文章类型：论文|刊号（ISSN）：2630-5178 /（中图刊号）：380GL019

面向真实课堂的学生行为数据集

刘怡萱[1] 夏永滢[1*] 孟凡爱[2] 樊荣妍[1] 张森[1]
1 华北理工大学经济管理学院
2 华北理工大学人工智能学院
DOI:10.32629/mef.v9i4.20354

[摘 要] 针对现有课堂行为数据集规模小、视角单一等问题,本文构建了一套来源于真实大学课堂的大规模学生行为数据集。该数据集包含2282张高分辨率图像和133586个标注实例,单图平均目标58.54个,并细粒度划分10类典型行为。标注采用“YOLO预标注+人工修正+双重质检”人机协同闭环机制,标注一致性达到0.85。基于YOLOv8的实验结果表明,模型在Precision和Recall等指标上均优于现有数据集,在复杂光照和高遮挡场景下表现出更强的泛化能力,为智慧教育研究提供了可靠的数据支撑。
[关键词] 课堂行为识别；数据集构建；高密度场景；多目标检测
中图分类号：G424.21 文献标识码：A

A student behavior dataset for real classrooms

Yixuan Liu[1] Yongying Xia[1*] Fan'ai Meng[2] Rongyan Fan[1] Miao Zhang[1]
1 School of Economics and Management, North China University of Science and Technology
2 School of Artificial Intelligence, North China University of Science and Technology,
[Abstract] To address the limitations of small scale and single perspective in existing classroom behavior datasets, this paper constructs a large-scale student behavior dataset derived from real university classrooms. The dataset contains 2,282 high-resolution images and 133,586 annotated instances, with an average of 58.54 targets per image, and defines 10 fine-grained behavior categories. A human-in-the-loop annotation framework, including YOLO-based pre-annotation, manual refinement, and dual quality inspection, is adopted to ensure high annotation consistency (Kappa = 0.85). Experimental results based on YOLOv8 demonstrate that the model outperforms existing datasets in terms of Precision and Recall, showing stronger generalization ability under complex lighting and heavy occlusion. The proposed dataset provides reliable support for research in smart education.
[Key words] classroom behavior recognition; dataset construction; high-density scenes; multi-object detection; smart education

1 引言

课堂行为分析是评估学生学习状态、优化教学策略的重要手段。随着智慧教育的不断发展,基于计算机视觉的课堂行为识别技术为实现客观、高效的教学评估提供了新的技术路径。[1]学生的身体姿态与面部表情作为反映学习状态与参与程度的重要外在表征,对其进行精准识别与建模,有助于实现个性化教学支持与课堂互动效果提升[2]。

然而,现有课堂行为识别技术高度依赖大规模、高质量且具备场景针对性的标注数据,而当前公开数据集在实际应用中面临多重瓶颈 。首先,采集场景缺乏真实性,现有数据多源于受控实验或单一视角的线上教学,难以还原真实大学教室中复杂空间布局及非中心视野的特征。[3]其次,数据规模与粒度难以支撑深度模型需求,现有数据集普遍体量较小且行为类别单一(通常仅涵盖4-5类),无法全面表征真实课堂中丰富多样的行为模式[4],尤其在处理长尾分布场景时表现欠佳。基于上述问题,本文构建了一套来源于真实大学课堂场景的学生行为数据集,主要贡献如下：

(1)构建大规模真实课堂数据集：数据集包含2282张高分辨率图像,全部采集自真实大学教室环境,能够真实反映高目标密度、复杂空间布局以及类别不均衡等典型课堂特征。(2)提供高密度多目标标注：数据集共包含133586个行为标注实例,平均每张图像包含58.54个目标,单张图像最多达138个目标,为高密度场景下的群体行为分析提供了重要数据支撑。(3)构建细粒度行为类别体系：结合真实课堂场景,定义了10类典型学生行为,同

时覆盖学习行为与异常行为,为细粒度行为分析及长尾分布研究提供了良好基础。

2 相关工作

现有的教育行为分析研究主要依赖于受控实验室环境采集的数据集或通用情感数据库(如JAFFE[5],CK+[6],FER2013[7])。尽管这些数据在基础情绪识别任务中表现良好,但它们缺乏真实教室特有的空间约束、高密度遮挡以及复杂光照条件。近期的一些现场数据采集尝试(如DFEW[8])往往侧重于宏观课堂氛围或教师行为,忽视了对学生细粒度状态的捕捉。Li等人[9]指出,由于非正面视角和严重遮挡引起的域偏移(Domain Shift),在理想化数据上训练的模型在实际部署时性能会显著下降。因此,目前极度缺乏能够捕捉非受控大学环境中真实、长期学生行为的大规模数据集。

基于视频的行为分析面临的核心挑战是如何在信息保留与计算效率之间取得平衡。静态图像数据集(如AffectNet[10])丢弃了关键的时序上下文,导致动态行为分类存在歧义;而直接处理原始视频流[11],则引入了巨大的冗余,相邻帧往往包含相同的状态信息,导致模型过拟合静态背景而非行为动态[12]。此外,现有的标注流程难以扩展。纯人工标注对于长期录像来说成本过高,而完全自动化的方法(如使用预训练YOLO)在复杂场景中会引入显著噪声。尽管已存在“人在回路”的框架,但鲜有研究针对高密度教育场景建立结合模型预标注、时序上下文验证及严格质量控制的标准化协议。

综上所述,现有研究在真实场景覆盖度、时序数据去冗余策略以及高效高精标注体系三个维度均存在显著缺口。针对上述不足,本文采集真实教室长时序视频,构建课堂行为数据集。通过定间隔抽帧以平衡信息量与冗余度,引入YOLO模型进行自动化预标注以大幅提升初始效率,并设计了包含人工修正、双重校验及视频回溯的严格质控流程,以填补真实复杂场景下高质量课堂数据集的空白,为后续的智慧教育算法研究提供坚实可靠的数据基石。

3 课堂行为数据集构建

3.1数据采集与场景设置

数据采集是构建高质量数据集的基础。为真实反映智慧教育场景的复杂性,我们在采集策略上重点考虑了场景真实性、设备稳定性、环境多样性及行为丰富性。

不同于实验室受控环境,本数据集所有视频均采集自真实大学教室,包括传统阶梯教室、多媒体互动教室及普通矩形教室。这些场景天然包含了真实课堂的复杂性:高密度人群带来的严重相互遮挡问题;黑板、幕布、桌椅等复杂背景干扰;学生在自然状态下被采集,确保了行为的真实性和自然度,避免了“表演式”偏差。

为模拟真实智慧教室的监控逻辑,本研究采用固定机位长时序录制,视角配置涵盖教室后方、左前侧、右前侧等非中心布局,引入侧脸、大角度低头及光照不均等极端视觉挑战,以全面检验识别算法的泛化能力。采集过程覆盖了早间至傍晚的多种自然光照,并跨越理论讲授型与互动讨论型等课程,以全面表征学生在不同认知负荷下的行为模式。

3.2标注类别与初步统计

本数据集专为真实大学课堂场景设计,经过严格的清洗与筛选,最终数据集共包含2,282张图像,这些图像源自真实的课堂视频录制。数据集中共标注了133,586个学生行为实例,充分反映了大学教室人员密集的典型特点。

统计显示,本数据集平均每张图像包含约58.54个标注目标,单帧最大目标数量高达138人。这种高密度场景引入了严重的相互遮挡、小目标检测困难以及背景杂乱等挑战,这些问题在COCO或MPII等通用数据集中较少见。因此,本数据集为评估拥挤教育场景下的多目标检测与行为识别算法提供了极具价值的基准。数据集涵盖了9种细粒度的课堂行为类别,旨在全面捕捉主流学习活动及特定的异常状态。如下表1所示:

表1 标注行为类别分布统计

编号	类别名称	描述	实例数量	占比
0	Bowing-Head	低头看桌面或记笔记	67,448	50.49%
1	Listening	专注听讲	31,696	23.73%
2	Reading-and-Writing	阅读材料或书写	21,816	16.33%
3	UsingPhone	操作或观看手机	5,911	4.42%
4	Null	状态不清或无有效行为	2,382	1.78%
5	Discussing	与同学互动或交谈	1,612	1.21%
6	Sleeping	趴桌睡觉或打瞌睡	1,005	0.75%
7	Turn-Around	转身向后看或互动	990	0.74%
8	Stand	起立(如回答问题)	651	0.49%
9	Eating/Drinking	吃东西或喝水	74	0.06%
总计			133,586	100%

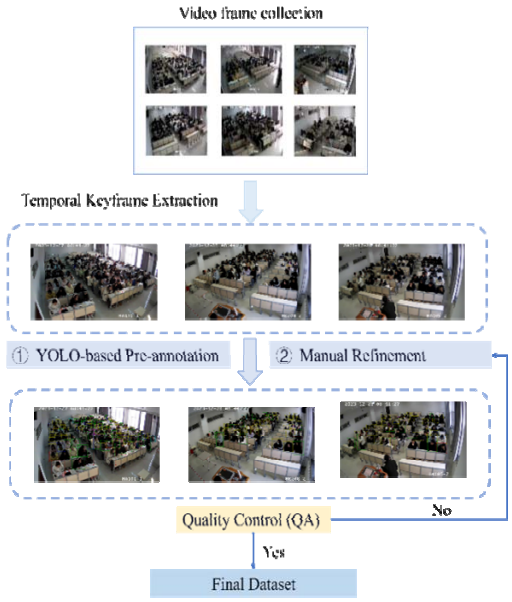

图1 数据集构建流程

如表1所示,前三类主要行为(“低头”、“听讲”、“读写”)占据了总实例数的90%以上,反映了标准课堂中学生的主导状

文章类型：论文|刊号（ISSN）：2630-5178 /（中图刊号）：380GL019

态。相比之下，“饮食”、“站立”和“回头”等关键但低频的行为占比不足1%。这种不平衡性真实模拟了课堂行为的自然规律，但会给模型训练带来挑战，需采用重采样或损失函数加权等技术防止模型偏向多数类。这一特性使本数据集在研究教育场景下的类别不平衡学习方面具有独特价值。

3.3人机协同标注闭环流程

高质量标注是监督学习成功的关键所在。面对海量的视频数据，纯人工标注成本过高且一致性难以保证，而纯自动标注噪声过大。参考Northcutt等人提出的置信学习理念及Settles的主动学习框架，我们设计了一套“模型预标注+人工修正+多重质检”的人机协作标准化流程。

如图1所示，本研究采用一套结构化、闭环式的“人在回路”标注流程，以在保证大规模数据处理效率的同时，确保标注结果的高精度与语义一致性。

整个流程始于原始课堂视频帧的采集，随后进入三阶段处理：首先，在①YOLO-based Pre-annotation阶段，我们利用预训练的YOLOv8模型对每张图像进行自动检测与初步分类，完成约90%的边界框生成任务，大幅降低人工负担；接着，在②Manual Refinement阶段，专业标注员基于源视频的时序上下文，对模型输出进行精细化修正——包括调整边界框位置、纠正易混淆类别、并补全因遮挡或模糊导致的漏检目标；最后，所有修正后的样本进入Quality Control(QA)环节，采用双人独立复核+专家仲裁机制，针对标注分歧或稀有类样本进行最终裁定，确保标签可靠性。该流程形成闭环反馈(虚线箭头表示可回溯修正)，最终产出高质量、高一致性的Final Dataset。经统计，本流程使标注者间一致性(Cohen’s Kappa)达到0.85，充分验证了其在复杂真实场景下的有效性与可扩展性。

3.4数据组织与标准化

本数据集采用“图像-标注分离但严格同步”的架构。原始图像存储于img/目录，涵盖多样化课堂场景下的光照、角度及行为模式；结构化标注文件存储于json/目录，与图像一一对应，包含边界框坐标、类别标签及置信度信息。每张图像文件(I_i)均配有一个JSON标注(A_i)，形成清晰的配对机制，便于批量解析与模型训练。

在标注完成后，执行了统一的数据清洗与标准化流程：对所有JSON文件进行格式校验，剔除异常样本；移除损坏图像、模糊样本及重复标注，降低噪声干扰；对类别标签进行数值化编码，建立全数据集一致的标签体系。

通过上述refine过程，最终构建了一个结构完整、标注规范、类别定义清晰的课堂行为图像数据集。

4 跨数据集基准评估

4.1实验设置与评价指标

本研究基于YOLOv8[14]目标检测框架开展实验。为验证本文所构建数据集的有效性，选取多个公开数据集作为对比对象，在相同模型结构、训练参数及优化策略下分别进行训练，并在统一测试集上进行评估，实验结果如表2所示。

表2 实验结果表

Dataset	Precision	Recall	mAP@50	mAP@50-95
A-1 ours	0.7852	0.7121	0.6982	0.6150
A-2 university	0.5884	0.4592	0.5322	0.5109
B-2 Structure	0.6733	0.5277	0.5585	0.5194

在实验结果对比显示，本文构建的A-1 ours数据集在各项评价指标上均显著优于其他对比数据集，相比之下，A-2 university数据集整体性能最低，主要归因于其数据多来源于理想化场景，缺乏复杂遮挡、光照变化及真实课堂环境特征；而B-2 Structure数据集虽引入了一定真实场景因素，性能优于university，但在高密度人群及复杂视角覆盖上仍存在局限。该对比结果有力证明了本文数据集通过充分考虑多视角分布、密集目标及复杂背景干扰等真实因素，能够显著提升模型在复杂课堂环境下的检测精度与特征表示的鲁棒性。

4.2不同数据集性能对比与分析

为验证本文构建数据集的有效性，选取不同数据集进行对比实验。在相同模型结构与训练策略下，分别基于不同数据集进行训练，并在统一测试集上进行评估，实验结果如图2所示。

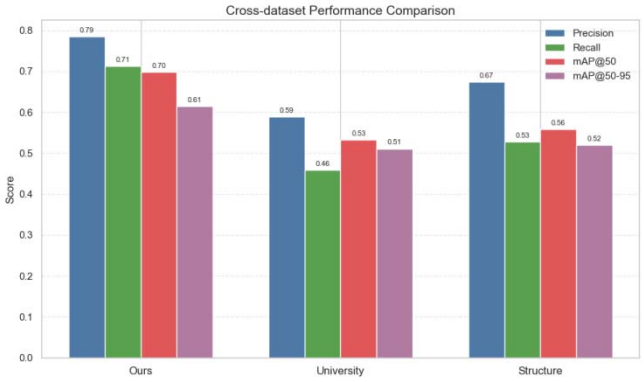

图2 不同数据集性能对比柱状图

可以看出，不同数据集训练得到的模型在各项评价指标上存在明显差异。其中，University数据集整体性能较低，Precision和Recall分别为0.5884和0.4592，说明该数据集多来源于理想化场景，缺乏复杂遮挡、光照变化以及真实课堂环境特征，导致模型在实际场景中的泛化能力较弱。

Structure数据集在各项指标上较University有所提升，表明其在一定程度上引入了真实场景因素，使模型具备一定的泛化能力。然而，由于其在复杂视角及高密度人群场景方面覆盖不足，整体性能仍存在一定局限性。

相比之下，本文构建的数据集在所有评价指标上均取得最优结果，其中Precision达到0.7852，Recall达到0.7121，mAP@50和mAP@50-95分别达到0.6982和0.6150，均显著优于对比数据集，说明该数据集能够有效提升模型的检测性能与定位精度。

该结果表明，本文数据集在构建过程中充分考虑了真实课堂环境的复杂性，包括多视角分布、密集目标场景以及复杂背景干扰等因素，使模型能够学习到更加鲁棒的特征表示，从而在复

杂环境中实现更高的检测精度和更低的漏检率。

4.3复杂场景定性结果展示

为进一步验证模型在真实课堂环境中的检测效果，本文从测试集中选取若干典型场景进行可视化分析，结果如图3所示。

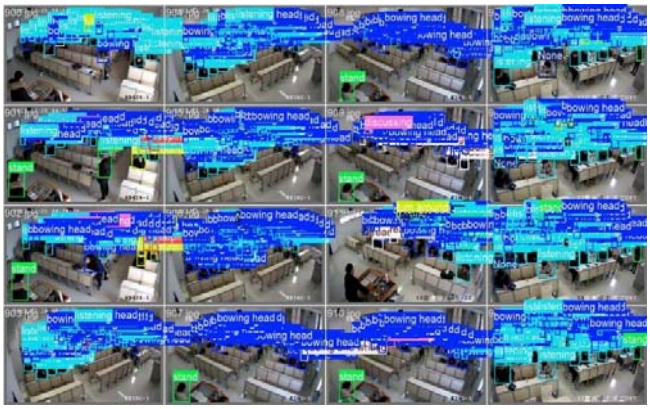

图3 课堂行为检测结果可视化

可视化结果表明，模型能够在高密度学生环境中同时检测多种行为类别，包括听讲、读写、使用手机以及睡觉等行为，体现出良好的多目标检测能力与复杂场景适应能力。在复杂课堂环境中，小目标检测与遮挡问题a仍然是当前方法需要进一步优化的重要方向。

5 结语

本文针对真实课堂行为分析中数据匮乏的痛点，构建了一个具备高生态效度的大规模数据集。该数据集突破了现有基准在受控环境与低密度场景下的局限，真实还原了高密度遮挡、复杂光照及非配合行为等挑战。通过提出“人在回路”的标准化标注范式，我们实现了效率与精度的平衡（Kappa=0.85），并揭示了行为分布的长尾特性。基线实验证实，基于本数据训练的模型在真实部署场景中显著优于传统合成数据模型，验证了高质量真实数据对提升算法鲁棒性的关键作用[13]。

[基金项目]

大学生创新创业训练计划项目(项目名称：复杂教室场景下基于超图关系推理的学生3D姿态重建,项目编202510081008)；(项目名称：基于视觉Mamba模型的大学生课堂参与度研究,项目编号：X2025125)；华北理工大学2025年度校级教育教学改革研究与实践项目：L2431生成式AI赋能智慧教学模式的创新与实践——以《信息系统项目管理》课程为例；教育部产学合作协同育人项目(项目名称：未来技术背景下的智慧实验室建设项目编号：2021102119023)。

[参考文献]

[1]LIN L, YANG H, XU Q, et al. Research on student classroom behavior detection based on the real-time detection transformer algorithm[J].Applied Sciences,2024;14(14):6153.

[2]HE Q C.Research and implementation of classroom behavior recognition system for teachers and students based on deep learning[D].Shihezi: Shihezi University,2025.

[3]HAN P C,REN J, SHEN Z P, et al. Overview of deep learning driven automatic classroom behavior recognition technology and method[J].Modern Information Technology,2025;9(20):109-114.

[4]LIU Q, JIANG X, JIANG R. Classroom behavior recognition using computer vision: A systematic review[J].Sensors,2025;25(2):373.

[5]GUTTA S,WECHSLER H,PHILLIPS P J.Gender and ethnic classification of face images [C]// Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition.Nara,Japan:IEEE,1998:194-199.

[6]LUCEY P,COHN J F,KANADE T,et al.The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression [C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. San Francisco,CA,USA:IEEE,2010:94-101.

[7]ZAHARA L,MUSA P, PRASETYO E, et al.The Facial Emotion Recognition(FER-2013)dataset for prediction system of micro-expressions face using the convolutional neural network CNN algorithm based Raspberry Pi [C]// 2020 Fifth International Conference on Informatics and Computing(ICIC).IEEE,2020:1-9.

[8]JIANG X,ZONG Y,ZHENG W,et al.DFEW:A large-scale database for recognizing dynamic facial expressions in the wild[EB/OL].https://arxiv.org/abs/2008.05924,2020-08-11.

[9]LI S,DENG W,DU J.Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2017:2852-2861.

[10]LI S, DENG W. Deep Facial Expression Recognition: A Survey[J].IEEE Transactions on Affective Computing,2022,13(3):1195-1215.

[11]MOLLAHOSSEINI A,HASANI B, MAHOOR M H. AffectNet: A database for facial expression, valence, and arousal computing in the wild[EB/OL].https://arxiv.org/abs/1708.07262, 2017-08-23.

[12]CARREIRA J,ZISSERMAN A.Quo vadis, action recognition? A new model and the Kinetics dataset[EB/OL].https://arxiv.org/abs/1705.07750,2018-02-22.

[13]DAI J Q,XIE A M, YU D Y. Classroom behavior recognition through joint improvement of Faster RCNN algorithm under the construction of smart education[J].Journal of Digital Contents Society,2025,26(2):485-494.

作者简介：

刘怡萱(2006--),女,汉族,河北保定人,本科,研究方向：深度学习。