakul, T. Srikaewsiew, S. Pongnumkul, "Deep learning-based human body pose estimation in providing feedback for physical movement: A review", Heliyon, Vol. 10, No. 17, 2024, p. e36589. https://doi.org/10.1016/j.heliyon.2024.e36589

[5] G. Du, K. Wang, S. Lian, K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review", Artificial Intelligence Review, Vol. 54, No. 3, 2021, pp. 1677-1734. https://doi.org/10.1007/s10462-020-09888-5

[6] R. Xu, F. J. Chu, C. Tang, W. Liu, P. A. Vela, "An affordance keypoint detection network for robot manipulation", IEEE Robotics and Automation Letters, Vol. 6, No. 2, 2021, pp. 2870-2877. https://doi.org/10.1109/LRA.2021.3062560

[7] P. K. Murali, A. Dutta, M. Gentner, E. Burdet, R. Dahiya, M. Kaboli, "Active visuo-tactile interactive robotic perception for accurate object pose estimation in dense clutter", IEEE Robotics and Automation Letters, Vol. 7, No. 2, 2022, pp. 4686-4693. https://doi.org/10.1109/LRA.2022.3150045

[8] H. Matsuyama, S. Aoki, T. Yonezawa, K. Hiroi, K. Kaji, N. Kawaguchi, "Deep learning for ballroom dance recognition: A temporal and trajectory-aware classification model with three-dimensional pose estimation and wearable sensing", IEEE Sensors Journal, Vol. 21, No. 22, 2021, pp. 25437-25448. https://doi.org/10.1109/JSEN.2021.3098744

[9] X. Cai, T. Wang, R. Lu, S. Jia, H. Sun, "Automatic generation of Labanotation based on human pose estimation in folk dance videos", Neural Computing and Applications, Vol. 35, No. 35, 2023, pp. 24755-24771. https://doi.org/10.1007/s00521-023-08206-8

[10] H. Zhao, B. Du, Y. Jia, H. Zhao, "DanceFormer: Hybrid transformer model for real-time dance pose estimation and feedback", Alexandria Engineering Journal, Vol. 121, 2025, pp. 66-76. https://doi.org/10.1016/j.aej.2025.02.014

[11] Y. Liu, T. Zhang, Z. Li, L. Deng, "Deep learning-based standardized evaluation and human pose estimation: A novel approach to motion perception", Traitement du Signal, Vol. 40, No. 5, 2023, pp. 2313-2320. https://doi.org/10.18280/ts.400549

[12] H. Kao, "Multi-person dance tiered posture recognition with cross progressive multi-resolution representation integration", PLoS One, Vol. 19, No. 6, 2024, p. e0300837. https://doi.org/10.1371/journal.pone.0300837

[13] Y. Miao, "Dance pose capture and recognition based on heterogeneous sensors", Procedia Computer Science, Vol. 228, 2023, pp. 171-184. https://doi.org/10.1016/j.procs.2023.11.021

[14] W. Fan, X. An, "A deep learning based framework for music-synchronized dance choreography with pose quantization and motion prediction for activity recognition", Scientific Reports, Vol. 15, No. 1, 2025, p. 37248. https://doi.org/10.1038/s41598-025-21266-1

[15] H. Liu, T. Liu, Z. Zhang, A. K. Sangaiah, B. Yang, Y. Li, "ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human-computer interaction", IEEE Transactions on Industrial Informatics, Vol. 18, No. 10, 2022, pp. 7107-7117. https://doi.org/10.1109/TII.2022.3143605

[16] S. Dubey, M. Dixit, "A comprehensive survey on human pose estimation approaches", Multimedia Systems, Vol. 29, No. 1, 2023, pp. 167-195. https://doi.org/10.1007/s00530-022-00980-0

[17] Y. Bai, S. Mao, J. Zhou, B. Zhang, "Clustered tomato detection and picking point location using machine learning-aided image analysis for automatic robotic harvesting", Precision Agriculture, Vol. 24, No. 2, 2023, pp. 727-743. https://doi.org/10.1007/s11119-022-09972-6

[18] X. Wang, G. Li, F. Liu, "HRDS: A High-Dimensional Lightweight Keypoint Detection Network Enhancing HRNet with Dim-Channel and Space Gate Attention Using Kolmogorov-Arnold Networks", Electronics, Vol. 14, No. 10, 2025, p. 2038. https://doi.org/10.3390/electronics14102038

[19] G. Marullo, L. Tanzi, P. Piazzolla, E. Vezzetti, "6D object position estimation from 2D images: a literature review", Multimedia Tools and Applications, Vol. 82, No. 16, 2023, pp. 24605-24643. https://doi.org/10.1007/s11042-022-14213-z

[20] M. Huang, J. Gao, L. Ma, "Enhanced keypoint recognition framework via multi-scale feature characteristics", Scientific Reports, Vol. 15, No. 1, 2025, p. 40136. https://doi.org/10.1038/s41598-025-23831-0

[21] J. Yang, Y. Feng, "An optimization high-resolution network for human pose recognition based on attention mechanism", Multimedia Tools and Applications, Vol. 83, No. 15, 2024, pp. 45535-45552. https://doi.org/10.1007/s11042-023-16793-w

[22] M. Piazza, M. Maestrini, P. Di Lizia, "Monocular relative pose estimation pipeline for uncooperative resident space objects", Journal of Aerospace Information Systems, Vol. 19, No. 9, 2022, pp. 613-632. https://doi.org/10.2514/1.I011064

[23] S. Dubey, M. Dixit, "A comprehensive survey on human pose estimation approaches", Multimedia Systems, Vol. 29, No. 1, 2023, pp. 167-195. https://doi.org/10.1007/s00530-022-0980-0

[24] L. Lonini, Y. Moon, K. Embry, R. J. Cotton, K. McKenzie, S. Jenz, A. Jayaraman, "Video-based pose estimation for gait analysis in stroke survivors during clinical assessments: a proof-of-concept study", Digital Biomarkers, Vol. 6, No. 1, 2022, pp. 9-18. https://doi.org/10.1159/000520732

[25] J. Komorowski, M. Wysoczanska, T. Trzcinski, "Egonn: Egocentric neural network for point cloud based 6dof relocalization at the city scale", IEEE Robotics and Automation Letters, Vol. 7, No. 2, 2021, pp. 722-729. https://doi.org/10.1109/LRA.2021.3133593

[26] L. Roda-Sanchez, C. Garrido-Hidalgo, A. S. García, T. Olivares, A. Fernández-Caballero, "Comparison of RGB-D and IMU-based gesture recognition for human-robot interaction in remanufacturing", The International Journal of Advanced Manufacturing Technology, Vol. 124, No. 9, 2023, pp. 3099-3111. https://doi.org/10.1007/s00170-021-08125-9

[27] S. Choudhuri, S. Adeniye, A. Sen, "Distribution alignment using complement entropy objective and adaptive consensus-based label refinement for partial domain adaptation", Artificial Intelligence and Applications, Vol. 1, No. 1, 2023, pp. 43-51. https://doi.org/10.47852/bonviewAIA2202524

[28] L. Tong, R. Liu, L. Peng, "LSTM-Based lower limbs motion reconstruction using low-dimensional input of inertial motion capture system", IEEE Sensors Journal, Vol. 20, No. 7, 2020, pp. 3667-3677. https://doi.org/10.1109/JSEN.2019.2959639