

基于TAR-YOLO的轻量化学生课堂行为检测算法

李金龙¹, 郑霖¹, 王浩宇¹, 季伟东^{1,2}

(1. 哈尔滨师范大学 计算机科学与信息工程学院;
2. 哈尔滨师范大学 未来教师教学与发展智能实验室, 黑龙江 哈尔滨 150025)

摘要: 实时、准确地对学生课堂行为进行检测存在许多挑战, 教室场景中存在的大量遮挡和教学图像中后排的小目标是阻碍检测精度提高的重要因素。同时, 教室中性能受限的摄像头限制了可部署目标检测模型的计算量和参数量。针对以上问题, 提出一种基于改进YOLOv8的学生课堂行为检测模型TAR-YOLO。首先, 提出一个结构新颖的任务对齐动态检测头(TAPH), 其中共享卷积和任务对齐结构的设计增强了模型识别小目标的能力。其次, 在特征融合部分引入层次化特征聚合的部分卷积(MSFAP)来改造YOLOv8特征提取模块, 并使用重参数化的空洞残差卷积(R-DWR)来增强模型的多尺度特征能力, 有效降低了模型的计算量和参数量。最后, 构建一个用于检测7种学生课堂行为和3种教师课堂行为的数据集STCB, 包含4 242张图像和74 571个标注。在该数据集上, TAR-YOLO模型的平均精度(mAP50)达到了79.5%, 同时参数量和计算量只有7.0 G和1.8 M。为进一步验证模型在其他场景中的小目标检测能力, 在PASCAL VOC数据集上进行实验, 验证了该方法的泛化性能。

关键词: 目标检测; 学生课堂行为识别; 轻量化模型; 深度学习

DOI: 10.11907/rjtk.251357

开放科学(资源服务)标识码(OSID):

中图分类号: TP391.41

文献标识码: A

文章编号: 1672-7800(2025)009-0189-10



Lightweight Student Classroom Behavior Detection Algorithm Based on TAR-YOLO

LI Jinlong¹, ZHENG Lin¹, WANG Haoyu¹, JI Weidong^{1,2}

(1. College of Computer Science and Information Engineering, Harbin Normal University;
2. Intelligent Laboratory for Teaching and Development of Future Teachers, Harbin Normal University, Harbin 150025, China)

Abstract: Real-time and accurate detection of student behavior in classrooms faces many challenges. Significant occlusion in classroom scenes and small targets in the back rows of teaching images are major obstacles to improving detection accuracy. Additionally, resource-constrained cameras in classrooms limit the computational cost and parameter size of deployable object detection models. To address these issues, we propose a student classroom behavior detection model based on an improved YOLOv8, named TAR-YOLO. Firstly, we introduce a novel Task Aligned Prediction Head (TAPH), where the use of shared convolution and task-alignment design enhances the model's ability to detect small objects. Secondly, we integrate Multi-Scale Feature Aggregation Partial Convolution (MSFAP) into the feature fusion stage to enhance the YOLOv8 feature extraction module, and adopt a Re-parameterized Dilation-Wise Residual Convolution (R-DWR) to improve the model's multi-scale feature capability, effectively reducing both computation and parameter load. Finally, we construct a dataset named STCB, which includes 4 242 images and 74 571 annotations for detecting 7 types of student behaviors and 3 types of teacher behaviors. On this

收稿日期: 2025-05-22

扫描二维码阅读全文:

基金项目: 黑龙江省自然科学基金项目(PL2024F007)

作者简介: 李金龙(2000-), 男, 哈尔滨师范大学计算机科学与信息工程学院硕士研究生, 研究方向为计算机视觉、目标检测; 郑霖(1998-), 男, 哈尔滨师范大学计算机科学与信息工程学院硕士研究生, 研究方向为计算机视觉、头部姿势检测; 王浩宇(2002-), 男, 哈尔滨师范大学计算机科学与信息工程学院硕士研究生, 研究方向为知识追踪、目标检测; 季伟东(1978-), 男, 博士, CCF会员, 哈尔滨师范大学计算机科学与信息工程学院教授, 研究方向为知识追踪、计算机视觉、知识图谱。本文通讯作者: 季伟东。



dataset, the proposed TAR-YOLO model achieves a detection accuracy of 79.5% mAP, with only 7.0G FLOPs and 1.8M parameters. To further validate the model's performance in small object detection under different scenarios, we conduct experiments on the PASCAL VOC dataset, demonstrating the generalization capability of our method.

Key Words: object detection; student classroom behavior recognition; lightweight model; deep learning

0 引言

学生课堂行为检测是对课堂场景中学生行为进行识别与分析的一项关键技术,广泛应用于课堂质量评估、教学反馈和学习状态分析等方面。随着智能教育的发展,该技术在教育信息化过程中发挥着越来越重要的作用^[1]。然而,学生课堂行为检测仍面临诸多挑战:①教室中存在大量遮挡现象,例如桌椅、书本、其他学生以及教师讲台等,这些遮挡物可能遮挡学生的头部、上半身或动作细节,导致部分或全部目标区域缺失,使检测模型难以准确识别和定位学生行为;②学生在教室中的空间分布不同,尤其是后排学生在图像中所占像素较少,目标尺度小,进一步加大了检测难度;③教学场景所用摄像设备普遍存在存储与计算资源受限问题,而长时间课堂视频的存储对硬件资源要求较高。这使得模型需具备较低的计算复杂度和较小的参数量,进一步限制了检测性能的提升。这些问题使现有方法在复杂教学场景中难以实现高效、准确的行为检测。

随着深度学习技术的迅速发展,基于深度学习的学生课堂行为检测方法研究取得了令人瞩目的成就。苏超等^[2]提出一种基于人体骨骼关节点的学生课堂行为识别方法,首先使用目标检测模型检测出目标区域,然后使用改进的 OpenPose 模型从检测图像中提取骨骼关节点坐标信息,并使用 ST-SVM 对骨骼关键点信息进行训练,从而实现对课堂行为的分类。Lin 等^[3]提出一种基于姿态估计和人体检测技术的学生课堂行为检测方法,从骨架数据中提取关节位置、关节距离和骨骼角度等特征用于深度神经网络训练,并通过匹配骨骼点与人体检测框的纠错算法减少骨骼数据中的错误连接,从而提升检测精度。尽管基于骨骼数据和姿态估计的方法具有较高准确性,但该方法不适用于复杂的教室场景。首先,姿态估计需要清晰的骨架信息,但教室中往往存在严重的遮挡现象。其次,姿态估计需要更多的计算资源和复杂的模型来实现,相比传统的目标检测方法,算法的复杂度更高,难以部署在教学资源受限的教室中。为了提升学生课堂行为识别的实时性能,许多研究人员使用 YOLO (You Only Look Once) 系列^[4]、SSD (Single Shot MultiBox Detector)^[5]等轻量化目标检测模型进行学生课堂行为检测。Ren 等^[6]提出 YOLOV4-Bi,其借鉴双向特征金字塔网络 (Bidirectional Feature Pyramid Network, BiFPN)^[7]的网络结构对 YOLOV4 的路径聚合网络 (Path Aggregation Network, PANet) 进行调整,通过去除 PANet 中仅具有单一输入功能且缺乏特征融

合作用的节点,实现特征融合过程的简化,并在输入和输出节点之间添加跳跃边,从而融合更多特征,实现了比 YOLOV4 更高的检测精度。Wang 等^[8]在 YOLOV5 中结合了压缩与激励注意力机制 (Squeeze-and-Excitation Attention, SE),通过降低教学图像中背景信息的权重来减少其对课堂行为识别过程的干扰,从而提高模型在多目标检测场景下的检测精度。然而,SE 注意力忽视了位置信息,这是教学图像中定位目标的关键。Zhao 等^[9]在 YOLOV5 中结合了通道注意力,通道注意力在通道特征中结合了位置信息,以此提高模型的定位性能。Qiu 等^[10]提出一种 CM (Command Attention Mechanism Integration Module) 模块,结合了 Command 注意力和级联操作,通过增强 YOLOV7 的特征融合来提升模型在小目标检测中的表现。徐韬等^[11]在 YOLOv7 中引入 SE 注意力机制,以增强通道特征关联性,提升检测精度,并结合 SPPFCSPC 结构降低计算量,从而提高模型的实时性能,同时显著提高了智慧课堂场景下的学生课堂行为检测精度。Cao 等^[12]提出一种基于多维特征优化的课堂行为检测算法 YOLO-AMM,通过引入 AEF 模块提升语义特征融合能力,设计 MFFN 结构加强多尺度和多层级特征关联,显著增强了模型在复杂课堂环境下的实时检测性能。这些研究推动了学生课堂行为识别技术的发展,但在小目标检测精度和模型的轻量化表现上还有改进的空间,需要更充分地考虑教学场景的复杂性,从而对检测模型作出针对性的优化。

基于以上分析,学生课堂行为检测目前的主要挑战是小目标检测和检测模型的轻量化。为了解决上述问题,本文提出了 TAR-YOLO 模型。本文的主要贡献如下:

(1) 提出一个新颖的任务对齐动态检测头 (Task Aligned Prediction Head, TAPH),参照任务对齐的思想,在检测头上设计了任务对齐的结构^[13]。通过任务对齐结构,分类和定位任务可以共享特征表示,从而增强模型对遮挡目标的认知能力。使用共享卷积提取共性特征,在多个任务之间形成有效协同,即便某一任务受遮挡影响较大,也能从其他任务中获得辅助信息。在减少模型参数量的同时,有效提高了模型从遮挡严重图像中提取小目标特征信息的能力。

(2) 在 YOLOV8 模型的特征融合部分引入层次化特征聚合的部分卷积 (Multi-Scale Feature Aggregation PConv, MSFAP),该模块能够从输入中提取多种尺度的特征信息,但其并不是在所有通道上都进行这种操作,而是部分地进行,从而提高了计算效率。

(3) 在 YOLOV8 模型主干中引入重参数化的空洞残差卷积 (Re-Parameterized Dilation-Wise Residual, R-DWR),

通过不同膨胀率的膨胀卷积来提取多尺度的上下文特征,提升模型识别小目标的能力,并使用重参数化技术减少模型计算量。

(4) 构建一个用于学生课堂行为检测和分析的数据集 STCB(Students and Teachers Classroom Behavior dataset),重点关注课堂中典型的7种学生行为和3种教师与学生的互动行为。

1 TAR-YOLO 轻量级学生课堂行为目标检测网络

1.1 模型结构

YOLOv8 网络结构由主干网络(Backbone)、特征融合(Neck)和检测头(Head)3部分组成。YOLOv8的主干网络采用了类似 CSPDarknet 的结构,其通过多个卷积层提取输入图像的低层次和高层次特征。特征融合部分采用 PANet(Path Aggregation Network)将低层次的细节信息和高层次

的语义信息结合在一起,从而提升了检测精度。检测头负责最终的目标检测,利用特征融合后的多尺度特征进行目标分类、边界框回归和置信度预测。尽管 YOLOv8 模型已经相当高效,但对小目标的检测精度较低。小目标的识别依赖浅层特征来捕捉边缘和细节信息,但 YOLOv8 的检测头依赖较深层的特征图,导致浅层信息在小目标检测中容易被忽视,难以准确地检测教学图像中坐在后排或受遮挡的学生。

针对上述问题,本文提出了 TAR-YOLO,针对小目标检测和模型轻量化问题对 YOLOv8 进行改进。模型的整体框架如图1所示。在模型主干网络与特征融合部分引入 MSFAP 和 R-DWR 模块,用于降低模型的参数量和计算量,增强模型的多尺度特征提取能力。图右侧为改进后的任务对齐动态检测头,通过动态分配定位特征和分类特征在检测中的权重,以增强定位任务和分类任务之间的交互性。

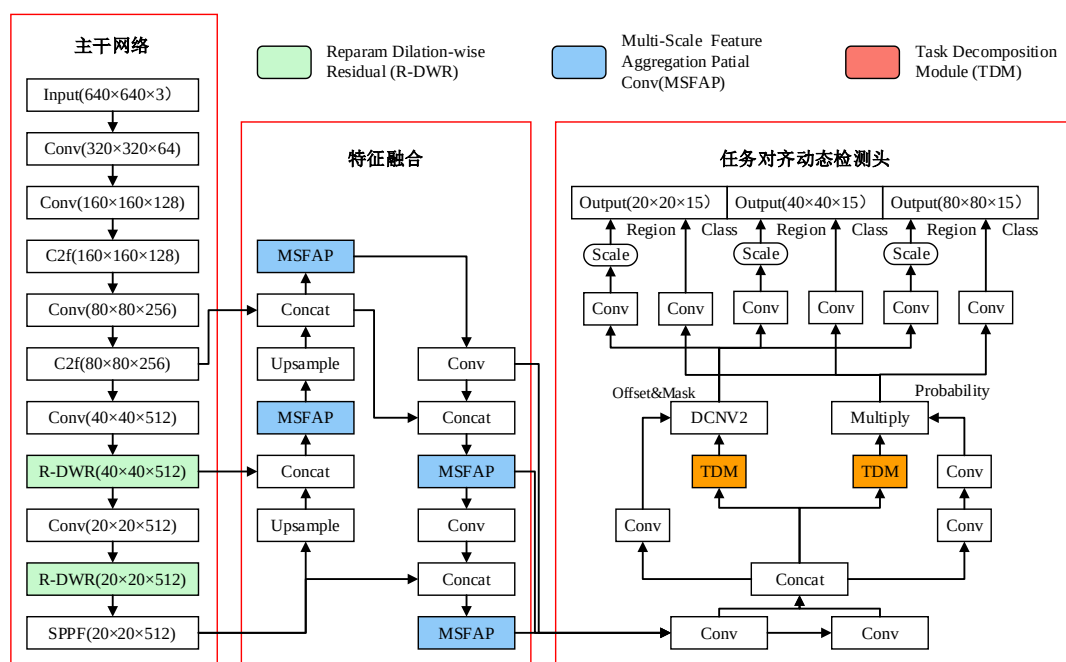


Fig. 1 Overall network structure of TAR-YOLO model

图1 TAR-YOLO 模型总体网络结构

1.2 任务对齐动态检测头

检测头的主要作用是定位和分类图像中的目标对象,通过生成边界框给出目标对象在图像中的位置,使用分类器给出目标对象的类别。现有的检测头通常使用独立的定位和分类分支,而忽视了定位任务和分类任务在学习策略上的差异,这可能导致两个任务之间的预测出现一定程度的空间错位,影响预测的准确性。Feng 等^[13]针对以上问题提出任务对齐的思想,通过计算任务交互特征来增加两个任务之间的交互,使用任务对齐得分来动态选择正样本,以提升检测头的定位和分类性能。此外,传统检测头

通常将目标划分为大、中、小3类,并为不同尺度的目标设计独立的特征提取器。尽管该方法可以实现针对性的优化,但在应对复杂遮挡场景和识别小目标时效果有限。这是因为独立的特征提取器缺乏对不同尺度目标间共性特征的提取能力,导致小目标及被遮挡目标的特征表达不足。为此,在检测头中引入共享卷积机制,以统一处理来自不同尺度的特征图,使同一组卷积核在多个尺度上进行操作,从而提取出一致性更强的语义特征。该方式不仅有效减少了模型参数和计算开销,而且增强了不同尺度特征之间的信息共享,提升了模型对小目标的感知能力。更重

要的是,共享卷积通过增强上下文建模能力,提高了模型对局部遮挡目标的鲁棒性,在后排学生、课桌、教具等遮挡物频繁出现的教室场景中,仍能保持较好的检测性能。综合上述思想,本文设计了一个结构新颖、计算效率高且具备较强鲁棒性的检测头 TAPH。该模块通过任务对齐机制与共享卷积结构的融合,有效提升了模型在复杂遮挡和小目标场景下的检测能力。

TAPH 由 4 部分组成,分别是特征提取、任务分解、边界预测和类别预测,具体细节如图 1 所示。特征提取器使用两个 3×3 的共享卷积从网络特征融合输出的融合特征中学习任务交互特征,之后将从模型 Neck 中学习到的任务交互特征经过融合操作后输入到任务分解模块 TDM (Task Decomposition Module, TDM) 中,将任务交互特征分解为定位预测特征和分类预测特征。定位预测特征和分类预测特征分别通过 Multiply 操作与 DCNV2 计算得到定位结果及分类结果。定位结果和分类结果经过 1×1 的卷积进行特征降维与特征融合。最后,通过计算任务对齐得分来动态选择正样本,确保选到的是分类和定位的最优位置,输出最终的预测框和分类结果。

将特征提取器学习到的任务交互特征输入到任务分解模块中,以对齐分类和定位任务。任务分解模块被用于任务分解,通过层注意力机制动态调整特征表示,以更好地对齐分类和定位任务。具体结构如图 2 所示。

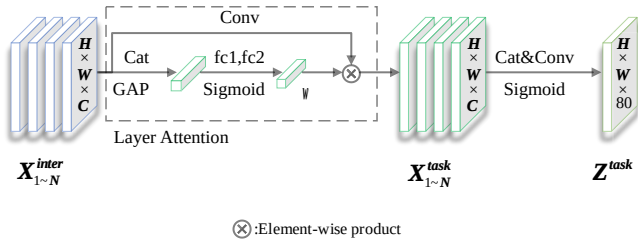


Fig. 2 Task decomposition module

图 2 任务分解模块

如式(1)所示,TDM 针对每个分类或定位任务分别计算特定任务的特征。 X^{inter} 表示特征提取器从网络 Neck 中提取到的任务交互特征, X^{task} 为 X^{inter} 在层级级别上针对每个分类或定位任务单独计算得到的特定任务的特征。其中, ω_k 表示从层注意力中学习到的第 k 个元素, X_k^{inter} 表示第 k 个 Conv_GN 从网络 Neck 输入中学习到的任务交互特征。

$$X_k^{\text{task}} = \omega_k \cdot X_k^{\text{inter}}, \forall k \in \{1, 2, \dots, N\} \quad (1)$$

ω 由跨层任务交互特征计算得到,能够捕获层之间的依赖关系,其计算公式如式(2)所示。 σ 、 δ 分别表示 sigmoid 函数和 ReLU 函数, fc_1 和 fc_2 是两个全连接层, x^{inter} 通过对 X^{inter} 进行全局平均池化得到。

$$\omega = \sigma(fc_2(\delta(fc_1(x^{\text{inter}})))) \quad (2)$$

各个 X_k^{task} 组合得到联合特征 X^{task} , X^{task} 用于计算定位或分类的结果 Z^{task} ,如式(3)所示。随后使用 sigmoid 函数将 Z^{task} 转换为分类预测特征 P 和边框预测特征 B 。

$$Z^{\text{task}} = \text{conv}_2(\delta(\text{conv}_1(X^{\text{task}}))) \quad (3)$$

P 、 B 分别用于分类预测分支和边界预测分支计算,以得到最终的预测结果。如式(4)所示,分类预测分支使用空间概率图来学习每个空间位置的任务一致性。其中, C 代表空间概率图。

$$P^{\text{align}} = \delta(\text{conv}_{\text{class}}(\sqrt{P \times C})) \quad (4)$$

C 由 X^{inter} 计算得到,如式(5)所示。 P 和 B 的元素乘积结果经过一个卷积层降维,得到最后的分类图 P^{align} 。

$$C = \sigma(\text{conv}_3(\delta(\text{conv}_2(X^{\text{inter}})))) \quad (5)$$

边界预测分支通过一个卷积层从 X^{inter} 中计算得到掩码 M 和空间偏移 O ,如式(6)、式(7)所示。

$$O = \delta(\text{conv}_1(X^{\text{inter}})) \quad (6)$$

$$M = \sigma(\delta(\text{conv}_1(X^{\text{inter}}))) \quad (7)$$

M 和 O 在可变形卷积中用于调整每个位置的预测边界框,如式(8)所示。

$$y(p) = \sum_{k=1}^K \omega_k \cdot x(p + O_k + \Delta O_k) \cdot \Delta M_k \quad (8)$$

x 、 y 分别表示输入和输出的特征图, $y(p)$ 表示输出特征图在位置 p 的值。 Ω_k 、 O_k 分别表示第 k 个位置的权重和卷积核相对于中心点指定的偏移,其中 ω_k 使用 GN 归一化方法得到。 ΔO_k 、 ΔM_k 分别表示第 k 个位置的可学习偏移和调制标量,其都通过对特征图 x 应用一个卷积层得到,输入特征图在偏移位置 $p + O_k + \Delta O_k$ 的值使用双线性插值进行计算,如式(9)所示。其中, G 表示双线性插值权重, q 表示特征图上与 $p + O_k + \Delta O_k$ 最接近的 4 个点。

$$x(p + p_k + \Delta p_k) = \sum_q G(q, p + p_k + \Delta p_k) \cdot x(q) \quad (9)$$

输出特征图 y 首先按照分类预测分支的方法经过卷积处理,再使用 Scale 层调整卷积层后的特征尺度,从而适应共享卷积的应用。最终得到边框图 B^{align} ,如式(10)所示。

$$B^{\text{align}} = \text{Scale}(\delta(\text{conv}_{\text{bbox}}(y))) \quad (10)$$

最后,分类图和边框图用于计算其高阶联合表示任务对齐得分,如式(11)所示。其中, s 、 u 分别表示分类分数和 IoU 值, α 、 β 用于控制两个任务对对齐指标的影响。

$$t = s^\alpha \times u^\beta \quad (11)$$

总体而言,本文提出的检测头使用共享卷积以提取不同尺度目标的共性特征,并设计了任务对齐结构,以提高分类任务和定位任务的交互性,从而有效提升了模型检测头的性能。

1.3 多尺度特征聚合部分卷积模块

针对小目标的检测一直以来都是目标检测领域的一个重点。教学图像中后排的学生由于座椅和其它学生的遮挡,其所占像素较小,检测更加困难。与中大型目标相比,小目标更难被检测和定位^[14]。小目标检测的难点体现在多个方面,一个主要难点是小目标的特征信息和上下文信息不足。深度 CNN 架构提供了层次结构特征图,导致不同层的特征图包含不同的空间分辨率。后层特征图中的分辨率较低且包含更强的语义信息,这对于识别大物体

很有效,但不足以检测小物体^[15]。小目标分辨率较低,且在整张图像中占比较小。因此其所包含的上下文信息较少,使其更难以被精确定位和识别。

为了加强对教学图像中小目标的检测,本文提出多尺度特征聚合部分卷积(MSFAP)模块。MSFAP模块是一种结合了逐层特征提取与多尺度特征融合结构的卷积模块,如图3所示。

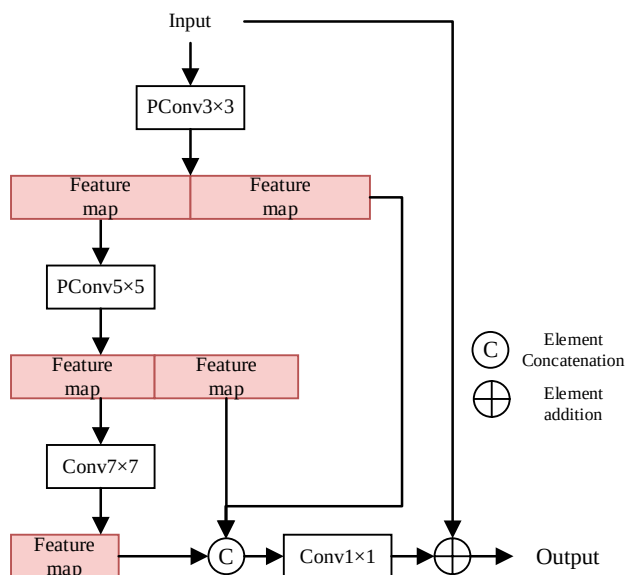


Fig. 3 MSFAP module

图3 多尺度特征聚合部分卷积模块

MSFAP模块的特征提取分为3层进行,分别使用卷积核大小为 3×3 、 5×5 和 7×7 的卷积来捕捉不同感受野的多尺度特征。在前两层的特征提取中,使用PConv对输出的特征图进行拆分,一半的特征图输入下一层的卷积作进一步特征提取,剩余部分则直接用于最后的特征融合,具体结构如图4所示。经过逐层特征提取后,最后一层使用 7×7 的卷积来提取高级语义特征,通过拼接操作将其输出的特征图与前两层未处理的特征图融合到一起。最后,通过 1×1 的深度可分离卷积压缩融合特征通道,并通过残差连接保持原始信息流,以提升模型的稳定性和表达能力。

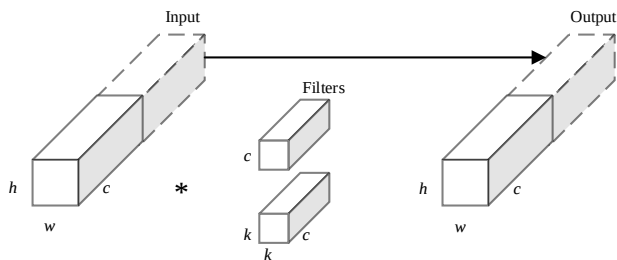


Fig. 4 PConv module

图4 PConv模块

MSFAP模块是一种渐进式特征聚合模块,其从深层到浅层逐步提取特征并进行融合,以生成具有层次结构的特征表示。该设计有利于减少计算成本,每层只处理部分

特征图,比直接处理整张特征图计算量更小,从而减少冗余,节省计算资源。此外,未处理的特征图在特征融合中保留了更多的高分辨率特征,从而增强了模型对小目标上下文信息的捕获能力。

1.4 重参数空洞残差卷积

模型主干中使用了重参数空洞残差卷积模块(R-DWR),R-DWR是一个高效的多尺度特征提取模块,采取了多分支特征提取的设计结构,如图5所示。R-DWR按照残差的风格设计,将传统的单步法分解为两步法,以降低提取多尺度特征的难度。多尺度特征提取过程被分为区域残差化和语义残差化两个步骤。在区域残差化部分,一个结合了批量归一化(Batch Normalization, BN)和线性整流函数(Rectified Linear Unit, ReLU)的 3×3 的卷积被用于初步的特征提取。之后,在语义残差部分,使用多个不同扩张率的膨胀卷积来提取多尺度特征。两步法的多尺度特征提取方法使模型避免了从表达复制的特征图中获得复杂的上下文信息,只需调整卷积的膨胀率,并对表达简洁的特征图进行简单的形态滤波,从而使特征学习过程更加有序。

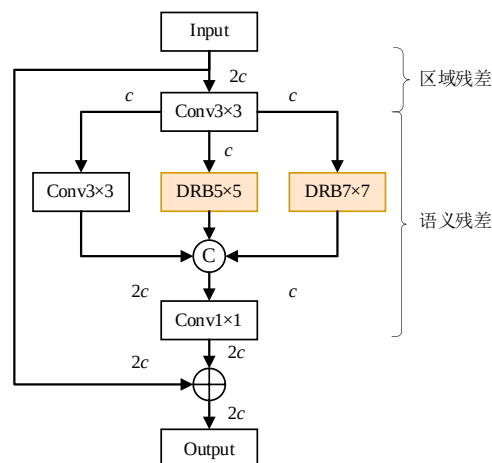


Fig. 5 R-DWR module

图5 重参数空洞残差卷积模块

在语义残差化中,使用重参数膨胀卷积(Dilated Reparam Block, DRB)替换原有的膨胀卷积^[16]。采用DRB旨在提升大内核卷积对稀疏模式的捕获能力,其通过一个转置的卷积将多个膨胀卷积等效转换为内核更大的非膨胀卷积,从而消除额外扩张的卷积层推理成本。等效转换如式(12)所示。

$$W' = \text{conv_transpose2d}(W, I, \text{stride} = r) \quad (12)$$

其中, W 、 W' 分别表示转换前和转换后的卷积核, r 表示卷积步长。通过等效转换,内核为 $k\times k$ 的膨胀卷积能够转换为内核为 $((k-1)r+1)$ 的非膨胀卷积。如图6所示,给定一组 $k = (5, 5, 3, 3)$, $r = (1, 2, 3, 4)$ 的膨胀卷积,通过等效转换将这组膨胀卷积转换为内核为9的非膨胀卷积,但参数量比直接使用内核为9的非膨胀卷积更少。DRB实现了使用非扩张小核和多个扩张小核层来增强非扩张大核

卷积层。与单独使用大内核卷积相比,该方法同时使用大内核卷积与并行小内核卷积,有利于模型在训练过程中提取小目标特征^[17]。在提取多尺度特征并融合所有特征图

后,对融合结果进行批量标准化。之后,采用逐点卷积合并特征,形成最终的残差。最后,将输入特征图加到最终的残差中,以构建更强、更全面的特征表示。

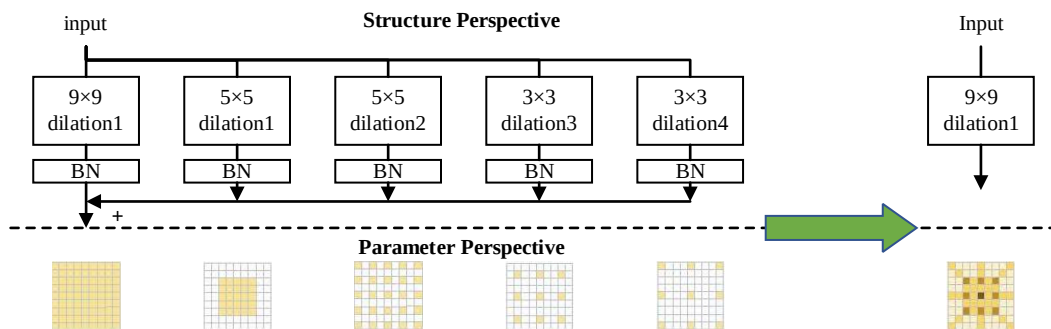


Fig. 6 Reparameterization process of DRB

图6 重参数膨胀卷积的重参数化过程

在R-DWR中,采用两步法的特征提取方法和多种膨胀率的膨胀卷积,可有效提升模型对多尺度特征的提取能力。重参数化技术在膨胀卷积转换过程中减少了该模块的参数量,与原有的膨胀卷积相比进一步提升了对小目标的检测能力。

2 实验与结果分析

2.1 数据集

鉴于现有公开数据集中针对学生课堂行为的资源有限,本文基于161个小学课堂视频构建了STCB数据集,以此评估模型性能。这些原始视频数据是从国家公共资源平台下载的,在数据收集之前,教师和学生都同意进

行视频拍摄。本文所有作者承诺,该数据集将仅用于学术研究。

在收集和整理视频数据后,使用Ffmpeg对视频数据进行自动切帧,从中筛选出与学生课堂行为高度相关的图像进行标注。图像标注在教学经验丰富的教师指导下进行,借助Labelimg工具完成。图像中的学生课堂行为被标注为7类,分别为听讲、阅读、写字、站立、举手、小组合作和睡觉。考虑到教学场景的需求,该数据集还标注了3种教师课堂行为,以更好地研究学生与教师之间的互动行为,分别为监督、讲课和指导。这些行为的分类标准如表1所示。经过标注之后,得到一个由4 242张图像和74 571个标注组成的用于检测10种典型学生与教师课堂行为的数据集。

Table 1 Classification description of student and teacher classroom behaviors

表1 学生与教师课堂行为分类说明

类别	行为描述	分类标准	标记数量
听讲	学生们上课时听教师讲课	学生们端坐,看着教师或黑板	49 929
阅读	学生阅读教科书或其他课程材料	学生们低着头,看着桌上的书	6 313
写字	学生使用笔来记笔记或做作业	学生们微微倾斜着头,手握笔	7 139
站立	学生站起来回答问题	学生们从座位上站起来	1 656
举手	学生举手以引起教师注意,以便回答或提问	学生举手向教师示意	4 130
小组合作	学生以小组形式合作,讨论和分享想法来完成任务	学生们分成小组紧密地坐在一起,互相看着对方或专注于自己的工作	3 243
睡觉	学生因疲劳而在课堂上休息	学生将手臂放在桌子上,头靠在手臂上	424
监督	教师在教室里巡视,监督学生的学习进度	教师在教室里走来走去,观察学生的状态	807
讲课	教师向学生传授知识	教师站在教室前面,看着学生或黑板	390
指导	教师在学生完成任务时为其提供指导	教师站在学生旁边,弯腰	540

图7展示了数据集中的部分样本图片。每个子图中用方框标注了一个对应行为的目标:(a)听讲;(b)读;(c)写;(d)站立;(e)举手;(f)小组合作;(g)睡觉;(h)监督;(i)讲课;(j)指导。

为进一步地验证所提出模型的有效性和泛化能力,还在公开数据集PASCAL VOC上进行了实验。PASCAL VOC数据集包含了20种日常生活中常见的物体图像,比如人、鸟、飞机、椅子、电视等,是一个广泛使用的目标检测基准数据集。

实验所使用的数据集下载地址如下:①STCB(<https://github.com/JinlongLi2025/STCB-dataset>);②PASCAL VOC(<http://host.robots.ox.ac.uk/pascal/VOC/voc2007/>)。

2.2 实验设置

所有实验在一个搭载Intel Xeon Gold 6230R CPU和NVIDIA GeForce 2080Ti(11 G)的GPU的服务器上进行,使用Pytorch1.13.0、CUDA 11.7作为深度学习训练框架,Python版本为3.8。

在模型训练之前,以8:1:1的比例划分训练集、验证

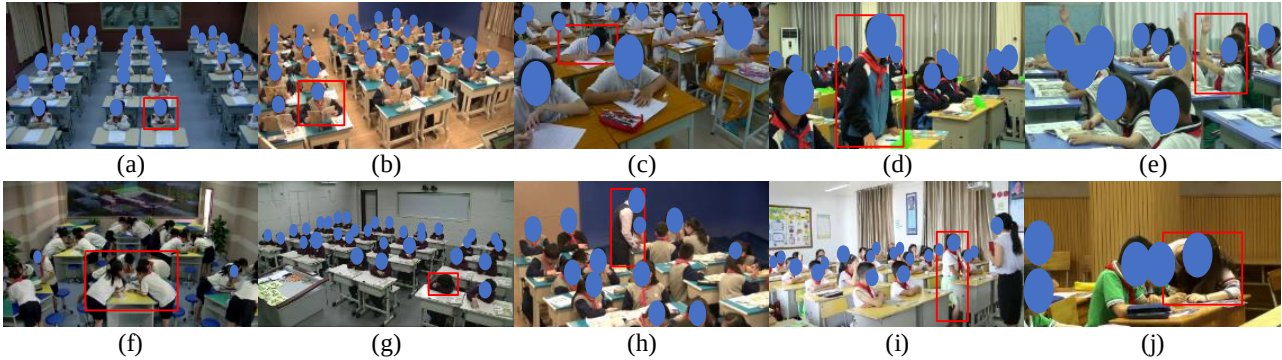


Fig. 7 Sample images of the STCB dataset
图 7 数据集 STCB 样本图片

集和测试集,输入图像尺寸统一设定为 640×640。在主干网络部分,特征图的嵌入维度依次按 128、256、512 的顺序逐层递增。为了使模型得到充分的训练,epoch 设置为 500,同时使用早停法来避免训练过程中可能发生的过拟合问题,并将对应的 patience 设置为 75,表示当模型在验证集上的指标连续 75 个周期没有得到提升后会停止训练。采用动量随机梯度下降优化器进行训练,初始学习率为 0.01,动量设置为 0.937。训练和其他实验中的批次大小均设置为 32。

在目标检测领域,通常将准确率、召回率和 mAP 作为评估模型的重要指标。准确率和召回率计算如式(13)、式(14)所示。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

其中,TP 表示真正率,FP 表示假正率,FN 表示假负率。以召回率为 x 轴,准确率为 y 轴,可以绘制准确率—召回率曲线(Precision-Recall Curve, P-R)。平均精度(Average Precision, AP)通过计算 P-R 曲线下面积得到,能够有效反映模型在不平衡数据集上的表现,mAP 则是对所有类别 AP 的平均值。具体计算公式如式(15)、式(16)所示。

$$\text{AP} = \int_0^1 P(R)d(R) \tag{15}$$

$$\text{mAP} = \frac{\sum_{i=1}^N \text{AP}(i)}{N} \tag{16}$$

其中,AP(i)表示对应第 i 个类别的 AP, N 是类别总数。mAP50 和 mAP50:95 是 mAP 的两种形式,mAP50 是 IoU(Intersection over Union)的阈值固定为 0.5 时的 mAP,mAP50:95 则是 IoU 在 0.5 ~ 0.95 阈值之间时计算的 mAP 平均值。除 mAP 外,还使用 FLOPs 和 Params 来衡量模型的轻量化表现。FLOPs 表示浮点运算次数,用来衡量模型计算量。Params 表示模型参数量,其决定了模型的存储大小。

2.3 STCB 数据集实验结果

将本文提出的 TAR-YOLO 模型与一些先进的轻量化模型进行对比, GhostNet^[18]、GhostNetV2^[19]、MobileNetV2^[20]、MobileNetV3^[21]、ShuttleNetV2^[22] 和 FasterNet^[23]

被用于替换 YOLOv8n 模型的主干,与 YOLOv8n 的检测头组合构成一个完整的检测器。此外,还对比了 YOLOv5、YOLOv11 模型。表 2 列举了以上模型的实验结果,每项用于评估模型指标的最好结果使用粗体和下划线标出。

Table 2 Experiment results comparison of different models on STCB dataset

表 2 不同模型在 STCB 数据集上的实验结果比较

Model	mAP50/%	mAP50: 95/%	AP _s /%	FLOPs/		Layer
				G	M	
GhostNet	77.7	53.0	25.3	6.8	3.8	30
GhostNetV2	76.7	52.2	23.8	7.3	2.5	22
MobileNetV2	78.0	53.0	18.7	5.6	6.4	30
MobileNetV3-Small	75.7	50.6	23.9	6.6	2.9	24
ShuffleNetV2	74.4	49.4	18.9	5.7	2.0	<u>20</u>
FasterNet	78.6	53.7	21.8	11.3	4.3	46
YOLOV5n	77.4	53.6	22.5	7.7	2.6	24
YOLOv8n	76.7	53.1	22.3	8.9	3.2	22
YOLOV11n	77.5	54.3	23.4	6.5	2.6	23
Our model	79.5	54.9	25.5	7.0	<u>1.8</u>	28

表 2 的结果表明,TAR-YOLO 模型的综合性能优于其他模型。TAR-YOLO 的 mAP50 与 mAP50:95 分别达到了 79.5% 和 54.9%,相比 YOLOv8n 提升了 2.8% 和 1.8%。在轻量化性能方面,TAR-YOLO 的网络深度为 28 层,在 YOLOv8n 检测头的基础上新增了任务交互增强模块。该模型的参数量最为紧凑,仅包含约 1.8 M 参数。虽然 TAR-YOLO 的计算量并非最低,但相比 YOLOv8n 仍减少了约 1.9 G 的计算开销。同时,在综合检测性能方面,TAR-YOLO 优于多个计算量更低的对比模型,展现出良好的轻量化与性能平衡能力。值得注意的是,MobileNet 系列模型采用深度可分离卷积,并在网络结构中大量使用了 1×1 卷积核。该方式虽然大大减少了参数量,但也降低了模型的特征表达能力。高度复杂的网络设计提高了模型部署难度。

同时,为进一步验证所提出的模块在提升模型检测小目标能力的效果,对模型识别小、中、大目标的 AP 进行了比较。如表 3 所示,所提出模块的加入使模型在对小目标的检测精度上得到显著提升,且提升幅度大于大目标,但略低于中目标。其中,TAPH 的提升最多,AP_s 达到了 24.4%,相比 YOLOv8 提升了 2.1%,表明该模型可以很好地

提取多尺度特征。

为了进一步对比 TAR-YOLO 和 YOLOv8n 的可视化表现,如图 8 所示,给出了 4 张测试集图像的原始版本、YOLOv8n 检测版本、TAR-YOLO 检测版本和真实标注。从图

Table 3 The improvement effect comparison of small object detection capability of TAR-YOLO model by different modules

表 3 不同模块对 TAR-YOLO 模型小目标检测能力的提升效果比较%

Model	mAP50	mAP50:95	AP _s	AP _m	AP _l
YOLOv8n	76.7	53.1	22.3	48.2	60.3
+TAPH	78.4	54.5	24.4	51.0	61.6
+MSFAP	77.9	54.1	23.8	50.8	61.8
+R=DWR	77.6	53.9	23.0	50.6	61.3

8 可以看到,本文模型在分类和定位任务上均优于 YOLOv8n。首先,如图 8 中前 3 行所示,本文模型正确地分类了举手、睡觉的学生和上课的教师,与 Ground Truth 相同。但 YOLOv8n 模型将其错误地分类为听讲和站立,说明本文模型在类别预测上比 YOLOv8n 更精确。其次,图 8 中第 4 行展示了本文模型成功定位和分类了图像左侧正在读书的学生,但 YOLOv8n 没有检测到这个较小的目标,表明本文模型在小目标检测方面优于 YOLOv8n。此外,与 Ground Truth 相比,本文模型和 YOLOv8n 都给出了多余的标注,将正在进行小组合作的学生检测为读书、听讲等其他课堂行为,因此在定位和分类上还有改进空间。



Fig. 8 Visual comparison of TAR-YOLO and YOLOv8n results on STCB dataset

图 8 STCB 数据集上 TAR-YOLO 与 YOLOv8n 结果可视化对比

总体而言,本文提出的 TAR-YOLO 模型在 STCB 数据集上的实验结果优于其他轻量化模型,在检测精度和轻量化的综合表现上更好。尤其是对小目标的检测中,TAR-YOLO 与基准模型相比效果显著提升。基于以上分析可以得出结论,TAR-YOLO 模型非常适合课堂行为检测任务,并且是一个易于部署在性能受限的设备上轻量化模型。

2.4 PASCAL VOC 数据集实验结果

为了验证本文提出的 TAR-YOLO 模型在不同场景中的泛化能力,在 PASCAL VOC 数据集上作进一步实验。所使用的对比模型与在 STCB 中使用的模型相同,训练周期设置为 300 个 epoch,其它设置均保持不变。如表 4 所示,TAR-YOLO 模型在 mAP 上的表现仍优于基准模型。mAP50 是所有模型中最高的,达到了 64.4%,相比 YOLOv8n 提高了 1.8%,mAP50:95 仅比最高的 MobileNetV2 模型低 0.4%。在对小目标的检测上,TAR-YOLO 的 AP 超过其它所有模型,相比第二高的 FasterNet 高出 1.4%。

Table 4 Experiment results comparison of different models on PASCAL VOC dataset

表 4 不同模型在 PASCAL VOC 数据集上的实验结果比较

Model	mAP50/%	mAP50:95/%	AP _s /%	FLOPs/G	Params/M
GhostNet	58.5	42.7	7.9	6.8	3.8
GhostNetV2	59.4	43.6	9.0	7.3	2.5
MobileNetV2	58.3	41.9	6.7	<u>5.6</u>	6.4
MobileNetV3-Small	63.4	<u>47.7</u>	6.6	6.6	2.9
ShuffleNetV2	50.3	34.7	4.3	5.7	2.0
FasterNet	63.3	46.4	9.8	11.3	4.3
YOLOV5n	63.5	43.6	8.8	7.7	2.6
YOLOv8n	62.6	46.4	9.5	8.1	3.2
YOLOV11n	63.7	46.8	9.7	6.5	2.6
Our model	64.4	47.3	<u>11.2</u>	7.0	<u>1.8</u>

以上实验结果表明,TAR-YOLO 不但能够胜任课堂行为检测任务,在其它场景中也有出色的表现,尤其是针对密集小目标的检测。TAR-YOLO 的计算量和参数量都很少,使其能更好地部署在性能受限的生活场景中。

2.5 消融实验

为了验证本文提出的关键模块对提升模型性能的贡献,依次将 TAPH、MSFAP、R-DWR 集成到基准模型 YOLOv8n 中,研究这些改进策略对模型小目标检测能力和轻

量化表现的提升效果。
实验在 STCB 数据集上进行,添加不同模块前后的对比如表 5 所示。

Table 5 Results of ablation experiment
表 5 消融实验结果

Model	mAP50/%	mAP50:95/%	AP _s /%	FLOPs/G	Params/M
YOLOv8n(Baseline)	76.7	53.1	22.3	8.9	3.2
+TAPH	78.5(↑1.8)	54.5(↑1.4)	24.4(↑2.1)	8.7((↓0.2)	2.4(↓0.8)
+MSFAP	77.9(↑1.2)	54.1(↑1.0)	23.8(↑1.5)	7.8((↓1.1)	3.0(↓0.2)
+R-DWR	77.6(↑0.9)	53.9(↑0.8)	23.0(↑1.0)	8.4((↓0.4)	2.8(↓0.4)
+TAPH+MSFAP	79.1(↑2.4)	54.2(↑1.1)	25.3(↑3.0)	7.6(↓1.3)	2.2(↓1.0)
+TAPH+MSFAP+R-DWR	79.5(↑2.8)	54.9(↑1.8)	25.5(↑3.2)	7.0(↓1.9)	1.8(↓1.4)

TAPH 使用共享卷积提取不同尺度目标的共性特征,其任务对齐的结构通过计算任务交互特征来提高分类任务和定位任务的交互性,从而提升检测头的定位和分类性能,mAP50 与 mAP50:95 相比基准模型分别提升了 1.8% 和 1.4%。与 YOLOv8n 所使用的检测头相比,TAPH 对不同尺度目标共性特征的提取增强了模型对不同尺度目标检测的泛化能力,因此对小目标的检测精度显著提升,AP_s与基准模型相比提升了 2.1%。同时,在检测头中采用共享卷积处理不同尺度的特征图,而不是为每个尺度配备独立的特征提取器。这种共享的特征提取结构减少了模型 0.8 M 的参数量,并降低了 0.2 G 的计算量。

R-DWR 采用两步法提取多尺度特征,提取多尺度特征的过程被分为区域残差和语义残差两个步骤。在区域残差部分进行特征的初步提取,在语义残差部分使用多个不同扩张率的膨胀卷积来提取多尺度特征。两步法使多尺度特征提取过程效率更高。DRB 被用于替换语义残差中原先使用的膨胀卷积,DRB 使用重参数化技术,通过将多个膨胀卷积等效转换为内核更大的非膨胀卷积,从而消除额外扩张的卷积层的推理成本。增加 R-DWR 模块后,与基准模型相比,mAP50 提升了 0.9%,mAP50:95 提升了 0.8%,AP_s提升了 1.0%。模型的参数量降低了 0.4 M,计算量减少了 0.4 G。

在 TAPH 的基础上加入 MSFAP,模型计算量和参数量分别减少了 1.1 G 和 0.2 M,mAP50 提升了 0.6%,mAP50:95 降低了 0.3%,但模型对小目标的检测效果得到提升,AP_s达到了 25.3%。最后,加入 R-DWR。R-DWR 同样提升了模型的检测精度,mAP50 提升了 0.4%,mAP50:95 提升了 0.7%,模型提取多尺度特征的能力进一步增强,AP_s提升至 25.5%。与此同时,重参数化技术在 DRB 中的使用再次减少了模型 0.4 M 的参数量和 0.4 G 的计算量。

综上,本文模型在精度和轻量化表现上都优于 YOLOv8n 基准模型,证明了各个模块在提升模型检测精度、减少模型计算量和参数量方面的有效性。

3 结语

本文提出一种面向课堂行为识别的轻量化目标检测网络 TAR-YOLO,针对教室场景中的遮挡与小目标识别难题,设计了结构新颖的检测头 TAPH,通过共享卷积提取多尺度共性特征,并引入任务对齐机制强化分类与定位之间的信息交互,从而显著提升了小目标检测精度。为适配教室中算力受限的部署环境,TAR-YOLO 集成了轻量化特征融合模块 MSFAP,通过部分卷积降低计算量,并逐层提取多感受野特征,进一步提升了检测性能。在主干网络中,提出两阶段多尺度特征提取模块 R-DWR,结合区域残差与语义残差化结构,有效提升了特征表达能力。同时引入 DRB 模块,用重参数化大核卷积替代传统膨胀卷积,提升模型对稀疏目标的感知能力,并降低推理成本。最后,构建 STCB 数据集,涵盖 7 类学生行为与 3 类教师行为,用于评估模型在真实课堂环境下的表现。实验结果表明,TAR-YOLO 在 STCB 与 PASCAL VOC 数据集上的精度和计算效率方面均表现优异,优于多种现有方法。

总体来看,TAR-YOLO 在小目标检测和模型轻量化方面取得了显著进展,具备良好的教育场景适应性,有望应用于智慧教室,实现课堂行为的自动识别与分析,为教学反馈、课堂管理和教育决策提供数据支持。STCB 数据集为相关研究提供了基础资源,可推动人工智能与教育的融合发展。未来工作将进一步优化 TAPH 结构以提升实时性,并扩展 STCB 数据集的规模与类别,持续提升模型的泛化能力与应用价值。

参考文献:

[1] LI Y T, QI X, SAUDAGAR A K J, et al. Student behavior recognition for interaction detection in the classroom environment[J]. Image and Vision Computing, 2023, 136: 104726.
[2] SU C, WANG G Z. Research on student behavior recognition based on improved OpenPose [J]. Computer Applications and Research, 2021, 38 (10): 3183-3188.
苏超,王国中. 基于改进 OpenPose 的学生行为识别研究[J]. 计算机应

- 用研究, 2021, 38(10): 3183–3188.
- [3] LIN F C, NGO H H, DOW C R, et al. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection[J]. *Sensors*, 2021, 21(16): 5314.
- [4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779–788.
- [5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//*2016 European Conference on Computer Vision*, 2016: 21–37.
- [6] REN X K, YANG D Y. Student behavior detection based on YOLOv4-Bi[C]//*2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering*, 2021: 288–291.
- [7] TAN M X, PANG R M, LE Q V. Efficientdet: scalable and efficient object detection[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 10781–10790.
- [8] WANG Z F, YAO J L, ZENG C Y, et al. YOLOv5 enhanced learning behavior recognition and analysis in smart classroom with multiple students[C]//*2022 International Conference on Intelligent Education and Intelligent Research*, 2022: 23–29.
- [9] ZHAO J H, ZHU H Y. CBPH-Net: a small object detector for behavior recognition in classroom scenarios[J]. *IEEE Transactions on Instrumentation and Measurement*, 2023, 72: 1–12.
- [10] QIU W K, WU X M, LIU X Z. Research on classroom behavior target detection algorithm based on YOLOv7[C]//*2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference*, 2024: 1433–1437.
- [11] XU T, CHEN M Y, CHENG Y L. Smart classroom behavior recognition algorithm based on YOLOv7-SSW model[J/OL]. *Journal of Chongqing Technology and Business University (Natural Sciences Edition)*, 2025–06–05. <http://kns.cnki.net/kcms/detail/50.1155.n.20250527.0949.002.html>.
徐韬, 陈孟元, 程云麟. 基于YOLOv7-SSW模型的智慧课堂行为识别算法[J/OL]. *重庆工商大学学报(自然科学版)*, 2025–06–05. <http://kns.cnki.net/kcms/detail/50.1155.n.20250527.0949.002.html>.
- [12] CAO Y, CAO Q, QIAN C S, et al. YOLO-AMM: a real-time classroom behavior detection algorithm based on multi-dimensional feature optimization[J]. *Sensors*, 2025, 25(4): 1142.
- [13] FENG C J, ZHONG Y J, GAO Y, et al. Tood: task-aligned one-stage object detection[C]//*2021 IEEE/CVF International Conference on Computer Vision*, 2021: 3490–3499.
- [14] CHEN G, WANG H T, CHEN K, et al. A survey of the four pillars for small object detection: multiscale representation, contextual information, super-resolution, and region proposal[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, 52(2): 936–953.
- [15] LIU Y, SUN P, WERGELES N, et al. A survey and performance evaluation of deep learning methods for small object detection[J]. *Expert Systems with Applications*, 2021, 172: 114602.
- [16] DING X H, ZHANG Y Y, GE Y X, et al. Unireplknet: a universal perception large-kernel convnet for audio video point cloud time-series and image recognition[C]//*2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 5513–5524.
- [17] DING X H, ZHANG X Y, HAN J G, et al. Scaling up your kernels to 31x31: revisiting large kernel design in CNNs[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 11963–11975.
- [18] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 1580–1589.
- [19] TANG Y H, HAN K, GUO J Y, et al. GhostNetV2: enhance cheap operation with long-range attention[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 9969–9982.
- [20] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 4510–4520.
- [21] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 1314–1324.
- [22] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//*Proceedings of the European Conference on Computer Vision*, 2018: 116–131.
- [23] CHEN J R, KAO S H, HE H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 12021–12031.

(责任编辑: 黄健)