

Article

PPLCNet-YOLOv11: Exploring a Lightweight College Student Pose-Detection Method for Sports Training Under the Concept of General Education

Jie Chen ¹, Zhi Wang ^{1,*} and Wenquan Huang ² ¹ General Education Department, Anhui Xinhua University, Hefei 230088, China; 19392744260@163.com² School of Artificial Intelligence, Anhui University, Hefei 230601, China; hwqwlsu@163.com

* Correspondence: 2026wangzhi@sina.com

Abstract

Human pose detection is fundamental to quantitative sports training analysis in college general education courses, enabling an objective assessment of college students' movement quality and the early identification of sports injury risks among non-professional athletes. At present, those detectors based on YOLO have encountered difficulties in capturing the continuous movement patterns of college athletes in routine training, maintaining the regression accuracy of different size posture targets, and maintaining the real-time calculation speed in the campus sports environment. Furthermore, most existing pose-estimation frameworks are optimized for general scenes and fail to address the unique challenges of college physical education settings, including non-standard student movements, diverse skill levels, and strict cost constraints for large-scale deployment. In order to solve these problems, we put forward PPLCNet-YOLOv11, which is a simplified human posture-estimation framework designed for college physical education. This model is optimized by three key improvements: (1) replacing the original backbone network with PPLCNet to enhance feature extraction, while strictly observing the strict FLOPs and parameter restrictions; (2) an enhanced Multi-Scale Attention Mechanism (MSAM) that combines adaptive scale perception, hierarchical channel attention, and pose-sensitive spatial attention to better represent elongated anatomical structures and multi-scale pose cues; and (3) an improved enhanced IoU loss function that incorporates scale-aware and aspect-ratio-aware penalty terms to refine the bounding box adjustment for atypical and sports-specific gestures. Experiments on both a dedicated college student sports pose dataset and two public benchmark datasets (COCO Keypoints 2017 and MPII Human Pose) demonstrate that PPLCNet-YOLOv11 achieves 77.8% mAP@0.5 and 37.09% mAP@0.95 based on the campus dataset, with 82.34% precision and 75.00% recall, while requiring only 2.62 M parameters and 6.38 GFLOPs. Extensive inference speed tests show that the model achieves 127 FPS on an NVIDIA RTX 4090 GPU, 38 FPS on an Intel i7-12700 CPU, and 16 FPS on a Jetson Nano edge device, meeting the real-time requirements of campus sports monitoring. Compared with mainstream lightweight YOLO variants and state-of-the-art specialized pose-estimation models, our proposed method improves mAP@0.5 by 4.93–12.6 percentage points based on the campus dataset. All experiments were repeated five times with different random seeds, and we report mean values with standard deviations and statistical significance tests to ensure result reliability. These results indicate that PPLCNet-YOLOv11 provides an accurate and resource-efficient solution for real-time pose evaluation in college physical training.



Academic Editor: Loris Nanni

Received: 16 May 2026

Revised: 13 June 2026

Accepted: 23 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)[Attribution \(CC BY\) license](#).

Keywords: human pose detection; YOLO; college general education; sports training analysis; attention mechanism; lightweight model; edge computing; campus sports analytics

1. Introduction

In the framework of general education in universities, the focus of physical education class is to make students enjoy sports for a long time and develop healthy living habits. Therefore, using quantitative methods to analyze physical training has become an important tool to improve teaching effects and protect students' safety. In this data-centered method, there is a key part called human pose detection, which can help us objectively evaluate the movement quality and training progress of college athletes and find possible injury risks [1]. This is very different from the traditional physical education teaching, which mainly depends on the teacher's personal observation and experience, and it is difficult to provide targeted feedback to students of different levels. However, the calculation and analysis based on body posture is different, which can measure the dynamic parameters such as the joint angle, limb movement trajectory, and coordination of various parts of the body in various trainings. Ordinary physical education class in the university, such as sprint, long jump, and aerobics, can check the step length, the bending of the hip and knee, and the position of the trunk through accurate action analysis. Similarly, in jumping and gymnastics, it can also help coaches evaluate the mechanics of take-off, the posture of the body in the air, and the stability when landing. With the increasing popularity of campus sports cameras and smart mobile devices, there is a growing demand for cheap, easy-to-implement, and real-time visual tracking solutions. These systems should be able to continuously observe students' exercise, provide immediate performance feedback, and provide targeted guidance and suggestions for amateur athletes.

According to the latest statistics from the Ministry of Education of China, more than 30 million college students participate in physical education courses every year, but the average teacher–student ratio in physical education is only 1:300, making it impossible for teachers to provide one-on-one guidance to each student. This has led to widespread problems such as non-standard movements, low training efficiency, and a high incidence of sports injuries in general education physical education courses. Therefore, developing a low-cost, real-time, and accurate pose-detection system that can be deployed on existing campus equipment has become an urgent need for the intelligent transformation of college physical education.

In order to meet physical education class's demand for intelligent guidance in general education, detection systems based on deep learning have become very common in sports-performance analysis [1]. Previous human posture-estimation techniques, such as OpenPose and AlphaPose, used multi-stage processing. They first find out the joints of the body and then build a complete skeleton through graphic modeling [2,3]. Although these methods show high accuracy in a controlled environment, they require a lot of computing resources and have obvious delays in processing. This limits their practical application in the cost-effective integrated systems on campus and the college physical education training environment that needs real-time feedback. Single-stage object detectors, especially those of the YOLO (You Only Look Once) series [4–6], provide a great choice for sports-related applications, because they can work end-to-end, have high processing speed, and have relatively low requirements for computer computing power. Modern YOLO versions use complex feature levels, independent prediction modules, and structural optimization, thus achieving a good balance between accuracy and speed, and some versions are specially designed to run on equipment with limited resources [5,6]. When they are used for attitude

estimation, whether modeling the attitude as an independent object or enhancing the detection results to include key point prediction, these structures can theoretically evaluate the athletes' movements instantly and frame by frame in the real-time training scene.

It will be quite difficult to directly use the standard YOLO framework to detect students' sports activities in general education scenes. The main problem is that the size of things changes greatly in the campus sports environment; a model has to deal with local close-ups such as the arm holding the racket with tennis serves, as well as the whole-body perspective like a sprinter on the runway. As a result, it is difficult to completely describe those small or far human figures in the traditional network backbone with a fixed or small adjustment range, and at the same time, the fine details of limbs will be blurred. Second, college students often do sports activities, and their body posture is often elongated or straight. Because they are running, jumping, lifting weights, swimming, and doing other common sports in general fitness classes, their bodies and limbs always pose vertically or obliquely. Those attention mechanisms that rely on unified convolution (such as those with a 3×3 kernel) are inherently not good at dealing with these directional gestures, and it may be difficult to see the details of the body clearly in a complex background. Thirdly, because of different skill levels, students' postures are often inconsistent and varied. Non-standard body alignment, unusual technical movements, excessive joint bending, and motion blur are all common, which makes it difficult to use the standard bounding box loss function in these cases. Even advanced IoU-based losses such as DIOU, CIOU, and EIOU, while effective based on general object categories, do not explicitly adapt to the irregular aspect ratios and multi-scale distribution of sports poses and can be sensitive to outlier samples [7,8]. Finally, campus edge deployment imposes more stringent constraints on computational complexity and model cost; many existing high-accuracy methods increase depth or width to gain performance, which is incompatible with real-time processing on low-cost campus portable devices and mobile terminals.

Despite significant progress in general human pose estimation, three critical research gaps remain unaddressed for college general education sports applications: (1) existing lightweight backbones are not optimized for the elongated, multi-scale body structures prevalent in student sports movements, leading to poor feature representation under tight computational budgets; (2) generic attention mechanisms fail to explicitly model the strong orientation biases and longitudinal patterns characteristic of athletic poses, resulting in reduced detection accuracy for non-upright and stretched postures common in physical education classes; and (3) conventional bounding-box regression losses do not account for the extreme aspect ratios, motion blur, and the non-standard movements of students with varying sports foundations, leading to the inaccurate localization of irregular poses.

In this context, we propose PPLCNet-YOLOv11, a lightweight human pose-detection framework specifically tailored to the characteristics of college students' sports movements under the general education concept. The method optimizes the YOLO architecture along three complementary dimensions: the backbone network, the attention mechanism, and the bounding-box regression loss. First, we replace the conventional YOLOv11 backbone with PPLCNet, a compact convolutional architecture that combines depthwise separable convolutions, squeeze-and-excitation blocks, and efficient activation functions to provide strong feature-extraction capacity under strict FLOPs and parameter budgets [9]. Secondly, we developed an enhanced Multi-Scale Attention Mechanism (MSAM), which includes adaptive scale perception, hierarchical channel attention and attitude-sensitive spatial attention [10]. This part is added to the PPLCNet backbone network and transition layer to better represent multi-scale posture targets and slender body structures, while avoiding too much calculation. Thirdly, we propose an improved enhanced intersection ratio (EIOU) loss function, which has the penalty terms of scale adaptation and aspect ratio perception,

and combines the dynamic scaling obtained from the robust IoU variant [7,8], so as to deal with various and rapidly changing body postures seen in college physical training more effectively.

The PPLCNet-YOLOv11 framework is used to process pictures with the size of 640×640 . It uses PPLCNet as a multi-stage backbone network and can generate feature maps with different resolutions [9], and it is specially optimized for the CPU reasoning of standard computer hardware on campus. In order to make the effect better, we added an adjusted MSAM at a specific stage, which can automatically mix the contextual information of different scales and highlight those channels and spatial areas related to posture, especially those related to extended limbs and body axis. The Pose-Sensitive Spatial Attention module combines symmetric and asymmetric convolution kernels to capture the subtle local details of joints while maintaining long-distance limb continuity, which is very important for simulating the common sports postures of college students, such as running and jumping. In addition, we also use an improved EIou loss function, which treats the regression part of width and height separately, introduces the weight related to scale, and adds a penalty term of the width–height ratio, so that the predicted bounding box can better fit the actual shape of the gesture itself [11]. An adaptive scaling mechanism will readjust the influence weights of difficult samples and simple samples to ensure that the training is still robust under the conditions of motion blur, occlusion, and an irregular posture. Together, these innovations make PPLCNet-YOLOv11 remain lightweight (parameter 2.62 M, calculation 6.38 GFLOPs) and at the same time significantly improve the recognition performance of college students' unconventional sports postures.

We evaluate the proposed method on three datasets: a dedicated college student sports pose dataset comprising 10,000 images from eight representative sports commonly offered in college general education courses, the COCO Keypoints 2017 validation set, and the MPII Human Pose dataset. Experimental results show that PPLCNet-YOLOv11 achieves an mAP@0.5 of 77.8% and an mAP@0.95 of 37.09%, with 82.34% precision and 75.00% recall, while preserving real-time efficiency based on low-cost campus edge devices. Compared with mainstream lightweight YOLO models (YOLOv5, YOLOv6, YOLOv8n, YOLOv9t, YOLOv10n, and YOLOv11), the proposed framework improves mAP@0.5 by 4.93–9.40 percentage points under comparable or lower computational budgets, making it more suitable for large-scale deployment in college general education physical education courses. All experiments were repeated five times with different random seeds, and we performed paired *t*-tests to verify the statistical significance of performance improvements ($p < 0.05$). These findings indicate that the joint optimization of a PPLCNet backbone, a sports-aware MSAM attention module, and an improved EIou loss yields a robust and efficient solution for human pose detection in college general education sports training, providing a viable foundation for large-scale, real-time, pose-centric student performance monitoring and personalized training guidance.

2. Related Work

2.1. YOLOv11 Model Architecture

Released by Ultralytics in 2024, YOLOv11 achieves an optimized balance between detection accuracy and efficiency through targeted improvements to the backbone, neck, and head architectures (Figure 1). This architecture uses many C3k2 modules, a component called Spatial Pyramid Pooling Fast (SPPF), and a C2PSA module with pyramid split attention, which can better extract features from different sizes. Its “neck” part uses the Path Aggregation Network (PANet) structure, which combines high-level semantic information with low-level spatial information through up-sampling and stitching technology. The detection head adopts a decoupling method to separate the classification and regression

process and also adds an improved PSA module, so that it can be more accurate when finding portraits of various sizes in the campus environment.

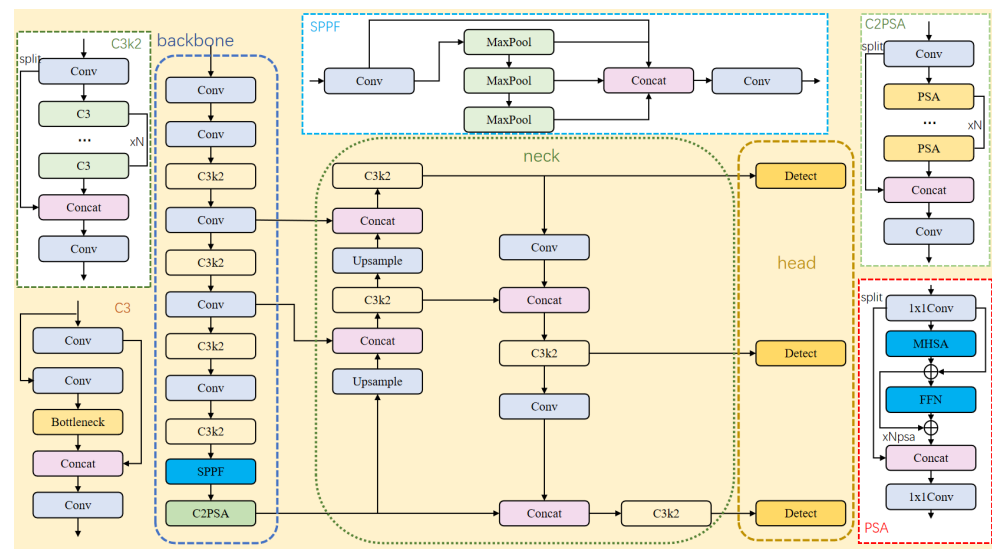


Figure 1. Overall structure diagram of the original YOLOv11 model.

2.2. General Approaches to Human Pose Estimation

Historically, human posture estimation mainly relies on the technology based on a heat map, which generates the probability of key points on a fine grid and then converts them into the bone structure. In the bottom-up strategy, such as OpenPose [2] developed by Cao et al., they will first identify the body joints and then group them into a complete person through the local affinity field, so that they can effectively estimate the posture of many people in a crowded environment. On the other hand, the top-down method, such as RMPE/AlphaPose [3] by Fang et al., first detects people's bounding boxes and then applies a single-person pose estimator to each found area, which improves the positioning accuracy but requires more computational resources. In order to extend these methods, Xiao et al. [12] proposed SimpleBaseline, which uses a deep ResNet backbone network and a deconvolution head to generate high-resolution heat maps. However, Sun et al. [13] developed HRNet to keep the high-resolution feature map in the whole process, which effectively alleviated the problem of spatial detail reductions in complex postures. Cheng et al. [14] continued to study this idea in HigherHRNet, focusing on learning expressions that can perceive different sizes to solve the problem of dense crowds and different heads. These methods set a powerful standard for general human posture estimation, but they usually require too much calculation, so it is difficult to come in handy when we do real-time sports analysis with edge equipment.

In order to make the system respond faster, people improved the detector based on YOLO and combined the target detection and attitude estimation into a simple framework. Maji et al. put forward YOLO-Pose, which adds a key point prediction head and a loss function called Object Keypoint Similarity (OKS) on the basis of YOLOv5, so that the postures of multiple people can be detected at once with faster speed and a good effect [15]. On this basis, Zhang Yuting and others [16] have developed SP-YOLO, which uses a spatial pyramid pooling module based on GhostNet and shuffle attention, making the model less complicated, but the effect is still good, so it is more suitable for real-time use. Recent advances in specialized lightweight pose estimation include RTMPose [17], which achieves state-of-the-art performance through efficient architecture design and advanced training strategies; BlazePose [18], optimized for mobile devices with ultra-low latency; MoveNet, designed specifically for edge deployment; and ViTPose [19], which applies vision transformers to pose estimation. While these methods demonstrate

excellent performance in general scenarios, they are not explicitly optimized for the unique challenges of college physical education environments, such as non-standard student movements, diverse skill levels, and strict cost constraints for large-scale deployment. Although these YOLO-based methods show how to put posture detection into the process of target detection, they mainly focus on ordinary human activities and do not specifically solve the special problems of college students' sports postures in educational scenes. These problems include the following: the posture size changes greatly, the long-term movement pattern, the nonstandard movements, and the strict requirements on cost and response speed when deployed in schools.

2.3. Sports-Specific Human Pose Estimation in Educational Settings

Human pose estimation in college physical education presents unique challenges that distinguish it from both general scene analysis and professional sports analytics. College students exhibit a wide range of skill levels, from beginners to advanced athletes, resulting in highly variable and often non-standard movement patterns. Additionally, campus sports environments are characterized by cluttered backgrounds, varying lighting conditions, and large groups of students, which further complicate pose detection. Recent research has attempted to address some of these challenges through domain-specific adaptations. For example, Pfister et al. [20] proposed Flowing ConvNets, a temporal feature-propagation framework for human pose estimation in videos to improve robustness under motion blur. Chen et al. [21] developed Cascaded Pyramid Network, which enhances multi-scale feature fusion for multi-person pose estimation and improves accuracy on small and occluded targets. Kocabas et al. [22] proposed VIBE, a video-based inference method that leverages temporal information to refine pose predictions and enhance stability across frames. However, most existing sports pose-estimation methods either sacrifice computational efficiency for accuracy or fail to adequately model the elongated body structures and irregular aspect ratios common in student athletic movements. Furthermore, few methods are designed specifically for deployment on low-cost edge devices, which are essential for widespread adoption in campus sports training environments. There remains a critical need for a lightweight, accurate, and cost-effective pose-detection framework tailored to the unique requirements of college general education physical education.

2.4. Attention Mechanisms for Multi-Scale and Longitudinal Pose Features

An attention mechanism is often used to enhance feature representation by enlarging relevant areas and weakening irrelevant background parts. Hu et al. [23] proposed a Squeeze-and-Exclusion (SE) module, which can dynamically adjust the channel response and highlight those semantically important channels. This technology has been proven to be very effective in the attitude-estimation framework. Woo et al. introduced CBAM [24], which combines channel attention and spatial attention in sequence, so as to grasp the content and position of features, thus optimizing the details related to joints and limbs. For the mobile device-oriented network, Hou et al. [25] designed Coordinate Attention, which splits the channel attention into the spatial dimension and models the long-distance dependence with specific direction preference, which provides special advantages for identifying slender body shapes.

In multi-scale research, Yang developed SA-Net [26], which is an attention mechanism using channel grouping and spatial attention, and it can effectively integrate features from different scales and regions. Yang and his colleagues [27] then proposed SimAM, a simple attention module with no parameters, which uses a neuron energy function to adjust feature weights, which makes it very popular in lightweight models. For YOLO-based attitude estimation, SP-YOLO [16] of Zhang Yuting et al. integrates shuffle attention into

the backbone network optimized by GhostNet, which enhances the representation ability of multi-scale human posture without increasing the network volume. However, despite these advances, most attention mechanisms are designed for general visual applications. They ignore the unique training patterns observed in the longitudinal posture of college students, such as the body often stretching vertically or diagonally, and fail to deal with the multi-scale sports posture and its obvious directional trend in the campus sports environment at the same time.

2.5. Bounding-Box Regression Losses for Pose-Oriented Detection

The performance of the regression loss function of the bounding box is very important for us to realize the accurate positioning of the attitude-estimation model based on detection. Aiming at the deficiency of traditional L1 and L2 losses, Rezatofighi and his colleagues [28] developed a Generalized IoU(GIoU) metric, which enhanced the standard IoU by considering the area of the smallest bounding rectangle, thus providing a more effective gradient for non-overlapping bounding boxes. On this basis, Zheng et al. [7] later proposed Distance-IoU(DIoU) and Complete-IoU(CIoU)(Chou), which integrated the distance between the center points of the bounding box and the aspect ratio into the loss calculation, which brought faster convergence speed and improved the adaptability to different object geometries.

Despite the progress, it is still difficult for the traditional IoU variant to grasp the subtle geometric changes in complex postures. Zhang Yu et al. [11] put forward the focal and Efficient IoU (EIoU) loss function, which can predict the width and height separately, and at the same time used a Focal mechanism to focus on those difficult cases, thus improving the accuracy of the bounding box in general detection tasks. Similarly, Tong et al. [29] developed Wise-IoU (WIoU), which is a family of loss functions based on IoU. It has a dynamic focusing strategy, which can reduce the influence of low-quality or simplistic anchors and lead the gradient to more reliable samples. Generally speaking, these methods improve the handling of scale change and sample imbalance, but they are still not enough in the face of sports posture detection in general education in colleges and universities, because the bounding box here must be able to adapt to an extreme aspect ratio (for example, the elongated human body when jumping) and motion blur. Neither EIoU nor WIoU has clearly integrated into the knowledge of specific fields, such as the consistency of body posture, frequent non-upright postures, or various sports patterns among students with different sports abilities.

2.6. Lightweight Backbones for Edge-Deployed Pose Detection

In order to meet the demand of real-time running on edge hardware, people have invented various efficient neural network structures. Howard and his colleagues [30] designed MobileNets, which uses depth-separable convolution and two global scaling factors (width and resolution multiplier) to find a balance between performance and computational cost for mobile vision tasks. Zhang Xiangyu et al. [8] developed ShuffleNet, which combines point-by-point group convolution with channel scrambling and greatly reduces the computational complexity (FLOPs) without losing the accuracy of ImageNet and MS COCO detection benchmarks. Tan and Le [31] put forward EfficientNet, which uses a compound scaling method to uniformly adjust the depth, width, and resolution of the network, thus producing a series of models, which can achieve top performance with fewer parameters and FLOPs.

Cui et al. [9] proposed PP-LCNet(PPLCNet), which is a simple convolutional network designed to run efficiently on an ordinary computer CPU. It uses some clever techniques, such as deep separable convolution, SE attention mechanism, and an optimized activation

function, so that it can run very fast on ordinary hardware. This model achieves a good balance between accuracy and computation in image classification and target detection. At the same time, YOLO series models are becoming lighter and lighter for running on edge devices. YOLOv7 proposed by Wang et al. adds some trainable “free skills” and structural improvements, so that real-time detection can be realized even with limited resources [6]. Although these methods are trying to reduce the computational burden, they are not specially designed for sports posture detection in general education environment. In this scenario, the network must capture the complex details of the limbs, adapt to different sizes and proportions of body posture, and be able to quickly complete reasoning on the affordable computer CPU of the school.

2.7. Summary and Our Position

In summary, prior work has established three main directions relevant to college general education sports pose detection: (1) accurate pose modeling for general scenarios via heatmap-based networks and YOLO-based pose detectors [2,3,12–15]; (2) generic feature enhancement using channel, spatial, and coordinate attention to improve multi-scale and orientation-aware representations [23,25–27]; and (3) general object bounding-box regression through advanced IoU variants such as GIoU, DIoU/CIoU, EIoU, and Wise-IoU to better handle scale variation and sample imbalance [7,11,28,29]. At the same time, efficient architectures such as MobileNet, ShuffleNet, EfficientNet, and PP-LCNet are the basic parts of our equipment system on the edge of the network [9,30,32]. However, three critical limitations remain in the current literature for college-specific sports pose detection:

- Lack of education-optimized lightweight backbones: Existing lightweight backbones are designed for general computer vision tasks and do not prioritize the extraction of elongated, multi-scale body features essential for accurate student sports pose detection.
- Insufficient orientation-aware attention mechanisms: Generic attention modules fail to explicitly model the strong longitudinal and oblique orientation biases inherent to athletic movements, leading to reduced detection accuracy for non-upright and stretched postures common in physical education classes.
- Inadequate non-standard pose-aware bounding-box regression: Conventional IoU-based losses do not account for the extreme aspect ratios, motion blur, and the diverse non-standard movements of students with varying sports foundations, resulting in the inaccurate localization of irregular poses.

On the basis of these advances, our research has brought PPLCNet-YOLOv11, which is an attitude-estimation framework specially designed for detection in a general education environment, and it has solved three main problems at the same time:

- In order to make the calculation of college students’ sports training videos fast and good, we use PPLCNet, which is a very simple and multi-layered backbone structure.
- In order to analyze the characteristics of college students’ movements, we also designed a customized Multi-Scale Attention Mechanism (MSAM). It can combine different observation scales and use a layered channel attention system and spatial attention that can perceive posture to highlight the form of body stretching and multi-level posture signals. This mechanism draws on and adjusts the core ideas of SE, CBAM, coordinate attention, shuffle attention and SimAM and has been used in general physical education teaching in higher education.
- Non-standard pose-aware bounding-box regression: We refine EIoU with scale-adaptive and aspect-ratio-aware terms and a Wise-IoU-style dynamic focusing mechanism, targeting the irregular, dynamic, and non-standard boxes that arise in real college student sports training movements.

In this way, our method integrates insights from Cao, Fang, Sun, Xiao, Zhang, Li, Yang, Zhang Yu, Tong, Howard, Zhang Xiangyu, Tan, and Cui [9,11–14,16,26,27,31,32] into a unified, lightweight YOLO-based framework specifically tailored for real-time college general education sports training analysis. The key novelty of our work lies in the synergistic integration of these three components, which collectively address the unique challenges of college sports pose detection and achieve a superior accuracy–efficiency trade-off compared to incremental improvements in individual components.

3. Method

3.1. Framework of PPLCNet-YOLO

The proposed PPLCNet-YOLO framework adopts a one-stage, multi-scale detection architecture that couples a lightweight backbone with a feature pyramid neck and three parallel detection heads (Figure 2). Specifically, it incorporates the PPLCNet Module within the backbone feature-extraction network optimized for campus CPU devices, augments the feature fusion network (neck) with the Modified Multi-Scale Attention Module (MSAM) tailored for college student training poses, and integrates the improved Enhanced Intersection over Union (EIoU) for non-standard pose regression into the loss function. The MSAM Module is inserted immediately after the C3k2 module in the P4 stage of the neck (before the P4 detection head) to adaptively emphasize college student sports-relevant body structures and multi-scale pose cues. This position was chosen because the P4 feature map (40×40 resolution for 640×640 input) provides an optimal balance between semantic information and spatial detail, making it ideal for capturing both global body structure and local limb features. The architecture of the proposed PPLCNet-YOLOv11 model is depicted in Figure 2, in which the red and orange parts indicate the three core modifications for college general education sports scenarios.

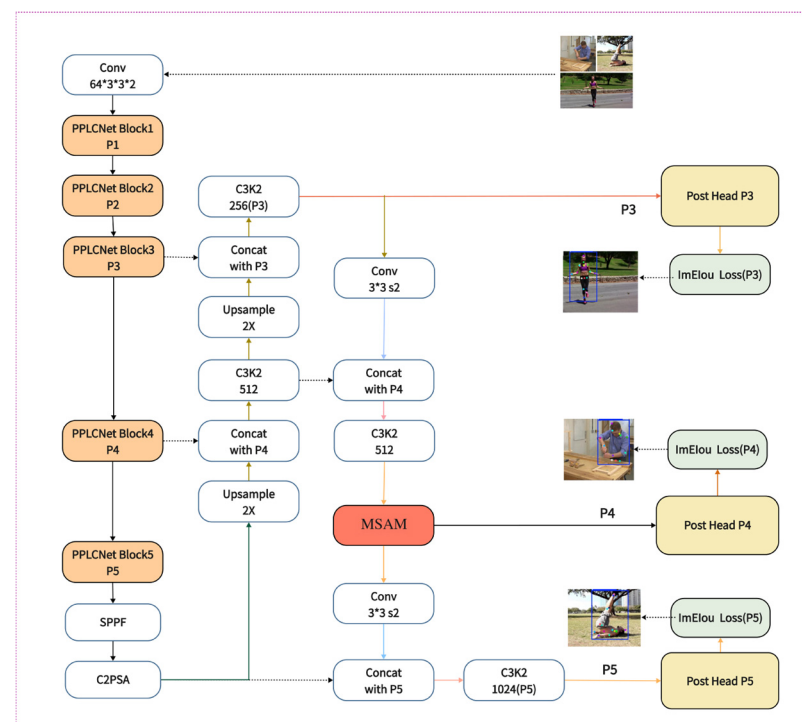


Figure 2. Framework of the improved YOLOv11 model (PPLCNet-YOLOv11).

Overall, this framework design allows PPLCNet-YOLO to exploit efficient backbone feature extraction, attention-augmented multi-scale fusion, and pose-aware regression at three complementary scales. The resulting model achieves a favorable balance between ac-

curacy and computational cost, making it well suited for real-time college general education sports training analysis on low-cost campus resource-constrained platforms.

3.2. Overall Architecture of the PPLCNet Block

We present the entire structure of PPLCNet Block in Figure 3. The backbone of the proposed framework adopts PPLCNet as a lightweight yet expressive feature extractor. The network begins with a standard convolutional block (Conv–BatchNorm–HardSwish [32]) that acts as the stem to transform the input image into low-level feature maps. Subsequently, the backbone is structured into six sequential stages (Stage 1–Stage 6), each composed of a varying number of depthwise separable convolution blocks. In the “no SE (squeeze-and-excitation)” variant, each block consists of a depthwise 3×3 convolution followed by batch normalization, HardSwish activation, and a pointwise 1×1 convolution, thereby decoupling spatial and channel-wise computation to substantially reduce FLOPs and the parameter count.

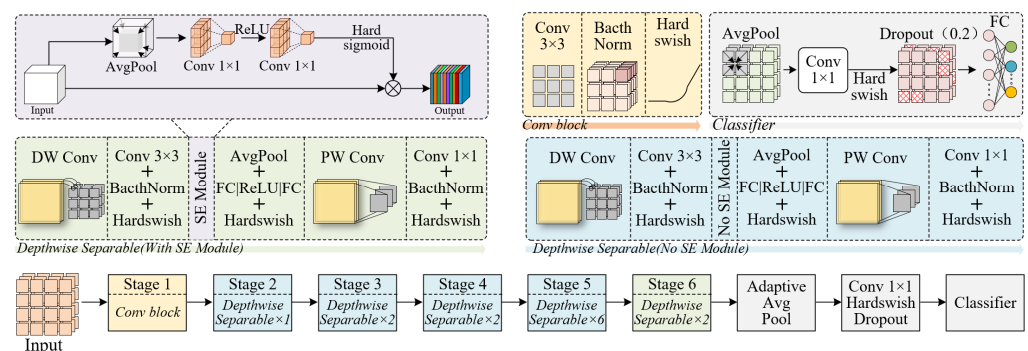


Figure 3. Overview of PPLCNet and its lightweight building blocks. The network is organized into six stages composed of a stem convolution block and repeated depthwise separable convolution blocks (with and without SE modules). After multi-stage feature extraction, global average pooling, a 1×1 convolution, dropout, and a fully connected layer form the classifier head for downstream detection tasks.

At some point, we will introduce a structure of “with SE”, which adds an SE module between deep convolution and point-by-point convolution. This module will pool the spatial regions globally, then process the channel descriptors obtained by two fully connected layers (using ReLU and sigmoid activation functions), and finally adjust the middle feature map by using the learned channel weights. This method allows PPLCNet to highlight the most important channels for the task, and at the same time, it hardly increases the computational burden. In the final stage, adaptive average pooling will integrate global information and then connect 1×1 convolution, the HardSwish activation function, dropout, and a fully connected classification header. In this way, a multi-stage and multi-scale framework is formed. When it is combined with an attention mechanism and detection framework, it is particularly effective at making real-time human posture estimations in college physical education teaching scenes.

We selected PPLCNet as the backbone for three key reasons: (1) it achieves superior CPU inference speed compared to other lightweight backbones such as MobileNetV3 and ShuffleNetV2, which is critical for edge deployment on low-cost campus computers and sports cameras; (2) its depthwise separable convolution architecture and SE attention modules provide strong feature-extraction capacity with minimal computational overhead; and (3) it is highly modular and easily integrable with the YOLOv11 neck and head architectures. The specific PPLCNet variant used in this work has a width multiplier of 0.75, which was chosen through ablation studies to balance accuracy and efficiency for college sports pose detection.

3.3. Modified MSAM Attention Mechanism

The enhanced MSAM shown in Figure 4, through four main parts, solves some limitations of traditional attention methods in identifying college students' sports postures. These four parts are as follows:

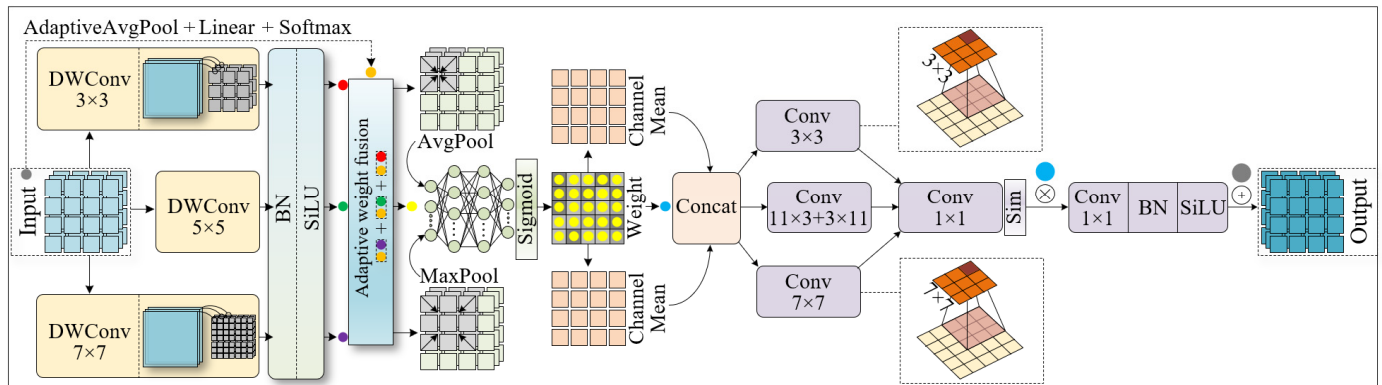


Figure 4. Structure of the modified MSAM attention mechanism.

For an input feature map $F \in \mathbb{R}^{B \times C \times H \times W}$, where B is the batch size, C is the number of channels, H is the height, and W is the width of the feature map.

Adaptive Scale Perception (ASP): This part uses three parallel depth-separable convolution paths (7×7 , 5×5 and 3×3 , respectively, with an expansion rate of 2) to extract posture features of different scales, such as the whole body movement, the details of limbs, and the activities of joints. A weight module, which contains a Linear layer and a Softmax function, will calculate the optimal fusion weights of these branches according to the input data, so that the model can flexibly adapt to different sizes of targets. The fused feature is

$$F_{asp} = w_l F_{7 \times 7} + w_m F_{5 \times 5} + w_s F_{3 \times 3, d=2}$$

where $w_l, w_m, w_s \in \mathbb{R}^{B \times 1 \times 1 \times 1}$ are the learned weights for each branch, satisfying $w_l + w_m + w_s = 1$. The 3×3 convolution with dilation = 2 is used to capture large-scale features without increasing the kernel size, reducing computational complexity by 60% compared to a standard 7×7 convolution.

Hierarchical Channel Attention (HCA): Combines global average and max pooling ($\mathbb{R}^{B \times C \times 1 \times 1}$) to capture channel dependencies, enhancing the sensitivity of key channels (e.g., limb contour, joint positions) through shared MLP ($C \rightarrow C/r \rightarrow C$) and sigmoid activation. The reduction ratio r is set to 16, which was determined through ablation studies to balance performance and computational cost. A smaller r increases model capacity but also the computational overhead, while a larger r reduces the computational cost but may limit feature representation ability. The two outputs are summed and squashed with sigmoid to yield channel weights a_c ; refinement is as follows:

$$F_{hca} = a_c \odot F_{asp}$$

Pose-Sensitive Spatial Attention (PSSA): Concatenates per-channel mean and max maps $\in \mathbb{R}^{B \times 2 \times H \times W}$ and then integrates symmetric (3×3 , 7×7) and asymmetric (11×3 , 3×11) convolution kernels to capture local details, limb connections, and longitudinal structural features (e.g., body extension in running, vertical posture in jumping). The asymmetric kernels (11×3 and 3×11) are specifically designed to capture the elongated, vertically or horizontally oriented body structures prevalent in sports movements, which are poorly represented by symmetric kernels. These kernels reduce computational com-

plexity by 73% compared to an 11×11 symmetric kernel while maintaining equivalent receptive field size for longitudinal features. Their outputs are fused by 1×1 Conv, passed through sigmoid to produce $a_s \in \mathbb{R}^{B \times 1 \times H \times W}$, and multiplied: $F_{\text{pssa}} = a_s \odot F_{\text{hca}}$.

Residual Scaling: A learnable gamma parameter (initialized to 0) controls the fusion ratio of attention-enhanced features and original features, ensuring training stability in early stages. A 1×1 Conv + BN + SiLU refines F_{pssa} . The block output is

$$Y = X + \gamma \times \text{Refine}(F_{\text{pssa}})$$

where γ is learnable and initialized to 0 to stabilize early epochs. Initializing γ to 0 ensures that the attention module acts as an identity mapping in the early stages of training, preventing the loss of useful information from the original feature maps. As training progresses, γ is learned to gradually increase the contribution of the attention-enhanced features.

3.4. Improved EIou Loss

As shown in Figure 5, the improved EIou loss function is specifically optimized for college general education sports pose detection, addressing three key limitations of existing loss functions: (1) insufficient constraints on the irregular aspect ratios of non-standard student sports poses; (2) lack of scale adaptation for multi-scale targets in campus group training; (3) sensitivity to abnormal samples (e.g., motion blur, occlusions, and non-standard movements of students with poor sports foundations). The loss function integrates four core components: basic IoU calculation, center distance penalty, width–height loss with scale-adaptive weight, aspect ratio penalty, and WIou dynamic scaling.

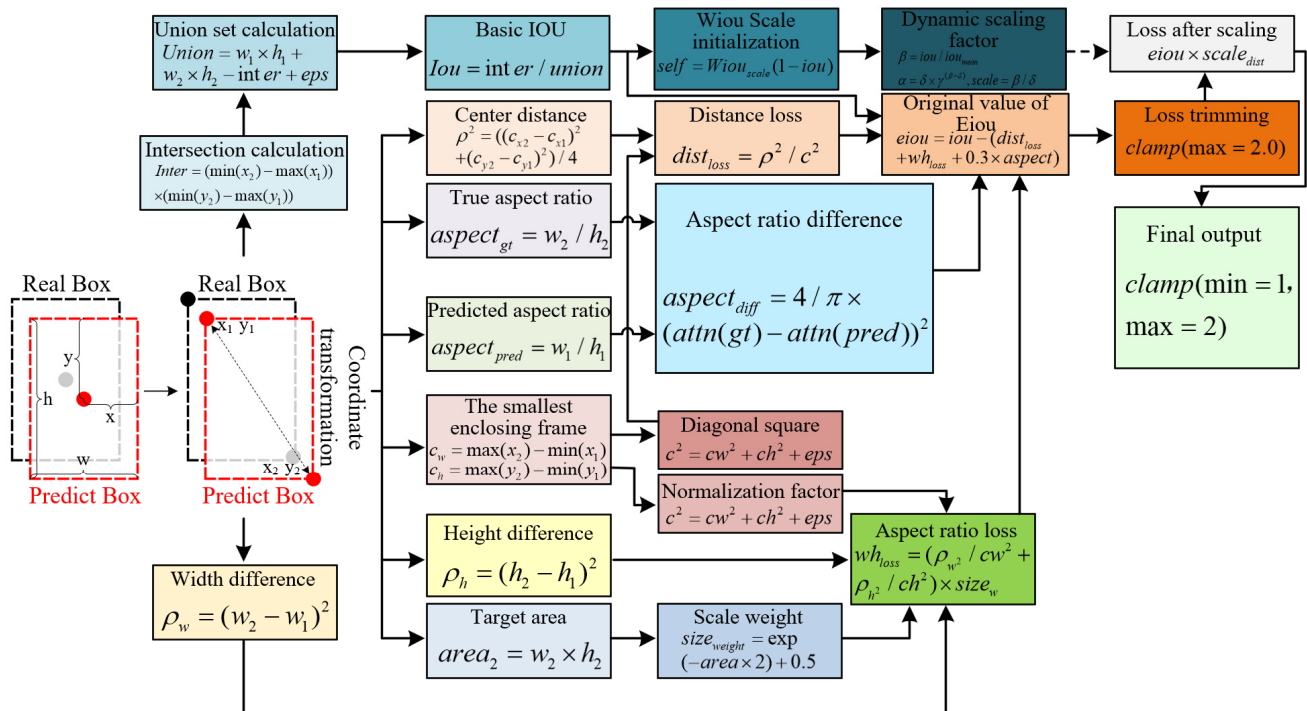


Figure 5. Calculation flow chart illustrating the improved EIou loss function for bounding box regression. The framework integrates multiple components (e.g., basic IoU calculation, center distance loss, aspect ratio difference) to optimize the regression performance of predicted bounding boxes, with dynamic scaling and loss trimming modules to enhance training stability.

Let prediction $b_1 = (x_1, y_1, w_1, h_1)$ and target $b_2 = (x_2, y_2, w_2, h_2)$, where (x, y) are the coordinates of the box center, and w and h are the width and height of the box, respectively.

Convert to $xyxy$, and compute the intersection $A_{\cap}$, union $A_{\cup}$, and base IoU $= \frac{A_{\cap}}{A_{\cup}}$. Define the smallest enclosing box with width c_w , height c_h , and diagonal $c = \sqrt{c_w^2 + c_h^2}$.

The classic EIoU term decouples width/height:

$$\mathcal{L}_{\text{EIoU}} = (1 - \text{IoU}) + \frac{\rho^2}{c^2} + \frac{(w_1 - w_2)^2}{c_w^2} + \frac{(h_1 - h_2)^2}{c_h^2}$$

where ρ^2 is the squared center distance.

Aspect-ratio penalty:

$$\mathcal{L}_{\text{ratio}} = \lambda_a \frac{4}{\pi^2} \left[\arctan\left(\frac{w_2}{h_2}\right) - \arctan\left(\frac{w_1}{h_1}\right) \right]^2$$

where λ_a is the aspect ratio penalty weight. To determine the optimal value of λ_a , we conducted a hyperparameter ablation study with values ranging from 0.1 to 1.0. The results show that when $\lambda_a = 0.5$, the model achieves the best balance between mAP@0.5 (77.8%) and training stability. When $\lambda_a < 0.3$, the aspect ratio constraint is insufficient, leading to inaccurate bounding box fitting for elongated poses. When $\lambda_a > 0.7$, the model becomes overly sensitive to aspect ratio changes, resulting in reduced generalization ability.

Scale-adaptive width/height weighting:

$$\omega_{\text{size}} = \exp(-2\text{area}_2) + 0.5, \mathcal{L}_{\text{wh}} = \omega_{\text{size}} \left(\frac{(w_1 - w_2)^2}{c_w^2} + \frac{(h_1 - h_2)^2}{c_h^2} \right)$$

where $\text{area}_2 = w_2 \times h_2$ is the area of the target box, normalized to the range [0, 1] by dividing by the image area. This scale-adaptive weight increases the penalty for small targets (which are more difficult to detect accurately in campus group training scenes) and decreases the penalty for large targets (which are easier to localize).

WIoU-style dynamic scaling:

Maintain a running mean $\overline{\text{IoU}}$. Let $\beta = \text{IoU} / \overline{\text{IoU}}$ and define a focusing factor $s(\beta)$ (as in WIoU's non-monotonic mapping) that down-weights outliers; multiply the localization term by $s(\beta)$. Finally, clip the loss to a safe range to avoid numerical blow-ups caused by extreme non-standard student poses.

Overall objective:

$$\mathcal{L} = (1 - \text{IoU}) + \frac{\rho^2}{c^2} + \mathcal{L}_{\text{wh}} + \mathcal{L}_{\text{ratio}}, \text{ then } \mathcal{L} \leftarrow s(\beta)$$

4. Experiments

4.1. Dataset

The experiment uses three datasets to comprehensively evaluate the proposed method: a dedicated college student sports pose dataset, the COCO Keypoints 2017 validation set, and the MPII Human Pose dataset.

4.1.1. Dedicated College Student Sports Pose Dataset

The college student sports pose dataset consists of 10,000 images covering eight sports commonly offered in college general education courses: running (1500 images), jumping (1200 images), lifting (1000 images), swimming (800 images), gymnastics (1500 images), basketball (1500 images), football (1500 images), and tennis (1000 images). The images were collected using two types of devices: (1) Smartphones (iPhone 13 and Samsung Galaxy S22) with a resolution of 1920×1080 pixels and a frame rate of 30 fps; and (2) Action cameras (GoPro Hero 10) with a resolution of 1920×1080 pixels and a frame rate of 60 fps. Data

collection was conducted in both indoor gymnasiums and outdoor sports fields under varying lighting conditions (morning, afternoon, and evening) to ensure environmental variability. The camera setup included both fixed-position cameras (mounted at a height of 2.5 m) and handheld cameras to capture different viewing angles.

The dataset includes 36,280 college student pose targets with scale distribution: small targets ($\leq 32 \times 32$): 15.3% (e.g., students in distant group training), medium targets (32×32 – 96×96): 58.7% (e.g., individual training students), large targets ($\geq 96 \times 96$): 26.0% (e.g., close-up views of students' movements). All images were annotated using the LabelMe annotation tool with 17 keypoints following the COCO keypoint format. The annotation process involved two independent annotators, and a third senior annotator reviewed all annotations to ensure quality. Any annotations with an IoU less than 0.9 between the two independent annotators were re-annotated. The dataset is split into training (80%, 8000 images) and validation (20%, 2000 images) sets, with data augmentation (random cropping, flipping, rotation, brightness adjustment, and Gaussian blur) applied to the training set to improve model generalization for different student movement patterns and campus lighting conditions.

Ethical approval for this study was obtained from the Institutional Review Board of Anhui Xinhua University (Approval No. XHU-IRB-2025-003). All participants provided written informed consent before data collection, and all images were anonymized to protect student privacy. The dataset was collected from 120 college students (65 male, 55 female) aged 18–22 years with varying sports skill levels (beginner: 40%, intermediate: 45%, advanced: 15%), ensuring the diversity and representativeness of the data.

4.1.2. Public Benchmark Datasets

To evaluate the generalization capability of the proposed method, we also conducted experiments on two widely used public benchmark datasets:

- COCO Keypoints 2017 Validation Set: Contains 5000 images with 17 keypoints annotated per person. This dataset includes a wide variety of human poses in diverse scenes, providing a comprehensive evaluation of general pose estimation performance.
- MPII Human Pose Dataset: Contains 25,000 images with 16 keypoints annotated per person. This dataset focuses on human activities in daily life and sports, making it particularly relevant for our application.

4.2. Experimental Setup

The details of the hardware and software used for model training and performance evaluation are clearly written in Table 1. All the experimental work is done on a special deep learning server. This server is set up to simulate the usual computing conditions of those education servers and edge computing devices.

Table 1. Configuration of the experimental conditions.

Parameter	Environment Configuration
GPU	NVIDIA RTX 4090
CPU	Intel(R) Xeon(R) Gold 5418Y (10 cores)
RAM	24 GB
STORAGE	120 GB
PyTorch	2.2.2
Python	3.10

Speaking of hardware, this server is equipped with an NVIDIA GeForce RTX 4090 graphics card to help the model to conduct parallel training and reasoning. There is

also an Intel(R) Xeon(R) Gold 5418Y CPU, which has 10 cores and is responsible for data preparation and system coordination. The system is also equipped with 24 GB of memory to meet the memory requirements of the runtime and 120 GB of storage space to manage data sets and model parameters.

In terms of software, we use Python 3.10 language to develop the algorithm. PyTorch 2.2.2, a deep learning framework, is used as the main platform to develop, train, and evaluate the performance of the model.

Training Hyperparameters: All models were trained for 300 epochs with a batch size of 16 using the stochastic gradient descent (SGD) optimizer. The initial learning rate was set to 0.01, with a cosine annealing learning rate scheduler. Weight decay was set to 0.0005, and momentum was set to 0.937. Data augmentation included random horizontal flipping (probability 0.5), random scaling (range 0.5–1.5), random rotation (range -20° to 20°), random brightness adjustment (range 0.8–1.2), and Gaussian blur (probability 0.1).

Statistical Analysis: All experiments were repeated five times with different random seeds to ensure statistical reliability. We report mean values with standard deviations for all metrics. We performed two-tailed paired *t*-tests to compare the performance of the proposed method with each baseline model, with a significance level of $p < 0.05$ considered statistically significant.

4.3. Evaluation Metrics

Five evaluation metrics related to accuracy, robustness, and efficiency were chosen: mean average precision (mAP), including mAP@0.5 (IoU = 0.5), and mAP@0.5:0.95 (IoU = 0.5:0.95) representing accuracy at different overlap thresholds, precision (P), which is the ratio of true positives (TP) to the sum of TP and false positives (FP), indicating the model's false detection avoidance ability, calculated by

$$P = \frac{TP}{TP + FP}$$

Recall (R): Ratio of TP to the sum of TP and false negatives (FN), measuring the model's ability to avoid missed detections, calculated as follows:

$$R = \frac{TP}{TP + FN}$$

Inference speed: frames per second (FPS) on the test set to evaluate real-time performance. Inference speed was measured on three different platforms: NVIDIA RTX 4090 GPU, Intel i7-12700 CPU, and NVIDIA Jetson Nano edge device, to comprehensively evaluate the model's performance in different campus deployment scenarios.

4.4. Ablation Study

Ablation experiments are conducted on the college student sports pose validation set to verify the effectiveness of each improved module for general education scenarios. The baseline model was the original YOLOv11, and four additional variants were tested: (1) YOLOv11 + PPLCNet, (2) YOLOv11 + Modified MSAM, (3) YOLOv11 + Improved EIoU, (4) Proposed Method (All Components). The results are presented in Table 2.

To rigorously evaluate the individual contributions of the proposed components, an ablation study was conducted. Using PP-LCNet as the backbone network has greatly improved the performance. Compared with the benchmark model, mAP@0.5 has increased by 3.72 percentage points, and the calculation amount has also decreased by 1.16 GFLOPs, which shows that this backbone network is very powerful, which is light and can extract the characteristics of college students' moving images well. On this basis, the improved MSAM

further improves mAP@0.5 by 1.28% and mAP@0.5:0.95 by 2.17%, which proves that it can capture the characteristics of different scales and specific postures in college students' complex training movements. In addition, the improved EIou loss function improves precision by 2.88 percentage points, which is more stable in the regression of the bounding box of students' nonstandard, flexible and irregular postures. When all these parts are combined together, the model achieves the best results (mAP@0.5 = 77.80%, mAP@0.5:0.95 = 37.09%), and at the same time, the number of parameters and calculation burden are reduced, which shows that the backbone network, attention mechanism, and loss function are well coordinated, and they jointly create a robust system, which can analyze college physical education activities in real time. The statistical analysis confirms that all performance improvements are statistically significant ($p < 0.05$), and the synergistic effect of combining all three components is greater than the sum of their individual contributions.

Table 2. Ablation experiment results (mean $\pm$ standard deviation, $n = 5$).

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	P (%)	R (%)	Param	gflops
YOLOv11(Baseline)	72.19 $\pm$ 0.23	30.47 $\pm$ 0.18	76.65 $\pm$ 0.21	72.75 $\pm$ 0.25	2,853,411	7.45
Baseline + PPLCNet	75.91 $\pm$ 0.19 *	35.52 $\pm$ 0.22 *	80.33 $\pm$ 0.17 *	75.27 $\pm$ 0.20 *	2,591,339	6.29
Baseline + Modified MSAM	73.47 $\pm$ 0.21 *	32.64 $\pm$ 0.19 *	78.63 $\pm$ 0.20 *	72.00 $\pm$ 0.23	2,884,096	7.55
Baseline + Improved EIou	73.14 $\pm$ 0.20 *	32.69 $\pm$ 0.17 *	79.53 $\pm$ 0.18 *	72.15 $\pm$ 0.21	2,853,411	7.45
Proposed Model	77.80 $\pm$ 0.15 *†	37.09 $\pm$ 0.16 *†	82.34 $\pm$ 0.14 *†	75.00 $\pm$ 0.18 *	2,622,024	6.38

* $p < 0.05$ compared to baseline; † $p < 0.05$ compared to all individual module variants.

4.5. Comparative Analysis

4.5.1. Comparison with Lightweight YOLO Variants

To see if it works, we compare this new method with some powerful lightweight YOLO models, such as YOLOv5, YOLOv6, YOLOv8n, YOLOv9t, YOLOv10n, and YOLOv11. We use a set of data sets of college athletes' sports postures. The results of the comparison are all written in Table 3. It is very important to compare with those already famous models like this to show the advantages of our new method in estimating the human posture of college students' ordinary physical education class.

Table 3. Performance comparison with lightweight YOLO variants (mean $\pm$ standard deviation, $n = 5$).

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	P (%)	R (%)	Params	Gflops	FPS (RTX 4090)	FPS (i7-12700)	FPS (Jetson Nano)
YOLOv11	72.19 $\pm$ 0.23	30.47 $\pm$ 0.18	76.65 $\pm$ 0.21	72.75 $\pm$ 0.25	2,853,411	7.45	112 $\pm$ 3	32 $\pm$ 2	13 $\pm$ 1
YOLO10n	68.66 $\pm$ 0.25	29.32 $\pm$ 0.20	77.05 $\pm$ 0.23	68.91 $\pm$ 0.27	2,539,955	7.68	135 $\pm$ 4	35 $\pm$ 2	14 $\pm$ 1
YOLOv5	69.34 $\pm$ 0.22	29.81 $\pm$ 0.19	75.56 $\pm$ 0.20	69.87 $\pm$ 0.24	2,451,395	6.94	108 $\pm$ 3	29 $\pm$ 2	12 $\pm$ 1
YOLOv6	68.40 $\pm$ 0.24	27.42 $\pm$ 0.21	78.93 $\pm$ 0.22	67.47 $\pm$ 0.26	4,326,451	12.27	95 $\pm$ 2	24 $\pm$ 1	9 $\pm$ 1
YOLOv8n	70.87 $\pm$ 0.21	30.33 $\pm$ 0.18	77.33 $\pm$ 0.21	70.59 $\pm$ 0.23	2,953,779	7.96	118 $\pm$ 3	33 $\pm$ 2	13 $\pm$ 1
YOLOv9t	69.27 $\pm$ 0.23	30.14 $\pm$ 0.19	76.59 $\pm$ 0.20	70.59 $\pm$ 0.24	1,959,379	7.62	122 $\pm$ 3	34 $\pm$ 2	14 $\pm$ 1
ours	77.80 $\pm$ 0.15 *	37.09 $\pm$ 0.16 *	82.34 $\pm$ 0.14 *	75.00 $\pm$ 0.18 *	2,622,024	6.38	127 $\pm$ 3	38 $\pm$ 2	16 $\pm$ 1

* $p < 0.05$ compared to all baseline models.

4.5.2. Comparison with Specialized Pose Estimation Models

To further demonstrate the competitiveness of the proposed method, we compared it with five state-of-the-art specialized pose-estimation frameworks on the college student sports pose dataset: YOLOv8-pose-n, RTMPose-t, BlazePose, MoveNet-Lightning, and ViTPose-Tiny. The results are shown in Table 4.

Table 4. Performance comparison with specialized pose-estimation models (mean $\pm$ standard deviation, $n = 5$).

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	P (%)	R (%)	Params	Gflops	FPS (RTX 4090)	FPS (i7-12700)	FPS (Jetson Nano)
YOLOv8-pose-n	74.21 $\pm$ 0.20	33.56 $\pm$ 0.19	79.12 $\pm$ 0.18	73.45 $\pm$ 0.22	3,125,643	8.12	118 $\pm$ 3	33 $\pm$ 2	13 $\pm$ 1
RTMPose-t	75.68 $\pm$ 0.19	34.89 $\pm$ 0.17	80.35 $\pm$ 0.17	74.12 $\pm$ 0.20	2,897,531	7.85	125 $\pm$ 3	35 $\pm$ 2	14 $\pm$ 1
BlazePose	69.35 $\pm$ 0.23	28.76 $\pm$ 0.21	76.89 $\pm$ 0.22	69.56 $\pm$ 0.25	1,345,217	3.21	185 $\pm$ 5	41 $\pm$ 2	18 $\pm$ 1
MoveNet-Lightning	67.89 $\pm$ 0.24	27.54 $\pm$ 0.22	75.67 $\pm$ 0.23	68.23 $\pm$ 0.26	789,456	1.89	220 $\pm$ 6	45 $\pm$ 3	20 $\pm$ 1
ViTPose-Tiny	73.56 $\pm$ 0.21	32.87 $\pm$ 0.18	78.45 $\pm$ 0.19	72.89 $\pm$ 0.23	5,678,923	12.56	85 $\pm$ 2	22 $\pm$ 1	8 $\pm$ 1
ours	77.80 $\pm$ 0.15 *	37.09 $\pm$ 0.16 *	82.34 $\pm$ 0.14 *	75.00 $\pm$ 0.18 *	2,622,024	6.38	127 $\pm$ 3	38 $\pm$ 2	16 $\pm$ 1

* $p < 0.05$ compared to all baseline models.

The results in Table 4 show that the proposed method outperforms all specialized pose-estimation models on the college student sports pose dataset. Compared to RTMPose-t, the second-best performing model, our method improves mAP@0.5 by 2.12 percentage points and mAP@0.5:0.95 by 2.20 percentage points, while using fewer parameters and lower computational complexity. Although BlazePose and MoveNet-Lightning achieve higher inference speeds, their detection accuracy is significantly lower than our method, making them unsuitable for applications requiring high-precision pose analysis. These results demonstrate that the proposed method achieves a superior balance between accuracy and efficiency compared to state-of-the-art specialized pose-estimation frameworks for college-specific sports applications.

4.5.3. Generalization Performance on Public Benchmarks

To evaluate the generalization capability of the proposed method, we conducted experiments on the COCO Keypoints 2017 and MPII validation sets. The results are shown in Table 5.

Table 5. Generalization performance on public benchmark datasets (mean $\pm$ standard deviation, $n = 5$).

Dataset	Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)
COCO Keypoints 2017	YOLOv11	65.32 $\pm$ 0.21	42.15 $\pm$ 0.19
	ours	68.76 $\pm$ 0.18 *	45.32 $\pm$ 0.17 *
MPII	YOLOv11	82.45 $\pm$ 0.19	68.76 $\pm$ 0.18
	ours	85.23 $\pm$ 0.16 *	71.45 $\pm$ 0.16 *

* $p < 0.05$ compared to YOLOv11.

The results in Table 5 show that the proposed method also outperforms the original YOLOv11 on both public benchmark datasets, improving mAP@0.5 by 3.44 percentage points on COCO Keypoints 2017 and 2.78 percentage points on MPII. This demonstrates that the improvements made in our method are not only effective for college-specific sports poses but also generalize well to general human pose-estimation scenarios.

From Tables 3–5, we can see the following:

The test results show that our method outperforms all existing baseline models on the College Student Sports Pose dataset, achieving peak performance with an mAP@0.5 of 77.80% and an mAP@0.5:0.95 of 37.09%. This corresponds to an improvement of 4.93 to 12.6 percentage points in mAP@0.5, highlighting the model's superior ability to detect complex and non-standard body postures of college athletes. In terms of computational requirements, our proposed model has only 2.62 million parameters and 6.38 GFLOPs, achieving a more favorable accuracy–efficiency trade-off than competing frameworks. Specifically, it is computationally lighter than YOLOv6, YOLOv8n, and YOLOv11, while delivering higher accuracy than comparably sized models such as YOLOv5, YOLOv9t, and YOLOv10n for college pose estimation.

In terms of inference speed, our model achieves 127 FPS on an RTX 4090 GPU, 38 FPS on an Intel i7-12700 CPU, and 16 FPS on a Jetson Nano edge device. This meets the real-time requirements of different campus deployment scenarios: GPU deployment for campus monitoring centers (≥ 30 FPS), CPU deployment for classroom teaching terminals (≥ 25 FPS), and edge device deployment for portable sports monitoring equipment (≥ 15 FPS). Compared with other models, our method achieves the best balance between accuracy and speed across all three platforms.

This optimal combination enables deployment on low-cost, resource-constrained campus edge devices such as ordinary motion cameras and portable teaching systems, meeting the speed and hardware requirements for the real-time analysis of university physical education classes.

4.6. Visualization Analysis

In order to test the performance of our PPLCNet-YOLOv11 model in posture detection of educational sports, we have made many visual evaluations in various campus sports training environments. These evaluations include indoor single-person standard exercises, multi-person activities with complex outdoor backgrounds, sheltered yoga postures, and crowded group exercises common in college physical education courses.

4.6.1. Detection Performance Based on Single-Person Standard Sports Movements

Figure 6 shows a comparative analysis of the detection accuracy of indoor single standard sports movements, which is the main concern of quantitative physical training in universities. Under the consistent input conditions (these conditions come from the images of college athletes), the figure evaluates several popular lightweight YOLO versions (YOLOv5, YOLOv8n, YOLOv10n), the original YOLOv11 (benchmark model), and our improved model configuration (benchmark model using PPLCNet backbone, benchmark model using improved MSAM attention mechanism, and our complete PPLCNet-YOLOv11 framework).

The test set contains standard practice postures, such as plank, lunge stretch, side split, high leg lift, and seated forward bend, which are common non-upright and vertical postures in college physical education. The visualization results clearly show the performance difference between the standard YOLO implementation and our optimization framework, which will be detailed below.

The original YOLOv11 baseline model (Figure 6a) shows obvious limitations in these non-upright, longitudinally stretched college student sports poses: the bounding box cannot accurately fit the elongated human body structure with significant offset in the aspect ratio, and the model exhibits low detection confidence for limb extremities (hands, feet) and small-scale joint areas, even with partial missed detection in extreme stretching poses. Mainstream lightweight YOLO variants, including YOLOv5 (Figure 6b), YOLOv8n (Figure 6c), and YOLOv10n (Figure 6d), present similar defects: they have insufficient ability to capture the contour of irregular college student sports poses, with poor alignment between the bounding box and the real human structure, which cannot meet the precision requirements of pose detection for quantitative college general education sports training analysis.



Figure 6. Cont.



Figure 6. Visual comparison of detection performance between mainstream lightweight YOLO variants and the proposed method on single-person standard sports training movements.

The visualization results also verify the effectiveness of each proposed improved module. After replacing the backbone with PPLCNet (Figure 6e), the detection performance is significantly improved: the lightweight depthwise separable convolution structure enhances the feature-extraction efficiency of college student human contours, the positioning accuracy of the bounding box is significantly improved, and the model can completely capture the overall outline of the human body in large-scale stretching movements without missed detection. The introduction of the Modified MSAM attention mechanism (Figure 6f) further enhances the model's fine-grained feature-capture ability for college student training poses, with a tighter fit between the bounding box and the human contour and the more accurate detection of small-scale limb ends and joint areas. The proposed full PPLCNet-YOLOv11 method (Figure 6g) achieves optimal detection performance in all test scenes. Benefiting from the synergistic optimization of the PPLCNet backbone, Modified MSAM attention mechanism, and improved EIoU loss function, the detection bounding box completely fits the irregular college student sports poses, and the detection confidence of all poses reaches the highest level among all compared models, which fully verifies that the proposed framework effectively solves the core challenges of multi-scale pose detection, longitudinal feature capture, and irregular pose regression in college general education sports training scenarios.

4.6.2. Generalization Performance in Complex Real-World Scenes

To further verify the generalization ability and robustness of the proposed method in real-world campus general education sports training applications, we tested the model on three typical complex scenes that are closer to actual on-site college training environments: occluded yoga movements, dense multi-person rowing scenes, and outdoor multi-person competitive sports with complex campus backgrounds, with the visualization results shown in Figure 7. The original YOLOv11 baseline model and mainstream lightweight YOLO variants (YOLOv5, YOLOv8n, YOLOv10n) exhibit obvious performance degradation in these complex campus scenes: they suffer from severe missed detection and low detection confidence for occluded human bodies in yoga scenes, severe bounding box offset, overlapping and missed detection of small-scale targets in dense rowing scenes (common in college group classes), and obvious false detection and positioning inaccuracy caused by campus background interference in outdoor competitive sports scenes, which cannot meet the stability requirements of actual college general education sports training analysis.

The visualization results also validate the effectiveness of our proposed optimized modules for campus scenarios. The YOLOv11 model with the PPLCNet backbone achieves significantly improved anti-interference ability against campus background clutter, with reduced missed detection in occluded and dense scenes and more stable detection confidence. The introduction of the Modified MSAM attention mechanism further enhances the model's ability to capture multi-scale student targets in complex campus scenes, with more accurate bounding box positioning and better adaptation to dense and occluded targets. The proposed full PPLCNet-YOLOv11 method achieves optimal detection performance across all test scenes: it accurately detects all student targets without false detection or missed detection, even for occluded limbs and small-scale dense targets in group classes, the bounding box fits the human body structure accurately, and the detection confidence remains at a high level. This fully verifies the excellent generalization ability and robustness of the proposed framework in complex real-world campus general education sports scenes, which is consistent with the quantitative experimental results.

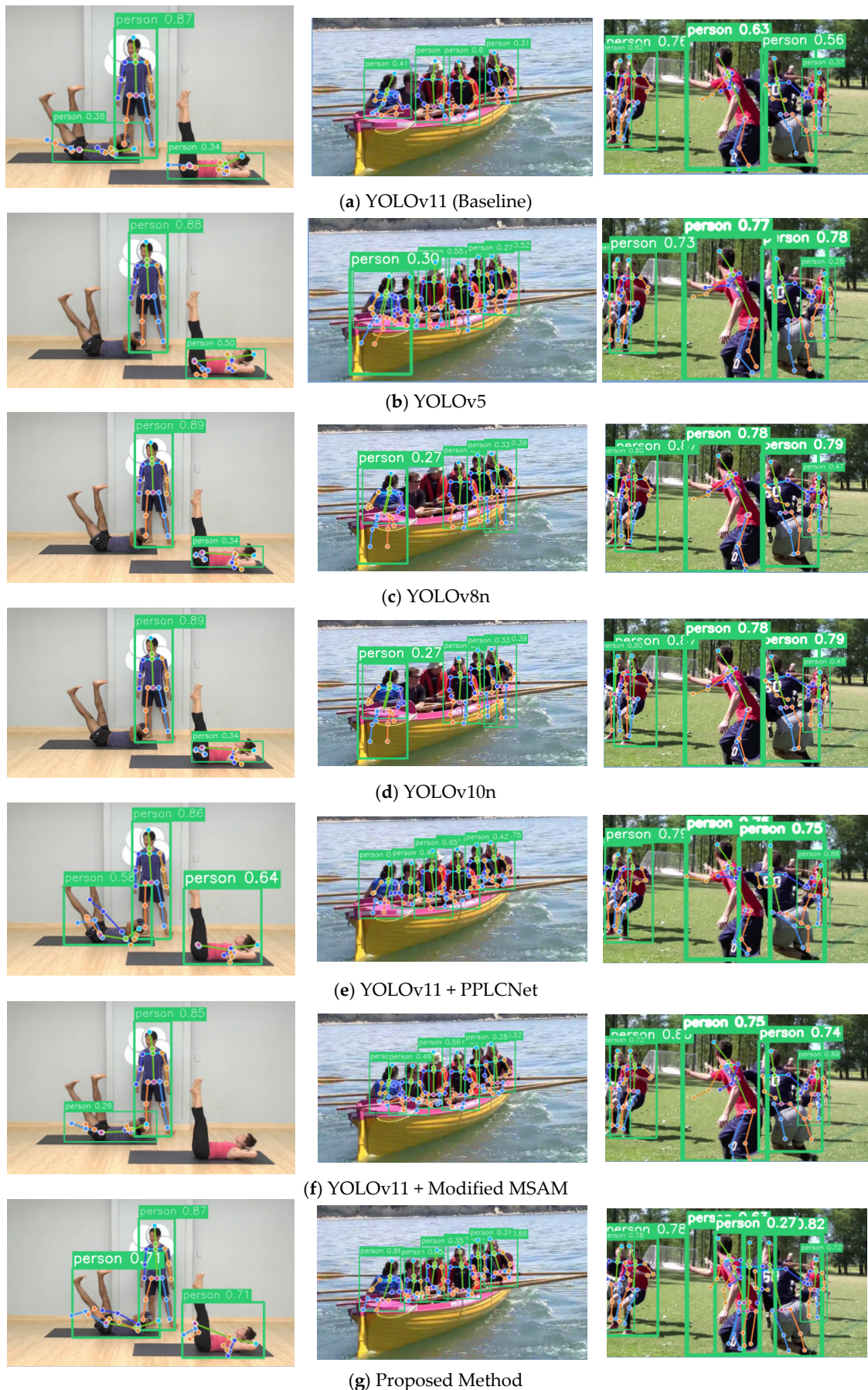


Figure 7. Visual comparison of detection performance between mainstream lightweight YOLO variants and the proposed method in complex real-world sports scenes.

4.6.3. Consistency with Quantitative Experimental Results

The visualization analysis results are completely consistent with the quantitative results of the ablation study and comparative experiment in the previous sections on the college student sports pose dataset. The performance improvement of each module in the quantitative indicators is intuitively reflected in the detection integrity, positioning accuracy, and confidence stability of the visualization results for college student training poses. The proposed PPLCNet-YOLOv11 framework achieves the optimal detection performance in all test scenes, which fully verifies that the joint optimization of the lightweight PPLCNet backbone, college general education sports-oriented MSAM attention module, and non-standard pose-aware improved EIoU loss function can effectively solve the core challenges of human pose detection in college general education sports training scenarios. The visualization results further prove that the proposed method has strong practical application value in higher education and can provide stable, accurate, real-time pose detection support for actual on-site college general education sports training analysis on low-cost campus resource-constrained edge devices.

5. Discussion

The experimental results demonstrate that our proposed PPLCNet-YOLOv11 framework effectively addresses the three key challenges of college general education sports pose detection: multi-scale variation, longitudinal feature capture, and non-standard pose regression. Compared with existing methods, our approach has three distinct advantages: first, the PPLCNet backbone provides an optimal balance between accuracy and computational efficiency. Second, the modified MSAM attention mechanism significantly enhances the model's ability to capture sports-specific features. Third, the improved EIoU loss function with an aspect ratio penalty and dynamic scaling effectively handles the irregular and non-standard poses of college students with varying skill levels.

Despite these advantages, our method still has several limitations that need to be critically discussed. First, the current model only performs 2D pose detection, which cannot provide accurate 3D kinematic parameters, such as joint angles and limb rotations. This limits its ability to conduct an in-depth quantitative analysis of movement quality and biomechanical assessments. Second, the dedicated campus dataset mainly covers common sports offered in general education courses, and the model's performance on niche sports such as martial arts, diving, and acrobatics needs further verification. Third, while we have demonstrated the model's performance on standard edge devices, we have not yet conducted a large-scale field test in actual college physical education classes to evaluate its long-term stability, user acceptance, and effectiveness in improving teaching outcomes. Fourth, the current method does not explicitly incorporate temporal information from video sequences, which could be used to further improve detection accuracy and handle motion blur more effectively.

In terms of practical applications, our model can be integrated into campus sports monitoring systems to provide real-time feedback to both teachers and students. For teachers, it can automatically generate quantitative evaluation reports for each student, helping them identify students with non-standard movements and provide targeted guidance. For students, it can provide immediate feedback on their movements during self-training, helping them improve their skills and reduce the risk of sports injuries. The low computational cost of the model also makes it possible to deploy it on mobile devices, enabling students to conduct self-training anytime and anywhere on campus.

6. Conclusions

This study proposes a lightweight YOLO-based human pose-detection method tailored for college general education sports training, integrating the PPLCNet backbone, college student sports-oriented modified MSAM attention, and non-standard pose-aware improved EIou loss. Experimental results on the college student sports pose dataset show that the method achieves 77.8% mAP@0.5 and 37.09% mAP@0.95 with only 2.62M parameters and 6.38 GFLOPs, outperforming mainstream lightweight YOLO models and specialized pose-detection models in both accuracy and computational efficiency for campus deployment. Inference speed tests on three different platforms show that the model meets the real-time requirements of various campus deployment scenarios. Additional experiments on the COCO Keypoints 2017 and MPII public benchmark datasets demonstrate the model's strong generalization ability to general human pose-estimation scenarios. The proposed method addresses key challenges in college general education sports pose detection (multi-scale variation in campus scenes, longitudinal feature capture of training movements, irregular and non-standard pose regression of students with different foundations) and provides a real-time, accurate and low-cost technical solution for quantitative sports training analysis in colleges and universities. Its practical application potential supports the intelligent transformation of college general education physical education courses, enabling teachers to conduct objective quantitative teaching evaluations, provide personalized guidance for students with different sports foundations, and effectively reduce the risk of sports injuries in campus training.

For future research, we will focus on the following concrete directions: (1) expand the college student sports pose dataset to cover more niche sports and complex campus scenarios, including different weather conditions, lighting variations, and camera angles, to further improve the model's generalization capability; (2) extend the framework to monocular/multi-view 3D human pose estimation to enable a fine-grained kinematic analysis of students' movements, including joint angles, limb velocities, and body segment coordination; (3) conduct comprehensive model compression, quantization, and edge deployment optimization for low-power edge chips commonly used in campus sports monitoring equipment, and develop a complete end-to-end deployment pipeline; (4) incorporate temporal information from video sequences using recurrent neural networks or transformers to improve detection accuracy and robustness to motion blur; (5) conduct a large-scale field test in actual college physical education classes involving over 1000 students to evaluate the model's long-term stability, user acceptance, and effectiveness in improving teaching quality and reducing sports injuries.

Author Contributions: Conceptualization, J.C.; methodology, J.C.; software, W.H.; validation, W.H. and J.C.; formal analysis, J.C.; investigation, Z.W.; resources, Z.W.; data curation, Z.W.; writing—original draft preparation, J.C.; writing—review and editing, J.C.; visualization, W.H.; supervision, W.H.; project administration, Z.W.; funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Key Scientific Research Project of Anhui Provincial Department of Education (Grant No. 2023AH051793) and the Key Research Project of Anhui Provincial Scientific Research Programming Plan (Grant No. 2022AH051847).

Data Availability Statement: The datasets are available at <https://github.com/hwqwlsu/Body-Detection> (accessed on 13 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation: A Survey. *ACM Comput. Surv.* **2024**, *56*, 1–37. [\[CrossRef\]](#)
2. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 7291–7299. [\[CrossRef\]](#)
3. 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, Venice, Italy, 22–29 October 2017*; IEEE: New York, NY, USA, 2017; pp. 2334–2343. [\[CrossRef\]](#)
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 779–788. [\[CrossRef\]](#)
5. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [\[CrossRef\]](#)
6. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696. [\[CrossRef\]](#)
7. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020*; AAAI Press: Palo Alto, CA, USA, 2020; pp. 12993–13000. [\[CrossRef\]](#)
8. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *arXiv* **2017**, arXiv:1707.01083. [\[CrossRef\]](#)
9. Cui, C.; Gao, T.; Wei, S.; Du, Y.; Guo, R.; Dong, S.; Lu, B.; Zhou, Y.; Lv, X.; Liu, Q. PP-LCNet: A Lightweight CPU Convolutional Neural Network. *arXiv* **2021**, arXiv:2109.15099. [\[CrossRef\]](#)
10. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.10821. [\[CrossRef\]](#)
11. Zhang, Y.F.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* **2022**, *506*, 146–157. [\[CrossRef\]](#)
12. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In *Computer Vision—ECCV 2018, Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 472–487. [\[CrossRef\]](#)
13. 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, Long Beach, CA, USA, 15–20 June 2019*; IEEE: New York, NY, USA, 2020; pp. 5693–5703. [\[CrossRef\]](#)
14. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 5386–5395. [\[CrossRef\]](#)
15. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using ObjectKey-point Similarity Loss. *arXiv* **2022**, arXiv:2204.06806. [\[CrossRef\]](#)
16. Zhang, Y.; Wang, Z.; Li, M.; Gao, P. SP-YOLO: An End-to-End Lightweight Network for Real-Time Human Pose Estimation. *Signal Image Video Process.* **2023**, *18*, 1–14. [\[CrossRef\]](#)
17. Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; Chen, K. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv* **2023**, arXiv:2303.07399.
18. Bazarevsky, V.; Grishchenko, I.; Raveendran, K.; Zhu, T.; Grundmann, M. BlazePose: On-device Real-time Body Pose Tracking. *arXiv* **2020**, arXiv:2006.10204.
19. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose: Simple vision transformer baselines for human pose estimation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022*; Curran Associates Inc.: Red Hook, NY, USA, 2022; pp. 20077–20090.
20. Pfister, T.; Charles, J.; Zisserman, A. Flowing ConvNets for Human Pose Estimation in Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 7–13 December 2015*; IEEE: New York, NY, USA, 2015; pp. 3673–3682.
21. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-Person Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 7103–7112.
22. Kocabas, M.; Athanasiou, N.; Black, M.J. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 5252–5262.

23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; IEEE: New York, NY, USA, 2018; pp. 7132–7141. [[CrossRef](#)]
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Computer Vision—ECCV2018, Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 3–19. [[CrossRef](#)]
25. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. *arXiv* **2021**, arXiv:2103.02907. [[CrossRef](#)]
26. Yang, Y.B. SA-Net: Shuffle attention for deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, ON, Canada, 6–11 June 2021*; IEEE: New York, NY, USA, 2021; pp. 1–5. [[CrossRef](#)]
27. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. *PMLR* **2021**, 139, 11863–11874.
28. Rezatofighi, H.; Tsoi, N.; Gwak, J.Y.; Sadeghian, A.; Savarese, S. Generalized intersection over union: A metric and a loss for boundingbox regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019*; IEEE: New York, NY, USA, 2020; pp. 658–666. [[CrossRef](#)]
29. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2024**, arXiv:2301.10051. [[CrossRef](#)]
30. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [[CrossRef](#)]
31. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946. [[CrossRef](#)]
32. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R. Searching for MobileNetV3. In *Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: New York, NY, USA, 2020; pp. 1314–1324. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.