计算机科学
COMPUTER SCIENCE

CPViG-Net:基于局部跨阶段视觉图卷积的学生课堂行为识别

张浩鹏, 施铮, 刘峰, 宋婉茹

引用本文

张浩鹏, 施铮, 刘峰, 宋婉茹. CPViG-Net:基于局部跨阶段视觉图卷积的学生课堂行为识别[J]. 计算机科学, 2026, 53(2): 57-66.

ZHANG Haopeng, SHI Zheng, LIU Feng, SONG Wanru. CPViG-Net:Students' Classroom Behavior Recognition Based on Cross-stage Visual GraphConvolution [J]. Computer Science, 2026, 53(2): 57-66.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250500100

CPViG-Net:基于局部跨阶段视觉图卷积的学生课堂行为识别

张浩鹏[1]　施　铮[2]　刘　峰[1,2]　宋婉茹[2]
1 南京邮电大学通信与信息工程学院　南京 210023
2 南京邮电大学教育科学与技术学院　南京 210023
(1224014629@njupt.edu.cn)

摘　要　随着教育范式从“人机协同”向“人智协同共育”演进,课堂教学的智能化评价也面临着新的要求和挑战,其中以学生行为为出发点的任务近些年来获得了广泛的关注。针对真实课堂环境中存在的学生行为多样、遮挡频繁及背景干扰严重等问题,提出一种局部跨阶段视觉图卷积模型,旨在提升复杂课堂环境下的学生行为识别精度。该模型以经典目标检测算法为基准框架,通过融合视觉图卷积神经网络的动态特征建模能力,构建了局部最大相对图卷积模块(PMG)与局部跨阶段融合(CPF)模块。其中,PMG模块通过嵌入最大相对图卷积来捕捉节点间特征差异最大的邻域信息,进而针对性地解决局部区域遮挡引起的信息丢失问题,并结合了深度可分离卷积降低图卷积算法的计算开销;CPF模块利用全连接层重构特征结构,并通过C2f模块的跨阶段连接机制,实现多层级的特征融合,从而增强模型对小尺度目标的识别能力。此外,模型通过近邻K值优化,提出针对不同数据集的优化策略。在公开数据集SCB03-S上,CPViG-Net的mAP@50达到70.9%,较基准模型提升2个百分点;在多个公开数据集上的实验表明,该模型在处理真实课堂情境下学生行为识别面临的诸多问题中表现出较好的性能和较高的鲁棒性。

关键词:学生行为;最大相对图卷积;多尺度目标识别;遮挡;深度可分离卷积
中图分类号　TP391

CPViG-Net:Students' Classroom Behavior Recognition Based on Cross-stage Visual Graph Convolution

ZHANG Haopeng[1],SHI Zheng[2],LIU Feng[1,2] and SONG Wanru[2]
1 School of Communication and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
2 School of Educational Science and Technology,Nanjing University of Posts and Telecommunications,Nanjing 210023,China

Abstract　With the evolution of educational paradigms from “human-computer collaboration” to “human-intelligence collaborative co-education”,the intelligent evaluation of teaching is also facing new requirements and challenges. In recent years,the task that takes student behavior as the starting point has gained widespread attention. Aiming at the challenges of diverse student behaviors,heavy occlusions and severe background interference in real classroom environments,a cross-stage partial vision graph network(CPViG-Net) is proposed to enhance the accuracy of student behavior detection in complex classroom settings. Based on a classic object detection framework,the model integrates the dynamic feature modeling ability of the vision GNN and constructs the partial max-relative graph convolution(PMG) module and the cross-stage partial fusion(CPF) module. The PMG module captures the neighborhood information with the greatest feature differences between nodes by embedding maximum relative graph convolution,thereby specifically addressing the issue of information loss caused by local occlusions. It also incorporates depthwise separable convolution to reduce the computational cost of the graph convolution algorithm. The CPF module reconstructs the feature structure using fully connected layers and leverages the cross-stage connection mechanism of the C2f module to achieve multi-level feature fusion,thereby enhancing the ability of the model to recognize small-scale objects. In addition,the model proposes optimization strategies for different datasets through the optimization of nearest neighbor K values. On the public dataset SCB03-S,the mAP@50 of CPViG-Net reaches 70.9%,which is a 2 percentage points improvement over the baseline model. Experiments on multiple publicly available datasets demonstrate that the model exhibits good performance and high robustness in addressing the various challenges of student behavior recognition in real classroom scenarios.

Keywords　Student behavior,Max-relative graph convolution,Multi-scale object recognition,Occlusion,Depthwise separable convolution

到稿日期:2025-05-26　返修日期:2025-09-15
基金项目:国家自然科学基金(62307025,62177029);南京邮电大学2023教改项目(JG01723JX71)
This work was supported by the National Natural Science Foundation of China(62307025,62177029) and Project of EducationTeaching Reform of Nanjing University of Posts and Telecommunications(JG01723JX71).
通信作者:宋婉茹(songwanru@njupt.edu.com)

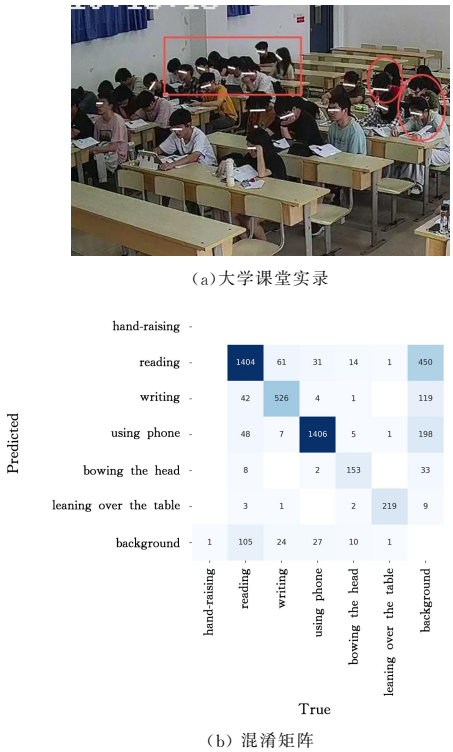
1 引言

随着科学技术的飞速发展，在国家政策的大力支持下，人工智能和大数据分析等新兴技术为教育事业注入了全新活力[1]。教育与技术的深度融合推动教学管理向智慧化转型，尤其是教育人工智能已步入人智协同阶段[2-3]。从底层数据支持的角度来说，教育人工智能的发展依托于对教育数据的挖掘，即从海量教育数据中提取有价值的信息，助力教育的智能化转型[4]。课堂教学质量的智能评估，是教育智能化的重要方向之一。在智能检测与识别等技术的加持下，对课堂教学效果实现自动评估，进而实现教学信息的及时反馈、教学方案的持续优化、课堂教学效果的不断提升这一闭环。传统教学质量评估依赖于人工评判，专家经验指导下的公开课或课堂视频观察与分析是常见的方式[5]。然而，这种评估方式存在诸多弊端，如耗费大量时间精力，评估维度单一，评价结果受主观影响很大，且无法量化分析等[6]。随着教育数字化的演进，智能化的技术方案被引入到课堂教学质量评价中。课堂教学过程中产生了大量图像、语音、文字等多模态数据，这些信息通过机器学习或深度学习算法的处理，可实现实时且自动的识别、分类和分析。其中，师生的课堂行为是反映课堂教学质量的重要标志。因此，以课堂行为为切入点，以技术手段实现师生课堂行为的识别和分析，已成为课堂教学质量智能化评价领域的研究热点之一[7-8]。

机器学习提供了广泛的聚类、分类算法，因此常被应用于课堂行为识别任务中[9-10]。而这些方法仍须人工预处理课堂视频等数据，无法满足课堂环境实时监测与分析的需求，仍需更高效、智能的解决方案。深度学习的发展为解决这一问题带来了新的思路和方法。以卷积神经网络（Convolutional Neural Network，CNN）为主要范式的方法将课堂教学质量评价落脚于课堂行为检测与识别[11-12]，再依据师生行为模型[13]进行分析。目前针对课堂行为检测与识别这一任务的典型解决方案，是将计算机视觉中的目标检测算法迁移过来。按照算法执行任务的方法与结构的不同，目前用于目标检测的模型主要分为两大类。1）两阶段算法，以 Faster R-CNN 为典型代表[14-16]。这类算法在第一阶段聚焦于生成候选区域，第二阶段专注于对这些候选区域进行分类和精确定位，具有精度高、处理时间长的特点。2）一阶段算法，以 Yolo 系列[17-8]和 SSD[19]为代表。该类算法执行端到端的处理，从整幅图像提取特征，直接输出位置和类别信息，具有速度快、效率高的显著优势。因此，一阶段的目标检测与识别算法更加符合课堂行为识别对实时性的需求。然而，这些算法依旧未能突破 CNN 在跨视频帧捕捉对象关系时的局限性[20]。

为了解决上述问题，鉴于图神经网络（Graph Neural Network，GNN）在动态建模时空交互方面具有独特优势，可以将其引入目标检测和行为识别任务中。GNN 自提出以来，便被广泛应用于社交网络分析、推荐系统和知识图谱等领域[21-22]。随着图神经网络可视化技术的不断发展，GNN 强大的图结构处理能力被进一步挖掘，为解决视觉任务提供了新的思路[23]。在行为识别领域，采用 OpenPose[24]等姿态估计算法与图卷积神经网络（Graph Convolutional Network，GCN）可以更好地提升模型的有效性[25-27]。此外，Vision GNN（ViG）[28]的出现为实现实时的检测任务提供了新的思路与可能性。如 Soudeep 等结合动态图卷积与 Yolo v11，实现了复杂路面小物体的目标检测与追踪[29]。在课堂行为检测与识别任务中，有研究针对教师课堂行为分析，提出采用多尺度特征 GCN 来更准确地对教师的教学行为进行捕获和分析[25]。

受到上述研究的启发，本文着眼于真实课堂教学环境，以学生为研究对象，尝试以 Yolo 为基准实现对学生课堂行为的检测与识别，联合 ViG 助力课堂教学质量的智能化评估。然而，如图 1(a)所示，真实的课堂环境具有较高的复杂性，如：1）拍摄设备分辨率有限，且拍摄角度固定；2）学生座位排布密集，存在相互遮挡的问题；3）课堂环境光线的明暗变化以及色彩的显著差异等。图 1(a)中，方框内的学生人脸因分辨率限制而模糊不清；椭圆框内的学生由于被遮挡，较难获取其完整的行为信息。因此，面向真实课堂环境的学生行为检测与识别，是一项极具挑战性的任务。此外，分析实际检测结果发现，Yolo 系列算法在处理跨视频帧的行为识别任务时，背景中冗余元素的干扰导致漏检与错检问题频发。图 1(b)展示了 Yolo v8 在学生行为数据集上的训练结果错误分析，可视化结果清晰显示，预测错误主要源于对背景物体的干扰性识别。上述结果，进一步佐证了 Yolo 算法在跨视频帧对象识别时存在背景语义混淆导致的检测性能局限。



(a)大学课堂实录

(b) 混淆矩阵

图 1 真实课堂情境下学生行为识别存在的问题

Fig. 1 Problems of student behavior recognition in real classroom situation

事实上，课堂行为识别任务场景复杂，不仅存在听讲、书写、走神等多种学生行为类别，还存在遮挡、人物尺度大小不一致、背景干扰等问题。受文献[25-26]提出的视觉任务图结构建模思想启发，本文将特征图分解为不重合的图像块作为

图的节点,将图像块的差异度定义为边,将 ViG 引入到本任务中。ViG 具备强大的能力来有效捕捉各类节点与边的特征信息,并实现对这些异构信息的深度融合。同时,鉴于课堂行为呈现为动态的时间序列过程,ViG 能够借助动态图结构,依据视频帧的持续输入,动态地更新图结构。这种动态更新机制,使得 ViG 在复杂的课堂场景中能够精准捕捉并分析时空交互目标,进而实现对课堂行为的良好检测。

综上所述,通过联合 Yolo 与 ViG,本文提出了一种局部跨阶段视觉图卷积神经网络(Cross-stage Partial Vision Graph Network,CPViG-Net)的框架。一方面,利用 Yolo 在快速目标检测与定位方面的高效性;另一方面,借助 ViG 弥补 Yolo 存在的不足,提升模型的准确性和鲁棒性。考虑到 Yolo v8 在目标检测和课堂行为分析等任务上的表现较好,CPViG-Net 采用 Yolo v8 的检测头来满足本研究对速度和精度的基本需求。网络的主干部分联合最大相对图卷积(Max-Relative Graph Convolution,MRConv)和 C2f(Cross Stage Partial Layer with Two Convolution)[30]构建了局部最大相对图卷积(Partial Max-relative Graph,PMG)模块和局部跨阶段融合(Cross-stage Partial Fusion,CPF)模块,将 C2f 与 ViG 网络提取的特征有效结合。C2f 的残差结构擅长从图像中提取局部空间信息;ViG 倾向于捕捉课堂图像中节点间的长距离依赖和全局结构,在挖掘潜在行为信息方面具有优势。CPViG-Net 结合二者特性,充分挖掘图像中的细节信息和节点间关系蕴含的潜在信息,进而有效应对真实课堂情境中学生行为识别面临的各种挑战。此外,CPViG-Net 在轻量化与模型参数方面进行了部分优化:1)考虑到图卷积的计算量较大,在 PMG 模块采用深度可分离卷积(Depthwise Separable Convolution,DWConv)替代部分标准卷积,该卷积结合了深度卷积和逐点卷积,在保证精度的基础上减少了运算量和参数量;2)通过对近邻 K 值的优化,探索 K 值大小与不同数据集的适配规律,进一步提升了模型准确率。本文的主要贡献总结为以下几个方面。

1)针对课堂行为识别任务,提出一种局部跨阶段视觉图卷积神经网络 CPViG-Net。CPViG-Net 以 Yolo v8 为主干网络,通过嵌入局部最大相对图卷积模块 PMG 和局部跨阶段融合模块 CPF,从空间维度与节点信息两个层面提升对学生行为特征的挖掘能力,有效缓解真实课堂环境面临的遮挡和背景干扰等问题。

2)设计了一种局部最大相对图卷积(PMG)模块,利用 MRConv 捕获邻近节点的特征,捕捉节点间信息,更好地融合该任务中的动态异构信息;将 DWConv 嵌入 PMG,通过减少卷积核数量,缓解图卷积带来的模型参数量显著增长问题。

3)设计了一种局部跨阶段融合模块 CPF,通过采用全连接层改进了 C2f 结构,实现了对 PMG 模块输出特征的重构,增强了其对空间特征的提取能力,实现了不同感受野特征的融合,提高了模型对不同尺度特征的识别能力。

4)在公开数据集 SCB03-S 上进行实验,结果表明,CPViG-Net 在参数量和准确率方面相较于现有 Yolo 系列算法具有一定优势,充分验证了其优越性。针对 SCB03-S 与 SCB03-U 数据集,提出近邻 K 值优化策略,研究图卷积在不同特征分布任务上的作用机制,为参数优化提供合理依据。

2 相关工作

课堂行为检测与分析是教学质量智能化评估的重要手段之一。根据技术手段的不同,其可以分为传统课堂行为分析方法、基于 Yolo 的课堂行为识别方法,以及基于图机器学习的视觉处理方法。

2.1 传统课堂行为分析方法

为了提高课堂行为分析的有效性,文献[9]采取改进型弗兰德斯编码理论对 23 个课堂实录视频编码,经编码类别统计后,实现基于机器学习的课堂师生互动风格分类,挖掘出不同模式的教学课堂存在的潜在问题。文献[31]提出了一种基于可解释性机器学习的课堂行为分析方法,利用随机森林算法对多模态数据集进行训练,再分析各学生行为指标之间的相关性,筛选出对教学效果影响显著的关键特征,以构建预测模型。文献[32]基于 BP 神经网络、K-means 聚类和 SVM 等多种机器学习算法,提出了一种基于机器学习的智能评估系统,实现了职业教育中的英语教学质量的自动化评估,为教学评估提供了新的角度。这些方法侧重于对已有特征的分类归纳,揭示特征间的关联性。然而,在特征检测方面,传统机器学习对图像的处理速度已不能满足课堂行为识别所需要的高精确度与及时处理能力的需求。

2.2 基于 Yolo 的课堂行为识别方法

基于 Yolo 的检测算法具有快速处理图像的能力,能够实时分析上课时学生的行为,因此被广泛地应用于课堂行为识别任务中[7-8,33]。文献[7]针对课堂图像中目标尺度不一致和遮挡的问题,在 Yolo 的基础上提出了一种轻量级非对称检测头网络。该网络使用分组卷积将输入特征图划分为多个子组,并结合多个大小不同的卷积与坐标注意力,在每个子组内单独进行操作,显著提高了对课堂中多尺度,特别是小尺度目标的识别能力。文献[8]结合了坐标注意力与 Yolo v9,构建了实时多尺度课堂行为识别系统。文献[33]以 Yolo v8 为基础,通过引入动态卷积改进了 C2f 模块,利用双向特征金字塔网络和全局局部空间聚合机制优化了 Neck 层,并在 Backbone 部分融入了高效局部注意力机制。此外,作者还对检测头进行了轻量化设计。这些改进在提升模型精确度的同时,维持了模型参数量不变,实现了模型轻量化与高性能的平衡。

尽管上述研究取得了进展,但是对于跨视频帧对象的检测,Yolo 仍然存在局限性。除此之外,在模型结构中堆叠许多不同的模块会增加计算复杂度,减缓运行速度。

2.3 基于图机器学习的视觉处理方法

GNN 在处理时空交互的目标时,通过将目标作为节点,空间信息作为边,并以递归的形式对目标的时空状态进行更新,更好地完成检测工作。文献[25]提出了一种基于 GCN 的多尺度特征图卷积网络与滞后序列分析法,先以 OpenPose 算法提取教师骨架序列数据,再通过图卷积网络识别行为,最后用滞后序列分析法评估,以此实现教师教学行为的识别与分析。文献[26]构建了一种基于信息瓶颈理论的骨架行为识别框架 InfoGCN,通过设计信息瓶颈目标来学习紧凑且信息

量最大的潜在表示，利用自注意力图卷积模块动态推断关节的内在拓扑结构，并引入多模态骨架表示方法以提供丰富的空间信息，显著提升了基于骨骼的人体动作识别的准确率。文献[34]将稀疏视觉图注意力机制嵌入到 MobileNet+ViG 的架构中，先利用 SVGA 固定图结构，以滚动操作替代 KNN 计算和输入重塑，降低计算开销；再将 SVGA 与最大相对图卷积、MobileNetv2 的倒残差块等结合，构建 MobileViG；最终，实现图像分类、目标检测和实例分割等移动视觉任务的高效处理。

在视觉领域，图机器学习与其他模型的融合已展现出巨大的潜力。因此，提出融合 Yolo 与 ViG 的方案，将 Yolo v8 高效准确的目标检测能力与 ViG 处理复杂视觉关系和多尺度特征的优势结合，旨在解决课堂行为图像中存在的光暗不均、遮挡严重等问题。

3 CPViG-Net 算法

首先，对 CPViG-Net 模型的网络结构进行介绍；接着，重点阐释图卷积特征提取模块 PMG，其通过构建图结构，将图片数据转换为节点与边进行处理；最后，分析 CPF 模块的多感受野特征融合能力。

3.1 CPViG-Net 模型概述

为了有效应对课堂行为识别的挑战，研究构建了 CPViG-Net。该网络由主干网络(Backbone)、颈部网络(Neck)和检测头(Head)3 部分组成，如图 2 所示。

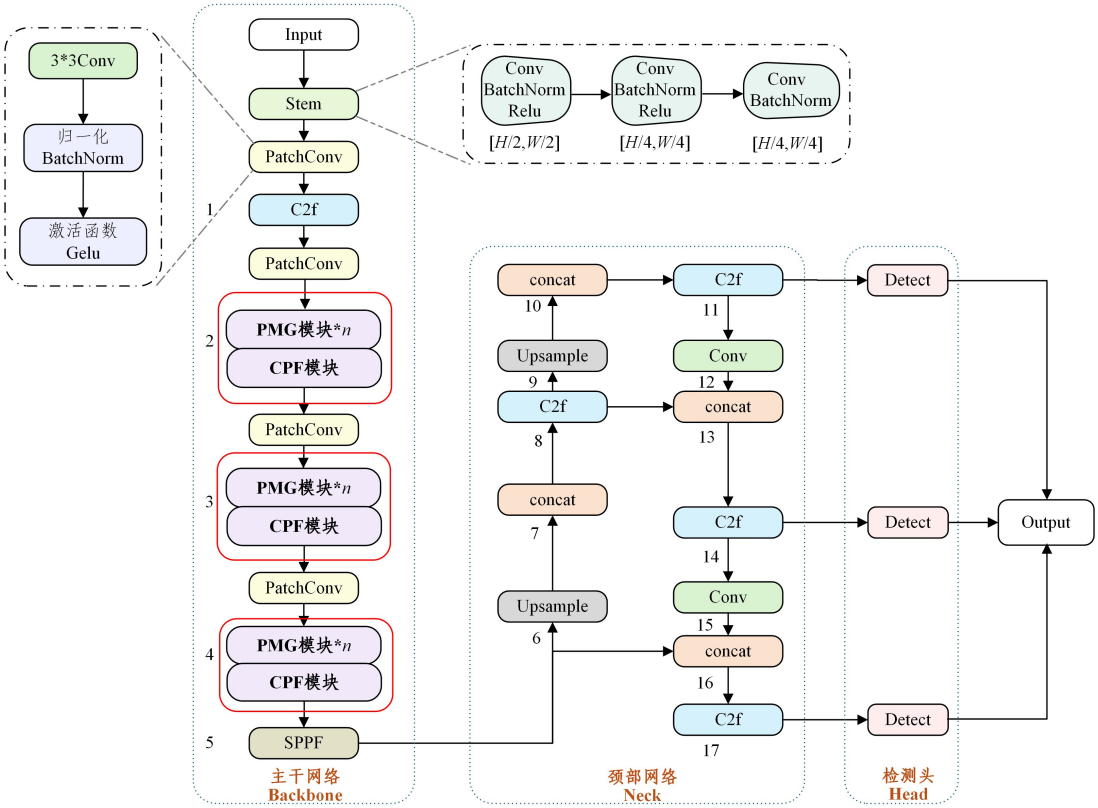


图 2 CPViG-Net 的结构

Fig. 2 Framework of CPViG-Net

1)主干网络由图像分割阶段 Stem、下采样阶段 PatchConv 以及特征提取阶段共同构成，其中特征提取阶段囊括了所提出的 PMG、CPF 以及经典的空间金字塔快速池化模块(Spatial Pyramid Pooling-Fast，SPPF)。Stem 包含二维卷积核和归一化层，输入图像经过 Stem 被分割成不重叠的图像块，每个图像块被转换为固定维度的嵌入向量，方便后续的特征提取。PatchConv 的作用是将图像大小从 $H\times W$ 压缩为 $H/2\times W/2$，再输入特征提取阶段。该阶段嵌入了所提出的两个模块 PMG 和 CPF。

值得注意的是，首个特征提取阶段直接采用了 Stem 联合 C2f 结构，这是因为若直接采用 PMG 模块，将显著增加计算复杂度与资源开销。PMG 模块以像素点作为基础节点，逐维度计算每个节点与其邻域节点的特征差异，从中筛选出最大差异的节点，进而在当前维度聚合特征。这种基于节点相似性的特征提取方式虽然能够有效捕捉局部特征信息，但计算开销大。假设有 n 个节点，每个节点的维度为 d，则特征矩阵的规模为 $n\times d$，其存储空间需求随 n 线性增长。由于节点间的关系不具备方向性，所构建的图结构为无向图，其邻接矩阵规模为 $n\times n$，当 n 增加时，邻接矩阵的元素数量呈平方级增长。尽管采用稀疏矩阵存储策略可有效减少零元素的冗余存储，但在第一阶段，每张图有约十万个节点，表征节点间关联的非零元素数量依然巨大，导致存储邻接矩阵所需的计算开销急剧上升。在模型训练阶段，当节点数量 n 增加时，计算梯度所涉及的节点与边的数量相应增多，不仅大幅提升了计算复杂度，还会产生大量中间梯度结果。这些中间结果在参数更新完成前须暂存于显存，进一步加剧了显存的消耗，对硬件存储能力提出了更高要求，并且对准确率没有很大影响，所以在第一阶段使用 C2f 模块进行特征提取。

主干网络的第二、第三、第四阶段使用了堆叠的 PMG 模块提取特征。该模块通过 MRConv，获取与各节点差异最大的邻居特征进行融合。随后，通过单个 CPF 模块提升特征增强与多维感受野特征的融合效率。PMG 模块与 CPF 模块的嵌入方式如图 2 所示，具体的介绍见 3.2 节和 3.3 节。最后，将特征输入 SPPF 模块，融合和提取尽可能多的高层语义特征。

2）颈部网络作为目标检测框架的核心组件，承担着聚合主干网络输出特征的关键任务。CPViG-Net 所采用的颈部网络架构继承自 Yolo v8 的路径聚合网络（Path Aggregation Network，PANet），通过独特的双路径结构，一方面借助自顶向下的路径，将深层特征图所蕴含的高层语义信息传递至浅层特征图，使浅层特征图具备更强的语义理解能力；另一方面，利用自底向上的路径，把浅层特征图中丰富的细节信息传递给深层特征图，让深层特征图能够捕捉到更多目标的细节特征。将 2，3，4 层提取的特征进行融合，构建多尺度、多维度的特征表示，有效整合细节与语义信息，提升模型对真实课堂这一复杂场景的适应性。

3）检测头网络通过 3 个并行检测头 Detect 实现多尺度特征处理。每个检测头基于级联的 3×3 与 1×1 卷积操作，通过多次遍历逐层融合不同特征层的信息。该结构在保证检测精度的前提下，通过高效的特征传递与聚合机制，增强了网络对多尺度目标的适应性，显著提升了模型的检测鲁棒性。

3.2 PMG 模块

针对课堂背景噪声对学生行为识别造成干扰的问题，联合 DWConv 与 MRConv 提出了一种 PMG 模块，其结构如图 3 所示。该模块通过构建轻量化的特征提取机制，实现了对时空交互目标的动态特征的有效捕捉，进而减少背景信息对行为特征的干扰，显著增强了模型在复杂教学场景下的鲁棒性，提升了课堂行为识别的精度。

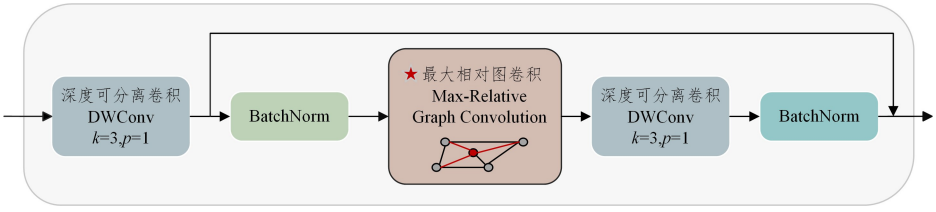


图 3 PMG 模块的结构

Fig. 3 Structure of PMG module

PMG 模块基于残差架构的设计，首先利用一个 3×3 DWConv 将输入通道数倍增，随后使用 MRConv 对归一化数据进行特征提取，最后通过第二个 3×3 DWConv 恢复原始通道维度，并强化特征。经过归一化与激活函数处理后，模块将输出特征与原始输入逐元素相加，构建残差结构，在避免梯度消失的同时，将原始特征与提取后包含长跨度信息的特征融合，增强了模型对全局与局部特征的表达能力。DWConv 将标准卷积拆成了逐通道卷积与逐点卷积，卷积核大小为 w 的 DWConv 与相同卷积核大小的标准卷积的计算量如式（1）和式（2）所示：

$$P_1 = C_{in} * w^2 + C_{in} * C_{out} \tag{1}$$

$$P_2 = C_{in} * w^2 * C_{out} \tag{2}$$

其中，C_{in} 代表输入通道数，C_{out} 代表输出通道数，P_1 表示 DWConv 的计算量，P_2 表示标准卷积的计算量，卷积核大小 $w=3$。因为 $C_{in} * C_{out}$ 的值远大于 $C_{in} * w^2$ 的值，所以可以推断出 P_1 远小于 P_2，采取 DWConv 可以有效减少计算量。

考虑到每张图片都由 Stem 划分为互不重叠的图像块，每个图像块可被看作一组特征向量 $\boldsymbol{X}=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3,\cdots,\boldsymbol{f}_n\}$，每一个特征向量 $\boldsymbol{f}_i$ 都可以表示为 $\boldsymbol{f}_i \in R^D$，D 是特征维度，$i=1,2,3,\cdots,n$。$\boldsymbol{f}_i$ 经过 DWConv 和归一化处理的过程可以表示为：

$$\boldsymbol{f}_i^* = BN(DW(\boldsymbol{f}_i)) \tag{3}$$

其中，$BN(\cdot)$ 表示对数据进行归一化，$DW(\cdot)$ 表示对数据进行 DWConv 操作。

接着，将输出的 $\boldsymbol{f}_i^*$ 输入 MRConv 进行特征提取，其原理图如图 4 所示。以图 4 中的图像块 $\boldsymbol{f}_3$ 为特征节点为例，MRConv 提取节点特征的流程如式（4）—式（6）所示：

$$\boldsymbol{f}_i' = RE(comb(\boldsymbol{f}_i^*, N(\boldsymbol{f}_i^*)), W_{co}, w_{re}) \tag{4}$$

$$comb(\cdot) = [\boldsymbol{f}_i^*, \max(\boldsymbol{f}_j^* - \boldsymbol{f}_i^* \mid \boldsymbol{f}_j^* \in N(\boldsymbol{f}_i^*))]W_{co} \tag{5}$$

$$RE(\cdot) = w_{re} * comb(\cdot) \tag{6}$$

其中，$RE(\cdot)$ 表示更新与中心特征聚合的邻居特征；$comb(\cdot)$ 表示通过聚合 K 个最近邻居的特征来增强中心节点特征的聚合操作；W_{co} 和 w_{re} 是用于聚合和更新的可学习权重；$N(\boldsymbol{f}_i^*)$ 为 $\boldsymbol{f}_i^*$ 的邻居节点，例如 $\boldsymbol{f}_3$ 的邻居节点为 $\boldsymbol{f}_1$，$\boldsymbol{f}_4$，$\boldsymbol{f}_6$，$\boldsymbol{f}_7$，$\boldsymbol{f}_9$。

MRConv 计算中心节点与其每个邻居的每维差异，并在每个维度上将最大差值和中心节点的特征拼接。如图 4 所示，$\boldsymbol{f}_1$，$\boldsymbol{f}_6$ 和 $\boldsymbol{f}_7$ 以不同的线连接中心节点，代表在不同的维度与中心节点进行特征融合；$\boldsymbol{f}_4$ 和 $\boldsymbol{f}_9$ 代表未与中心节点特征融合的节点。用计算最大差异代替计算空间位置差异，使得目标在局部区域被遮挡时，依然可以捕捉到未被遮挡部分的结构特征，有效解决了课堂遮挡严重的问题。

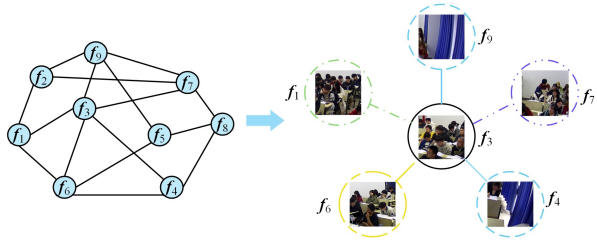


图 4 MRConv 的原理

Fig. 4 Principles of MRConv

3.3 CPF 模块

PMG 模块的输出成为下一个 PMG 模块的输入，或者在 n 次循环后输入 CPF 模块中。但近邻值为 K 的 PMG 模块输

出的是维度为 n 的特征向量，而 C2f 模块的输入为 $H\times W$ 大小的特征图。为了解决 MRConv 输出的数据格式与 C2f 模块所处理的数据格式不匹配的问题，在构建 CPF 模块时，引入了两个全连接层组成的残差结构，旨在重构数据格式，同时增强输出特征的表达能力，进而实现不同模块间数据的有效衔接与特征强化。

C2f 模块通过构建跨阶段连接机制，实现早期低层次细节特征与后期高层次语义特征的融合，显著提升了模型对多尺度目标的检测能力。同时，通过调整 K 的大小，可以扩大或缩小 PMG 模块的感受区域，为 C2F 模块提供更适配的输入特征，进一步优化其对复杂场景下目标特征的提取与表达效率。CPF 模块的流程可以总结为：

$$X'=C2f(BN(\partial(X*w_1)*w_2)+X) \tag{7}$$

其中，∂ 代表 Gelu 函数，w_1 和 w_2 是全连接层的权重。

在课堂环境中，人物尺度大小不一，C2f 模块通过跨阶段连接，实现浅层网络高分辨率细节特征与深层网络强语义特征的直接融合，增强小目标特征表达，在保留目标细节信息的同时，提升语义可区分性，从而显著提升复杂课堂环境下小尺度人物的检测精度。CPF 的输出在经过 PatchConv 采样后，成为下一个阶段的 PMG 模块的输入。使用 PMG+CPF 模块提取的特征既包含局部信息，又包含长期相关性，相比单独的 C2f 或 MRConv，能更有效地对图像特征进行处理。

4 实验及结果分析

4.1 数据集和实验设置

实验主要在公开数据集 SCB03-S 和 SCB03-U[35] 上进行。SCB03-S 收集了从幼儿园到高中的近千个视频，并从每个视频中选取了 3 到 15 个具体的行为视频帧，包括 4200 张图像，25000 个标签，覆盖举手、写字、阅读等典型中小学生课堂行为，并从多角度呈现不同的行为特征。SCB03-U 涵盖了举手、书写、阅读、使用手机、伏案学习和低头这 6 种大学课堂的典型行为，由 671 张图像和 19768 个标签组成。

模型使用 PyTorch 框架开展训练，其版本为 2.2.2。GPU 配备两张显存为 48GB 的 NVIDIA L20 显卡。实验参数设置如下：训练迭代周期（Epochs）设定为 300 轮，以促进模型充分收敛；单批次处理图像数量为 16，输入图像分辨率统一调整至 640×640；采用随机梯度下降算法（Stochastic Gradient Descent，SGD）作为优化器，初始学习率设为 0.01，权重衰减系数为 0.0005，学习率动量参数配置为 0.937。此外，为有效规避过拟合风险，实验引入早停机制：若验证集损失在连续 50 个 epoch 内未呈现显著下降趋势，训练进程将自动终止，并保存当前最优模型作为最终输出，以此平衡模型训练效率与泛化性能。

4.2 评价指标

本文实验主要通过以下 3 个指标来评估模型性能：平均精度（mean Average Precision，mAP）、参数量（Param）、浮点运算次数（GFLOPs）。mAP 由精确度（Precision，P）和召回率（Recall，R）求得。

$$P=\frac{TP}{TP+FP} \tag{8}$$

$$R=\frac{TP}{TP+FN} \tag{9}$$

$$AP=\int_0^1 P(R)\mathrm{d}r \tag{10}$$

$$mAP=\frac{1}{N}\sum_1^N AP_i \tag{11}$$

其中，TP 代表预测正确的正样本数量，FP 代表预测错误的正样本数量，FN 代表预测错误的负样本数量。

以 R 为横轴、P 为纵轴画出 PR 曲线，并对该曲线求积分，所得值即为每个类别的 AP 值。mAP 值便是所有类别 AP 值的均值。本文选取 mAP@50%来衡量 mAP 值，mAP@50%指的是在 IoU 阈值设定为 0.5 时计算得到的 mAP 值，它反映了在这一特定 IoU 水平下模型的识别能力。

Param 即模型可训练参数总数，训练时通过优化算法更新，以最小化损失函数。更多的参数赋予模型更强的特征捕捉能力，但也会提升过拟合风险，因此需要平衡模型参数量与泛化性能。

GFLOPs 为每秒的浮点运算次数，是衡量模型计算复杂度与硬件需求的重要指标。该数值直接反映模型运行时的计算负荷，其值越低，意味着模型运行所需的计算资源越少，对硬件的性能要求也相对较低，更易于在资源受限的设备上高效部署。

4.3 对比实验

为了验证算法的性能优势，在 SCB03-S 数据集上与 Yolo 系列代表算法进行了对比，包括 Yolo v5，Yolo v8，Yolo v10，Yolo v11 和 Yolo v12，且使用 FasterNet[36] 和 StarNet[37] 的主干分别改进 Yolo v8 的主干部分，与 CPViG-Net 对比。对比实验结果如表 1 所列。

表 1 对比实验结果

Table 1 Results of comparative experiments

Dataset	Algorithm	*mAP*@50	*GFLOPs*	*Param*
SCB03-S	Yolo v5	0.683	7.1	2.5×10^6
	Yolo v8(baseline)	0.689	8.1	3.0×10^6
	Yolo v8-FasterNet	0.687	10.7	4.1×10^6
	Yolo v8-StarNet	0.647	6.9	2.5×10^6
	Yolo v10	0.671	6.5	2.3×10^6
	Yolo v11	0.643	6.3	2.5×10^6
	Yolo v12	0.678	5.8	2.5×10^6
	CPViG-Net(K=11)	0.709	10.6	4.0×10^6

从表 1 中可以看出，Yolo v8 在课堂行为识别任务上拥有较好的性能表现。相较于新版本的 Yolo 模型，其准确率更高；不同于 Yolo v5 中 C3 结构，C2f 采用了多支路方式替代 C3 的两条支路方式，仅需堆叠一次即可获取多尺度特征，更契合本文的设计。因此，选择 Yolo v8 作为基准网络 baseline。实验结果表明，在针对课堂学生行为识别的任务中，CPViG-Net 展现了在性能与效率间的高效平衡。在精度检测方面，CPViG-Net 的 mAP 值达到了 70.9 %，较 Yolo v8 有 2 个百分点的提升，超越了 Yolo v5 的 68.3%及其他同类检测模型，展现出较强的行为识别能力；在计算速度方面，FasterNet通过引入部分卷积（PConv）技术，优化计算流程，实现了推理速度的提升，但是精度较 baseline 略有下降，且较 CPViG-Net 降低 2.2 个百分点。在轻量化方面，StarNet 利用

星型拓扑结构，大幅减少了参数量，却伴随显著的精度降低；相比之下，CPViG-Net 在参数量和精度方面取得了理想平衡，综合优势明显。此外，与 Yolo v10，Yolo v11 和 Yolo v12 相比，CPViG-Net 的准确率具有大幅度提升。上述实验结果表明了所提出的模型在真实课堂情景下的学生行为识别任务中的优越性。

4.4 消融实验

为深入剖析本文所设计模型各组成部分的有效性与合理性，本文开展了一系列消融实验。实验以 Yolo v8 为基准网络，在数据集上对各个改进模块分别进行训练与测试。为精准地评估每个模块对模型性能的贡献，本组实验引用了计算量(GFLOPs)、参数数量(Parameters)、mAP@50 和显存占用(GPU Memory)4 个指标进行比较。考虑到显存占用在实验过程中处于动态波动状态，因此实验结果中所呈现的显存占用数据为其均值。考虑到实验的计算开销与变量控制，选取邻居值 $K=9$，实验结果如表 2 所列。

表 2 消融实验结果

Table 2 Results of ablation experiments

MRConv	PMG (first stage)	C2f (first stage)	PMG (other stage)	CPF	*mAP*@50	*GFLOPs*	*Param*	*GPU_Mem*/GB
					0.689	8.1	3.0×10^6	2.54
√					0.695	14.5	5.7×10^6	63.55
	√		√	√	0.701	10.7	4.1×10^6	33.67
		√	√		0.689	13.3	5.1×10^6	11.70
		√	√	√	0.705	10.6	4.0×10^6	9.60

观察表 2 可知：

1)相较于基准网络，虽然直接使用 MRConv 提取特征可以将 mAP 值提高 0.6 个百分点；但是参数量和显存占用急剧增大，尤其是显存占用，在训练的 batchsize 仅为 4 的情况下，显存占用 63.55 GB，若 batchsize 取 16，显存占用将会大于 96 GB，超过两张 L20 的可训练范围。对于课堂行为识别来说，显然是不可取的方法。

2)引入 PMG 模块与 CPF 模块，在第一到第四阶段(Stage)均使用 PMG 模块与 CPF 模块结合的方式提取特征，mAP 较基准网络提高 1.2 个百分点，相较于直接使用 MRConv，显存占用减小接近 50%，这一结果验证了 PMG 模块和 CPF 模块的有效性。然而，显存占用仍处于较高水平，故不适合普遍推广。

3)采用本文提出的组织架构，在 2)的基础上将第一阶段的特征提取方式改为 C2f，即最大的特征图不使用 PMG 处理，仅在第二、第三、第四阶段使用 PMG 模块与 CPF 模块结合提取特征，所得到的 *mAP* 值对比基准网络提高 1.6 个百分点，且显存占用仅为使用 MRConv 方法的 1/6。

4)为了验证 CPF 模块存在的必要性，实验删除了 CPF 模块，并将 PMG 模块的堆叠次数加 1，对比 CPViG-Net，准确率明显下降，且参数量和 GFLOPs 分别上涨 1.1×10^6 和 2.7 GB。

上述实验充分表明了 CPViG-Net 模型在课堂行为识别场景下的有效性，且其符合识别普及应用的需求，为智慧课堂的构建和教育智能化推广提供了技术支持。

4.5 参数 *K* 的优化

在模型构建过程中，PMG 模块输出的数据是特征维度为 n 的特征向量，而 CPF 模块的输入、输出数据均是结构为 $H\times W$ 的特征图，二者连接使用需要重构特征。重构过程仅将长度为 n 的向量转换为大小为 $H\times W$ 的矩阵，由于缺乏明确的空间映射关系与语义约束，目前难以判断重构后的特征是否真实反映了现实场景中的图像信息。针对这些问题，本文提出了对近邻 K 值的优化，利用 K 值的变化，调整 PMG 模块获取的语义信息，减小重构时造成的语义损失，并在 SCB03-S 与 SCB03-U 上进行测试，结果如表 3 和表 4 所列。

表 3 SCB03-S 数据集上 K 值的优化结果

Table 3 Results of K-value optimization on SCB03-S

Dataset	K	*mAP*@50	*GFLOPs*	*Param*
SCB03-S	7	0.693	10.6	4.0×10^6
	9	0.705	10.6	4.0×10^6
	10	0.702	10.6	4.0×10^6
	11	0.709	10.6	4.0×10^6
	12	0.690	10.6	4.0×10^6
	13	0.698	10.6	4.0×10^6
	9,10,11	0.699	10.6	4.0×10^6
	11,10,9	0.689	10.6	4.0×10^6

表 4 SCB03-U 数据集上 K 值的优化结果

Table 4 Results of K-value optimization on SCB03-U

Dataset	K	*mAP*@50	*GFLOPs*	*Param*
SCB03-U	4	0.923	10.6	4.0×10^6
	5	0.934	10.6	4.0×10^6
	6	0.942	10.6	4.0×10^6
	7	0.936	10.6	4.0×10^6
	8	0.934	10.6	4.0×10^6
	9	0.856	10.6	4.0×10^6
	5,6,7	0.938	10.6	4.0×10^6
	7,6,5	0.938	10.6	4.0×10^6

由表 3 和表 4 可得，改变 K 值并没有引起参数量的改变。这是因为，改变 K 值仅改变了每个节点所选取的最近邻居节点数量，虽然 K 值的增大会增加计算距离或者索引等张量间的运算，但并未引入新的网络层或改变现有层结构。实验结果进一步表明，K 值与语义信息间不存在线性关系。对于 SCB03-S，当 K 从 7 增大到 11 时，mAP 值逐渐增大；当 $K\geqslant12$ 时，mAP 值陡然下降。而对于 SCB03-U，当 K 从 4 增大到 6 时，mAP 值呈上升趋势；当 $K\geqslant7$ 时，mAP 逐渐下降。由此可见，并不存在一个万能的 K 值适配所有任务。

对 SCB03-S 和 SCB03-U 数据集的标签分布统计显示，前者聚焦中小学生课堂，学生座位密集，行为趋同；后者以大学生课堂为主，学生分布松散，行为差异显著。PMG 模块的中心节点在学习特征时，保留与邻居节点每个维度的最大差异。当中心节点与邻居节点行为特征趋同时，较大的 K 值可

以帮助中心节点学习到较为完备的特征。例如，SCB03-S 数据集中，K 取 7～11 的阶段的准确率与 K 值呈正相关。但是随着 K 值的持续增大，邻居节点中课桌、黑板和墙壁等背景信息的占比会持续增多，背景特征与行为特征的差异远超学生行为特征间的差异，导致模型过度学习背景信息，准确率显著下降。当 $K \geqslant 12$ 时，背景信息过量引入，使得模型对 SCB03-S 的识别准确率较 $K=11$ 时下降 1.9 个百分点，验证了 K 值过大会导致模型识别性能衰减。

SCB03-U 的实验结果同样印证了上述规律。由于学生分布松散，行为特征离散，因此当 $K=6$ 时，模型准确率达到最大值，该结果进一步佐证了近邻 K 值与场景特征耦合适配的重要性。

4.6 SCB03-S 可视化结果与背景影响分析

为了验证本文模型对减少背景干扰具有一定有效性，现对 $K=11$ 的 CPViG-Net 的可视化结果进行分析。可视化结果如图 5 所示。

图 5 CPViG-Net 在 SCB03-S 上的可视化结果

Fig. 5 Visualization results of CPViG-Net on SCB03-S

从图 5 中可以看到，CPViG-Net 对 hand-rising，reading 和 writing 的目标具有一定的识别能力，大部分目标都被框定并标注了相应类别；但也存在部分标注框可能没有完全精准贴合目标人物的情况，比如部分举手动作的标注框边界与人物实际位置有偏差。在教室场景中，桌椅和黑板等背景元素多，部分背景区域被误标，例如第二列第三行的检测图，对黑板标注了 hand-rising，说明背景干扰依旧存在，模型虽能一定程度区分背景与目标，但在背景复杂、特征混淆区域仍有误检，处理背景干扰的能力须提升。

随后，就背景干扰项与 Yolo v8 进行比较。SCB03-S 选取的场景为中小学生课堂，教师对学生有较大的约束，学生的课堂行为较为统一，所以背景中的学生极易造成误判。本文选取背景中被误判的项为指标来分析模型的抗背景干扰能力。比较结果如图 6 所示。

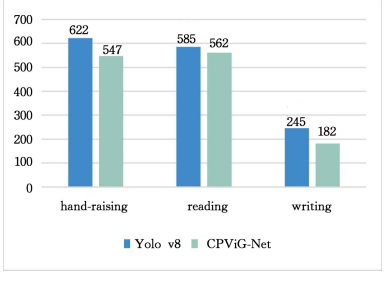


图 6 背景干扰的对比

Fig. 6 Comparison of background interference

图 6 中的数据代表真值为背景，被误判为 hand-rising，reading 和 writing 的样本分布情况。由图 6 可得，相较于 Yolo v8，CPViG-Net 在上述 3 类误判场景中展现出显著优势，误判率分别降低 12%，3.9% 和 25.7%，尤其在 hand-rising和 writing 的误判抑制上效果突出。由于阅读姿态与自然低头动作存在视觉相似性，因此该类别的识别仍面临较大挑战。即便如此，CPViG-Net 在全场景下的误判率显著下降，充分验证了该框架在复杂背景下的强大识别能力。

4.7 泛化性测试

为了进一步测试 CPViG-Net 的泛化性，在公开数据集 VisDrone2019[38] 上，将其与当前领域内的多个先进目标检测算法进行了对比实验，包括 Faster R-CNN[13]，YOLOX[39]，RTDETR[40]，D-Fine[41]，Yolo v10，Yolo v11 和 Yolo v12。对比结果如表 5 所列。

表 5 VisDrone2019 上的泛化性实验结果

Table 5 Results of generalization experiments on VisDrone2019

Algorithm	*mAP@50*	*GFLOPs*	*Param*
Yolo v8(baseline)	0.259	8.1	3.000×10^6
Faster R-CNN	0.329	208	4.139×10^7
YOLOX-Tiny	0.278	7.58	5.000×10^6
RTDETR	0.333	60	2.000×10^7
D-Fine-N	0.334	7.13	3.730×10^6
Yolo v10	0.261	6.5	2.300×10^6
Yolo v11	0.258	6.3	2.500×10^6
Yolo v12	0.259	5.8	2.500×10^6
CPViG-Net	0.344	10.6	4.000×10^6

从表 5 中可以看出，相较于 Yolo 系列的算法，CPViG-Net 在识别被遮挡物体与小目标时展现了更加优越的性能。

在参数量差距不大的情况下，CPViG-Net 的 mAP50 值比新兴的实时监测算法 RTDETR 提高了 1.1 个百分点，表明模型有较强的泛化能力，能够胜任不同环境的检测任务。

结束语 针对真实课堂场景中人物尺度差异大、遮挡普遍及背景干扰强等问题，本文基于视觉图卷积神经网络和目标检测算法，提出了一种融合节点关系特征与空间多尺度特征的学生行为检测模型 CPViG-Net，来缓解复杂课堂环境下背景语义混淆导致的误检问题，并增强对多尺度人物目标的检测能力。模型将最大相对图卷积引入 PMG 模块，通过 MRConv 捕捉节点间的最大差异特征，在有效应对遮挡问题的同时，采用 DWConv 降低计算开销；设计了 CPF 模块，通过结合全连接层和 C2f 模块，增强模型对多尺度目标的识别能力。公开数据集上的实验结果表明，本文方法有效减少了复杂背景干扰，提升了学生行为检测能力，精度优于许多先进的目标检测算法。

在未来工作中，可以进一步扩充相关数据集，以增强模型的泛化能力。这是因为目前公开的学生行为数据集较少，CPViG-Net 虽在 SCB03-S 和 SCB03-U 数据集上表现良好，但仍需在一些不同的课堂环境进行进一步的测试与调整。

参 考 文 献

[1] ZHU Z T, HAN Z M, HUANG C Q. Educational Artificial Intelligence(eAI): A new paradigm of human-centered artificial intelligence[J]. e-Education Research, 2021, 42(1): 5-15.

[2] SINGH H, MIAH S J. Smart education literature: A theoretical analysis[J]. Education and Information Technologies, 2020, 25(4): 3299-3328.

[3] HUANG T, ZHANG Z M, LIU S Y. Coexistence for Symbiosis: How Human-Intelligence Collaborative Co-education is Possible [J]. Educational Research, 2025, 46(1): 147-159.

[4] ROMERO C, VENTURA S. Educational data mining: a review of the state of the art[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2010, 40(6): 601-618.

[5] YU M, XU J, ZHONG J, et al. Behavior detection and analysis for learning process in classroom environment[C]// 2017 IEEE Frontiers in Education Conference(FIE). IEEE, 2017: 1-4.

[6] ZHANG X F. The Transformation of Traditional Educational Evaluation: Educational Evaluation Based on the Theory of Multiple Intelligences[J]. Educational Science Research, 2002(4): 28-30.

[7] ZHAO J, ZHU H. Cbph-net: A small object detector for behavior recognition in classroom scenarios[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-12.

[8] TAN S Y, WANG Z X, HE G D. Real-time Panoramic Multi-scale Classroom Behaviors Recognition Based on CA-YOLOv9 Network [J]. Modern Educational Technology, 2024, 34(7): 123-130.

[9] LIAO S B, QI F. Research on machine analysis of classroom teacher-student interaction behavior [J]. Journal of Central China Normal University(Natural Sciences), 2024, 58(2): 279-285.

[10] WANG D Q, LIU H, QIU M L. Analysis Method and Application Verification on Teacher Behavior Data in Smart Classroom [J]. China Educational Technology, 2020(5): 120-127.

[11] YADAV D K, KUMARI N, HARRON S. Advances in Convolutional Neural Networks for Object Detection and Recognition [C]// 2024 International Conference on Optimization Computing and Wireless Communication(ICOCWC). IEEE, 2024: 1-6.

[12] XIAO H, LIU X D. Real-time acquisition and dynamic analysis of learning state based on hybrid intelligence[J]. Journal of Jilin University(Engineering and Technology Edition), 2025, 55(7): 2402-2408.

[13] LI X W, YE J W, ZHANG Q H. On the Index Model of Teacher-Student Interaction Behaviors Under the Background of "Internet+Teaching"[J]. Research in Higher Education of Engineering, 2020(3): 157-162

[14] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[15] ZENG C, YAN K, WANG Z, et al. Abs-CAM: a gradient optimization interpretable approach for explanation of convolutional neural networks[J]. Signal, Image and Video Processing, 2023, 17(4): 1069-1076.

[16] WANG Z, WANG Z, ZENG C, et al. High-quality image compressed sensing and reconstruction with multi-scale dilated convolutional neural network[J]. Circuits, Systems, and Signal Processing, 2023, 42(3): 1593-1616.

[17] 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.

[18] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection [J]. arXiv: 2004.10934, 2020.

[19] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// Computer Vision-ECCV 2016: 14th European Conference. Springer, 2016: 21-37.

[20] CHU Q, OUYANG W, LI H, et al. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4836-4845.

[21] CHENG J, TANG Y, HE C, et al. Rethinking Variational Bayes in Community Detection From Graph Signal Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(5): 2903-2917.

[22] UTTARKABAT S, NAYAK S, CHAUDHURI S P, et al. e-Framework for m-Health Detection and Control Using GNN [C]// IECON 2023－49th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2023: 1-6.

[23] HUANG X, HUANG C. NGD: Filtering graphs for visual analysis[J]. IEEE Transactions on Big Data, 2016, 4(3): 381-395.

[24] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2d

pose estimation using part affinity fields[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:7291-7299.
[25] LI J N, LI R Y, ZHAO Z F, et al. Recognition and Analysis of Teaching Behavior Based on Multi-scale GCN[J]. Computer Science, 2024, 51(10):135-143.
[26] CHI H, HA M H, CHI S, et al. Infogcn: Representation learning for human skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:20186-20196.
[27] YANG W, ZHANG J, CAI J, et al. HybridNet: Integrating GCN and CNN for skeleton-based action recognition[J]. Applied Intelligence, 2023, 53(1):574-585.
[28] HAN K, WANG Y, GUO J, et al. Vision gnn: An image is worth graph of nodes[J]. Advances in Neural Information Processing Systems, 2022, 35:8291-8303.
[29] SOUDEEP S, MRIDHA M F, JAHIN M A, et al. DGNN-YOLO: Dynamic Graph Neural Networks with YOLO11 for Small Object Detection and Tracking in Traffic Surveillance[J]. arXiv: 2411.17251, 2024.
[30] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020:390-391.
[31] YAN Q G, ZHAO J, CHA X G, et al. The Evaluation of Teaching Effect based on Interpretable Machine Learning[C]// 2021 8th International Conference on Dependable Systems and Their Applications(DSA). IEEE, 2021:712-715.
[32] JIANG H, WANG H. Designing and Implementing an Intelligent Machine Learning-Based Evaluation System for Assessing English Teaching Quality in Vocational Education[C]// 2024 International Conference on Interactive Intelligent Systems and Techniques(IIST). IEEE, 2024:36-40.
[33] WANG X Y, GAO D H, NING Y W, et al. Research on Lightweight Student Behavior Detection Method Based on Improved YOLO Algorithm[J/OL]. Computer Science, 1-15.
[34] MUNIR M, AVERY W, MARCULESCU R. Mobilevig: Graph-based sparse attention for mobile vision applications[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:2211-2219.
[35] YANG F, WANG T, WANG X. Student classroom behavior detection based on YOLOv7+ BRA and multi-model fusion[C]// International Conference on Image and Graphics. Cham: Springer, 2023:41-52.
[36] CHEN J, KAO S, 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.
[37] MA X, DAI X, BAI Y, et al. Rewrite the stars[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024:5694-5703.
[38] ZHU P, WEN L, DU D, et al. Detection and tracking meet drones challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11):7380-7399.
[39] GE Z, LIU S, WANG F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv:2107.08430, 2021.
[40] ZHAO Y, LYU W, XU S, et al. Detrs beat yolos on real-time object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16965-16974.
[41] PENG Y, LI H, WU P, et al. D-FINE: redefine regression Task in DETRs as Fine-grained distribution refinement[J]. arXiv: 2410.13842, 2024.

ZHANG Haopeng, born in 2001, postgraduate. His main research interests include image processing and AI for education.

SONG Wanru, born in 1992, Ph.D. Her main research interests include image processing, pattern recognition and AI for education.

(责任编辑:柯颖)