基于改进 YOLOv11 的学生课堂行为检测算法

李 美[1]
LI Mei

摘 要 学生课堂行为是衡量学习效果的直接观测指标，基于深度学习的行为检测方法可辅助教师优化教学方式、调整教学进度，进而提升教学效率。为实现精准的学生课堂行为检测与教学效率优化，文章提出一种基于 YOLOv11 模型的课堂行为识别方法。该方法通过引入 MLCA 模块，增强模型对小目标的特征提取能力，提升复杂课堂环境下的行为检测性能；同时，采用 MFM 模块替换模型原有简单拼接模块，强化多尺度特征融合效果，进一步提高检测精度。实验结果表明，所提方法对课堂常见行为识别的 mAP@50 达到 76.4%，相较于原始 YOLOv11 模型提升 2.2%，能够更高效、精准地识别学生课堂行为，为精准教学分析与教学效率优化提供技术支撑。

关键词 目标检测；计算机视觉；课堂行为识别；YOLOv11；注意力机制

doi：10.3969/j.issn.1672-9528.2026.03.012

0 引言

近年来，随着智慧教育理念不断深入，课堂作为教学活动的核心场景，其教学质量备受关注。而学生的课堂行为是反映学生的学习状态、检验教学效果的重要依据，因此收集并分析大量的课堂图像数据，将其与行为检测相结合，可广泛应用于教学质量评估、学生学习状态监测、个性化教学指导等方面。由此可见，对学生课堂行为检测技术的研究具有重要的现实意义与广阔的应用前景[1]。

目前，随着深度学习和计算机视觉技术持续发展，课堂行为检测在检测精度、实时性等方面取得了一定进展。但在实际应用中，如课堂中对于学生举手、抬头、低头、写字等行为的检测，仍面临精度与速度的双重挑战，其核心难点主要体现在以下几个方面：

（1）检测目标的形状差异较大。课堂中学生座位分布存在前后、左右差异，同一行为在图像中呈现的尺寸差异明显，前排学生的“举手”动作可能占据大区域，而后排学生的相同动作可能仅表现为较小区域，由于小目标的形状表征能力较弱，无法准确反映局部细节信息，导致部分学生的行为检测效果不佳。

（2）图像分辨率限制。因课堂拍摄分辨率有限，加之覆盖整间教室，摄像头距离远，因此，一位学生在画面中占比较小。这对动作幅度较小和后排学生来说，无论是抬手、低头、写字，对应像素区域较小，容易被光线干扰或模糊掩盖，难以清晰分辨。

（3）背景环境复杂。课堂环境中存在桌椅、书本及其多种元素。学生的行为容易被周围环境干扰，当学生的行为动作与周围环境的视觉特征相近时，难以准确定位，增加了检测难度。

教室是密集场所，一间教室可能超过 50 名学生。另外，教室不同位置的学生在图片中的尺寸差别较大，这些都给检测的准确率带来了挑战。为解决上述问题，本文在 YOLOv11 的基础上，通过在主干网络上添加 MLCA 模块，引入 MFM 模块来代替模型中原来的简单拼接操作，增强了特征提取的能力以及提高了更有效的多尺度特征融合，从而更好地还原输入图片的轮廓信息[2]。

1 相关工作

1.1 YOLO 算法

YOLO（you only look once）系列算法是基于深度学习的单阶段目标检测算法，自 2016 年首次提出以来，凭借“端到端”的检测模式，在特征图上同时完成目标分类与边界框回归，无需单独生成候选区域，成为实时目标检测领域的主流方向。其发展历程始终围绕“精度与速度的平衡”持续优化：初代 YOLO 通过将目标检测转化为回归问题，实现了远超同期算法的检测速度，但存在小目标检测精度不足的问题；YOLOv2 引入批量归一化、锚框机制等改进，提升了定位精度与召回率；YOLOv3 采用多尺度特征图检测策略，并使用残差网络作为 backbone，进一步平衡了精度与速度；YOLOv4 融入 Mosaic 数据增强、CIoU 损失函数等技术，在复杂场景下的检测性能显著提升；YOLOv5 通过动态锚框

1. 太原师范学院 山西太原 030000

计算、自适应图片缩放等创新，首次实现模型轻量化设计（提供 n/s/m/l/x 等不同规模模型）；YOLOv6、YOLOv7 则分别在网络结构和训练策略上优化，进一步提升推理速度；YOLOv8、YOLOv11 等版本，通过引入更高效的特征融合模块、注意力机制及轻量化设计，在小目标检测精度、实时性等方面持续突破，使其在自动驾驶、安防监控、智慧教育等场景中具备更强的实用价值。

1.2 目标检测

目标检测是计算机视觉领域的核心任务之一，旨在从图像或视频中定位并识别出指定的目标（如行人、车辆、物体等），输出目标的类别及边界框坐标，其技术发展历经了从传统方法到深度学习方法的迭代：早期传统方法以手工设计特征结合分类器（如 SVM）为主，例如在 2005 年提出的基于 HOG+SVM 行人检测框架，虽然奠定了目标检测的基础，但受限于手工特征的表达能力，在复杂场景下精度较低、泛化能力弱；2012 年，深度学习兴起后，基于卷积神经网络的目标检测方法逐渐占领主导地位；2014 年，R-CNN 的提出标志着两阶段检测框架的诞生——通过选择性搜索生成候选区域，再用卷积网络提取特征并分类，大幅提升了检测精度，但速度较慢；随后 Fast R-CNN、Faster R-CNN 等改进模型不断优化两阶段算法的效率，使其在精度上保持优势 [3]。与此同时，单阶段算法也快速发展：2016 年，YOLO 将检测转化为端到端的回归问题，实现了实时检测，SSD 则结合多尺度特征提升小目标检测能力，单阶段算法以速度优势填补了实时场景需求。

近年来，两阶段与单阶段算法不断地持续融合优化，如 YOLO 系列（从 YOLOv1 到 YOLOv11）通过改进网络结构、特征融合和注意力机制平衡精度与速度，目前目标检测技术在精度、速度、泛化能力上仍在不断突破。

2 改进 YOLOv11 的学生课堂行为检测

在现有的研究基础上，考虑到课堂场景中学生行为检测面临的目标尺寸多变、检测精度低等问题，本文选取性能优异的 YOLOv11 作为基础模型进行改进。改进后的模型结构图如图 1 所示。

先对包含学生举手、抬头、低头、写字等行为的课堂原始图像进行尺寸缩放、像素值归一化等标准化预处理；随后送入主干网络，先通过前两层卷积生成特征图并提取基础视觉特征，再在关键层引入 C3k2_MLCA 模块替换原有基础卷积块，通过捕捉局部特征的上下文关联与多尺度信息，提升细微特征的提取能力，经多轮交替运算得到强化后的特征图；再经 SPPF 模块完成多尺度池化融合，并通过 C2PSA 模块进一步增强特征表达能力。

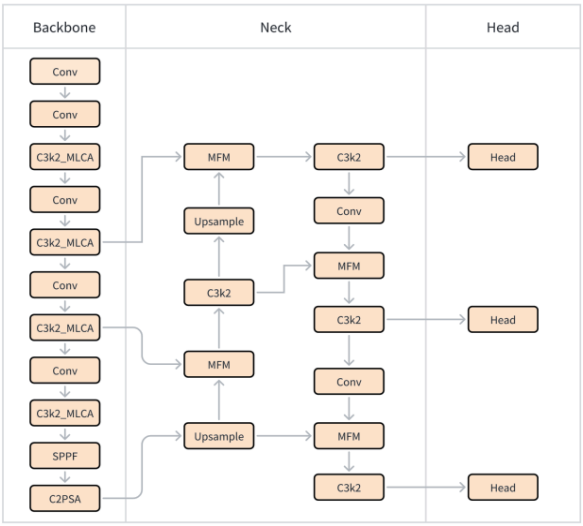

图 1 改进后的框架图

接下来在颈部网络开展多尺度特征融合：从高层特征出发，经上采样后与对应尺度特征图通过 MFM 模块融合，借助精细的特征加权与通道交互机制，整合不同层级特征，更完整地还原图像整体轮廓信息；融合后的特征经 C3k2 模块细化，生成三组适配不同尺寸目标的融合特征图。

最后将 3 个尺度的融合特征图输入 Detect 检测层，同步完成类别预测与边界框回归，输出包含行为类别、位置坐标、置信度等信息的检测结果 [4]。

2.1 MLCA 模块

大多数的通道注意力机制仅包含通道特征明显而忽略空间特征信息，导致模型表达效果或对象检测性能不佳，空间注意力模块通常复杂且计算成本高。

MLCA（mixed local channel attention）能够同时整合通道信息和空间信息，以及局部信息和全局信息来提高网络的表达效果。如图 2 所示，对输入的 $C\times W\times H$ 形状的特征图，一方面，借局部平均化在特征图上划分局部区域，提取图像中学生个体周边的小范围，包含细节纹理的局部特征；另一方面，依靠全局平均化遍历整个特征图，抓取教室场景整体布局、人群分布趋势等宏观语义。双路径并行，使模型既不会丢失小目标的局部细节，又能理解全局场景上下文，区分课堂不同区域行为 [5]。

局部的特征通过“乘法”操作，让有用的特征信息强化，弱化没用的特征信息，具体而言，先通过 Reshape 操作重塑特征维度，再经 1×1 卷积与 Reshape 的处理回到原来的尺寸，依据特征自身重要性，为不同通道赋予差异化权重。如在学生课堂行为场景里，当识别到举手，写字这类关键行为特征时，对应通道的权重会被显著提升，使其响应得到强化；而像桌椅纹理、背景装饰等无关的冗余特征，所在通道权重则

被降低，实现弱化。然后将全局上下文信息经上采样与局部特征融合，最终输出特征图。在学生课堂行为中，对于人员密集、检测目标较小的情况，MLCA 能关注更多信息，在捕获关键特征时，以轻量级设计实现高效计算，既保证模型性能，又避免了过多的参数和计算量增加。

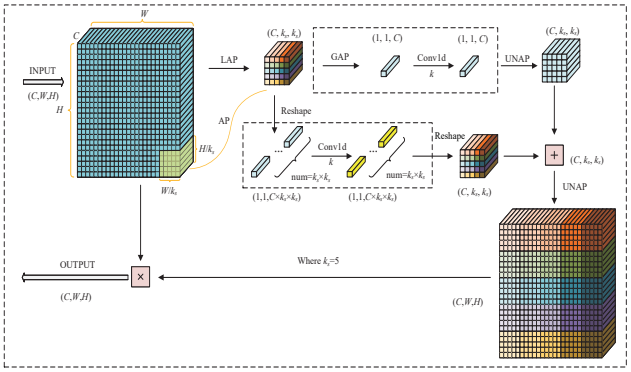

图 2 MLCA

2.2 MFM 模块

该模块通过动态权重调整能够区分前景目标与背景区域，通过对输入的多源特征进行重要性评估，生成与特征空间维度匹配的权重矩阵 [6]。这种动态调整机制使融合后的特征更贴合任务需求，增强特征的判别性与表达能力。如图 3 所示，MFM 结构主要包含三个核心部分：特征输入与初步融合、权重生成、动态加权融合。

对于复杂场景中存在的小目标、遮挡目标或低对比度目标，MFM 可自适应提升其边缘、纹理等判别性特征的权重，同时抑制冗余背景信息的干扰，使融合后的特征更聚焦于目标本身，从而增强模型对目标的定位精度与类别判别能力，尤其在特征差异微弱的场景中能有效提升检测的鲁棒性。

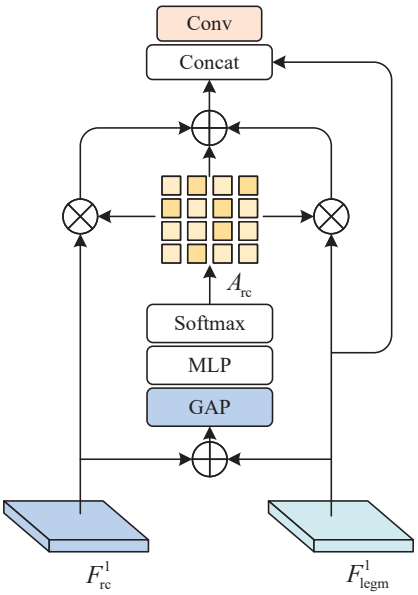

图 3 MFM

3 实验结果与分析

3.1 实验环境

本文的实验环境使用 Intel Xeon E5-2683 v4 处理器、60 GB 内存、NVIDIA RTX 4060 28 GB GPU；Ubuntu 20.4 系统，基于 Python 3.10、PyTorch 2.5.0 框架，通过 CUDA 12.4 实现 GPU 加速。实验参数使用 SGD 优化器进行优化，输入图像分辨率为 640 px×640 px，epochs=250，batch_size = 32，学习率 0.01。

3.2 数据集

本文实验所使用的数据集是由两部分来构成的：一部分是从网上获取的公开的上课视频素材；另一部分是在获得同学们知情同意后的监控视频。每 30 s 截取一张图片，最终形成包含 5 015 张图片的数据集。图片覆盖如普通教室、多媒体教室等不同的教室环境，包含不同的课堂人数规模和不同天气影响下的多种场景，同时涵盖了学生低头、抬头、举手、写字等多种课堂行为 [7]。为满足模型训练和评估的需求，本文按照 8:1:1 的比例，将数据集划分为训练集、测试集和验证集，以此为后续实验提供数据支撑。

3.3 评价指标

实验采取平均精度 mAP、参数量 Params 和浮点运算次数 GFLOPs 作为核心评价指标，全面衡量算法性能。其中，mAP@0.5 指在交并比阈值为 0.5 的条件下，所有待检测类别的平均精度，能够直观反映算法对每个类别的识别与定位能力。mAP 计算过程如式（1）~（4）所示，涉及召回率（Recall）和精确率（Precision），其中 TP 指预测正确的样本；FP 指预测错误的样本；FN 指漏检的样本；AP 指以 Recall 为横坐标、以 Precision 为纵坐标所围曲线的面积 [8]。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{AP} = \sum_{i=1}^{n-1} (r_{i+1} - r_i) p(r_{i+1}) \tag{3}$$

$$\text{mAP} = \frac{1}{m} \sum_{i=1}^{m} \text{AP}_i \tag{4}$$

3.4 实验结果

3.4.1 对比实验

使用当前流行的 YOLO 系列的目标检测模型进行对比实验，进一步验证本文改进算法的有效性。在同样的实验环境配置和同样的参数的情况下，对同一个数据集进行检测，模型对比的实验结果如表 1 所示。从不同的评价指标分析，本

文改进的模型比其他的模型的平均精度均有所增加，且参数量和浮点运算次数增加较少，同时对于在密集场景下容易产生漏检和误检的问题，本文改进算法对这一问题也有一定的改善。

表 1 不同模型之间的结果对比

算法名称	mAP@0.5	P	R	F_1	Params/10^6	GFLOPs
YOLOv8	0.70	0.664	0.684	0.674	3.01	8.1
YOLOv9	0.735	0.689	0.707	0.698	7.17	26.7
YOLOv10	0.694	0.649	0.678	0.663	2.70	8.2
YOLOv11	0.742	0.692	0.713	0.702	9.41	21.3
Ours	0.764	0.735	0.725	0.730	9.60	22.1

3.4.2 消融实验

为了评估改进模块的具体性能，实验以 YOLOv11 作为基准模型开展消融实验。为确保实验的公平性与结果的可靠性，所有消融实验均在完全统一的环境配置（包括硬件设备、软件框架）和参数设置（如训练轮次、批次大小、学习率等）下进行，实验结果如表 2 所示。

表 2 消融实验结果对比

方法	MLCA	MFM	mAP@0.5	Params/10^6	GFLOPs
YOLOv11	×	×	0.742	9.41	21.3
YOLO11-MLCA	√	×	0.757	9.52	21.8
YOLO11-MFM	×	√	0.747	9.45	21.4
Ours	√	√	0.764	9.60	22.1

消融实验结果清晰显示，添加了 MLCA 和 MFM 模块对 YOLOv11 在课堂行为检测任务中的性能均有积极提升。仅添加 MLCA 模块时，mAP@0.5 提升 1.5%，参数和计算量小幅增加，其通过整合通道与空间、局部与全局信息增强了特征表达；仅添加 MFM 模块时，mAP@0.5 提升 0.5%，参数和计算量增幅更小，其通过动态权重调整优化了前景与背景的区分能力；而两者结合的模型实现 2.2% 的提升，在控制参数和计算量增长的同时，有效增强了模型对课堂复杂场景的检测鲁棒性 [9]。

4 结语

在对复杂环境中的目标检测方面，提出了改进 YOLOv11 的方法，引入 MLCA 模块并将其与 C3k2 结合来整合局部与全局的信息，将图像中丰富的特征进行融合，提高检测的准确性，并引入 MFM 模块，通过动态权重调整优化了前景与背景的区分能力，提高对小目标对象检测的敏感度 [10]。实验结果相比改进前的 YOLOv11 算法的检测效果有了明显的提升。在以后的工作中，将继续对模型进行优化，提高检测的精度。同时，在检测速度方面也更进一步的提高，以达到更加轻量、更加效率、更加准确地为学生行为的检测提供一定的技术支持。

参考文献：

[1] 高陈强 , 陈旭 . 基于深度学习的行为检测方法综述 [J]. 重庆邮电大学学报 (自然科学版), 2020, 32(6): 991-1002.

[2] 张喆 . 学生课堂行为检测算法研究 [D]. 西安 : 西安理工大学，2022.

[3] 赵奕博 , 杨琼 . 基于 YOLOv8 的课堂教学行为目标检测技术应用研究 [J]. 移动信息 ,2024,46(6):118-120.

[4] 白雨亭 . 基于视频的学生动作识别方法研究 [J]. 仪器仪表用户 , 2020,27(1):10-12.

[5] 张小妮 . 基于深度学习的课堂环境下学生行为检测与分析 [D]. 信阳 : 华北水利水电大学 ,2023.

[6] 闫兴亚 , 匡娅茜 , 白光睿 , 等 . 基于深度学习的学生课堂行为识别方法 [J]. 计算机工程 ,2023,49(7):251-258.

[7]WAN D H, LU R S, SHEN S Y, et al. Mixed local channel attention for object detection[J]. Engineering applications of artificial intelligence, 2023, 123(Part C): 106442.

[8]LIU Q, LIU Y, LIN D. Revolutionizing target detection in intelligent traffic systems: YOLOv8 snake vision[J]. Electronics, 2023, 12(24):4970.

[9]BHARATI P, PRAMANIK A.Deep learning techniques: R-CNN to mask R-CNN: a survey [C]//Computational Intelligence in Pattern Recognition.Berlin:Springer,2020:657-668.

[10]LOU H T, DUAN X H, GUO J M,et al. DC-YOLOv8: small-size object detection algorithm based on camera sensor[J]. Electronics, 2023, 12(10): 2323.

【作者简介】

李美（2001—），女，山西运城人，硕士研究生，研究方向：目标检测。

（收稿日期：2025-08-12 修回日期：2026-03-10）