Journal of Intelligent Computing and Networking
https://www.ffspub.com/index.php/jicn/index
ISSN 3079-9228 (print)
ISSN 3107-2798 (online)
E-mail: jicn.office@ffspub.com

Article

Body Keypoint Detection Algorithm Based on Channel Attention Mechanism

Shaojun Yu[1], Wenhao Huo[1,†], Yuping Lu[1], Hanqing Zhao[2], Yilin Wang[2], Lili Wang[2], Muhammad Rizwan Anjum[3]
[1] *School of Computer Science & Technology, Beijing Jiaotong University, Beijing 100044, China*
[2] *School of Physical Education, Beijing Jiaotong University, Beijing 100044, China*
[3] *Department of Electronic Engineering, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan*
[†] *E-mail: 24115053@bjtu.edu.cn*

Received: April 8, 2026 / Revised: May 2, 2026 / Accepted: May 7, 2026 / Published online: May 25, 2026

Abstract: With the implementation of national strategies aimed at building a leading sporting nation and promoting nationwide fitness, physical fitness assessment has gained increasing attention as a crucial metric for evaluating students' physical condition and motor abilities. Concurrently, advancements in computer vision have enabled body keypoint detection technology to gradually replace traditional manual measurement methods, demonstrating significant potential for application in automated assessment systems. Accurate recognition of keypoints serves as the fundamental support for intelligent physical fitness testing and smart sports. However, existing keypoint detection algorithms often suffer from drifting of extremity keypoints, such as those of the hands and feet keypoints, in physical fitness test scenarios, thereby compromising the accuracy of the assessment. To address this challenge, this paper proposes Channel Attention BlazePose(CA-BlazePose), a body keypoint detection algorithm based on a channel attention mechanism, specifically designed for count-based physical fitness test scenarios, namely sit-ups and pull-ups. To tackle the issue of keypoint drift in motion detection, CA-BlazePose aims to enhance keypoint detection accuracy. It employs a two-stage network architecture consisting of heatmap training and regression fine-tuning, incorporating a channel attention module. This module strengthens the feature extraction process for extremity keypoints such as hands and feet, thereby improving recognition accuracy during detection.Experimental results demonstrate that, compared to mainstream keypoint detection algorithms such as OpenPose and BlazePose, the proposed CA-BlazePose algorithm achieves improvements in the PCK on two representative motion datasets, Common Objects in Context(COCO) and Leeds Sports Pose Extended(LSPET). Specifically, it shows an approximate increase of 7% for hand and foot keypoints and 8% for overall keypoints. Furthermore, in real-time detection tests for sit-ups and pull-ups captured from various viewing angles, CA-BlazePose demonstrates superior performance in handling frames with missing or drifting keypoints compared to existing algorithms, exhibiting more stable recognition performance under identical detection conditions.

Keywords: Body Keypoint Detection; Channel Attention Mechanism; BlazePose; Motion Analysis; Physical Fitness Assessment
https://doi.org/10.64509/jicn.21.98

1 Introduction

Physical fitness assessment serves as a critical indicator for evaluating students' physical condition and motor abilities, and has long garnered extensive attention from both the state and society. Traditional physical fitness testing often relies on on-site supervision by instructors and manual recording, which is not only labor-intensive but also susceptible to subjective bias. Furthermore, constrained by testing resources, the overall efficiency of such assessments remains low. In recent years, with the rapid advancement of computer vision technologies, vision-based motion detection methods have gradually been applied to fitness testing and training scenarios, offering a feasible pathway toward automated fitness assessment [1–3]. As a key component, human keypoint detection technology enables keypoint localization and motion analysis in count-based exercises such as sit-ups and pull-ups [4–6], thereby providing more objective data support for counting and evaluation, as well as process-oriented references for subsequent training improvements.

[†] Corresponding authors: Wenhao Huo
[*] Academic Editor: Chunxiao Jiang

The development of human keypoint detection has evolved from traditional model-driven methods to data-driven deep learning approaches [7]. Early studies primarily relied on geometric models and template matching, combined with handcrafted features such as edges and textures for pose estimation [8, 9]; however, these methods were sensitive to pose variations and occlusions, and exhibited limited generalization capability.

With the advancement of machine learning, research gradually shifted toward feature learning-based frameworks. For instance, Shotton *et al.* [10] proposed learning local pose parts using decision trees, which improved recognition accuracy to some extent but relied heavily on manual annotations and suffered from limited computational efficiency. In recent years, breakthroughs in deep learning have propelled human keypoint detection into a stage emphasizing both end-to-end learning and high-precision localization. Newell *et al.* [11] introduced the Stacked Hourglass Networks, which captures multi-scale information through symmetric encoder-decoder structures, significantly improving localization accuracy at the cost of large model parameters and low inference efficiency. Gines Hidalgo *et al.* [12] proposed OpenPose, which detects multi-person 2D poses directly in images via part affinity fields, enabling real-time multi-person pose estimation with favorable efficiency and scalability. However, its bottom-up grouping strategy is prone to keypoint association errors in scenarios with severe occlusions or overlapping. To address the quantization error and regression difficulty in heatmap-based approaches, Sun *et al.* [13] proposed integral pose regression, which obtains continuous coordinates by integrating heatmaps, thereby further enhancing localization accuracy. Subsequently, Fang *et al.* [14] introduced AlphaPose, a regional multi-person pose estimation framework that reduces errors caused by detection box inaccuracies through modules such as symmetric spatial transformer networks and pose-guided proposal generators, improving robustness and accuracy in multi-person pose estimation. Nevertheless, instability in recognition may still occur in scenarios involving rapid motion or drastic pose changes.

Although the aforementioned pose estimation methods have achieved promising results on general-purpose datasets and conventional pose scenarios, their direct application to physical fitness testing scenarios—particularly in count-based rapid motion exercises such as sit-ups and pull-ups—still faces significant challenges, as summarized in Table 1. Current mainstream human keypoint detection algorithms exhibit notable differences in model scale and the number of detectable keypoints. In terms of model scale, algorithms such as High-Resolution Network (HRNet), SimpleBaseline, and OpenPose involve large parameter counts or high computational overhead. Although they perform well in general keypoint detection tasks, they impose considerable inference burdens in real-time fitness testing scenarios, making it difficult to meet the demands of high frame rates and low latency required for practical deployment.

Meanwhile, regarding the number of detectable keypoints, most general-purpose keypoint detection algorithms still adhere to the keypoint definitions of datasets such as COCO and MPII Human Pose Dataset, with the maximum number of detectable keypoints typically limited to 16 or 17. Examples include HRNet, Cascaded Pyramid Network (CPN), PoseNet, SimpleBaseline, and Stacked Hourglass Networks. While such methods can accomplish conventional human keypoint detection tasks, they provide insufficient support for fine-grained body part information required for tasks such as movement standardization assessment and violation detection in fitness testing scenarios, thus failing to meet the demands for higher-precision evaluation.

Notably, BlazePose demonstrates strong application potential in Table 1. With approximately 5 million parameters, this method supports 33 keypoint detection, outperforming most general-purpose keypoint detection models in terms of both lightweight design and keypoint coverage. However, fitness testing scenarios typically employ monocular cameras for data acquisition, and exercises such as sit-ups and pull-ups often involve rapid movements, changes in viewing angles, and variations in body scale. These factors can easily lead to localization drift or temporary loss of extremity keypoints, such as those of the hands and feet, thereby reducing the accuracy and stability of counting results and movement determination. Therefore, how to balance keypoint coverage while ensuring real-time performance, and further improve the localization accuracy of extremity keypoints and robustness across varying viewing angles, remains a critical issue to be addressed in pose estimation tasks for physical fitness assessment.

Table 1: Comparison of keypoint detection algorithms

Model	Parameters	Max Keypoints
HRNet[15]	∼100M	17
CPN[16]	∼30M	17
PoseNet[17]	∼10M	17
SimpleBaseline[18]	∼50M	17
Stacked Hourglass Networks[11]	∼10M	16
AlphaPose[14]	∼30M	17
OpenPose[12]	∼50M	25
BlazePose[19]	∼5M	33

To address the aforementioned issues, this paper proposes a keypoint detection method based on a channel attention mechanism. By introducing a channel attention mechanism, the proposed method adaptively recalibrates features across different channels, thereby enhancing the feature representation capability for extremity keypoints such as those of the hands and feet. This effectively alleviates the problems of localization drift and temporary loss of extremity keypoints, while improving overall keypoint detection accuracy. On this basis, the proposed method is further applied to motion recognition and assessment tasks in physical fitness testing scenarios, validating its effectiveness and robustness in practical applications. The main contributions of this paper are as follows:

(1) To address the issues of extremity keypoint localization drift and temporary loss in physical fitness testing scenarios, a keypoint detection algorithm based on a channel attention mechanism is proposed. This algorithm improves

the recognition accuracy of extremity keypoints, such as those of the hands and feet, while maintaining overall recognition accuracy.

(2) To validate the performance of the proposed method, experiments are conducted on publicly available motion datasets, including Leeds Sports Pose-Leeds Sports Pose Extended (LSP-LSPET) and COCO , as well as in real-world testing scenarios. Experimental results demonstrate that the proposed method not only improves the recognition accuracy of extremity keypoints but also exhibits better recognition robustness under varying viewing angles.

2 Proposed Method

During sit-up and pull-up detection, extremity keypoints such as the hands and feet occupy a small proportion in the image, resulting in weak responses in their corresponding feature channels. In addition, these keypoints are easily overwhelmed by strong responses from background or torso channels in the heatmap stage, leading to low recognition accuracy. While simply deepening the network or increasing the kernel size can enhance feature extraction, it significantly increases computational cost and is not suitable for real-time detection. Spatial attention mechanisms can focus on keypoint locations but cannot directly strengthen specific semantic channels. In contrast, the channel attention mechanism can adaptively learn the importance of different channels, suppress irrelevant channels, and highlight channel responses sensitive to extremity keypoints, making it well-suited to mitigate the local feature drowning problem. Therefore, to address the issue of low recognition accuracy for extremity keypoints such as the hands and feet during sit-up and pull-up detection, this paper proposes CA-BlazePose, a keypoint detection method based on a channel attention mechanism. It builds channel attention modules within a two-stage network architecture to simultaneously improve both coarse heatmap localization and fine-grained regression adjustment. The keypoint detection network is primarily divided into two stages: heatmap training and regression fine-tuning.

In the heatmap training stage, the network employs an encoder-decoder structure to generate corresponding heatmaps and offsets for 33 body keypoints, leveraging the heatmaps to effectively supervise feature embedding. Meanwhile, the network embeds channel attention modules in multiple sampling branches of the heatmap branch and adds an additional channel attention module before heatmap output. Global average pooling is used to extract the global response of each channel, followed by a two-layer fully connected network to model nonlinear relationships across channels, generating corresponding channel weights. These weights are used to adaptively recalibrate the original feature maps along the channel dimension, highlighting channels with strong keypoint responses while suppressing background and noise channels. Before outputting heatmap features, the high-resolution features are reweighted to improve recognition accuracy for extremity keypoints such as those of the hands and feet.This mechanism effectively improves the coarse localization accuracy of extremity keypoints in the heatmap branch.

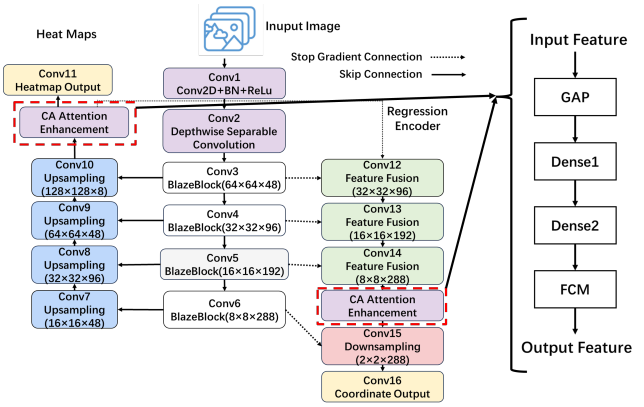
In the regression fine-tuning stage, the features learned in the heatmap stage are fed into a regression encoder to predict precise keypoint coordinates, while also outputting keypoint visibility confidence scores to determine whether a keypoint is occluded. Since feature propagation from the heatmap to the regression stage may incur losses, which can still lead to coordinate drift, a channel attention module is also introduced in the regression branch to reweight high-level features used for coordinate prediction, thereby enhancing the fine-grained coordinate adjustment capability for extremity keypoints.

In the attention enhancement module, input features first undergo global average pooling (GAP) to compress spatial information from each channel into a single value. The features are then passed through a Dense1 fully connected layer with a Rectified Linear Unit (ReLU) activation function to reduce dimensionality and capture correlations across channels. Subsequently, a Dense2 fully connected layer with a Sigmoid activation function restores the original channel dimension and generates attention weights. Finally, feature recalibration is performed in the Fuzzy C-Means (FCM) clustering module, where the weights are multiplied channel-wise with the original input feature maps to enhance responses from critical channels and suppress redundant information. The architecture of the CA-BlazePose keypoint detection network is shown in Figure 1(a).

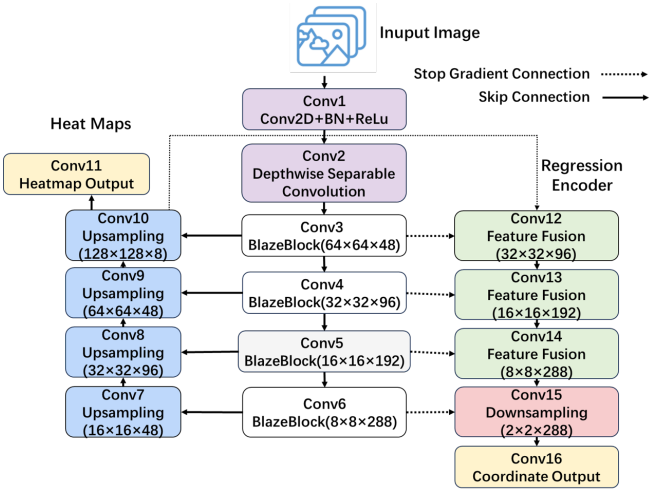
Regarding the loss function, the heatmap branch employs binary cross-entropy or mean squared error for pixel-wise supervision of Gaussian keypoint heatmaps, while the regression branch constructs a smooth L1 loss based on the Euclidean distance of joints. Evaluation is conducted using the Percentage of Correct Keypoints (PCK) accuracy metric. This approach ensures convergence during network training while improving recognition accuracy for extremity keypoints such as those of the hands and feet.

In comparison, the BlazePose keypoint detection method also adopts a hybrid architecture combining heatmap and regression components [19, 20], as illustrated in Figure 1(b). However, in real-time detection tasks for typical count-based physical fitness test exercises such as sit-ups and pull-ups, BlazePose exhibits severe recognition drift for extremity keypoints such as those of the hands and feet, leading to considerable deviations in movement determination. In our analysis, the response of extremity keypoints in the heatmap stage is weak and easily overwhelmed by background noise. If a channel attention module is introduced only in the heatmap branch, although the heatmap quality can be improved, the regression branch may still suffer from coordinate drift due to feature propagation loss. If a channel attention module is introduced only in the regression branch, the localization bias in the heatmap stage inherently limits the upper bound of regression accuracy. Therefore, channel attention must be introduced in both stages simultaneously: the attention in the heatmap branch enhances the coarse localization response of extremity keypoints, and the attention in the regression branch further refines the coordinates based on that. Both are indispensable. To address these issues, CA-BlazePose introduces channel attention modules in both the heatmap training and regression fine-tuning stages. By weighting the features used for keypoint prediction, the proposed method improves the

recognition accuracy of keypoint coordinates. While maintaining overall keypoint detection performance, it enhances recognition accuracy for extremity keypoints such as those of the hands and feet, thereby improving the accuracy of movement assessment.



(a) CA-BlazePose keypoint detection network architecture



(b) BlazePose keypoint detection network architecture

Figure 1: Network architectures of BlazePose and CA-BlazePose for keypoint detection

Specifically, the mathematical formulation of the channel attention mechanism is as follows.

Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, global average pooling is first applied to extract the global response of each channel:

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c}, \quad c = 1,\ldots,C \tag{1}$$

where H, W, and C denote the height, width, and number of channels of the feature map, respectively, and $z \in \mathbb{R}^{1\times 1\times C}$ represents the channel descriptor vector. Subsequently, a two-layer fully connected network models the nonlinear relationships between channels:

$$s = \sigma(W_2\delta(W_1 z)) \tag{2}$$

where $W_1 \in \mathbb{R}^{C/r\times C}$ is the dimensionality reduction matrix, r is the reduction ratio, $\delta(\cdot)$ denotes the ReLU activation function, $W_2 \in \mathbb{R}^{C\times C/r}$ is the dimensionality increasing matrix, and $\sigma(\cdot)$ denotes the Sigmoid activation function. The resulting channel weights $s \in \mathbb{R}^{1\times 1\times C}$ are then used for channel-wise recalibration of the original feature map:

$$\tilde{X}_{i,j,c} = s_c \cdot X_{i,j,c} \tag{3}$$

where $\tilde{X}_{i,j,c}$ represents the value at spatial position (i,j) of the c-th channel after recalibration.

In the heatmap branch, the Mean Squared Error (MSE) loss is employed to fit Gaussian heatmaps:

$$\mathscr{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4}$$

In the regression branch, to improve keypoint localization accuracy, a smoothed L1 loss is applied to the coordinate error distance $d_k = \|p_k^{\text{pred}} - p_k^{\text{gt}}\|_2$ for each keypoint k:

$$\mathscr{L}(d_k) = \begin{cases} \frac{1}{2}d_k^2, & |d_k| < \delta \\ \delta|d_k| - \frac{1}{2}\delta^2, & |d_k| \geq \delta \end{cases} \tag{5}$$

where p_k^{pred} denotes the predicted coordinate vector of the k-th keypoint, p_k^{gt} denotes the corresponding ground truth coordinate vector, and δ is the threshold parameter of the L1 loss, set to $\delta = 5.0$.

Finally, the model evaluates keypoint detection accuracy using the PCK (Percentage of Correct Keypoints) metric:

$$\text{PCK}_i^k = \frac{\sum_p \mathbb{I}\left(\frac{d_k}{d^{\text{def}}} \leq T_k\right)}{\sum_p 1} \tag{6}$$

where d_k represents the Euclidean distance between the predicted and ground truth values for the k-th keypoint, d^{def} is the normalization factor, and T_k denotes the preset threshold.

The operational framework of the CA-BlazePose keypoint detection method primarily consists of a face detection module, an image processing module, and a keypoint detection module. The specific framework diagram is shown in Figure 2. In real-time detection tasks for count-based exercises, when detection begins, the model obtains the region of interest (ROI) parameters of the human body through a face detector. After image processing, the data are fed into the keypoint detection network to obtain keypoint coordinates, and the ROI parameters are updated accordingly. The obtained keypoint coordinates are then input into the motion assessment module to generate the output image. In subsequent recognition frames, it is not necessary to rerun the face detector. Instead, recognition for the current frame is performed using the keypoint coordinates and ROI parameters from the previous frame, combined with the input image, thereby improving operational efficiency. The pseudocode corresponding to this operational framework is presented in Algorithm 1, and the parameter definitions are provided in Table 2.

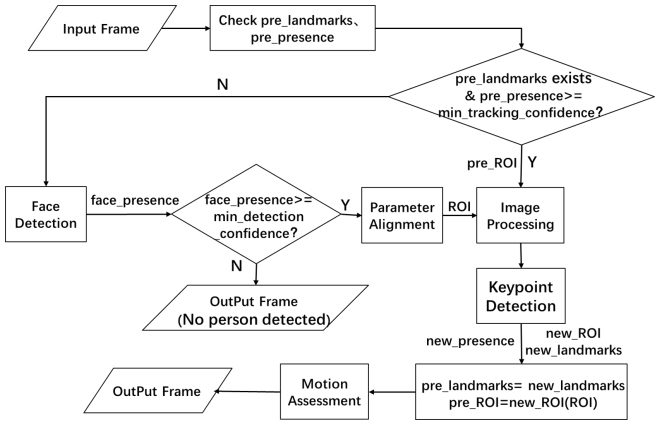


Figure 2: CA-BlazePose keypoint detection framework

Algorithm 1 CA-BlazePose keypoint detection method

Require: Input image frame_0
Ensure: Output image frame_2
// Check if it is the first frame
if $pre\_landmarks = \textbf{None}$ **or** $pre\_presence < min\_tracking\_confidence$ **then**
// Face detector & parameter alignment
$presence', ROI \leftarrow face\_detection(frame_0)$
if $presence' < min\_detection\_confidence$ **then**
output: No person detected in the image
else if $presence' \geq min\_detection\_confidence$ **then**
// Image processing
$frame_1 \leftarrow frame\_processing(frame_0, ROI)$
end if
//Keypoint detection
$new\_landmarks, new\_ROI, new\_presence \leftarrow landmarks\_detection(frame_1)$
$pre\_landmarks, pre\_ROI, pre\_presence \leftarrow new\_landmarks, new\_ROI, new\_presence$
else if $pre\_landmarks \neq \textbf{None}$ **and** $pre\_presence \geq min\_tracking\_confidence$ **then**
$frame_1 \leftarrow frame\_processing(frame_0, pre\_ROI)$
$new\_landmarks, new\_ROI, new\_presence \leftarrow landmarks\_detection(frame_1)$
$pre\_landmarks, pre\_ROI, pre\_presence \leftarrow new\_landmarks, new\_ROI, new\_presence$
end if
//Assess and output image
$frame_2 \leftarrow evaluation\_function(new\_landmarks)$

3 Experiments

To comprehensively evaluate the effectiveness and applicability of the proposed method in human keypoint detection and real-world physical fitness testing, we present extensive experimental results covering six aspects: Section 3.1 experimental setup, Section 3.2 datasets and evaluation metrics, Section 3.3 comparison with different baseline methods , Section 3.4 ablation experiment, Section 3.5 performance analysis of extremity keypoints under different viewpoints, Section 3.6 computational complexity analysis .

Table 2: Parameter definitions for CA-BlazePose keypoint detection framework pseudocode.

Parameter	Definition
$pre\_landmarks$	Keypoint queue from the previous frame
$pre\_presence$	Keypoint visibility confidence from the previous frame
pre_ROI	Region of Interest (ROI) of the human body from the previous frame
$min\_tracking\_confidence$	Minimum visibility confidence threshold for keypoints
$presence'$	Face visibility confidence
ROI	Region of Interest (ROI) of the human body
$min\_detection\_confidence$	Minimum visibility confidence threshold for face detection
$frame_1$	Frame after image processing
$new\_landmarks$	Keypoint queue for the current frame
new_ROI	Region of Interest (ROI) of the human body for the current frame
$new\_presence$	Keypoint visibility confidence for the current frame

3.1 Experimental Setup

All experiments are conducted on a server equipped with an NVIDIA Tesla T4 GPU (16 GB memory). The operating system is Ubuntu 20.04, and the software environment includes PyTorch 1.10 and CUDA 11.3. Input images are uniformly resized to 256 × 256 pixels. The model is trained using the Adam optimizer with an initial learning rate of 1e-3, which is decayed using a cosine annealing strategy. The batch size is set to 32, and the total number of training epochs is 200. In both the heatmap and regression branches, the reduction ratio of the channel attention module is set to 16. For fair comparison, all baseline methods are either retrained under the same experimental setup or evaluated using officially provided pre-trained models.

3.2 Datasets and Evaluation Metrics

For standard dataset evaluation, COCO [21] and LSP-LSPET [22] public human keypoint datasets were selected as training and testing datasets. The COCO dataset contains human samples from diverse scenes and poses, enabling comprehensive performance evaluation. The LSP-LSPET datasets primarily consist of athletic poses, better reflecting the model's keypoint localization capability under complex posture conditions. keypoint detection accuracy was evaluated using the Percentage of Correct Keypoints (PCK) metric. Additionally, average error metrics including AED, MSE, and RMSE were introduced to quantitatively analyze the model's keypoint regression error.

In the standard dataset evaluation, we selected the COCO and LSP-LSPET public human keypoint datasets as training

and testing datasets. As summarized in Table 3, the COCO dataset contains 1,200 training images and 800 validation images, each annotated with 17 human keypoints, while the LSP-LSPET dataset contains 1,200 training images and 800 testing images, each annotated with 14 human keypoints. To adapt to the 33 keypoint outputs of BlazePose, we mapped the annotations of COCO and LSP-LSPET to the keypoint indices of BlazePose according to publicly available correspondences, and only used the common keypoints (e.g., shoulders, elbows, wrists, hips, knees, ankles) for evaluation.

Keypoint detection accuracy is evaluated using the Percentage of Correct Keypoints (PCK) metric. For the k-th keypoint, the PCK is calculated as:

$$\mathrm{PCK}_k = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\frac{\|\mathbf{p}_{i,k}^{\mathrm{pred}} - \mathbf{p}_{i,k}^{\mathrm{gt}}\|_2}{d_{\mathrm{def}}} \le T_k\right) \tag{7}$$

where $\mathbf{p}_{i,k}^{\mathrm{pred}}$ and $\mathbf{p}_{i,k}^{\mathrm{gt}}$ denote the predicted and ground-truth coordinates of the k-th keypoint in the i-th sample, respectively; $\|\cdot\|_2$ is the Euclidean distance; d_{def} is the normalization factor, which is taken as the head length (distance from the top of the head to the neck) in COCO and as the torso diameter (distance from the midpoint of the shoulders to the midpoint of the hips) in LSP-LSPET; T_k is a preset threshold, uniformly set to 0.2 in this paper; $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition holds and 0 otherwise; and N is the total number of samples. A higher PCK value indicates more accurate keypoint localization.

Table 3: Dataset summary of COCO and LSP-LSPET

Dataset	Training images	Test images	Keypoints per image
COCO	1,200	800	17
LSP-LSPET	1,200	800	14

In addition, three average error metrics are introduced to quantitatively analyze the keypoint regression error of the models:

Average Euclidean Distance (AED): The average Euclidean distance between predicted and ground-truth keypoints over all samples and keypoints, defined as:

$$\mathrm{AED} = \frac{1}{N\cdot K}\sum_{i=1}^{N}\sum_{k=1}^{K}\|\mathbf{p}_{i,k}^{\mathrm{pred}} - \mathbf{p}_{i,k}^{\mathrm{gt}}\|_2 \tag{8}$$

where N is the total number of samples, K is the total number of keypoints (in this paper, we use the common keypoints, so $K = 14$ or 17), and $\mathbf{p}_{i,k}^{\mathrm{pred}}$ and $\mathbf{p}_{i,k}^{\mathrm{gt}}$ have the same meanings as in the PCK formula.

Mean Squared Error (MSE): The average of the squared Euclidean distances between predicted and ground-truth keypoints over all samples and keypoints, with units of pixel squared (px^2), defined as:

$$\mathrm{MSE} = \frac{1}{N\cdot K}\sum_{i=1}^{N}\sum_{k=1}^{K}\|\mathbf{p}_{i,k}^{\mathrm{pred}} - \mathbf{p}_{i,k}^{\mathrm{gt}}\|_2^2 \tag{9}$$

The symbols have the same meanings as in the AED formula.

Root Mean Squared Error (RMSE): The square root of MSE, with the same units as the original coordinates (pixels), defined as:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{10}$$

Smaller values of AED, MSE, and RMSE indicate smaller keypoint regression error and more accurate localization.

3.3 Comparison with Different Baseline Methods

To validate the effectiveness of the proposed method, we compared the keypoint detection performance of BlazePose, OpenPose, and CA-BlazePose on the COCO and LSP-LSPET datasets, uniformly using PCK as the evaluation metric. The experimental results are shown in Figure 3 and Figure 4, Table 4, and Table 5. In Figure 3, and Figure 4, the horizontal axis represents the number of training epochs, while the vertical axis represents the overall PCK score of the trained model. The tables present the PCK scores for individual keypoints across different models.

On the LSP-LSPET dataset, the overall PCK values for BlazePose, OpenPose, and CA-BlazePose were 82.52%, 79.87%, and 88.23%, respectively. Regarding recognition accuracy for hand and foot keypoints, CA-BlazePose achieved improvements of 6.53% and 6.34%, respectively, compared to BlazePose, and improvements of 7.43% and 6.72%, respectively, compared to OpenPose. On the COCO dataset, the PCK values for BlazePose, OpenPose, and CA-BlazePose were 80.93%, 78.00%, and 86.25%, respectively. In terms of hand and foot keypoint detection accuracy, CA-BlazePose achieved improvements of approximately 7% and 6%, respectively, compared to BlazePose, and improvements of 4% and 8%, respectively, compared to OpenPose.

These results indicate that the proposed CA-BlazePose model improves the overall PCK by 5.71% on the LSP-LSPET dataset and by 5.32% on the COCO dataset compared to BlazePose. Meanwhile, on the COCO dataset, which contains multi-person samples with substantial pose variations, the proposed method increases the recognition accuracy for hand keypoints by approximately 7% compared to BlazePose and 4% compared to OpenPose, and for foot keypoints by approximately 6% compared to BlazePose and 8% compared to OpenPose. These results collectively validate the effectiveness of the proposed method in enhancing performance on complex athletic poses.

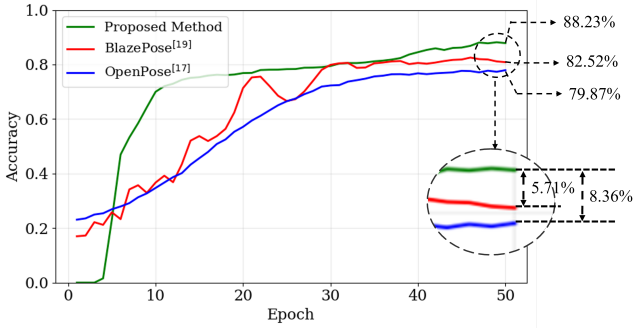


Figure 3: PCK comparison of BlazePose, OpenPose, and CA-BlazePose on the LSP-LSPET dataset.

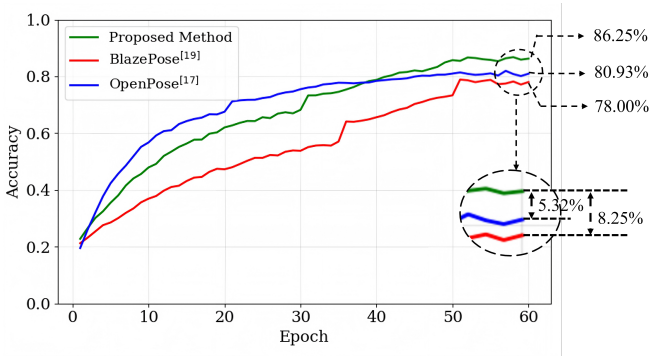


Figure 4: PCK comparison of BlazePose, OpenPose, and CA-BlazePose on the COCO dataset.

Table 4: PCK comparison of BlazePose, OpenPose, and CA-BlazePose on LSP-LSPET dataset for different keypoints.

Body Part	BlazePose [19]	OpenPose [12]	Proposed Method
Right ankle	74.18%	73.80%	**80.52%**
Right knee	86.23%	77.50%	**87.89%**
Right hip	84.67%	75.65%	**85.13%**
Left hip	84.67%	75.65%	**85.13%**
Left knee	86.23%	77.50%	**87.89%**
Left ankle	74.18%	73.80%	**80.52%**
Right wrist	71.45%	70.55%	**77.98%**
Right elbow	80.89%	75.65%	**81.34%**
Right shoulder	85.34%	77.50%	**85.76%**
Left shoulder	85.34%	77.50%	**85.76%**
Left elbow	80.89%	75.65%	**81.34%**
Left wrist	71.45%	70.55%	**77.98%**
Neck	89.12%	**87.72%**	89.67%
Head top	94.23%	92.88%	**94.78%**
Overall	82.52%	79.87%	88.23%

Table 5: PCK comparison of BlazePose, OpenPose, and CA-BlazePose on COCO dataset for different keypoints.

Body Part	BlazePose [19]	OpenPose [12]	Proposed Method
Right ankle	72.62%	68.80%	**78.50%**
Right knee	81.62%	77.50%	**94.17%**
Right hip	93.95%	76.65%	**94.32%**
Left hip	85.62%	78.21%	**94.50%**
Left knee	75.45%	83.50%	**93.33%**
Left ankle	69.62%	72.80%	**78.83%**
Right wrist	70.62%	72.55%	**76.50%**
Right elbow	**84.79%**	77.65%	84.67%
Right shoulder	90.12%	79.93%	**92.50%**
Left shoulder	91.12%	82.20%	**93.33%**
Left elbow	78.95%	**79.26%**	78.83%
Left wrist	67.29%	70.37%	**75.83%**
Neck	83.95%	**87.72%**	85.83%
Head top	87.29%	**88.58%**	87.17%
Overall	80.93%	78.00%	86.25%

To further analyze the keypoint prediction error of the models, we statistically compared the AED, MSE, and RMSE metrics of the three models on a unified test set. The results are shown in Table 6. As can be seen from the table, CA-BlazePose outperforms both BlazePose and OpenPose across all three error metrics, with a particularly notable reduction in the RMSE metric. This indicates that after introducing the channel attention mechanism, the overall error fluctuation during the keypoint regression process is effectively suppressed.

Table 6: Comparison of keypoint regression error metrics for BlazePose, OpenPose, and CA-BlazePose

Model	AED (px)↓	MSE (px^2)↓	RMSE (px)↓
OpenPose[12]	7.42	68.90	8.30
BlazePose[19]	6.37	54.10	7.72
Proposed Method	**5.96**	**45.20**	**6.72**

3.4 Ablation Experiment

To validate the effectiveness of the channel attention module design in the two-stage network architecture of CA-BlazePose, we conducted an ablation experiment on the LSP-LSPET dataset, comparing the keypoint detection accuracy under four configurations.

Table 7: Ablation experiment results of channel attention modules at different stages on the LSP-LSPET dataset

Body Part	None	Heatmap only	Regression only	Proposed Method
Right ankle	74.18%	76.5%	77.2%	**80.52%**
Right knee	86.23%	86.9%	86.6%	**87.89%**
Right hip	84.67%	85.1%	84.9%	**85.13%**
Left hip	84.67%	85.1%	84.9%	**85.13%**
Left knee	86.23%	86.9%	86.6%	**87.89%**
Left ankle	74.18%	76.5%	77.2%	**80.52%**
Right wrist	71.45%	73.8%	74.5%	**77.98%**
Right elbow	80.89%	81.4%	81.1%	**81.34%**
Right shoulder	85.34%	85.9%	85.6%	**85.76%**
Left shoulder	85.34%	85.9%	85.6%	**85.76%**
Left elbow	80.89%	81.4%	81.1%	**81.34%**
Left wrist	71.45%	73.8%	74.5%	**77.98%**
Neck	89.12%	89.5%	89.3%	**89.67%**
Head top	94.23%	94.6%	94.4%	**94.78%**
Overall	82.52%	84.03%	83.68%	**88.23%**

As shown in Table 7, the experimental results show that on the LSP-LSPET dataset, the overall PCK values for the configurations without attention, with attention only in the heatmap stage, only in the regression stage, and with dual-stage attention are 82.52%, 84.03%, 83.68%, and 88.23%, respectively. The hand keypoint PCK values are 71.45%,

73.21%, 72.85%, and 77.98%, respectively, and the foot keypoint PCK values are 74.18%, 76.09%, 75.64%, and 80.52%, respectively. Compared with the configuration without attention, the dual-stage attention module improves the overall PCK by 5.71%, hand PCK by 6.53%, and foot PCK by 6.34%. Compared with the heatmap-only attention, the improvements are 4.20%, 4.77%, and 4.43%, respectively. Compared with the regression-only attention, the improvements are 4.55%, 5.13%, and 4.88%, respectively. Thus, the dual-stage attention module significantly outperforms both the no-attention configuration and the single-stage attention configurations in terms of overall accuracy and extremity keypoint (hands and feet) recognition rates, validating the effectiveness of the proposed dual-stage channel attention design.

3.5 Performance Analysis of Extremity Keypoints under Different Viewpoints

To quantitatively evaluate the performance of the proposed method in real-world scenarios, we assess its hand and foot keypoint detection success rates under different camera views, including oblique front, high central, and low central views for sit-ups, and the front view for pull-ups. Specifically, we randomly sample 1,000 frames from sit-up detection videos captured from three camera views and 1,000 frames from a front-view pull-up video.

To evaluate the detection performance, the visibility and pixel coordinates of hand and foot keypoints are manually annotated as ground truth. A keypoint is considered successfully detected if its visibility confidence exceeds 0.5. We then compute the detection success rates for hand and foot keypoints for BlazePose and CA-BlazePose. Hand and foot keypoints are selected because these extremity keypoints are prone to missing or drifting in detection scenarios, which better reflects the recognition accuracy of the model in practical scenarios.

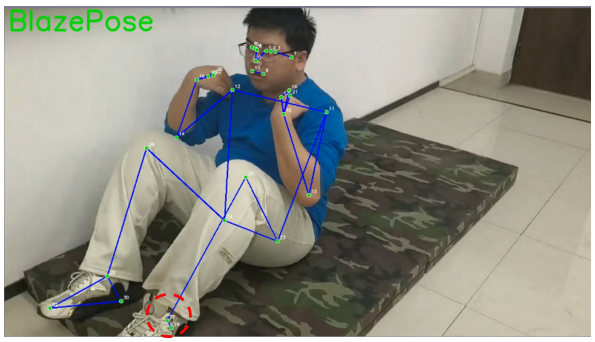
The experimental results are shown in Table 8. Under the oblique frontal view, both the proposed method and BlazePose achieve high detection success rates for hands and feet. Specifically, the hand detection success rates are 97.8% and 96.5%, and the foot detection success rates are 97.1% and 95.2%, respectively. However, under challenging views such as the high central, low central, and front pull-up views, due to significant occlusion or fast motion of the extremity keypoints, BlazePose exhibits a notable drop in detection success rates for both hands and feet: the hand rate drops to as low as 72.4%, and the foot rate drops to 75.7%. In contrast, under the same challenging views, the proposed method maintains hand detection success rates of at least 91.2% and foot detection success rates of at least 89.9%, achieving average improvements of approximately 20.2 percentage points for hands and 17.5 percentage points for feet over BlazePose. These results validate that the proposed method can adapt to different camera views in real-world deployment, reduce deployment difficulty, and offer greater practical application potential.

To more intuitively present the performance of the proposed method and the baseline methods in real-world testing scenarios, we further visualize the keypoint detection results, as shown in Figures 5–8.

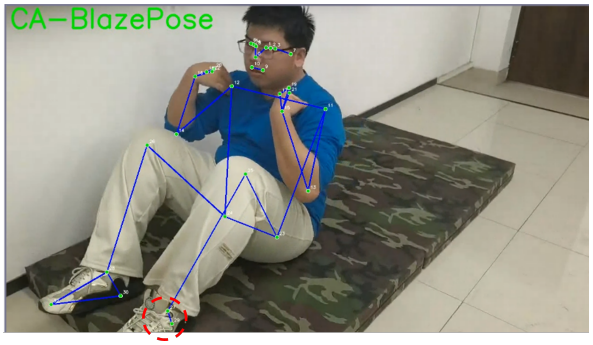
Table 8: Comparison of keypoint detection success rates of BlazePose and CA-BlazePose in physical fitness testing scenarios.

Viewpoint	Model	Hand success rate (%)	Foot success rate (%)
Oblique front (sit-up)	BlazePose	96.5	95.2
Oblique front (sit-up)	Proposed Method	**97.8**	**97.1**
Middle high (sit-up)	BlazePose	73.3	75.7
Middle high (sit-up)	Proposed Method	**93.6**	**91.8**
Middle low (sit-up)	BlazePose	78.2	76.5
Middle low (sit-up)	Proposed Method	**91.5**	**89.9**
Front (pull-up)	BlazePose	72.4	85.8
Front (pull-up)	Proposed Method	**91.2**	**94.5**

Figure 5a and Figure 5b visualize the keypoint detection results of the two models under the oblique frontal view of sit-ups. Since the detection success rates of both methods exceed 95% under this view, only subtle differences can be observed in the visualization results, as indicated by the red circles in the figures.



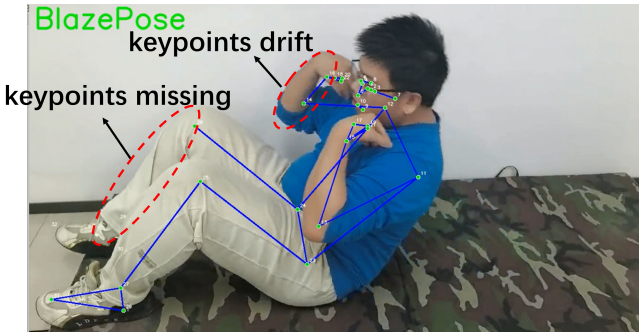
(a) BlazePose on sit-ups from oblique frontal view.



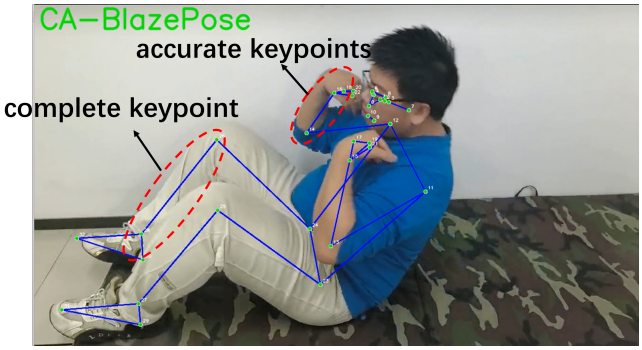
(b) CA-BlazePose on sit-ups from oblique frontal view.

Figure 5: Keypoint detection results on sit-ups from oblique frontal view.

Figure 6a and Figure 6b visualize the keypoint detection results of the two models under the high central position of sit-ups. Under this view, the detection success rates of BlazePose for hands and feet are 73.3% and 75.7%, respectively, with obvious drift or missing keypoints on the hands and feet, while those of CA-BlazePose are 93.6% and 91.8%, respectively. As shown in the red circle areas, compared with BlazePose, which suffers from missing leg keypoints and drifting of hand and elbow keypoints, the proposed method accurately identifies the keypoints of the legs and hands.



(a) BlazePose on sit-ups from high central position.



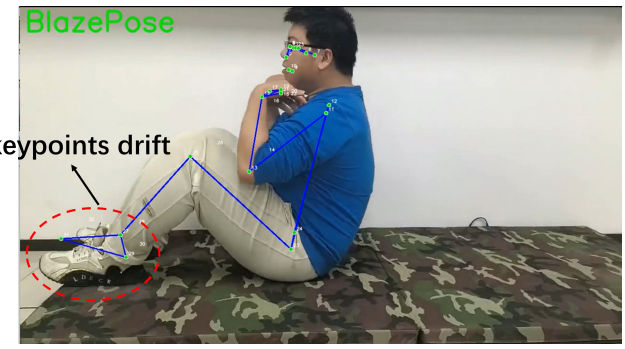
(b) CA-BlazePose on sit-ups from high central position.

Figure 6: Keypoint detection results on sit-ups from high central position.

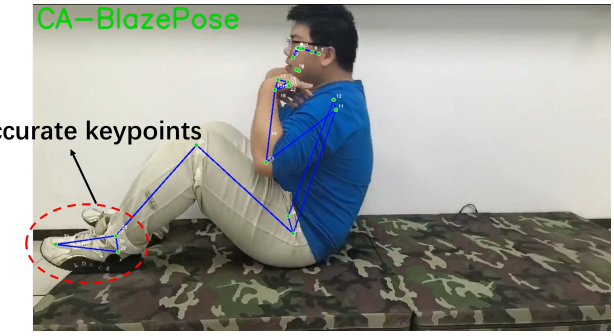
Figure 7a and Figure 7b visualize the keypoint detection results of the two models under the low central position of sit-ups. Under this view, the detection success rates of BlazePose for hands and feet are 78.2% and 76.5%, respectively, with obvious drift or missing keypoints on the feet, while those of CA-BlazePose are 91.5% and 89.9%, respectively. As shown in the red circle areas, compared with BlazePose, which suffers from drifting of foot keypoints, the proposed method accurately identifies the foot keypoints.

Figure 8a and Figure 8b visualize the keypoint detection results of the two models under the front view of pull-ups. Under this view, the detection success rates of BlazePose for hands and feet are 72.4% and 85.8%, respectively, with obvious drift on the hands and legs, while those of CA-BlazePose are 91.2% and 94.5%, respectively. As shown in the red circle areas, compared with BlazePose, which suffers from drift in hand, foot, elbow, and leg keypoints, the proposed method accurately identifies the corresponding keypoints.

These results demonstrate that in real-world motion detection under identical acquisition conditions, the keypoints detected by the CA-BlazePose model are closer to the actual human body positions compared to those detected by the BlazePose model, exhibiting stronger view adaptability. This highlights the significant application value of the proposed method for automated physical fitness testing systems, where stable and accurate keypoint detection is essential for reliable movement assessment and scoring.

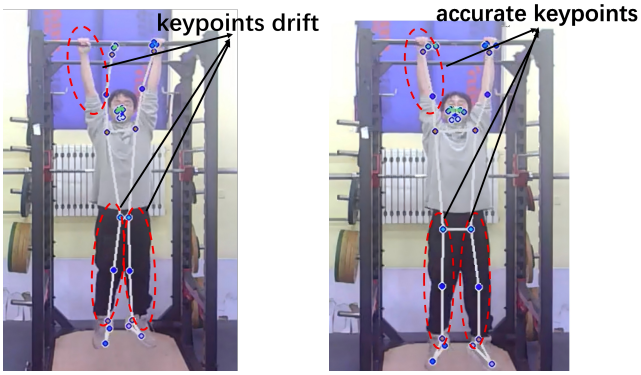


(a) BlazePose on sit-ups from low central position.



(b) CA-BlazePose on sit-ups from low central position.

Figure 7: Keypoint detection results on sit-ups from low central position.



(a) BlazePose on pull-ups from frontal position.

(b) CA-BlazePose on pull-ups from frontal position.

Figure 8: Keypoint detection results on pull-ups from frontal position.

3.6 Computational Complexity Analysis

To evaluate the computational efficiency and deployment feasibility of the proposed CA-BlazePose method, we used three metrics: number of parameters, floating-point operations (FLOPs), and inference time to assess the model

complexity, comparing CA-BlazePose with BlazePose and OpenPose. The number of parameters reflects the storage cost of the model, FLOPs measure the computational complexity, and inference time directly indicates the actual runtime speed. All tests were conducted on an NVIDIA Tesla T4 GPU under the same software environment as the training, i.e., PyTorch 1.10 and CUDA 11.3, with an input resolution of 256×256 and a batch size of 1.

Table 9 presents the complexity comparison results. OpenPose has approximately 50.0M parameters, 18.5G FLOPs, and an average inference time of 22.6 ms per frame, making it difficult to meet the high-frame-rate requirements of real-time physical fitness testing. BlazePose has 5.0M parameters, 0.98G FLOPs, and an inference time of 2.15 ms per frame. CA-BlazePose introduces a channel attention module in both the heatmap training stage and the regression fine-tuning stage, adding about 0.2M parameters. Its total parameters are approximately 5.2M, FLOPs increase to 1.05G, and inference time rises to 2.30 ms per frame. Although the inference speed of CA-BlazePose is slightly lower than that of BlazePose (an increase of about 0.15 ms, or 7%), its absolute latency is far lower than that of OpenPose and still meets the real-time requirement (less than 30 ms per frame). Meanwhile, as shown in Table 4 and Table 5, CA-BlazePose achieves the highest detection accuracy, significantly outperforming OpenPose and BlazePose in terms of overall PCK, extremity keypoint PCK, and regression error metrics. In summary, the proposed method achieves a significant improvement in keypoint detection accuracy and robustness with a very small additional computational cost (only 0.2M extra parameters and 0.07G extra FLOPs), and it offers the best overall performance in real-time physical fitness testing scenarios, demonstrating good practical value.

Table 9: Complexity comparison of OpenPose, BlazePose, and CA-BlazePose

Model	Params (M) ↓	FLOPs (G) ↓	Inference time (ms/frame) ↓
OpenPose [12]	50.0	18.5	22.6
BlazePose [19]	**5.0**	**0.98**	2.15
Proposed Method	5.2	1.05	**2.23**

4 Conclusion

This paper proposes CA-BlazePose, a keypoint detection method based on the channel attention mechanism. By incorporating channel attention into the network architecture, the model enhances the feature representation capability for specific keypoints such as the hands and feet, effectively improving keypoint detection accuracy and recognition stability under different camera views while maintaining overall recognition performance.

Experimental results on the COCO and LSP-LSPET datasets demonstrate that compared to BlazePose, the proposed model achieves approximately 7.1% and 6.5% improvement in recognition rates for hand and foot extremity keypoints, respectively, and an overall recognition rate improvement of approximately 5.5%. Compared to OpenPose, the proposed model achieves approximately 6.4% and 7.0% improvement in hand and foot keypoint detection rates, respectively, with an overall recognition rate improvement of approximately 8.2%. Furthermore, CA-BlazePose outperforms both BlazePose and OpenPose in terms of AED, MSE, and RMSE metrics, indicating that CA-BlazePose not only improves keypoint detection accuracy but also achieves more stable recognition performance relative to the baseline algorithms.

The ablation experiment further validates the necessity of the dual-stage attention design: when attention is introduced only in the heatmap stage or only in the regression stage, the overall PCK improves by only about 1.5% and 1.2%, respectively. In contrast, the dual-stage attention increases the overall PCK from 82.52% to 88.23%, an improvement of 5.7%, with hand and foot keypoint PCK improvements of 6.5% and 6.3%, respectively, significantly outperforming any single-stage configuration.

Further experiments in real-world physical fitness testing scenarios validate the recognition performance of CA-BlazePose and BlazePose under different camera views. The results show that under various viewing angles, the proposed method effectively mitigates the keypoint missing and drift problems observed in BlazePose, demonstrating better recognition robustness.

For a quantitative assessment of the performance differences in real-world scenarios, a total of 4,000 frames from four different capture views were evaluated. The results show that CA-BlazePose improves the hand and foot keypoint detection success rates by an average of approximately 20 percentage points compared to BlazePose, providing more reliable keypoint inputs for motion assessment in physical fitness test items such as sit-ups and pull-ups.

Moreover, the complexity analysis shows that CA-BlazePose has approximately 5.2M parameters, 1.05G FLOPs, and an inference time of 2.30 ms per frame. Compared to BlazePose, it adds only 0.2M parameters and 0.07G FLOPs, with a 7% reduction in inference speed, yet it still far exceeds the required 30 FPS for real-time applications and significantly outperforms OpenPose. All the above results demonstrate that the proposed method achieves a notable improvement in keypoint detection accuracy and robustness with very little additional computational cost.

Based on the above experiments, the CA-BlazePose keypoint detection method offers advantages in both human keypoint detection accuracy and recognition stability in practical physical fitness testing applications, providing significant application value for action recognition and assessment in computer vision-based automated count-based physical fitness testing.

Future work may explore extending the CA-BlazePose algorithm to other non-count-based physical fitness test items, such as standing long jump and sit-and-reach, to accommodate more diverse motion scenarios. Additionally, combining lightweight model design with edge computing deployment could promote real-time application of the algorithm on mobile or embedded devices, providing more efficient and accurate technical support for the development of smart sports.

Funding

National Natural Science Foundation of China (NSFC) (62571028 and 62071033); Changping Innovation Joint Fund of Beijing Natural Science Foundation (L234084).

Author Contributions

The manuscript was written with contributions from all authors. Conceptualization, Shaojun Yu, Wenhao Huo, Yuping Lu, Hanqing Zhao, Yilin Wang, Lili Wang and Muhammad Rizwan Anjum; methodology, Shaojun Yu, Wenhao Huo, Yuping Lu, Hanqing Zhao, Yilin Wang, Lili Wang and Muhammad Rizwan Anjum; software, Shaojun Yu; validation, Shaojun Yu, Wenhao Huo and Yuping Lu; formal analysis, Shaojun Yu, Wenhao Huo and Yuping Lu; investigation, Shaojun Yu; resources, Wenhao Huo; data curation, Shaojun Yu and Yilin Wang; writing—original draft preparation, Shaojun Yu and Wenhao Huo; writing—review and editing, Shaojun Yu, Wenhao Huo, Yuping Lu, Hanqing Zhao, Yilin Wang, Lili Wang and Muhammad Rizwan Anjum; visualization, Shaojun Yu; supervision, Wenhao Huo and Yuping Lu; project administration, Wenhao Huo; funding acquisition, Wenhao Huo. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

Data Available

The COCO dataset is publicly available at https://cocodataset.org/#download. The LSP-LSPET dataset is publicly available at https://www.uky.edu/engr/computer-vision/training/datasets/lsp/.

Ethical Approval

Informed consent for participation was obtained from all subjects involved in the study.

References

[1] Voeikov, R., Falaleev, N., Baikulov, R.: TTNet: Real-time temporal and spatial video analysis of table tennis. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3866–3874 (2020). https://doi.org/10.1109/CVPRW50498.2020.00450

[2] Ma, Y., Li, H., Yan, H.: Efficient real-time sports action pose estimation via EfficientPose and temporal graph convolution. IEEE Access **13**, 39901–39911 (2025). https://doi.org/10.1109/ACCESS.2025.3542240

[3] Colyer, S.L., Evans, M., Cosker, D.P., Salo, A.T.: A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. Sports Medicine - Open **4**(1), 24 (2018). https://doi.org/10.1186/s40798-018-0139-y

[4] Shi, Z., Zhao, H., Chen, J., Cheng, G.: Research on sit-up counting method and system based on human skeleton key point detection. Quality in Sport **24**, 55408 (2024). https://doi.org/10.12775/QS.2024.24.55408

[5] Song, Z., Chen, Z.: Sports action detection and counting algorithm based on pose estimation and its application in physical education teaching. Informatica **48**(10), 35–50 (2024). https://doi.org/10.31449/inf.v48i10.5918

[6] Guo, T., Yin, Q., Liu, X., Sun, Y., Qin, Z., Yu Han, Y., Lu, G.: Fitness exercise evaluation system based on improved DTW algorithm. Scientific Reports **15**(1), 19961 (2025). https://doi.org/10.1038/s41598-025-02535-5

[7] Lu, J., Yang, T., Zhao, B., Wang, H., Luo, M., Zhou, Y., Li, Z.: Review of deep learning-based human pose estimation methods. Laser Optoelectronics Progress **58**(24), 2400005 (2021). https://doi.org/10.3788/LOP202158.2400005

[8] Ramanan, D., Forsyth, D.A., Zisserman, A.: Strike a pose: Tracking people by finding stylized poses. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 271–278 (2005). https://doi.org/10.1109/CVPR.2005.335

[9] Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In Conference on Computer Vision and Pattern Recognition (CVPR 2011), pp. 1385–1392 (2011). https://doi.org/10.1109/CVPR.2011.5995741

[10] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R.: Real-time human pose recognition in parts from single depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), pp. 1297–1304 (2011). https://doi.org/10.1109/CVPR.2011.5995316

[11] Newell, A., Yang, K., Deng, J.: Stacked Hourglass Networks for Human Pose Estimation. In European Conference on Computer Vision, pp. 483–499 (2016). https://doi.org/10.1007/978-3-319-46484-8_29

[12] Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, **43**(1), 172–186 (2019). https://doi.org/10.1109/TPAMI.2019.2929257

[13] Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral Human Pose Regression. In Proceedings of the European Conference on Computer Vision (ECCV 2018), pp. 536–553 (2018). https://doi.org/10.1007/978-3-030-01231-1_33

[14] Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2353–2362 (2017). https://doi.org/10.1109/ICCV.2017.256

[15] Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5686–5696 (2019). https://doi.org/10.1109/CVPR.2019.00584

[16] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018). https://doi.org/10.1109/CVPR.2018.00742

[17] Kendall, A., Grimes, M., Cipolla, R.: PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2938–2946 (2015). https://doi.org/10.1109/ICCV.2015.336

[18] Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 472–487 (2018). https://doi.org/10.1007/978-3-030-01231-1_29

[19] Bazarevsky, V., Grishchenko, I., Raveendran, K., Grundmann, M., Zhang, F., Zhu, T.: BlazePose: On-device real-time body pose tracking. In CVPR 2020 Workshop on Computer Vision for Augmented and Virtual Reality, pp. 1–4 (2020).

[20] Hulleck, A.A., Alshehhi, A., El Rich, M., Khan, R., Katmah, R., Mohseni, M.: BlazePose-Seq2Seq: Leveraging regular RGB cameras for robust gait assessment. IEEE Transactions on Neural Systems and Rehabilitation Engineering **32**, 1715–1724 (2024). https://doi.org/10.1109/TNSRE.2024.3391908

[21] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In European Conference on Computer Vision (ECCV 2014), pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48

[22] Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In Proceedings of the British Machine Vision Conference, pp. 12.1–12.11 (2010). https://doi.org/10.5244/C.24.12