

FC-YOLO: 一种课堂行为检测的改进算法

高博文, 李焕哲*

河北地质大学信息工程学院, 河北 石家庄

收稿日期: 2026年9月7日; 录用日期: 2026年10月7日; 发布日期: 2026年10月15日

摘要

针对课堂场景中目标密集、行为差异细微、手机和书本等局部目标尺度较小以及复杂背景干扰等问题, 提出一种融合前景感知与坐标先验的课堂学生行为检测算法。该方法以YOLO11n为基础模型, 在特征融合阶段设计Foreground-Aware Scale Routing Concat模块, 简称FASR-Concat, 通过前景感知和尺度路由机制增强不同尺度特征的自适应融合能力; 在检测头部分设计Coordinate Prior Detection Head模块, 简称Coord-Detect, 通过注入归一化二维坐标信息增强模型对固定视角课堂图像中目标空间位置的感知能力。实验在公开POCO课堂行为数据集上进行, 结果表明, YOLO11n基线模型在验证集上的mAP50为0.812; 加入FASR-Concat后, mAP50提升至0.835; 采用FASR-Concat与Coord-Detect的完整改进模型后, mAP50达到0.850。与基线模型相比, 最终模型mAP50提升3.8个百分点, mAP50:95提升2.6个百分点。类别结果显示, 该方法对play_phone、look_book和turn_head等依赖局部区域或姿态细节的类别具有较明显提升。实验结果验证了所提方法在课堂学生行为检测任务中的有效性。

关键词

课堂行为检测, YOLO11, 目标检测, POCO, 特征融合

FC-YOLO: An Improved Algorithm for Classroom Behavior Detection

Bowen Gao, Huanzhe Li*

School of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: September 7, 2026; accepted: October 7, 2026; published: October 15, 2026

Abstract

To address the problems of dense student targets, subtle behavioral differences, small local objects such as mobile phones and books, and complex background interference in classroom scenes, this

*通讯作者。

文章引用: 高博文, 李焕哲. FC-YOLO: 一种课堂行为检测的改进算法[J]. 计算机科学与应用, 2026, 16(10): 24-36.
DOI: 10.12677/csa.2026.1610307

paper proposes a classroom student behavior detection algorithm integrating foreground awareness and coordinate priors. Based on YOLO11n, the proposed method designs a Foreground-Aware Scale Routing Concat module, abbreviated as FASR-Concat, in the feature fusion stage to enhance the adaptive fusion capability of multi-scale features through foreground awareness and scale routing. In the detection head, a Coordinate Prior Detection Head module, abbreviated as Coord-Detect, is introduced to inject normalized two-dimensional coordinate information, thereby improving the model's perception of target spatial positions in fixed-view classroom images. Experiments were conducted on the public POCO classroom behavior dataset. The results show that the YOLO11n baseline achieves an mAP50 of 0.812 on the validation set. After introducing FASR-Concat, the mAP50 increases to 0.835. With the complete improved model using both FASR-Concat and Coord-Detect, the mAP50 reaches 0.850. Compared with the baseline model, the final model improves mAP50 by 3.8 percentage points and mAP50:95 by 2.6 percentage points. Class-wise results show that the proposed method achieves clear improvements for categories relying on local regions or pose details, such as play_phone, look_book, and turn_head. The experimental results verify the effectiveness of the proposed method for classroom student behavior detection.

Keywords

Classroom Behavior Detection, YOLO11, Object Detection, POCO, Feature Fusion

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着智慧教育和课堂教学分析技术的发展, 学生课堂行为识别逐渐成为教育信息化和计算机视觉交叉领域的重要研究内容[1]。传统课堂考勤方式主要关注学生是否到课, 难以反映学生在课堂中的真实听课状态[2]。通过计算机视觉方法自动识别学生在课堂中的抬头、低头、转头、看书、玩手机、睡觉等行为, 可以为后续课堂专注度分析、学生状态分析和教学管理系统提供更细粒度的数据支撑[3]。

课堂学生行为检测通常可视为目标检测任务, 即在课堂图像中同时预测学生行为类别及其边界框位置。与一般目标检测场景相比, 课堂图像具有较强的场景特殊性[4]。一方面, 学生目标数量较多, 存在遮挡、姿态相似和背景复杂等问题; 另一方面, 手机、书本等相关目标尺度较小, 玩手机、看书、低头和弯腰等行为之间的局部差异较为细微, 导致模型容易出现漏检和误检[5]。此外, 固定摄像头或相对稳定拍摄视角使课堂图像中的座位、桌面和人体位置具有一定空间规律, 但普通卷积检测头对显式坐标信息的利用仍然有限。

近年来, YOLO 系列目标检测算法因检测速度快、结构简洁、部署便利, 在学生课堂行为识别等任务中得到广泛应用[6]。已有研究通常通过引入注意力机制、多尺度特征融合或损失函数改进等方式提升检测性能[7]。然而, 普通特征拼接操作往往直接在通道维度融合不同尺度特征, 缺少对前景区域和不同尺度贡献的自适应建模; 同时, 常规检测头主要依赖卷积特征进行分类和定位, 对固定视角下的空间坐标先验利用不足。这些问题限制了模型在课堂局部行为目标和细粒度姿态类别上的检测效果[8]。

为解决课堂行为检测中局部小目标难检、相似姿态类别区分困难以及固定视角空间信息利用不足等问题, 本文在 YOLO11n 基线模型的基础上, 提出一种融合前景感知与坐标先验的课堂学生行为检测算法[9]。本文的主要贡献如下:

- (1) 设计前景感知尺度路由特征融合模块 FASR-Concat。该模块在普通特征拼接基础上引入前景权重

图和尺度路由机制, 使模型能够根据课堂图像中的有效行为区域自适应调整不同尺度特征的贡献, 从而增强局部行为目标的特征表达能力。

(2) 提出坐标先验检测头 **Coord-Detect**。该模块在检测头前注入归一化二维坐标信息, 使检测分支在分类和定位前获得显式空间位置提示, 从而提高模型对固定视角课堂图像中目标位置分布规律的感知能力。

(3) 在公开 POCO 课堂行为数据集上进行实验验证。实验结果表明, YOLO11n 基线模型的 mAP50 为 0.812, 加入 FASR-Concat 后提升至 0.835, 采用 FASR-Concat 与 Coord-Detect 的完整改进模型后达到 0.850。最终模型参数量为 3.11 M, 计算量为 13.0 GFLOPs, 较基线模型在检测精度上取得了较大提升。

2. 相关理论研究

2.1. 课堂行为识别研究

课堂行为识别旨在通过图像或视频分析方法判断学生在课堂中的行为状态。早期方法多依赖人工观察、问卷调查或课堂互动分析模型, 存在效率低、主观性强和难以实时反馈等不足[10]。随着深度学习的发展, 研究者开始利用卷积神经网络、人体骨骼关键点、视频动作识别和目标检测方法进行课堂行为识别[11]。其中, 基于目标检测的方法可以同时输出行为类别和目标位置, 适合处理多学生共存的课堂图像场景[12]。

现有课堂行为识别研究通常围绕姿态识别、注意力评估和学生参与度分析展开。部分方法基于人体姿态或骨架序列建模学生动作, 能够较好捕捉动作变化, 但在学生密集、遮挡严重或分辨率较低的课堂图像中, 关键点检测稳定性容易受到影响[13]。另一些方法基于图像目标检测模型直接识别学生行为, 对数据标注和类别定义具有较强适应性, 但仍需解决小目标、遮挡和相似行为区分问题[14]。

2.2. YOLO 系列目标检测算法

YOLO 系列算法属于单阶段目标检测方法, 能够在一次前向传播中完成目标类别预测和边界框回归, 具有较好的实时性和工程应用价值。从 YOLOv3、YOLOv5 到 YOLOv8、YOLO11, YOLO 系列模型在网络结构、特征融合、检测头和训练策略等方面持续改进。由于课堂行为检测通常需要在多目标场景下兼顾准确率和推理速度, YOLO 系列模型成为该任务中的常用基线。

不过, 课堂场景中的目标尺度差异和行为类别差异对 YOLO 模型提出了更高要求[15]。普通 YOLO 特征融合结构虽然能够整合不同层级的语义信息与空间细节, 但在融合过程中通常缺少对不同尺度特征重要性的细粒度区分[16]。对于手机、书本等小目标以及低头、转头等姿态相关类别, 仅依赖常规多尺度特征融合可能难以充分表达局部关键区域[17]。

2.3. 多尺度特征融合与注意力机制

多尺度特征融合是目标检测任务中提升小目标和多尺度目标检测性能的重要手段。特征金字塔网络、路径聚合网络以及多尺度注意力结构通过整合浅层细节和深层语义信息, 提高模型对不同尺寸目标的适应能力[18]。在课堂行为检测中, 前排和后排学生目标尺寸差异明显, 手机、书本等目标又具有较小尺度, 因此多尺度特征融合对检测结果具有重要影响。

注意力机制能够帮助模型突出关键区域或重要通道, 抑制无关背景干扰[19]。已有研究将通道注意力、空间注意力、坐标注意力等结构用于课堂行为识别或目标检测任务, 取得了一定效果。然而, 普通注意力模块往往关注单一输入特征内部的信息增强, 而课堂行为检测还需要在不同尺度特征之间动态分配贡献[20]。本文提出的 FASR-Concat 重点针对特征拼接阶段进行改进, 在融合过程中同时考虑前景区域和

尺度路由权重。

2.4. 坐标先验与检测头改进

检测头负责将融合后的特征映射为类别概率和边界框位置。常规检测头主要通过卷积特征进行分类和定位,对显式坐标信息利用较少。在固定摄像头课堂场景中,座位、桌面、人体位置和局部目标分布具有一定空间规律,合理引入坐标先验有助于提高模型对目标位置的感知能力[21]。

现有坐标注意力或位置编码方法表明,显式空间信息能够增强模型对方向、位置和长程关系的表达能力[22]。受此启发,本文在检测头前设计 CoordStem 坐标注入结构,将归一化 x 、 y 坐标图与原始特征拼接,再通过 1×1 卷积融合回原通道数,使检测头在进行分类和回归前获得明确的位置提示[23]。

3. 融合前景感知与坐标先验的课堂学生行为检测算法

3.1. 整体网络结构

针对课堂行为检测场景中局部小目标难检、相似姿态类别区分困难以及固定视角下空间位置信息利用不足等问题,本文在 YOLO11n 的基础上提出一种融合前景感知与坐标先验的课堂学生行为检测算法。改进后的网络总体结构如图 1 所示。模型输入为课堂图像,输出为图像中学生行为或相关目标的类别及边界框位置。

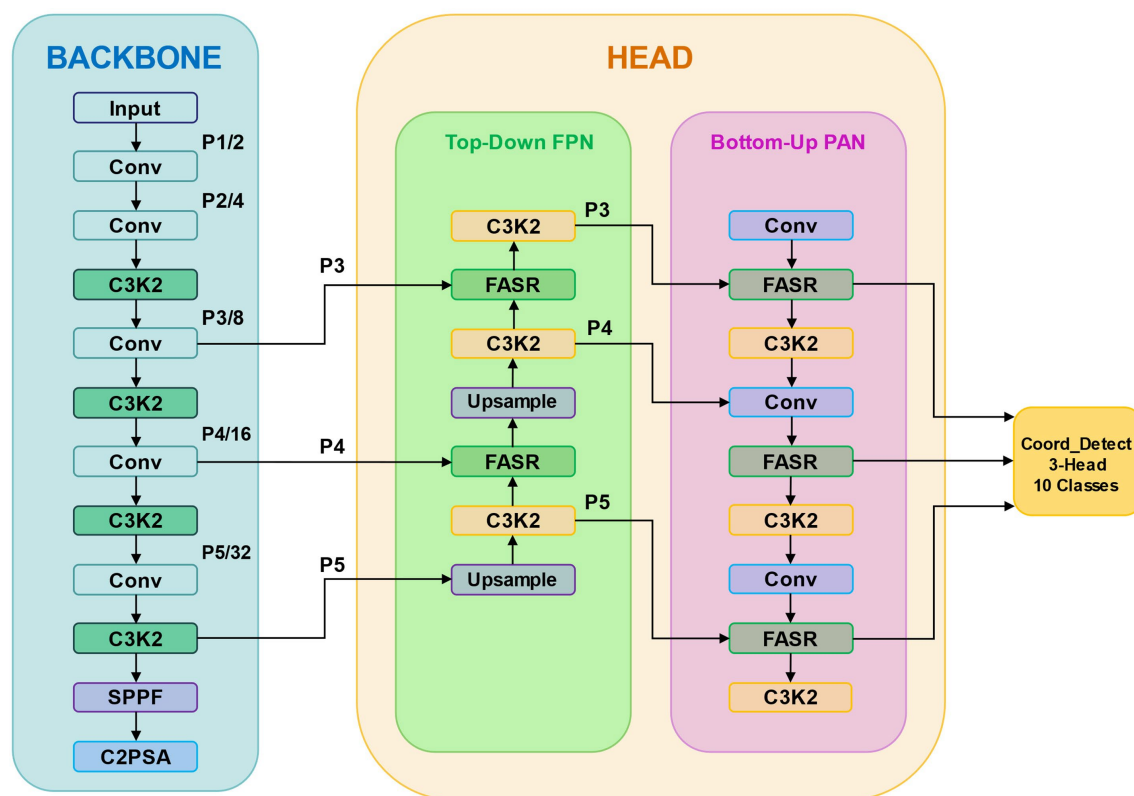


Figure 1. Overall network architecture of the proposed FC-YOLO

图 1. 本文提出的 FC-YOLO 总体网络结构

改进模型主要从特征融合和检测头两个方面进行优化:(1)在特征融合阶段引入前景感知尺度路由特征融合模块 FASR-Concat,替代普通特征拼接操作,使网络能够根据前景区域和不同空间位置自适应

调整多尺度特征贡献；(2) 在检测头部分设计坐标先验检测头 **Coord-Detect**，在分类和边界框回归前注入归一化二维坐标信息，增强模型对固定视角课堂图像中目标空间分布规律的感知能力。通过上述改进，模型能够更好地表达手机、看书、转头等依赖局部区域或姿态细节的行为类别，为后续课堂专注度分析提供行为检测基础。

3.2. FASR-Concat 前景感知尺度路由特征融合模块

在 YOLO 网络的特征融合阶段，普通 Concat 操作通常直接将不同层级或不同尺度的特征在通道维度进行拼接[24]。该方式虽然能够保留多尺度信息，但难以根据空间位置区分前景区域和背景区域，也难以自适应调整不同尺度特征的贡献。对于课堂行为检测任务而言，手机、书本、头部姿态和身体姿态等有效信息往往集中在局部前景区域，而桌椅、墙面和空座位等背景内容容易对检测结果产生干扰。为增强特征融合阶段对有效区域和尺度贡献的建模能力，本文设计前景感知尺度路由特征融合模块 **FASR-Concat**，用于替代普通特征拼接操作，其结构如图 2 所示。

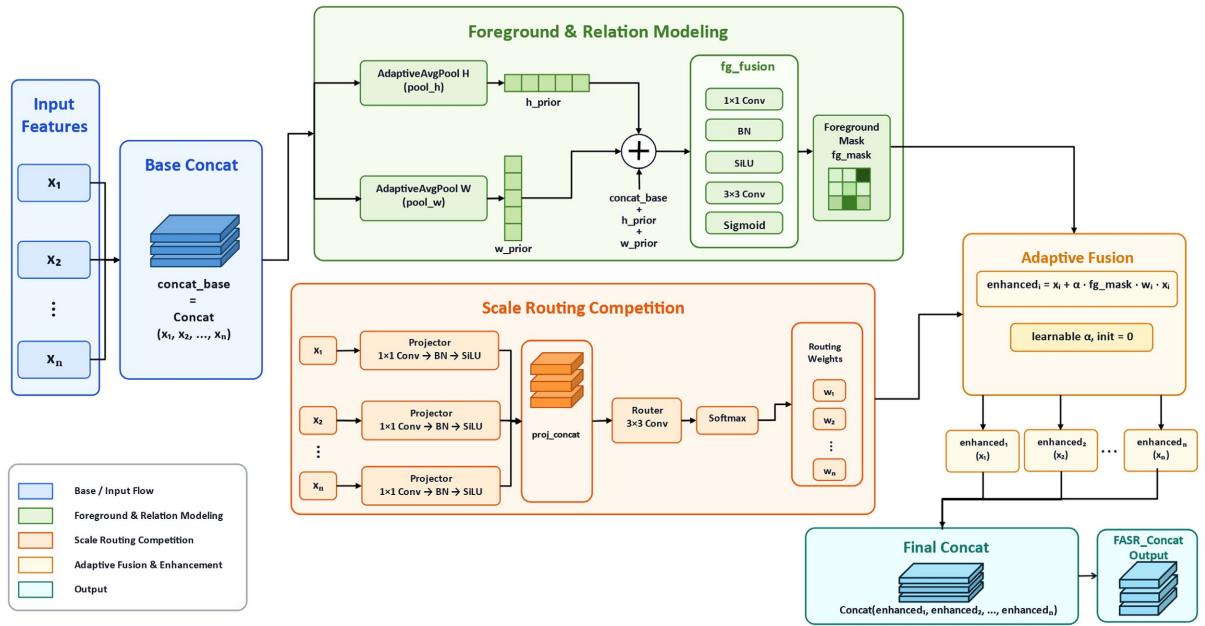


Figure 2. Structure of the FASR-Concat

图 2. FASR-Concat 模块的结构

设参与融合的输入特征集合为 $X = \{x_1, x_2, \dots, x_n\}$ ，其中 n 表示输入特征数量， x_i 表示第 i 个输入特征。经过尺寸对齐后，各输入特征在通道维度进行基础拼接，得到融合前的基础特征：

$$F_c = \text{Concat}(x_1, x_2, \dots, x_n) \quad (1)$$

为引入课堂图像中的空间结构信息，FASR-Concat 分别采用横向和纵向自适应平均池化对 F_c 进行建模。本文采用横向和纵向自适应平均池化，而不是直接采用全局平均池化，主要是考虑到课堂图像通常具有较稳定的空间布局，学生座位、桌面区域以及局部行为目标在水平方向和垂直方向上均具有一定分布规律。横向和纵向池化能够分别聚合不同方向上的上下文信息，在保留位置信息的同时扩大特征感受范围，有助于模型感知课堂场景中的行列结构和前后排尺度差异。与复杂自注意力结构相比，该池化方式计算量较低，能够在较小额外开销下增强前景区域建模能力，更适合轻量化课堂行为检测任务。池化

结果经过扩展后与基础特征相加, 并输入前景融合分支生成前景权重图:

$$M_f = f_{ig}(F_c + F_h + F_w) \quad (2)$$

其中, F_h 和 F_w 分别表示由纵向池化和横向池化得到的空间先验特征, $f_{ig}(\cdot)$ 表示由 1×1 卷积、批归一化、SiLU 激活、 3×3 卷积和 Sigmoid 函数组成的前景权重生成分支。前景权重图 M_f 的取值范围为[0, 1], 用于表示不同空间位置对课堂行为检测的贡献程度。

在尺度路由分支中, 模块首先将不同输入特征通过 1×1 卷积映射到统一隐藏维度, 得到投影特征 z_i 。随后, 将各投影特征拼接并输入路由卷积, 通过 Softmax 生成不同尺度在各空间位置上的竞争权重:

$$S = \text{Softmax}\left(f_r\left(\text{Concat}\left(z_1, z_2, \dots, z_n\right)\right)\right) \quad (3)$$

其中, $f_r(\cdot)$ 表示 3×3 路由卷积, S 表示尺度路由权重。第 i 个尺度对应的权重记为 s_i 。通过该分支, 模型能够根据不同空间位置自适应调整各尺度特征的贡献。在路由权重生成过程中, 本文采用 3×3 卷积作为尺度路由卷积。与 1×1 卷积相比, 3×3 卷积不仅能够建模通道间关系, 还能够引入局部空间邻域信息, 使路由分支在生成尺度权重时同时考虑目标区域及其周围上下文。对于 play_phone、look_book 和 turn_head 等细粒度类别, 其判别信息往往分布在手机、书本、头部姿态及相邻桌面区域等局部范围内, 因此引入局部空间上下文有助于模型更准确地判断不同尺度特征的贡献。相比更大卷积核或自注意力结构, 3×3 卷积在表达能力和计算成本之间具有较好的平衡。

最后, FASR-Concat 将前景权重和尺度路由权重共同作用于原始输入特征, 并采用残差形式进行增强:

$$\hat{x}_i = x_i + \alpha \cdot M_f \odot s_i \odot x_i \quad (4)$$

其中, $\odot$ 表示逐元素乘法, α 可学习参数, 初始化为 0。该设计使模块在训练初期近似等价于普通 Concat, 有助于降低对预训练特征分布的扰动。增强后的各尺度特征再次在通道维度拼接, 得到模块输出:

$$Y = \text{Concat}(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n) \quad (5)$$

通过上述设计, FASR-Concat 在普通特征拼接的基础上引入了前景感知和尺度路由机制。前景权重图用于突出与课堂行为相关的有效区域, 尺度路由权重用于根据不同空间位置动态调整各尺度特征的贡献。二者共同作用, 使融合特征能够更充分地表达手机、看书、转头等依赖局部区域或姿态细节的类别特征, 从而为后续检测头提供更具判别性的特征输入。

3.3. Coord-Detect 坐标先验检测头

在课堂行为检测场景中, 图像通常来自固定或相对稳定的摄像头视角, 学生座位、桌面区域以及手机、书本等局部目标在图像中具有一定空间分布规律。与一般自然场景目标检测任务不同, 课堂监控图像通常具有较强的固定视角特征, 学生座位、讲台、桌面以及人体区域在图像中的相对位置较为稳定。常规卷积操作具有一定平移等变性, 能够提取局部纹理和语义特征, 但对目标在整幅图像中的绝对空间位置表达相对有限。对于课堂行为检测任务而言, 不同行为类别往往与特定空间区域和局部上下文相关, 例如看书、玩手机等行为多发生在桌面附近, 转头、低头等行为则与头部区域的位置变化密切相关。因此, 在检测头前引入归一化二维坐标信息, 可以为分类和定位分支提供显式空间提示, 增强模型对固定视角课堂图像中目标位置分布规律的感知能力。普通检测头主要依赖卷积特征完成分类和边界框回归, 对显式空间坐标信息的利用较少[25]。为增强检测头对目标空间位置的感知能力, 本文设计坐标先验检测头 Coord-Detect, 在检测分支前引入轻量化坐标注入结构 CoordStem, 其结构如图 3 所示。

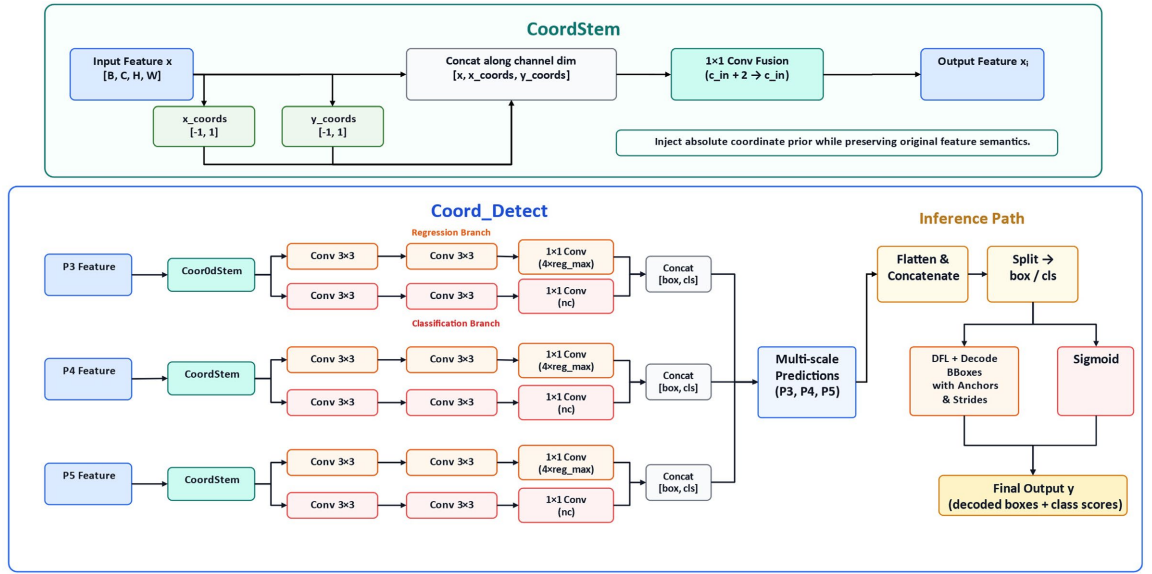


Figure 3. Structure of the coord-detect

图 3. Coord-Detect 模块的结构

设某一检测尺度的输入特征图为 x ，其空间尺寸为 $H \times W$ 。CoordStem 首先根据特征图尺寸生成水平和垂直方向的归一化坐标图。对于特征图中位置 (i, j) ，坐标图可表示为：

$$G_x(i, j) = \frac{2j}{W-1} - 1 \quad (6)$$

$$G_y(i, j) = \frac{2i}{H-1} - 1 \quad (7)$$

其中， G_x 和 G_y 的取值范围均为 $[-1, 1]$ ，分别表示特征图在水平方向和垂直方向上的归一化位置信息。通过该方式，检测头能够获得显式空间位置提示，而不完全依赖卷积特征隐式学习目标位置关系。随后，CoordStem 将原始输入特征与两个坐标图在通道维度进行拼接，得到包含坐标先验的特征：

$$F_{\text{coord}} = \text{Concat}(x, G_x, G_y) \quad (8)$$

为避免改变后续检测头的输入通道数，模块进一步采用 1×1 卷积对拼接后的特征进行融合，并将其映射回原始通道数：

$$\hat{x} = f_{1 \times 1}(F_{\text{coord}}) \quad (9)$$

其中， F_{coord} 表示拼接坐标后的特征， $f_{1 \times 1}(\cdot)$ 表示用于坐标融合的 1×1 卷积操作， $\hat{x}$ 表示注入坐标先验后的特征。该过程仅额外引入两个坐标通道，并通过轻量卷积完成融合，结构较为简洁。本文采用 1×1 卷积进行坐标融合，主要是为了在较低计算开销下完成坐标通道与原始语义特征之间的信息交互。由于 CoordStem 仅额外引入水平方向和垂直方向两个坐标通道，若直接改变后续检测头输入维度，可能破坏原 YOLO 检测头结构并增加额外适配成本；而 1×1 卷积能够在通道维度上对坐标信息和视觉特征进行线性组合，在保持特征图空间尺寸不变的同时恢复原始通道数，从而保证后续分类分支、回归分支和 DFL 解码流程不需要额外修改。该设计兼顾了空间先验注入效果和模型轻量化需求。

在 Coord-Detect 中，注入坐标先验后的特征分别送入边界框回归分支和类别预测分支。第 i 个检测尺度上的输出可表示为：

$$P_i = \text{Concat}(f_{\text{reg}}(\hat{x}_i), f_{\text{cls}}(\hat{x}_i)) \quad (10)$$

其中, $f_{\text{reg}}(\cdot)$ 表示边界框回归分支, $f_{\text{cls}}(\cdot)$ 表示类别预测分支, P_i 表示第 i 个检测尺度的预测结果。Coord-Detect 保留 YOLO 检测头原有的分类、回归和 DFL 解码流程, 仅在检测分支前加入坐标先验注入过程。

通过上述设计, Coord-Detect 在较小结构改动下为检测头提供了显式空间位置信息。对于课堂场景中具有固定视角和相对稳定空间分布的目标, 该模块有助于增强模型对局部目标和姿态相关行为的定位感知能力, 从而为 phone、play_phone、look_book 和 turn_head 等类别的检测提供更有效的空间特征支持。

4. 实验设计

4.1. 实验环境及参数设置

本文实验基于 Ultralytics YOLO 框架实现, 深度学习框架为 PyTorch。实验设备采用单张 NVIDIA Tesla T4 GPU, Python 版本为 3.11.1, PyTorch 版本为 2.4.1+cu121, Ultralytics 版本为 8.3.248。训练时首先根据改进后的 YAML 网络配置构建模型, 并加载 YOLO11n 官方预训练权重 yolo11n.pt, 随后在 POCO 数据集上进行训练和验证。

为保证不同模型之间对比的公平性, 本文在基线模型、加入 FASR-Concat 的模型以及完整改进模型中保持主要训练参数一致。输入图像尺寸统一设置为 640×640 , batch size 设置为 8, 训练轮数设置为 300 epoch, 训练设备设置为 device = 0。训练过程中采用 Ultralytics 默认训练策略和数据增强配置, 并在训练后期自动关闭 Mosaic 增强, 以提高模型收敛稳定性。除网络结构不同外, 各组实验的训练数据、输入尺寸、batch size、训练轮数和预训练权重均保持一致。

4.2. 数据集及评价指标

本文实验采用公开 POCO 数据集。POCO 数据集共包含 1903 张图像和 137,960 个标注实例, 按照原始划分包括训练集 1146 张、验证集 379 张和测试集 378 张图像。根据课堂行为检测任务需求, 本文使用其中 10 类课堂行为及相关目标作为检测类别, 分别为 raise_head、bow_head、turn_head、phone、play_phone、book、look_book、sleep、upright 和 bend。其中, raise_head、bow_head、turn_head、play_phone、look_book、sleep、upright 和 bend 主要表示学生行为或姿态状态, phone 和 book 表示与课堂行为相关的辅助目标类别。

本文采用目标检测任务中常用的 Precision、Recall、mAP50 和 mAP50:95 作为主要评价指标。其中, Precision 表示预测为正样本的检测框中正确检测框所占比例, Recall 表示真实目标中被正确检测出的比例; mAP50 表示 IoU 阈值为 0.5 时的平均精度均值, mAP50:95 表示 IoU 阈值从 0.5 到 0.95 时的平均精度均值, 可用于衡量模型在不同定位精度要求下的综合检测性能。

5. 结果分析

5.1. 消融实验

为了验证前景感知尺度路由特征融合模块 FASR-Concat 和坐标先验检测头 Coord-Detect 对模型检测性能的影响, 本文在 POCO 数据集上进行消融实验, 实验结果见表 1。各组消融实验均采用相同的数据集划分、输入图像尺寸、训练轮数、batch size 和预训练权重, 仅改变模型结构, 以保证实验结果的可比性。本文以 YOLO11n 作为 Baseline, 在此基础上依次加入 FASR-Concat 和 Coord-Detect, 以分析不同模块对课堂学生行为检测性能的影响。

由表 1 可知, Baseline 的 Precision、Recall、mAP50 和 mAP50:95 分别为 0.847、0.766、0.812 和 0.574。在加入 FASR-Concat 模块后, 模型的 mAP50 提升至 0.835, mAP50:95 提升至 0.582, Recall 由 0.766 提升至 0.788, 说明该模块能够增强模型对课堂行为目标的检出能力。该阶段 Precision 从 0.847 降至 0.827,

表明模型在检出更多目标的同时可能引入了一定误检。

在 FASR-Concat 的基础上进一步加入 Coord-Detect 后, 模型 Precision、Recall、mAP50 和 mAP50:95 分别达到 0.858、0.802、0.850 和 0.600。相比 Baseline, 完整改进模型的 mAP50 提升 3.8 个百分点, mAP50:95 提升 2.6 个百分点, Recall 提升 3.6 个百分点。实验结果表明, FASR-Concat 和 Coord-Detect 对模型性能提升具有一定互补作用。其中, FASR-Concat 通过前景感知和尺度路由机制增强多尺度特征融合能力, Coord-Detect 通过引入坐标先验增强检测头对目标空间位置的感知能力, 二者共同提升了模型在课堂学生行为检测任务中的检测性能。

Table 1. Step-by-step ablation analysis of the proposed modules on the POCO dataset

表 1. 本文提出模块在 POCO 数据集上的逐项消融实验分析

Model	FASR-Concat	Coord-Detect	Precision	Recall	mAP50	mAP50:95
1			84.7	76.6	81.2	57.4
2	√		82.7	78.8	83.5	58.2
3	√	√	85.8	80.2	85.0	60.0

5.2. 对比实验

为了进一步验证本文算法的有效性, 将 FC-YOLO 与 YOLOv5n、YOLOv8n、YOLOv9t、YOLOv10n 和 YOLO11n 等主流轻量级目标检测算法在 POCO 数据集上进行对比, 实验结果见表 2。各对比模型均采用相同的数据集划分、输入图像尺寸和训练轮数, 以保证实验结果的可比性。

由表 2 可知, YOLO11n 基线模型的 mAP50、mAP50:95、Precision 和 Recall 分别 81.2%、57.4%、84.7%和 76.6%; 本文提出的 FC-YOLO 分别达到 85.0%、60.0%、85.8%和 80.2%。与 YOLO11n 相比, FC-YOLO 的 mAP50 提升 3.8 个百分点, mAP50:95 提升 2.6 个百分点, Recall 提升 3.6 个百分点, 说明本文改进方法能够提升模型对课堂行为目标的整体检测能力, 尤其有助于减少密集课堂场景中的漏检问题。

与其他 YOLO 系列模型相比, FC-YOLO 同样取得了较好的检测性能。YOLOv5n、YOLOv8n、YOLOv9t、YOLOv10n 和 YOLO11n 的 mAP50 分别为 81.0%、82.0%、82.2%、81.3%和 81.2%, 而 FC-YOLO 的 mAP50 达到 85.0%, 均高于上述对比模型; 在 mAP50:95 指标上, FC-YOLO 达到 60.0%, 相比对比模型中表现较好的 YOLOv8n 提高 2.0 个百分点, 表明本文模型在不同 IoU 阈值下具有更好的综合检测性能。同时, FC-YOLO 的 Recall 达到 80.2%, 高于各对比模型, 说明其对课堂场景中学生行为目标的检出能力更强。

Table 2. Comparison of detection performance of different algorithms on the POCO dataset

表 2. 不同检测算法在 POCO 数据集上的检测性能对比

Model	Params (M)	GFLOPs	mAP50	mAP50:95	Precision	Recall	FPS
YOLOv5n	2.5	7.1	81.0	56.7	85.9	75.3	312.5
YOLOv8n	3.01	8.1	82.0	58.0	85.0	76.9	175.4
YOLOv9t	1.97	7.6	82.2	57.8	84.6	77.1	243.9
YOLOv10n	2.27	6.5	81.3	56.9	81.2	76.8	270.3
YOLO11n	2.58	6.3	81.2	57.4	84.7	76.6	227.3
FC-YOLO	3.11	13.0	85.0	60.0	85.8	80.2	166.7

从模型复杂度和推理速度来看, FC-YOLO 的参数量为 3.11 M, GFLOPs 为 13.0, 相比 YOLO11n 的 2.58 M 和 6.3 GFLOPs 有所增加。其中, 参数量增加 0.53 M, 增幅约为 20.5%; 计算量增加 6.7 GFLOPs, 约为 YOLO11n 的 2.06 倍。这主要是由于本文在特征融合阶段引入 FASR-Concat, 并在检测头中加入 Coord-Detect 坐标先验结构, 使模型具备更强的前景区域建模、多尺度特征路由和空间位置感知能力。相应地, FC-YOLO 的 FPS 由 YOLO11n 的 227.3 下降至 166.7, 下降约 26.7%。虽然 FC-YOLO 在计算成本和推理速度方面不如原始 YOLO11n 轻量, 但其检测精度和召回能力均有所提升, 说明改进模型在一定计算成本增加的情况下换取了更强的目标检出能力。

因此, FC-YOLO 更适用于对检测精度和漏检控制要求较高的课堂行为分析场景, 如固定摄像头课堂视频分析、课后教学状态统计以及部署在 GPU 服务器或具备一定算力的边缘设备上的实时检测任务。需要指出的是, FC-YOLO 的 GFLOPs 增长较为明显, 在低功耗嵌入式设备、移动端 CPU 或多路摄像头并发检测场景中可能面临一定部署压力。若应用场景更加关注极低延迟或低算力运行, 可在后续研究中结合模型剪枝、通道压缩、量化加速或 TensorRT 部署优化等策略, 进一步降低模型计算成本。

为进一步分析本文方法在不同类别上的检测效果, 本文对 YOLO11n 基线模型和 FC-YOLO 模型在 POCO 数据集上的各类别 AP50 和 AP50:95 结果进行统计, 结果见表 3。从类别结果来看, FC-YOLO 在多数类别上均取得了不同程度的提升。其中, play_phone 类别提升最为明显, AP50 由 67.6% 提升至 80.9%, 提高 13.3 个百分点, AP50:95 由 35.4% 提升至 43.4%, 提高 8.0 个百分点; phone 类别 AP50 提高 8.7 个百分点, book、look_book 和 turn_head 类别 AP50 分别提高 3.6、3.1 和 3.3 个百分点。上述类别通常与手机、书本、头部姿态和桌面区域等局部信息密切相关, 说明本文提出的 FASR-Concat 能够增强局部前景区域和多尺度特征的表达能力, Coord-Detect 能够进一步利用固定课堂视角下的空间位置先验, 从而提高模型对细粒度课堂行为类别的检测效果。同时, upright 和 bend 类别在 AP50:95 指标上略有下降, 但下降幅度较小, 分别为 0.5 和 0.1 个百分点, 整体仍保持较高检测水平。这说明改进模型在提升局部小目标和细粒度行为类别检测性能的同时, 对部分姿态类别的高 IoU 定位精度仍存在轻微波动。后续研究可进一步结合定位损失优化、数据增强和模型轻量化策略, 提高不同类别检测性能的均衡性。

Table 3. Class-wise AP comparison between YOLO11n and FC-YOLO on the POCO dataset

表 3. YOLO11n 与 FC-YOLO 在 POCO 数据集上的各类别 AP 对比

类别	YOLO11n mAP50	FC-YOLO mAP50	提升/%	YOLO11n mAP50:95	FC-YOLO mAP50:95	提升/%
raise_head	90.2	92.0	1.8	67.4	68.9	1.5
bow_head	95.1	95.6	0.5	67.3	69.0	1.7
turn_head	68.0	71.3	3.3	47.5	40.8	2.3
phone	49.2	57.9	8.7	26.6	30.5	3.9
play_phone	67.6	80.9	13.3	35.4	43.4	8.0
book	86.6	90.2	3.6	57.2	61.3	4.1
look_book	74.5	77.6	3.1	46.0	49.8	3.8
sleep	93.7	94.6	0.9	68.0	69.3	1.3
upright	97.6	98.0	0.4	82.8	82.3	-0.5
bend	91.1	92.4	1.3	76.2	76.1	-0.1

5.3. 可视化分析

如图 4 至图 5 所示, 基线 YOLO11n 模型与改进后的 FC-YOLO 在同一课堂场景下的检测结果存在明

显差异。该场景中中学生分布密集, 且存在大量低头、看书、玩手机等小尺度行为目标, 同时后排学生受距离、遮挡和桌椅背景干扰影响较大。基线模型虽然能够检测出部分前排学生的典型行为, 但在后排小目标、边缘区域以及手机、书本等辅助目标上仍存在一定漏检现象, 对密集课堂场景中的细粒度行为识别能力不足。相比之下, FC-YOLO 能够在更大范围内捕获学生行为区域, 对低头、玩手机、看书、睡觉等类别的检测更加充分, 尤其是在远距离学生、局部遮挡和复杂桌面背景下仍能保持较好的目标召回能力。

从检测数量上看, 在相同测试图像和阈值设置下, 基线 YOLO11n 模型共检测出 191 个目标, 而 FC-YOLO 检测出 207 个目标, 说明改进模型在课堂密集目标场景中具有更强的小目标感知能力和特征表达能力。可视化结果表明, FC-YOLO 通过增强特征融合与定位能力, 能够更有效地识别课堂场景中的细粒度行为目标, 减少远景、遮挡和小尺寸目标带来的漏检问题, 为后续学生行为统计和课堂专注度分析提供了更完整的检测基础。

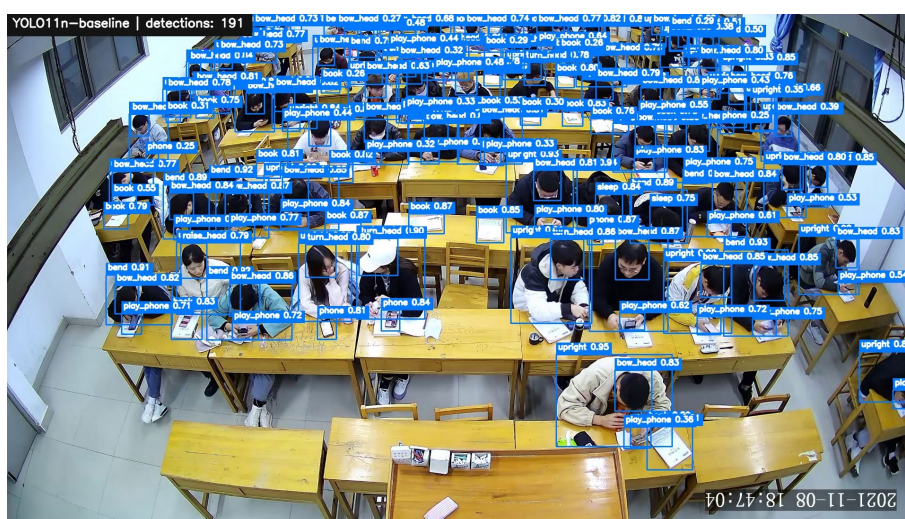


Figure 4. Detection results of the baseline model

图 4. 基线模型检测结果



Figure 5. Detection results with FC-YOLO

图 5. 加入 FC-YOLO 后的检测结果

6. 结论

本文针对课堂学生行为检测中目标分布密集、局部小目标难以识别以及细粒度姿态差异较明显等问题, 提出一种融合前景感知与坐标先验的课堂学生行为检测算法。该方法以 YOLO11n 为基础模型, 在特征融合阶段引入 FASR-Concat 模块, 通过前景感知和尺度路由机制增强学生行为相关区域的多尺度特征表达能力; 在检测头部分引入 Coord-Detect 模块, 通过归一化二维坐标注入增强检测头的位置感知能力, 从而提高模型对课堂场景中局部行为目标和姿态变化类别的检测效果。

在 POCO 数据集上的实验结果表明, YOLO11n 基线模型 mAP50 为 81.2%, mAP50:95 为 57.4%; 加入 FASR-Concat 后, 模型 mAP50 提升至 83.5%; 采用 FASR-Concat 与 Coord-Detect 的完整改进模型后, mAP50 达到 85.0%, 较基线模型提升 3.8 个百分点, mAP50:95 提升至 60.0%。类别实验结果显示, 最终模型在 phone、play_phone、look_book 和 turn_head 等局部行为或细粒度姿态类别上取得了较明显提升, 说明所提出的改进模块能够有效增强模型对课堂行为目标的检测能力。消融实验结果进一步表明, FASR-Concat 和 Coord-Detect 均能在不同程度上改善模型检测性能, 验证了本文算法设计的有效性。

与基线模型相比, 改进模型在保持课堂行为检测任务适用性的基础上提升了整体检测精度, 能够在一定程度上减少局部行为类别的漏检和误检现象。但本文方法仍存在不足, 例如 phone 等小目标类别的检测精度仍有进一步提升空间, 模型在不同课堂环境、不同摄像头视角下的泛化能力还需要进一步验证。后续研究将从小目标特征表达优化、多场景数据验证以及系统应用集成等方面展开, 并将行为检测结果进一步接入课堂专注度分析系统, 为学生课堂状态分析和教学管理提供更加稳定的数据基础。

参考文献

- [1] Liu, Q., Jiang, X. and Jiang, R. (2025) Classroom Behavior Recognition Using Computer Vision: A Systematic Review. *Sensors*, **25**, Article 373. <https://doi.org/10.3390/s25020373>
- [2] Huang, W., Li, N., Qiu, Z., Jiang, N., Wu, B. and Liu, B. (2020) An Automatic Recognition Method for Students' Classroom Behaviors Based on Image Processing. *Traitement du Signal*, **37**, 503-509. <https://doi.org/10.18280/ts.370318>
- [3] Hu, M., Wei, Y., Li, M., Yao, H., Deng, W., Tong, M., et al. (2022) Bimodal Learning Engagement Recognition from Videos in the Classroom. *Sensors*, **22**, Article 5932. <https://doi.org/10.3390/s22165932>
- [4] Uçar, M.U. and Özdemir, E. (2022) Recognizing Students and Detecting Student Engagement with Real-Time Image Processing. *Electronics*, **11**, Article 1500. <https://doi.org/10.3390/electronics11091500>
- [5] Lin, F.-C., Ngo, H.-H., Dow, C.-R., Lam, K.-H. and Le, H.L. (2021) Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection. *Sensors*, **21**, Article 5314. <https://doi.org/10.3390/s21165314>
- [6] Zhang, Y., Zhu, T., Ning, H. and Liu, Z. (2021) Classroom Student Posture Recognition Based on an Improved High-Resolution Network. *EURASIP Journal on Wireless Communications and Networking*, **2021**, Article ID: 140. <https://doi.org/10.1186/s13638-021-02015-0>
- [7] Tang, L., Xie, T., Yang, Y. and Wang, H. (2022) Classroom Behavior Detection Based on Improved YOLOv5 Algorithm Combining Multi-Scale Feature Fusion and Attention Mechanism. *Applied Sciences*, **12**, Article 6790. <https://doi.org/10.3390/app12136790>
- [8] Chen, H., Zhou, G. and Jiang, H. (2023) Student Behavior Detection in the Classroom Based on Improved YOLOv8. *Sensors*, **23**, Article 8385. <https://doi.org/10.3390/s23208385>
- [9] Chen, H. and Guan, J. (2022) Teacher-Student Behavior Recognition in Classroom Teaching Based on Improved YOLO-V4 and Internet of Things Technology. *Electronics*, **11**, Article 3998. <https://doi.org/10.3390/electronics11233998>
- [10] Liu, M., Meng, F., Wu, Q., Xu, L. and Liao, Q. (2021) Behaviour Detection in Crowded Classroom Scenes via Enhancing Features Robust to Scale and Perspective Variations. *IET Image Processing*, **15**, 3466-3475. <https://doi.org/10.1049/ipr2.12318>
- [11] Zhang, J., Guo, L. and Wang, X. (2025) Student Classroom Behavior Recognition Based on YOLOv8 and Attention Mechanism. *Information*, **16**, Article 934. <https://doi.org/10.3390/info16110934>
- [12] Zhang, X., Nie, J., Wei, S., Zhu, G., Dai, W. and Yang, C. (2024) A Study of Classroom Behavior Recognition

- Incorporating Super-Resolution and Target Detection. *Sensors*, **24**, Article 5640. <https://doi.org/10.3390/s24175640>
- [13] Zheng, Z.J., Liang, G.J., Luo, H.B. and Yin, H.C. (2025) Attention Assessment Based on Multi-View Classroom Behaviour Recognition. *IET Computer Vision*, **19**, e12146. <https://doi.org/10.1049/cvi2.12146>
 - [14] Shou, Z., Yan, M., Wen, H., Liu, J., Mo, J. and Zhang, H. (2023) Research on Students' Action Behavior Recognition Method Based on Classroom Time-Series Images. *Applied Sciences*, **13**, Article 10426. <https://doi.org/10.3390/app131810426>
 - [15] Jia, Q. and He, J. (2024) Student Behavior Recognition in Classroom Based on Deep Learning. *Applied Sciences*, **14**, Article 7981. <https://doi.org/10.3390/app14177981>
 - [16] Zhao, J., Zhu, H. and Niu, L. (2023) Bitnet: A Lightweight Object Detection Network for Real-Time Classroom Behavior Recognition with Transformer and Bi-Directional Pyramid Network. *Journal of King Saud University-Computer and Information Sciences*, **35**, Article 101670. <https://doi.org/10.1016/j.jksuci.2023.101670>
 - [17] Wang, C., Mohamed, A.S.A., Yang, X., Zhang, H., Li, X. and Noor, M.H.B.M. (2025) Enhancing Classroom Behavior Recognition with Lightweight Multi-Scale Feature Fusion. *Computers, Materials & Continua*, **85**, 855-874. <https://doi.org/10.32604/cmc.2025.066343>
 - [18] Li, Y., Chen, Y., Wang, N. and Zhang, Z. (2019) Scale-Aware Trident Networks for Object Detection. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, 27 October-2 November 2019, 6054-6063. <https://doi.org/10.1109/iccv.2019.00615>
 - [19] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., *et al.* (2021) Multiscale Vision Transformers. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, 11-17 October 2021, 6824-6835. <https://doi.org/10.1109/iccv48922.2021.00675>
 - [20] 陈晨, 保文星, 陈旭, 等. 改进 YOLOv8 的学生课堂行为识别算法: DMS-YOLOv8 [J]. 计算机工程与应用, 2024, 60(24): 222-234.
 - [21] 曾钰琦, 刘博, 钟柏昌, 等. 智慧教育下基于改进 YOLOv8 的学生课堂行为检测算法[J]. 计算机工程, 2024, 50(9): 344-355.
 - [22] 曹倩, 曹焱, 钱承山. 改进 YOLO11 的学生课堂行为检测算法[J]. 电子测量技术, 2025, 48(15): 185-198.
 - [23] 曹焱, 曹倩, 钱承山, 等. 改进 YOLO11 的高精度课堂行为检测算法[J]. 计算机科学与探索, 2025, 19(8): 2135-2148.
 - [24] Zhang, H., Zhou, G., He, W. and Deng, H. (2025) Classroom Behavior Detection Method Based on PLA-YOLO11n. *Sensors*, **25**, Article 5386. <https://doi.org/10.3390/s25175386>
 - [25] Peng, S., Zhang, X., Zhou, L. and Wang, P. (2025) YOLO-CBD: Classroom Behavior Detection Method Based on Behavior Feature Extraction and Aggregation. *Sensors*, **25**, Article 3073. <https://doi.org/10.3390/s25103073>