



OPEN

Curvelet-enhanced transformer architecture for blurred action fine-grained detection

Yuxiang Ren¹, Zhetao Guo²✉, Wei Zhang³, Yushi Shen⁴ & Ying Xing⁵

This study proposes a novel Multi Curvelet Transformer Network (MCTN) for fine-grained human behavior recognition in dynamic video scenarios. A key challenge in this field lies in accurately identifying human actions under adverse conditions such as motion blur, occlusion, and varying illumination. To address this, we introduce a motion blur restoration module leveraging the curvelet transform to enhance motion image clarity, thereby improving downstream behavior detection. Furthermore, we enhance the Transformer architecture by embedding curvelet-based multi-scale attention mechanisms, which significantly improve the model's ability to extract spatial-temporal features at different resolutions. The proposed network also adopts a multi-curvelet transform structure to deepen semantic representation. Experimental results on benchmark datasets, including an action recognition dataset and the MSCOCO dataset, demonstrate that MCTN achieves superior performance, reaching a mean average precision (mAP) of 0.822. These results underscore the potential of MCTN in real-time intelligent video analysis and human-computer interaction applications.

Keywords Motion blur restoration, Multi curvelet, Transformer, Human behavior recognition

Accurate recognition of human behavior states in dynamic environments is essential for a wide range of intelligent video applications, including sports performance analysis, human-computer interaction, and surveillance. With the rapid development of image processing and computer vision technologies, it has become feasible to analyze and interpret human postures and actions from video data with increasing precision^{1–5}. Such capabilities not only facilitate performance monitoring but also enable targeted guidance in both professional training and automated systems.

However, human behavior recognition in high-speed and complex environments remains a significant challenge^{6–8}. Fast-paced actions such as rapid gestures, sudden turns, and subtle limb movements often occur within fractions of a second, making them difficult to capture and classify accurately. Moreover, the variability in individual motion patterns, including unconventional postures and spontaneous reactions, cannot be effectively handled by rigid rule-based methods. External environmental factors—such as dynamic lighting conditions, motion blur, occlusions, and camera noise—further degrade the reliability of visual features^{9–11}. Additionally, in scenarios where multiple subjects or objects (e.g., players and sports equipment) interact simultaneously, real-time tracking and behavior recognition become even more complicated due to frequent occlusion and spatial-temporal interference.

Another critical limitation lies in the availability and quality of annotated datasets. For niche activities or fine-grained motion analysis, labeled data often require expert involvement, leading to high annotation costs and limited sample sizes^{12,13}. These constraints highlight the urgent need for robust models capable of handling complex visual scenes with limited supervision, adaptable to various dynamic human activities.

Around these difficulties, player behavior detection^{14–18} has been proposed one after another. Nicolai et al.¹⁹ proposed the DeepSORT algorithm, which is a strategy for object tracking by the object detection. It uses the Hungarian algorithm to associate the tracking box and the detection box, fuses the appearance information and the Mahalanobis distance to obtain the best matching value, and then predicts the position of the tracking box at the next moment according to the Kalman filter. Finally, it decides whether to update the detection box according to the result and the strategy. Wu et al.²⁰ used a residual convolutional neural network to estimate the continuous 2D upper body pose of a table tennis player and then used a recurrent long short-term memory network to learn the serving motion of the player and predict the landing point of the table tennis player.

¹Beijing Dianjing Ciyan Network Technology Co.,Ltd, Beijing 100124, China. ²Cloudspace Technology Co., Ltd, Beijing 100176, China. ³The Third Affiliated Hospital, Beijing University of Chinese Medicine, Beijing 100029, China. ⁴NovNet Computing System Tech Company, Beijing 100081, China. ⁵School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China. ✉email: gzt@ieee.org

Huang et al.²¹ used OpenPose as a mankindkey point detection to recognize posture, and corrected the exercise training elements through the index scores to reduce the sports injuries of athletes. Aiming at the blurring phenomenon of acquired images caused by high-speed motion, the MixSort tracker proposed by Cui et al.^{22–24} is based on ByteTrack, OC-SORT, and MixFormer for end-to-end connection. However, despite the performance improvement, the algorithm still faces the challenges of real-time and computational resource consumption, especially when dealing with complex scenes and a large number of targets. Han et al.²⁵ suggested a combined asymmetric net and triple loss function of the tracking, which can prove the effectiveness of the complex moving object in the guarantee and the number of cases.

These methods have achieved certain results in the field of athlete behavior state detection research, and they have high accuracy for the state recognition of a single frame. However, complex environments and large motion actions need to solve image problems such as motion blur. To solve these problems, we propose a Multi Curvelet Transformer Network for Athlete Behavior Detection. The main contributions of this paper are as follows:

- (1) We propose a curvelet transformer-based motion blur restoration method, which exploits the relation between consecutive video frames to compensate for the information lost due to motion blur.
- (2) By incorporating curvelet transform into the self-attention mechanism, we deepen the understanding of the relationship between video content and improve the accuracy of action detection. At the same time, we also innovatively design a multi-curvelet transform structure, which can capture image information at different scales and deeply mine the deep semantic features of images.

Related works

Recent advancements in athlete behavior state detection have seen significant contributions, especially in the field of human pose detection. In the domain of 2D state detection, Wang et al.²⁶ proposed an hourglass structure that adaptively extracts features at multiple scales to better accommodate diverse human poses. While this approach is promising, it struggles with high computational complexity during the convolution process, despite efforts to mitigate this through downsampling and upsampling operations. Chen et al.²⁷ utilized a multi-scale pyramid method, dividing the image into various scales and processing them separately. Although this method improves accuracy by handling different scales individually, it suffers from the limitations of static feature extraction and does not account for dynamic changes in the athlete's movement. Sun et al.²⁸ introduced HRNet, a model that retains high-resolution feature information across multiple resolutions to improve the accuracy and speed of attitude detection. However, while HRNet excels at capturing detailed feature maps, it may face challenges in real-time processing, especially in highly dynamic environments such as sports. Xu et al.²⁹ proposed a multi-person pose detection system that first detects the object in the input image, isolates the human body, and then applies a separate network for keypoint detection. While this method is effective in handling multiple subjects, it requires robust handling of occlusion and interaction between players, which may not be adequately addressed in some cases.

In the field of 3D state detection, Ji et al.³⁰ demonstrated the feasibility of using a deep neural network (CNN) to directly predict 3D human pose with acceptable accuracy. However, despite the network's ability to predict 3D coordinates, the method struggles to generalize across complex and varying postures, particularly in the context of fast-paced athletic movements. Heravi et al.³¹ employed a combined model of CNNs and RNNs to learn both the spatial structure of human poses and the relative positions of joints. While these hybrid methods offer an improvement in capturing temporal information^{32–34}, it can still be computationally intensive, making it less suitable for real-time applications in sports. Jiang et al.³⁵ introduced a domain-based 3D human pose distribution model that predicts more diverse human poses with greater complexity. However, despite its capacity to predict a variety of poses, this method can suffer from inaccuracies in predicting poses with extreme angles or fast movements, which are typical in sports environments.

While recent techniques have achieved notable success in athlete behavior detection^{36,37}, they still exhibit critical limitations when deployed in fast-paced scenarios such as niche sports. In particular, applications to sports like pickleball are challenged by frequent motion blur, rapid temporal transitions, and multi-object occlusion, which significantly degrade model performance. Moreover, the scarcity of labeled training data in such domains—where annotations often rely on domain experts—further restricts the effectiveness of data-driven approaches and hinders generalization.

Methods

Aiming at the serious motion blur problem in motion images, we propose a multi curvelet Transformer network for athlete behavior detection method. To deal with the challenge of image blur during sports, we specially design the motion blur recovery (MBR) module, and refine the internal structure of the Transformer. This innovative design enables our model to effectively detect the behavior of blurred samples. An overview of the entire network is present in Fig. 1.

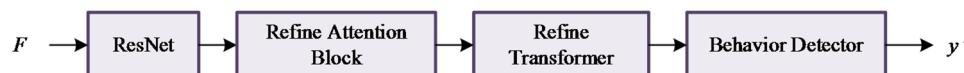


Fig. 1. The framework of MCTN for athlete behavior detection.

Motion blur restoration module based on curvelet transform

For any image F of a player, we first decompose it into a series of video frames, and represent these video frames by $\{f_i\}$ to form a video sequence. For the blurred frames, a homography model based on warp bundling is used to register multiple adjacent video frames to the blurred frame. To improve the registration, we employ the block method in the process of calculating the registration images of blurred video frames and their adjacent video frames. Specifically, we divide each video frame into several uniform blocks, and let the number of blocks in each video frame be Q , $I_{i,q}$ is a block in the video frame f_i , where q ranges from $[1, Q]$. To obtain the homography $G_{i,n,q}$, we use a warping-based motion model, which is described as Eq. (1):

$$\widehat{C}_{i,n,q} = G_{i,n,q} \bullet C_{i,n,q} \quad (1)$$

where $\widehat{C}_{i,n,q}$ and $C_{i,n,q}$ represent the position coordinates from image block $I_{i,n,q}$ to block $I_{i,q}$ before and after warping respectively, and the value range of n is $[0, N]$, where N denotes the account of adjacent video frames waiting for registration. The formula for calculating the registered image is Eq. (2):

$$I_{i,n,q \rightarrow i,q} = G_{i,n,q} \bullet I_{i,q} \quad (2)$$

where $G_{i,n,q}$ is further defined as the trainable homography, which represents the registration process from image block $I_{i,n,q}$ to block $I_{i,q}$, denoted as $I_{i,n,q \rightarrow i,q}$. In particular, when $n=0$, $I_{i,n,q \rightarrow i,q}$ represents the image block I itself. The adjacent video frame $f_{i,n,q \rightarrow i,q}$ after registration is composed of all the registered image blocks $I_{i,n,q \rightarrow i,q}$, where the value of q is in the range $[1, Q]$.

Equations (1) and (2) describe how each block of an adjacent video frame is geometrically warped to align with the corresponding block in the blurred reference frame. Conceptually, this can be understood as “shifting and stretching” small image patches so that overlapping structures (e.g., edges of limbs or equipment) match across frames. This alignment reduces inconsistencies caused by rapid motion. Once registered, the frames are processed in the frequency domain using the curvelet transform. Unlike the Fourier transform, which decomposes signals into sinusoidal waves, the curvelet transform provides multi-scale, multi-orientation representations, making it particularly adept at capturing directional features such as edges and contours.

Then, we process the registered video frames in the frequency domain. Given the continuous and large spatial span characteristics of motion actions, we introduce curvelet transform into MBR. The curvelet transform shows superior performance over the traditional transform in capturing edges and other exotic features. The curvelet transform contains three key parameters: scale (s), orientation (o) and position (p), which are used to accurately describe the characteristics of the transform. The basis function of the curvelet transform can be expressed as $\delta_{s,o,p}$. The construction of the curvelet transform is based on the radial window function and angle window function. The radial window function, designated as $\mu(w)$, is defined over the domain $w \in [1/2, 1]$, while the angular window function, denoted as $v(x)$, operates within the domain $x \in [-1, 1]$. Both functions must adhere to specific mathematical constraints, as outlined in Eqs. (3–4). However, since the frames of an image are discontinuous in both time series and pixel space, it is necessary to further convert them into discrete functional forms. Then, the curvelet transform is performed on the registered video frames to obtain the discrete curvelet coefficients $\partial'_{f_i}(s, o, p)$. The method of updating these coefficients is described by Eq. (9).

$$\partial'_{f_i}(s, o, p) = \sum_{n=0}^N W_{f_i}(s, o, p) \delta_{s,o,p}(y_{f_i}) \quad (3)$$

where $\partial'_{f_i}(s, o, p)$ denotes the discrete curvelet coefficient of video frame f_i , $\delta_{s,o,p}(y_{f_i})$ denotes the continuous curvelet coefficient of video frame f_i at position y , and $W_{f_i}(s, o, p)$ is the corresponding weight of the coefficient. The specific way to calculate the weight is given by Eq. (10).

$$W_{f_i}(s, o, p) = \frac{e^{\delta_{s,o,p}(y_{f_i})}}{\sum_{n=0}^N e^{\delta_{s,o,p}(y_{f_i})}} \quad (4)$$

where i in f_i is the set of low, middle and high frequency information of the image.

To facilitate understanding of the motion blur restoration pipeline, we provide a pseudo-code.

```

def motion_blur_restoration(frames):
    restored_frames = []
    for i, frame in enumerate(frames):
        # Step 1: Register adjacent frames using homography
        registered = []
        for adj in get_adjacent_frames(frames, i, N=5):
            H = estimate_homography(adj, frame) # Eq.(1)
            registered.append(apply_homography(adj, H)) # Eq.(2)

        # Step 2: Apply curvelet transform on registered frames
        coeffs = [curvelet_transform(r) for r in registered]

        # Step 3: Update coefficients (Eq.3-4)
        updated_coeffs = update_coefficients(coeffs)

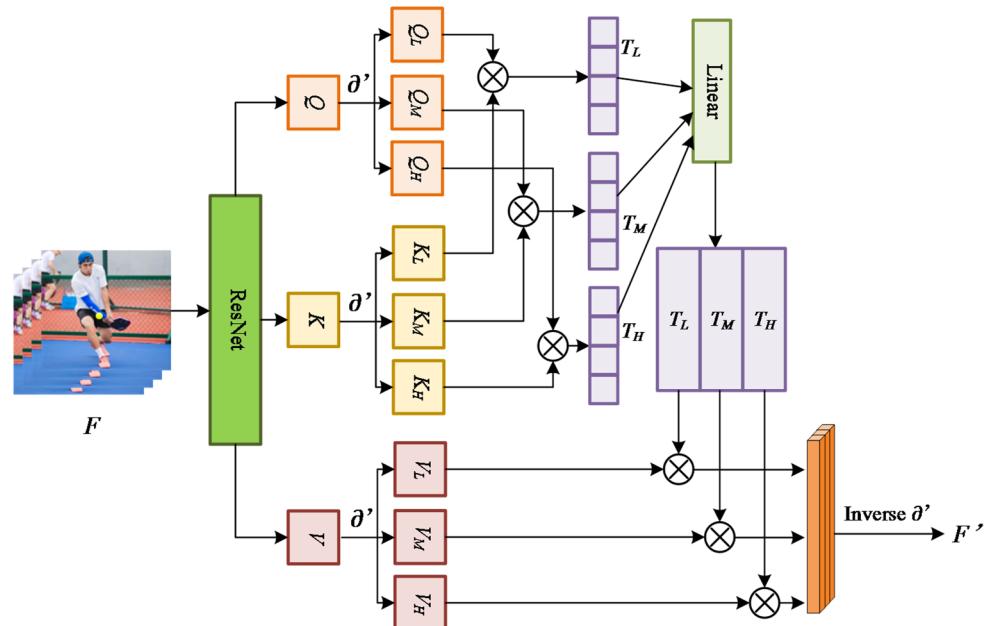
        # Step 4: Reconstruct restored frame
        restored = inverse_curvelet_transform(updated_coeffs)
        restored_frames.append(restored)
    return restored_frames

```

Algorithm 1: Pseudo-code for motion blur restoration using curvelet transform.

Refining transformer with MBR

Using the MBR module, we optimize and upgrade the Transformer to effectively deal with motion blur in images. We make a structural improvement to the self-attention mechanism by introducing MBR, as shown in Fig. 2. In this improved structure, we input the video frame F into the ResNet to obtain the corresponding features.

Fig. 2. The framework of refined attention block with MBR (θ').

Subsequently, these features are used as query (Q), key (K), and value (V), which are fed into the attention mechanism. The formula for the original self-attention mechanism (SA) is shown in Eq. (5) :

$$SA(F) = \frac{e^{\frac{(f_n W^Q)(f_m W^K)^T}{\sqrt{d}}}}{\sum_{k=1}^n e^{\frac{(f_n W^Q)(f_p W^K)^T}{\sqrt{d}}}} f_j W^V \quad (5)$$

Our process of using MBR to improve self-attention is shown in Eqs. (6–8):

$$f_i' = \partial \tau(f_i) \quad (6)$$

$$\sigma_{i,j} = \frac{(f_i' W^Q)(f_i' W^K)^T}{\sqrt{d}} \quad (7)$$

$$T_M = e^{\sigma_{i,j}} / \sum_{k=1}^n e^{\sigma_{i,j}} \bullet f_i' W^V \quad (8)$$

where M refers to a set that contains information in low, middle and high frequency. To integrate and utilize these features from different frequency bands, we adopt the strategy of concatenation and apply the inverse curvelet transform to the concatenation results, as shown in Eqs. (9–10) :

$$T = W([T_L, T_M, T_H]) \quad (9)$$

$$F' = \text{Inverse}(T) \quad (10)$$

where W is a trainable matrix, $[,]$ denotes the concatenation of features, and $\text{Inverse}()$ is the inverse curvelet transform calculation. We fully optimize the Transformer architecture by using MBR-improved self-attention (MSA). In addition, in the network architecture, we construct a Polycurvelet Transform (PCT) structure as shown in Fig. 3, which can deeply mine the content features in F_y .

We apply MCTN to the motion image processing of players. The proposed MCTN can not only effectively remove image noise, but also solve the problem of motion blur. Finally, we embed MCTN into various transformer-based behavior detection models to ensure accurate and efficient evaluation of the actions of players.

Equations (13)–(15) describe the integration of frequency-domain information into the self-attention mechanism. The intuition is that standard attention treats all pixel-level features uniformly, whereas the proposed MBR-enhanced attention assigns different importance to low-, mid-, and high-frequency components. Low-frequency signals capture global shape, mid-frequency signals capture texture and contour information, and high-frequency signals capture fine edges. By concatenating these multi-frequency features (Eqs. 16, 17) and performing an inverse curvelet transform, the model reconstructs feature maps that retain both global consistency and fine structural details. This enhancement enables the Transformer to more robustly attend to motion-relevant regions, even under severe blur or occlusion.

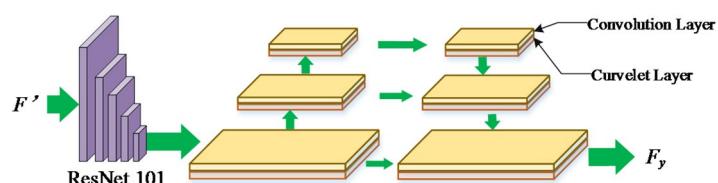


Fig. 3. The framework of multi curvelet transformation.

```

class MBR_Attention(nn.Module):
    def __init__(self, embed_dim, num_heads):
        super().__init__()
        self.attention = nn.MultiheadAttention(embed_dim, num_heads)
        self.W = nn.Linear(embed_dim * 3, embed_dim) # For concatenated
features

    def forward(self, features):
        # Decompose into frequency bands
        low, mid, high = curvelet_decompose(features) # Eq.(6-8)

        # Concatenate frequency bands
        concat_feats = torch.cat([low, mid, high], dim=-1)

        # Apply inverse curvelet transform (Eq.9-10)
        fused_feats = inverse_curvelet_transform(concat_feats)

        # Standard multi-head self-attention on enhanced features
        out, _ = self.attention(fused_feats, fused_feats, fused_feats)
        return self.W(out)

```

Algorithm 2: MBR-enhanced self-attention.

Experiment and analysis

Dataset and details

To evaluate the effectiveness of the proposed behavior state analysis framework, we conducted experiments on both domain-specific and general-purpose datasets. Specifically, we utilized the Pickleball Dataset (source) to assess the model's performance in a specialized sports context characterized by rapid and irregular movements. This dataset serves as a representative case for challenging real-world scenarios involving motion blur and fine-grained athlete behavior recognition. In parallel, we employed the widely-used MS COCO dataset³⁸ to validate the robustness and generalization capability of the model under diverse conditions. The COCO dataset includes annotations for 17 body keypoints per person, with over 1.5 million object instances spanning 80 object categories. It contains data from approximately 250,000 individuals, averaging 2 annotated persons per image, with some images featuring up to 13 individuals.

Parameters	Value
Initial learning rate	2×10^{-4}
Training rounds	40
Runtime per epoch	20.2 min
Batch-size	20
Decay	0.85
Optimizer	SGD
Loss functions	Cross entropy
CPU	Ryzen 7 9700X
GPU	RTX 4090
Image input size	512×512
Image feature dimension	1024

Table 1. Model parameter settings during training.

In the training pipeline, the model configuration parameters are shown in Table 1, covering the key elements of learning rate, training rounds, batch size, decay, and gradient descent. In the early stage of training, the model first uses the cross-entropy loss function for 20 rounds of basic training, aiming to obtain a relatively stable preliminary model. Subsequently, based on this, we further implemented 10 rounds of reinforcement learning training, aiming to deeply optimize the model through specific evaluation indicators. It is worth noting that the learning rate adopts a decreasing strategy during the training process. After every two rounds of training, the learning rate is reduced by 20% until the whole training process is completed. We adopt DETR³⁹ as the baseline model for our method.

In addition, to accurately evaluate the performance of MCTN, we use mAP (mean Average Precision) and OKS (Object Keypoint Similarity) as evaluation criteria, whose formulas are shown in Eqs. (11–13):

$$AP = P \times \sum_n (R_n - R_{n-1}) \quad (11)$$

$$mAP = \sum_i^N AP_i \quad (12)$$

$$OKS = \frac{1}{N} \sum_i^N v_i e^{-\frac{d_i^2}{2\sigma^2}} \quad (13)$$

where N refers to the total number of keypoints, d and v represent the Euclidean distance and visibility label of keypoints, respectively, and σ is a scale parameter.

Parameter experiments

Before conducting ablation experiments and comparison experiments, we need to determine some parameters of the model. Our parameter experiments, including the number of adjacent video frames N, the number of encod-decoder layers M of Transformer, and the number of CNN + Curvelet block layers L of PCT, are all performed in the complete MCTN.

Firstly, given that there is an interactive relationship between N and M, we decide to experiment on these two parameters simultaneously. The N parameter directly determines the number of relevant frames to be sampled, which has the influence on the curvelet transform. The M parameter is directly related to the output of MCTN. To ensure the accuracy of the parameter determination, we removed the PCT structure from the model during this experiment. The experimental results are reported in Fig. 4. After careful analysis, we find that when M is set to 6 and N is set to 5, the model can achieve the highest mAP of 0.856. This finding provides an important reference for us to optimize the model performance in the future.

Subsequently, we determined the parameter L. In the experiment, we adopted the optimal configuration determined earlier, that is, M is set to 6 and N is set to 5. The results are presented in Fig. 5. Through observation, we find that when the value of L is 4, the performance reaches the best, and the specific performance is AP⁵⁰, AP⁷⁵ and AP achieve 0.813, 0.912 and 0.837, respectively. We repeated the experiment several times and made sure that the conditions of each experiment were as consistent as possible. The results demonstrate that the model is stable and excellent under the condition of L = 4, which proves the effectiveness of the parameter configuration.

Ablative studies

To evaluate the effect of the three modules MBR, MSA, and PCT, ablative studies will be performed. Given that MBR is the basis of MSA and PCT, the specific setup of the experiment is shown in Table 2. In this table, we analyze the role of MBR in MSA and PCT in detail, and the results show that MBR improves the performance of MSA and PCT by 2.5% and 1.2% of AP scores, respectively. Further observed, MSA increased by 1.3% compared with the baseline model of AP scores, while PCT based on the MSA (w/MBR) of 1.2% of AP score again. Finally, when the three modules of MBR, MSA, and PCT are combined, the overall performance reaches the best, with the AP score of 0.813, AP⁵⁰ score of 0.912, AP⁷⁵ score of 0.837, AP^M score of 0.775 and AP^L score of 0.822.

As a base module, MBR is effective in the performance enhancement of subsequent modules. By introducing multi-scale and multi-direction curvelet transform, MBR enhances the model's ability to capture image features, and provides a richer information basis for the subsequent self-attention mechanism and Transformer. By improving the self-attention mechanism, MSA effectively uses the features provided by MBR, thereby improving the AP score, indicating that MSA can more accurately capture the key information and increase the accuracy of behavior detection. PCT further improves the AP score by 1.2%. This is due to the polycurvelet transform structure, which can deeply mine the deep semantic features of images at different scales. When the three modules of MBR, MSA and PCT are used together, the performance reaches the best, indicating that these three modules complement each other in function and jointly optimize the behavior detection. MBR provides a wealth of multi-scale and multi-direction features. MSA effectively utilizes these features by improving the self-attention mechanism. PCT further mines the deep semantic information of images. The collaborative work of these three modules enables the model to achieve significant improvement in multiple evaluation indicators.

Comparisons with state-of-the-art models

We compared the performance of our method with the SOTA (State-of-the-Art) methods, including SwinT⁴⁰, SimpleBaseline⁴¹, DERK⁴², HigherHRNet + SWAHR⁴³, AECA⁴⁴, EBA⁴⁵, TokenPose⁴⁶, RIFFormer⁴⁷, TransPose⁴⁸, HRNet²⁸, PRTR⁴⁹, BCIR⁵⁰, and SimCC⁵¹. To verify the broad applicability and extensibility of our method, experiments are implemented on the Pickleball Dataset and MSCOCO dataset respectively.

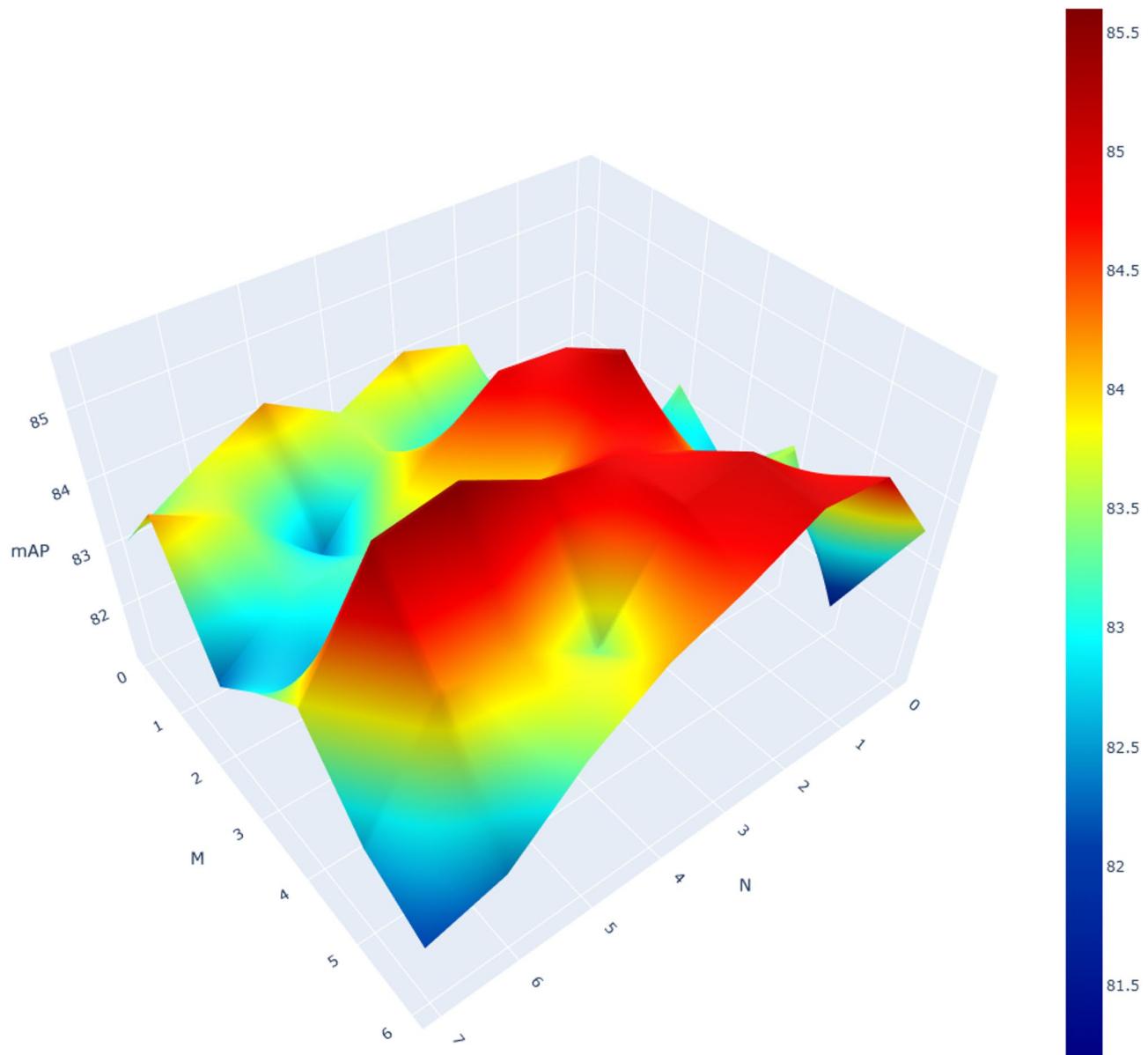
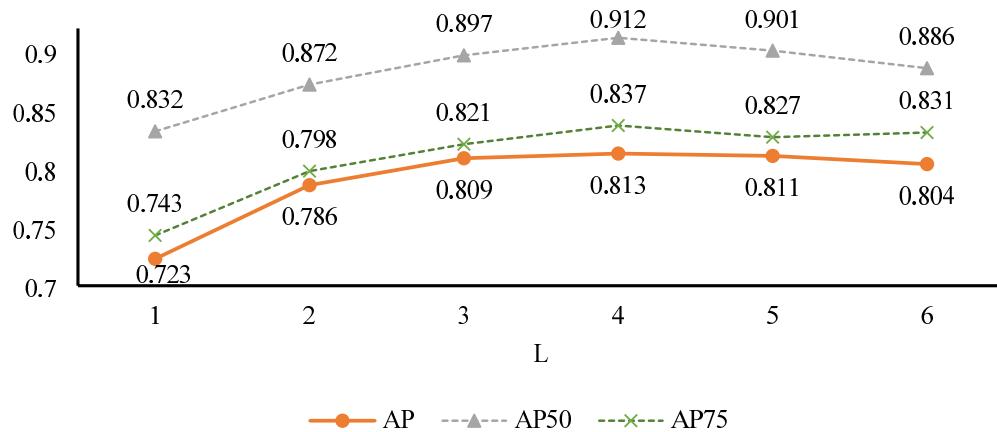


Fig. 4. The measurement of parameters M and N.

Comparisons on the pickleball dataset

First, we evaluate the performance of MCTN on the Pickleball Dataset. According to the data in Table 3, MCTN shows excellent performance, with an AP score of 0.813, an AP^{75} score of 0.837, an AP^M score of 0.775, and an AP^L score of 0.822, which surpasses all the methods participating in the comparison. Furthermore, we optimized and upgraded the backbone of MCTN by replacing the original DERT with the more advanced RT-DETRv3³². This change resulted in significant performance improvements, with the AP score jumping to 0.822, AP^{75} to 0.846, AP^M to 0.780, and AP^L to 0.844. This result strongly proves that MBR, MSA and PCT modules have excellent plug-and-play characteristics, which can be easily integrated into different backbones to improve performance. In MCTN, MBR enhances video frame clarity through deblurring algorithms, providing high-quality input for subsequent analysis. MSA strengthens the correlation of spatiotemporal features via multi-head parallel computation, enabling precise capture of athletes' dynamic motion patterns. PCT, through multi-scale and multi-directional geometric decomposition, breaks down complex motions into directionally sensitive feature sub-bands, enhancing the representation of details such as limb rotations and rapid movements. Together, these three components achieve full-pipeline optimization, spanning blur restoration, feature refinement, and high-dimensional semantic extraction. In addition, we also deeply study the influence of frame input size on model performance. By comparing the performance results of models with different frame sizes, we find that appropriately increasing the frame size can further improve the model to a certain extent.

Secondly, we conduct a comparative analysis of the running time, frame per second (FPS) and the parameters of the MCTN model, and the results are shown in Fig. 6. At the same time, Fig. 6 also visually shows the

**Fig. 5.** The choice of parameter L.

Methods	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L
Baseline	0.751	0.846	0.795	0.741	0.773
+MSA (w/o MBR)	0.764	0.864	0.799	0.742	0.786
+MSA (w/MBR)	0.789	0.869	0.812	0.756	0.798
++PCT (w/o MBR)	0.801	0.885	0.821	0.771	0.815
++PCT (w/MBR)	0.813	0.912	0.837	0.775	0.822

Table 2. Ablation experiments on pickleball dataset.

Methods	Backbone	Frame size	AP	AP ⁷⁵	AP ^M	AP ^L
SwinT	SwinT	256×192	0.755	0.795	0.723	0.801
SimpleBaseline	ResNet-50	256×192	0.748	0.768	0.704	0.781
DERK	HRNet-W32	512×512	0.751	0.776	0.743	0.802
HigherHRNet + SWAHR	HRNet-W32	512×512	0.786	0.799	0.771	0.793
AECA	ResNet-18	384×288	0.769	0.787	0.762	0.810
EBA	ResNet-18	256×255	0.799	0.821	0.756	0.809
TokenPose	TokenPose-L/D24	256×192	0.796	0.819	0.751	0.810
RIFormer	HRFormer-B	256×192	0.801	0.818	0.764	0.813
MCTN	DETR	256×192	0.813	0.837	0.775	0.822
MCTN	RT-DETRv3	256×192	0.822	0.841	0.778	0.846
MCTN	DETR	384×288	0.816	0.837	0.777	0.827
MCTN	RT-DETRv3	384×288	0.822	0.846	0.780	0.844

Table 3. Compare with methods on the pickleball dataset. In this experiment, we re-implement the SOTA methods on this dataset.

performance of each model in the form of a line chart. It can be concluded that the model of MCTN(w/ DETR) has 24.5 M parameters and the execution time is 96ms (10.4 FPS). The MCTN(w/ RT-DETRv3) model has more parameters, reaching 32.8 M, and its running time is a little longer, 126ms (7.9 FPS). Compared with other methods, it can be concluded that the improvement of the performance of MCTN depends not only on the increase of the parameters and running time, but also on the rationality of the model structure design and the effectiveness of the algorithm optimization. By introducing innovative modules such as MBR, MSA and PCT, MCTN realizes the efficient capture and accurate analysis of the key information of the image, to achieve performance improvement while maintaining low computational complexity.

Finally, we present the behavior analysis results of player images in the form of visualization, as shown in Fig. 7. For each athlete, the top row shows the input frames under four conditions—Original, Motion Blur, Low Light, and High Light—while the bottom row presents the corresponding pose estimation outputs generated by the proposed model.

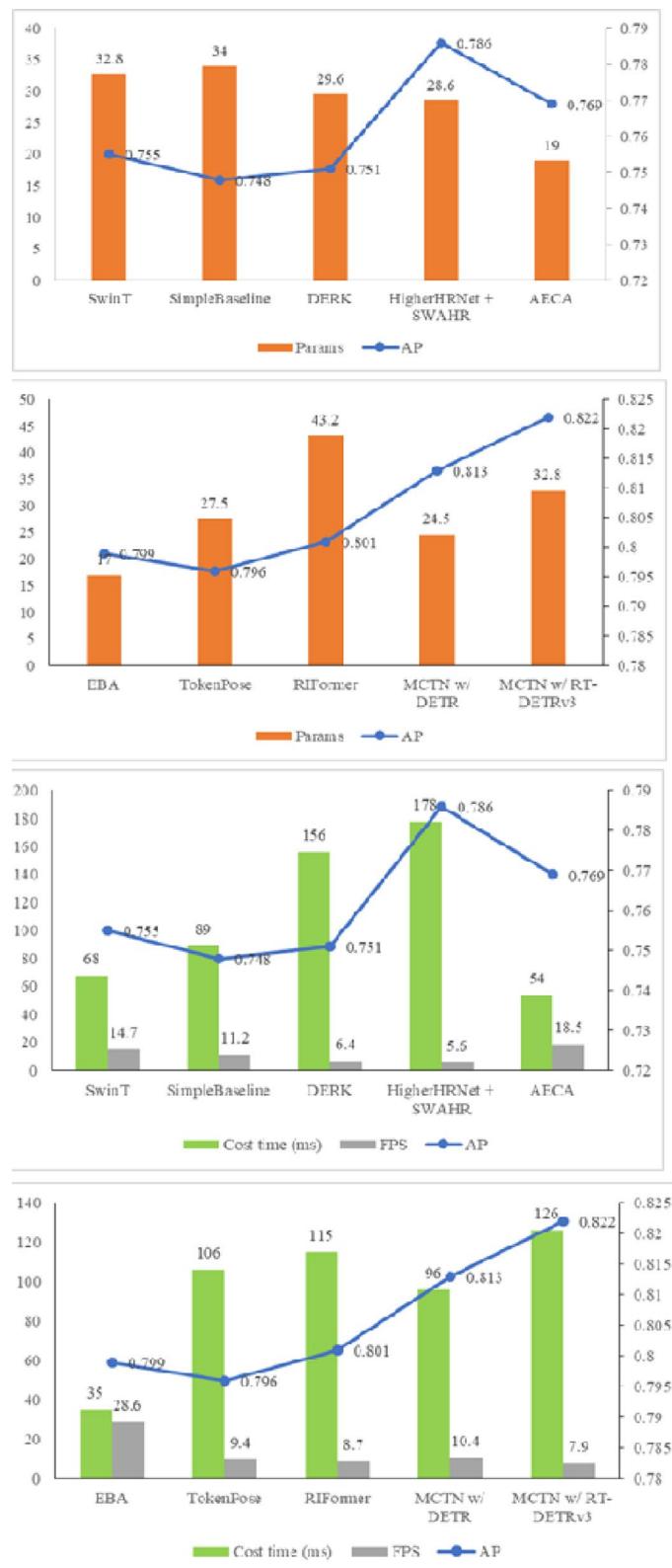


Fig. 6. Comparison with other methods on the Pickleball dataset in terms of cost time FPS and parameters.

It can be observed from the figure that MCTN can still stably and accurately capture the keypoints of the human body, such as shoulder, elbow, knee, etc., even when the athlete's movement changes rapidly and the image appears to be blurred to a certain extent. This is crucial for subsequent applications such as behavior recognition, action analysis, and athlete training feedback. In the original and moderately degraded conditions,

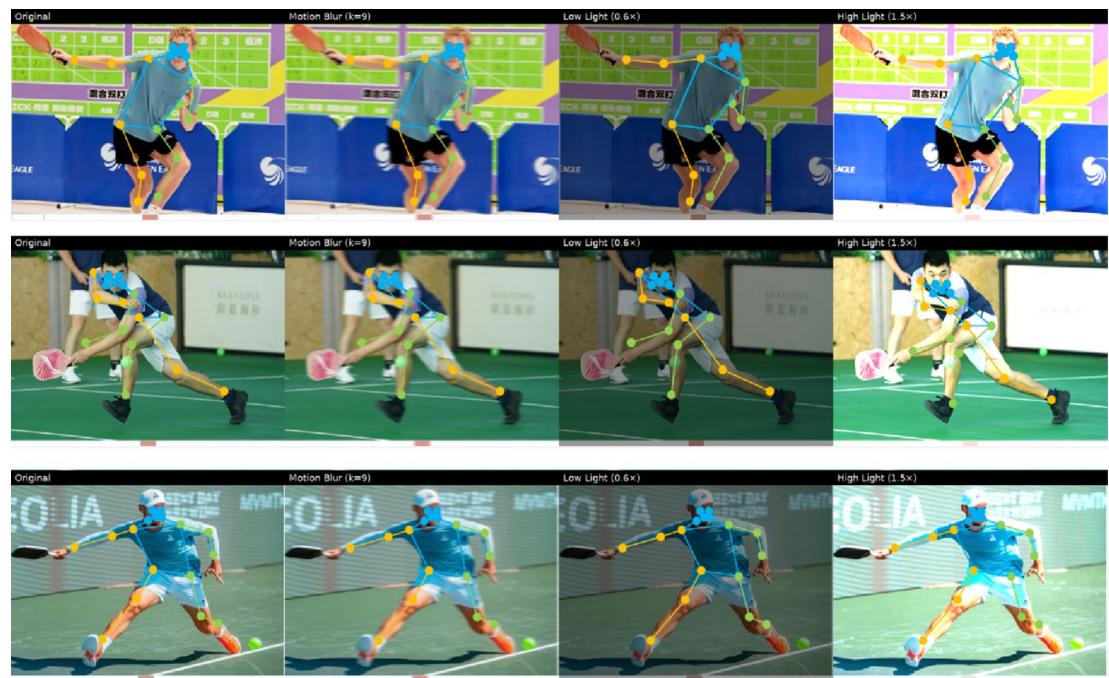


Fig. 7. Visualization of MCTN performance under different input conditions.

Methods	Backbone	Frame size	Parameters (M)	AP	AP ⁵⁰	AP ⁷⁵
TransPose	TransPose-H-A4	256 × 192	17.3	0.753	—	—
SimCC	ResNet-50	256 × 192	25.7	0.708	—	—
HRNet	HRNet-W32	256 × 192	28.5	0.734	0.895	0.807
PRTR	ResNet-50	384 × 288	41.5	0.682	0.882	0.752
EBA	ResNet-18	256 × 256	17.0	0.713	0.915	0.781
RIFformer	HRFormer-B	256 × 192	43.2	0.756	0.908	0.828
BCIR	ResNet-50	256 × 192	34.0	0.675	0.872	0.740
AECA	ResNet-18	384 × 288	19.0	0.745	0.925	0.814
MCTN	DETR	256 × 192	24.5	0.759	0.926	0.822
MCTN	RT-DETRv3	256 × 192	32.8	0.767	0.938	0.836
MCTN	DETR	384 × 288	34.6	0.761	0.922	0.828
MCTN	RT-DETRv3	384 × 288	40.4	0.766	0.941	0.833

Table 4. Comparison with other methods on the MSCOCO validation set.

the model consistently captures the global body structure and accurately localizes major joints, ensuring reliable pose estimation. Under motion blur, the curvelet-based motion restoration and frequency-aware attention enable the network to preserve critical edge details, which mitigates the loss of high-frequency information and supports stable detection of limb orientations. Similarly, in both low-light and high-light environments, the multi-scale curvelet representation provides enhanced adaptability by balancing global shape cues and local textures, thereby reducing the negative impact of illumination imbalance. Thanks to its internal MSA and PCT, MCTN can capture subtle changes in human posture at different scales, and uses time series information to enhance the understanding of motion patterns, to effectively deal with the challenge brought by motion blur. Nonetheless, the results also reveal limitations: in extreme cases of blur or strong lighting, distal joints such as the ankles or wrists occasionally deviate from their true positions, reflecting the difficulty of capturing fine-grained motion details under severe visual degradation. These findings confirm the effectiveness of MCTN in handling common real-world disturbances, while suggesting potential improvements through integration with temporal modeling or adaptive illumination normalization.

Comparisons on the MSCOCO dataset

To show the stability of MCTN, we conduct extended tests on the MSCOCO Dataset. The experimental results of the Validation set are shown in Table 4. It can be found that MCTN achieves the 0.766 AP score, 0.941 AP⁵⁰ score and 0.833 AP⁷⁵ score on the dataset. Meanwhile, MCTN with different backbones and different input sizes

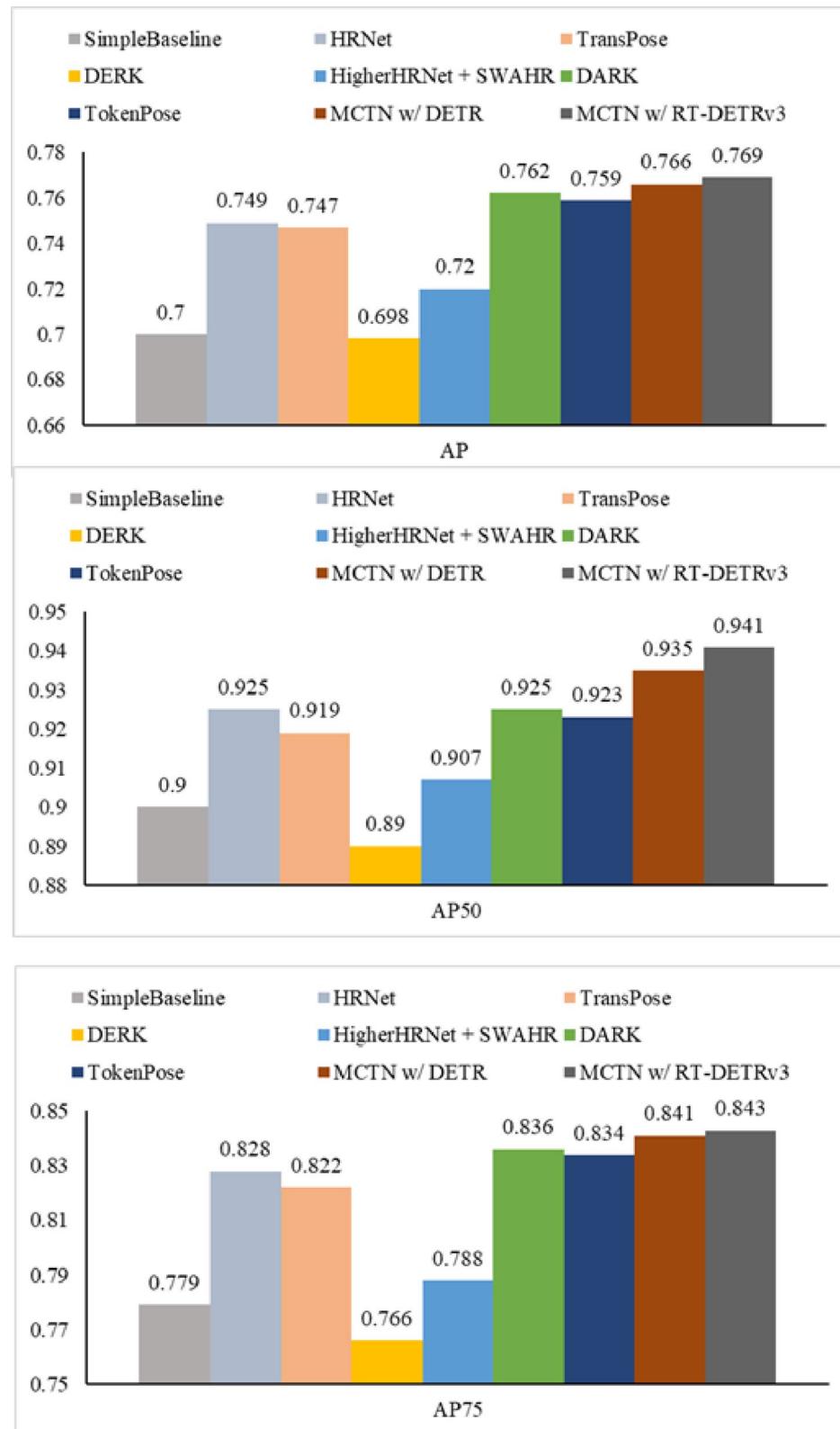


Fig. 8. Comparison with other methods on the MSCOCO test set.

achieves good performance results. The experiments on the test set with the input size of 384×288 are shown in Fig. 8. MCTN w/ DETR achieves 0.766 AP score, 0.935 AP⁵⁰ score and 0.841 AP⁷⁵ score. While MCTN w/ RT-DETRv3 achieves 0.769, 0.941, and 0.843 scores on these indicators, respectively.

To more comprehensively reflect the stability and generalization ability of MCTN, we not only conduct extended tests on the MSCOCO Dataset, but also carefully analyze the performance of the model under different configurations. The experimental results on the validation set and the test set clearly show that the MCTN model shows robust performance on this dataset. These results not only verify the advantages of MCTN in dealing with complex scenes and variable target poses, but also highlight its leading position in human keypoint detection tasks. Furthermore, we investigate the different backbone and different input sizes on the performance of MCTN. Experiments present that MCTN can maintain the excellent performance regardless of the backbone or input size, which further proves its strong adaptability and stability.

Discussion

The empirical evaluation of the MCTN reveals its robust performance in complex human behavior recognition tasks, particularly in scenarios characterized by fast motion, irregular postures, and degraded visual quality. Across both the domain-specific Pickleball dataset and the general-purpose MS COCO dataset, MCTN consistently demonstrates superior accuracy and robustness, validating the effectiveness of its architectural innovations. In high-speed motion contexts, conventional models often fail to capture subtle spatial transformations and lose critical features due to motion blur and occlusion. MCTN overcomes these limitations by introducing a hybrid architecture that integrates three functionally complementary modules—each designed to address a specific visual degradation or recognition bottleneck—thereby improving both the feature discriminability and the model's generalization ability.

At the module level, the MBR component plays a foundational role by recovering high-frequency texture details that are typically suppressed in blurred sequences. Unlike traditional deblurring approaches, which often rely on handcrafted priors or simplistic convolutional filters, MBR leverages the directional sensitivity and multi-resolution capacity of the curvelet transform to reconstruct motion-degraded inputs with minimal information loss. This facilitates more stable pose estimation and feature encoding in downstream tasks. The MSA module further enhances the model's ability to adapt to human body variations by allocating attention weights across hierarchical spatial scales. This mechanism not only strengthens the model's sensitivity to small-scale joint displacements and limb articulations, but also improves its robustness under pose deformation and partial occlusion. Finally, the PCT module introduces frequency-domain semantics into the Transformer architecture, allowing the network to capture structural motion features across orientations and scales. The experimental ablation studies confirm that the inclusion of these three modules leads to measurable performance gains, with each contributing uniquely to the final accuracy, especially in low-quality or cluttered input conditions.

Despite the strong empirical results, several limitations remain that open avenues for future research. First, the incorporation of curvelet-based processing increases the computational overhead of both training and inference stages, which may hinder real-time deployment, especially in resource-constrained environments. Future work could explore model compression techniques or fast approximation algorithms for curvelet transforms to address this challenge. Second, the current model requires high-quality labeled data for optimal performance, yet annotated datasets for fine-grained motion analysis—particularly in niche sports like pickleball—are often scarce and expensive to produce. To alleviate this limitation, future studies may consider semi-supervised or self-supervised learning frameworks that leverage unlabeled data through contrastive or generative mechanisms. Additionally, the static-frame-based architecture of MCTN could be extended to incorporate temporal dynamics through modules such as temporal attention, graph-based spatio-temporal modeling, or recurrent units, thereby enhancing its ability to model continuous motion and behavior evolution in video sequences. These enhancements would further expand MCTN's applicability in domains such as rehabilitation monitoring, intelligent coaching systems, and real-time interactive environments.

Conclusion

The MCTN proposed in this study addresses several critical challenges in human behavior recognition from video data, including motion blur, scale variation, and structural complexity of human posture. By incorporating curvelet-based restoration and multi-scale representation into a Transformer framework, the model effectively captures both low-level visual cues and high-level semantic features. The experimental results affirm the advantages of this integrated design, especially in scenarios involving rapid motion and occlusion. Beyond its performance gains, the modular architecture of MCTN offers flexibility for integration with other temporal or multimodal systems. This research not only contributes to the advancement of multi-scale representation learning in vision tasks but also provides a scalable foundation for future development in behavior understanding systems across various domains.

Data availability

All data and source codes supporting the findings of this study are available in the supplementary files. (<https://doi.org/10.3390/sports12090234>, https://doi.org/10.1007/978-3-319-10602-1_48).

Received: 21 April 2025; Accepted: 23 December 2025

Published online: 31 December 2025

References

- Du, Q., Tang, L. & Li, Y. SCB-YOLOv5: a lightweight intelligent detection model for athletes' normative movements. *Sci. Rep.* **14** (1), 8624 (2024).
- Born, Z. et al. The eye in the sky—A method to obtain on-field locations of Australian rules football athletes. *AI* **5** (2), 733–745 (2024).
- Muksimova, S. et al. GazeCapsNet: A lightweight gaze estimation framework. *Sensors* **25** (4), 1224 (2025).
- Song, Z. & Chen, Z. Sports action detection and counting algorithm based on pose Estimation and its application in physical education teaching. *Informatica* **48**(10). (2024).
- Liu, Y. et al. Identifying competition characteristics of athletes through video analysis. *Int. J. Adv. Comput. Sci. Appl.* **15**(5). (2024).
- Cui, Z. 3D-CNN-based action recognition algorithm for basketball players. *Informatica* **48**(13). (2024).
- Kausalya, K. & Kanaga Suba Raja, S. OTRN-DCN: an optimized transformer-based residual network with deep convolutional network for action recognition and multi-object tracking of adaptive segmentation using soccer sports video. *Int. J. Wavelets Multiresolut. Inf. Process.* **22** (01), 2350034 (2024).
- Yang, M. et al. Adapting short-term transformers for action detection in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18570–18579 (2024).
- Xu, J. et al. FineSports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21773–21782 (2024).
- Li, X., Luo, R. & Islam, F. U. Tracking and detection of basketball movements using multi-feature data fusion and hybrid YOLO-T2LSTM network. *Soft. Comput.* **28** (2), 1653–1667 (2024).
- Abbas, Y. & Jalal, A. Drone-based human action recognition for surveillance: a multi-feature approach. In *2024 International Conference on Engineering & Computing Technologies (ICECT)*, 1–6 (IEEE, 2024).
- Wang, J. & Zuo, L. Basketball technique action recognition using 3D convolutional neural networks. *Sci. Rep.* **14** (1), 13156 (2024).
- Zhang, D. et al. Poseflow: A deep motion representation for understanding human behaviors in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6762–6770 (2018).
- Fani, M. et al. Hockey action recognition via integrated stacked hourglass network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 29–37 (2017).
- Wu, J. et al. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 202, 12352–12361.
- Lei, F., Yan, J. & Wang, X. Human pose estimation of diver based on improved stacked hourglass model. In *Proceedings of the 3rd International Conference on Video and Image Processing*, 178–182 (2019).
- Nam, Y. et al. A hybrid framework to predict ski jumping forces by combining data-driven pose Estimation and model-based force calculation. *Eur. J. Sport Sci.* **23** (2), 221–230 (2023).
- Cheng, B. et al. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5386–5395 (2020).
- Mabrouk, A. B. & Zagrouba, E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Syst. Appl.* **91**, 480–491 (2018).
- Wu, E. & Koike, H. Futurepong real-time table tennis trajectory forecasting using pose prediction network. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8 (2020).
- Huang, L. & Liu, G. Functional motion detection based on artificial intelligence. *J. Supercomput.* **78** (3), 4290–4329 (2022).
- Cui, Y. et al. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13608–13618 (2022).
- Cui, Y. et al. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9921–9931 (2023).
- Cui, Y. et al. Mixformerv2: efficient fully transformer tracking. *Adv. Neural. Inf. Process. Syst.* **36**. (2024).
- Han, W. et al. *Asymmetric Convolution: an Efficient and Generalized Method To Fuse Feature Maps in Multiple Vision Tasks* (IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024).
- Wang, Y. et al. MS-HRNet: multi-scale high-resolution network for human pose estimation. *J. Supercomput.* 1–23 (2024).
- Chen, Y. et al. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7103–7112 (2018).
- Sun, K. et al. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703 (2019).
- Xu, X., Zou, Q. & Li, J. Rethinking the sparse end-to-end multiperson pose estimation. *IEEE Trans. Syst. Man Cybern. Syst.* (2025).
- Ji, A., Fan, H. & Xue, X. Vision-based body pose estimation of excavator using a transformer-based deep-learning model. *J. Comput. Civil Eng.* **39** (2), 04024064 (2025).
- Heravi, M. Y. et al. Deep learning-based activity-aware 3D human motion trajectory prediction in construction. *Expert Syst. Appl.* **239**, 122423 (2024).
- Zhou, X. et al. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, 398–407 (2017).
- Wu, G. et al. *Video Motion Blur Attack Via Grad-Weighted and Discrete-Fusion Based Perturbation Generation* (IEEE Transactions on Multimedia, 2025).
- Singh, H., Verma, M. & Cheruku, R. Novel dilated separable convolution networks for efficient video salient object detection in the wild. *IEEE Trans. Instrum. Meas.* **72**, 1–13 (2023).
- Jiang, Z. et al. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6142–6152 (2024).
- Sun, X., Wang, Y. & Khan, J. Hybrid LSTM and GAN model for action recognition and prediction of lawn tennis sport activities. *Soft. Comput.* **27** (23), 18093–18112 (2023).
- Singh, H., Verma, M., Cheruku, R. & DAMFNet Breaking computational barriers in video salient object detection with multi-scale deformable appearance-motion fusion. *Arab. J. Sci. Eng.* 1–17 (2025).
- Lin, T. Y. et al. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V* 13, 740–755 (Springer International Publishing, 2014).
- Carion, N. et al. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229 (Springer International Publishing, 2020).
- Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022 (2021).
- Xiao, B., Wu, H. & Wei, Y. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 466–481 (2018).
- Geng, Z. et al. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14676–14686 (2021).
- Luo, Z. et al. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13264–13273 (2021).
- Zakir, A. et al. AECA-PRNetCC: Adaptive efficient channel attention-based PoseResNet for coordinate classification in 2D Human Pose. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1–6 (IEEE, 2023).

45. Zakir, A. et al. EBA-PRNetCC: An efficient bridge attention-integration PoseResNet for coordinate classification in 2D human pose estimation. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (SCITEPRESS-Science and Technology Publications, 2024).
46. Li, Y. et al. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11313–11322 (2021).
47. Wang, J. et al. Riformer: Keep your vision backbone effective but removing token mixer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14443–14452 (2023).
48. Yang, S. et al. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11802–11812 (2021).
49. Li, K. et al. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1944–1953 (2021).
50. Gu, K. et al. Bias-compensated integral regression for human pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (9), 10687–10702 (2023).
51. Li, Y. et al. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, 89–106 (Springer Nature Switzerland, 2022).
52. Wang, S. et al. RT-DETRv3: Real-time End-to-End object detection with hierarchical dense positive Supervision. arXiv preprint arXiv:2409.08475 (2024).

Author contributions

Y.R. led the model implementation and conducted the main experiments. Z.G. proposed the original concept, supervised the overall project, and was responsible for manuscript revision and correspondence. W.Z. contributed to the design and evaluation of the motion blur restoration module and assisted in algorithm optimization. Y.S. handled the dataset preprocessing and contributed to the experimental analysis and visualization. Y.X. participated in the literature review, contributed to drafting the introduction, and supported the interpretation of results. All authors discussed the results, revised the manuscript critically for important intellectual content, and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-33985-6>.

Correspondence and requests for materials should be addressed to Z.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026