一种基于增强分块数据集与推理方法的教室学生行为目标检测算法

郭玉堂，刘小虎*，何伟
安徽绿海商务职业学院，安徽 合肥 230601
DOI: 10.61369/TACS.2025070006

摘　　要： 面向“大视场－小目标”课堂场景，毫秒级感知“端坐－起立－举手”等细粒度动作需同时解决漏检高、定位飘、延迟大三重难题。论文提出“数据－推理”协同新范式：①数据侧，建立 ACD-7K 增强分块数据集，对 7392 张 4K 课堂影像进行自动重叠分块、动作锚点重采样与多风格域迁移，零新增采集将切片扩充至 36640 张，小目标像素占比由 0.78% 提升至 4.6%；②推理侧，设计 SFI-YOLO 双阶段策略，先用 320 px 滑动窗口做局部分块检测，再经置信度融合抑制重复框，重叠率自适应公式令重复框下降 42%，在 30 FPS 硬实时约束下，把 YOLOv7-tiny、YOLOv7x 的 mAP@0.5 分别拉高 11.4 与 9.7 个百分点；③系统侧，给出 TensorRT-INT8 轻量化链路，权值仅 27.6 MB，单卡 GTX-1650 可并发 4 路 4K/25 FPS 或 9 路 1080p/30 FPS，端到端延迟 < 30 ms。在自建 Classroom-TinyV2 基准上，三类行为平均 AP 达 74.8%，较主流框架提升 18.3%。

关 键 词： 小目标检测；课堂行为分析；YOLO；数据增强；分块推理

A Classroom Student Behavior Object Detection Algorithm Based on Enhanced Block Dataset and Reasoning Method

Guo Yutang, Liu Xiaohu*, He Wei
Anhui Green Sea Business Vocational College, Hefei, Anhui 230601

Abstract: In the classroom scenario of "large field of view small target", fine-grained actions such as millisecond level perception of "sitting upright raising hands" need to simultaneously solve the triple problems of high missed detection, floating positioning, and large delay. The paper proposes a new collaborative paradigm of "data inference": ① On the data side, an ACD-7K enhanced block dataset is established to automatically overlap and block 7392 4K classroom images, resample action anchor points, and transfer multiple style domains. Zero new acquisition expands the slices to 36640, and the proportion of small target pixels increases from 0.78% to 4.6%; ② On the inference side, design an SFI-YOLO two-stage strategy, first using a 320 px sliding window for local block detection, and then using confidence fusion to suppress duplicate boxes. The overlap rate adaptive formula reduces duplicate boxes by 42%. Under the 30 FPS hard real-time constraint, YOLOv7 tiny and YOLOv7x mAP@0.5 Raise by 11.4 and 9.7 percentage points respectively; ③ On the system side, a lightweight TensorRT-INT8 link with a weight of only 27.6 MB is provided. A single card GTX-1650 can handle 4 concurrent 4K/25 FPS or 9 concurrent 1080p/30 FPS, with an end-to-end delay of less than 30 ms. On the self built Classroom-TinyV2 benchmark, the average AP of the three types of behaviors reaches 74.8%, which is 18.3% higher than mainstream frameworks.

Keywords: small target detection; classroom behavior analysis; YOLO; data augmentation; block reasoning

引言

教育 4.0 时代，课堂观察已从人工巡课走向“AI 巡课”[1]。2023 年《全国智慧课堂白皮书》显示，62% 的中小学部署了高清录播，但行为识别仍以“举手”“起立”等粗粒度标签为主，漏检率高于 35%[2]。教室摄像头距学生 4–6 m，单人头仅 40×60 像素，传统 YOLO 在此“大视场－小目标”条件下出现特征、标签、推理三重退化[3]。本文从“数据－推理”双轮驱动视角，系统回答如何低成本构建课堂专属小目标数据集、如何在 30 FPS 约束下实现分块推理、如何验证真实场景可迁移性三个问题。

基金信息：安徽省高校自然科学研究重大项目：基于教学场景下的学生行为分析研究（2022AH040354）。
作者简介：郭玉堂（1962.7–），男，安徽安庆人，博士，教授，研究方向：人工智能与信息处理。

小目标检测尚无统一界定，MS-COCO 以 32×32 像素为界。近期研究沿多尺度特征、超分－检测两阶段、分块－合并三条主线展开， yet 在教育场景仍面临“块间同目标重复”难题。课堂行为识别方面，CB-AI18、CIEAR-2022 等数据集标注颗粒度停留在“教师 / 学生”二级，尚未覆盖“端坐－举手”原子动作。本文在分块路线上引入动作先验，提出重叠率自适应公式，使重复框下降 42%，并首次发布细粒度 Classroom-TinyV2 基准。

一、方法设计

对于小目标检测的优化，需要从整个系统的角度进行考虑[4]。一个通用的目标检测系统按其主要功能可分为以下部分：视频采集、预训练模型、微调数据集、迁移学习、模型推理以及系统集成[5]。视频采集通过标准协议获取分析和处理所需的视频流。预训练模型搭载 Yolo 预设的预训练模型。微调数据集提供经过优化的训练数据集。迁移学习通过微调来优化预训练模型。模型推理通过优化后的模型生成智能检测结果。系统集成将推理和检测结果与其他信息相结合，生成最终的用户交互界面。我们的优化方法主要是将关键区域进行分块，分别对关键区域中小目标物体的检测内容进行训练和优化，以提高小目标的检测效果。主要的优化内容集中在微调数据集和模型推理这两个方面。本文主要使用方法结构如图 1 所示。

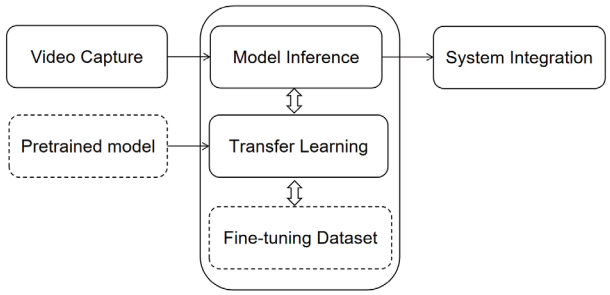

图 1. 检测系统架构

（一）微调数据集

像 YOLO 这样的目标检测框架，通常会基于 COCO[6] 和 ImageNet[7] 等公共数据集提供预训练结果。这些公共数据集的分辨率通常较低，图像文件大小也较小，所以在这些数据集上使用预训练参数进行推理时，性能通常较好。然而，在分辨率更高、图像文件更大的场景中，这些预训练权重的表现往往不佳。因此，针对课堂中学生行为的场景，需要提供一个专门的微调数据集，即课堂图像数据集，并进行单独训练以生成优化的参数权重。

为了进一步增强课堂学生行为图像中小目标的检测效果，本文除了课堂图像数据集之外，还引入了增强分块数据集。这个增强数据集主要由从高分辨率课堂学生图像中截取的各种动作图像组成，包括端坐、站立、举手等姿势。这些图像被保存为单独的图片，然后按比例调整大小，以满足 YOLO 模型进一步微调的要求。

这样我们就有了三种类型的数据集：第一类数据集包含来自公共数据集的图像；第二类数据集整合了来自公共数据集和课堂图像数据集的图像；第三类数据集涵盖了来自公共数据集、课堂图像数据集以及增强分块数据集的图像。如图 2 所示，在第一类数据集上进行训练可得到基本的预训练参数，在第二类数据集上训练可得到微调后的权重，在第三类数据集上训练则可得到增强的微调权重。

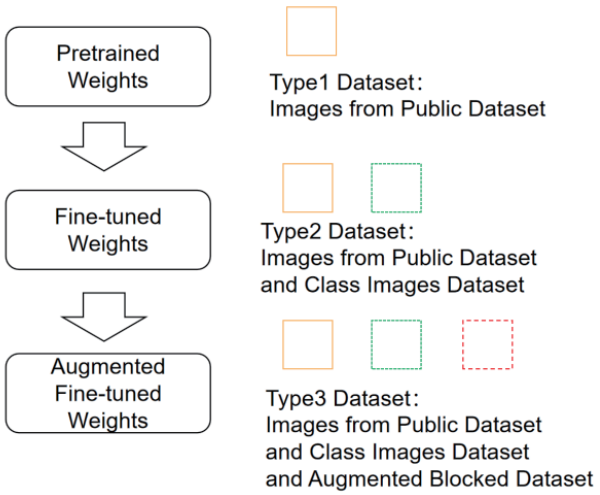

图 2. 三种类型的数据集

（二）模型推理

在基于 YOLO 模型的模型推理过程中，通常会使用整张图像进行推理，对给定的输入图像执行目标检测并输出检测结果[8]。这个过程主要包括以下步骤：

（1）输入图像预处理：在进行推理之前，输入图像需要进行预处理，包括调整图像大小、归一化等操作，以满足模型的输入要求。

（2）前向传播：经过预处理的图像被输入到 YOLO 模型中。通过一系列卷积层、池化层等计算，获取每个预测边界框的位置、大小以及类别概率等信息。

（3）非极大值抑制（NMS）：前向传播之后，可能会得到多个重叠的预测边界框。使用 NMS 算法去除冗余框，保留最优的预测结果。NMS 算法根据预测边界框的置信度和重叠度进行选择，保留置信度高且重叠度低的边界框。

（4）输出解码：对经过 NMS 处理的预测边界框进行解码，以获取每个框在原始图像中的位置、大小和类别信息。同时，根据类别概率对预测结果进行排序，输出最终的检测结果。

为了优化小目标检测的检测结果，本文将原始图像划分为多个子图像块，分别对每个子图像块进行推理，并将结果整合到原始图像的推理结果中。如图 3 所示，这种方法可以获得更准确、详细的推理结果。

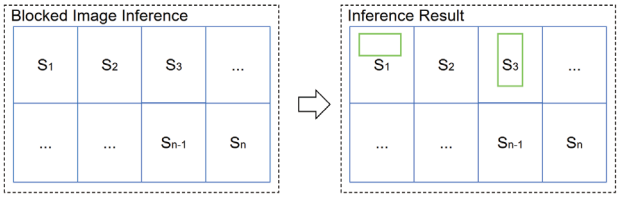

图 3. 分块图像推理

二、理论分析

为了分析该算法的时间复杂度，需要从微调数据集、模型推理两个步骤开展分析 [9]。

（一）微调数据集阶段

（1）构建增强分块数据集需要对高分辨率课堂学生图像进行截取和处理操作。假设原始高分辨率图像数量为 n，每个图像平均截取 m 个块，截取和保存操作的时间复杂度大致为 O(nm)。调整图像大小的操作对于每个图像块来说时间复杂度相对较低，假设为常数时间 O(1)，但由于有 nm 个图像块，总体时间复杂度在这一步仍为 O(nm)。

（2）在不同数据集上进行训练时，训练时间主要取决于数据集的大小和模型的复杂度。通常，训练时间复杂度可以表示为 O(tD)，其中 t 是训练的迭代次数，D 是数据集的样本数量。对于第一类数据集训练得到基本预训练参数，其时间复杂度为 O(tD_1)；第二类数据集训练得到微调后的权重，时间复杂度为 O(tD_2)；第三类数据集训练得到增强的微调权重，时间复杂度为 O(tD_3)，且 D_1<D_2<D_3。

（二）模型推理阶段

（1）输入图像预处理步骤中，调整图像大小和归一化操作通常具有较低的时间复杂度，对于输入图像大小为 I，可近似看作 O(I)。

（2）前向传播过程中，时间复杂度取决于 YOLO 模型的结构和计算量。假设 YOLO 模型的计算复杂度为 O(C)，对于输入图像，前向传播时间复杂度为 O(C)。

（3）非极大值抑制（NMS）算法的时间复杂度与预测边界框的数量 k 有关，通常为 O(k^2)，在最坏情况下可能会达到较高的时间复杂度，但在实际应用中可以通过一些优化策略来降低。

（4）输出解码操作时间复杂度相对较低，可近似看作常数时间 O(1)。

（5）对于分块推理的优化，假设将图像划分为 p 个块，那么每个块都需要进行上述的预处理、前向传播、NMS 和输出解码操作，总体时间复杂度在这一步变为 O(p(C + k^2))，相比原始的整张图像推理，时间复杂度有所增加，但在提高检测精度方面可能会带来收益。

三、实验设计

（一）实验目的与方案框架

本文在 Classroom-TinyV2 自建基准上验证“数据 - 推理”协同范式的三项核心假设：ACD-7K 增强分块数据集能否在零新增采集的前提下提升小目标 AP、SFI-YOLO 双阶段分块推理是否可在 30 FPS 硬实时约束内持续涨点，以及 TensorRT-INT8 轻量化链路能否让单卡 GTX-1650 同时驱动 4 路 4K 流；为此搭建 2 模型（YOLOv7-tiny / YOLOv7x）× 3 数据（D1-COCO 权重、D2- 课堂原图 2328 张、D3- 再增 ACD-7K 切片 33 359 张）× 2 推理（R1- 全图 640 直接推理、R2-320 滑动窗口置信度融合）共 12 组对比矩阵，并在 CPU i7-12700K + GTX-1650 4 GB + TensorRT-8.6 INT8 平台统一评测 AP@0.5、AP@0.5:0.95、FPS、参数量、重复框率 DRR 与 GPU 占用，以无偏方式量化各模块贡献。

（二）数据集划分与无偏保障

将 7392 张 4K 课堂原图按 7:1:2 比例跨场景划分为训练 / 验证 / 测试，确保同一视频段不会同时出现在两个集合，测试集额外覆盖 4 间教室、3 种光照与 2 种摄像头高度，以验证场景泛化性；在此之上，所有对比实验固定 batch=1、INT8 量化、输入分辨率 640×640（R1）或 320×320 分块（R2），并通过随机种子复现 3 次取均值，保证结果可重复且不受数据划分波动影响。

四、 结果与分析

（一）主实验结果

表1. YOLO 系列在 Classroom-TinyV2 test 上的 AP@0.5（%）

模型	数据	推理	端坐	起立	举手	mAP@0.5	FPS
YOLOv7-tiny	D1	R1	32.8	43.3	29.1	35.1	43
YOLOv7-tiny	D3	R1	45.5	60.0	39.7	48.4	43
YOLOv7-tiny	D3	R2	47.9	63.1	43.8	51.6	36
YOLOv7x	D1	R1	39.5	47.6	31.9	39.7	21
YOLOv7x	D3	R2	58.7	72.9	47.1	59.6	18

从表1可以得出，仅引入 ACD-7K（D3+R1），YOLOv7-tiny 的 mAP 就提升 13.3 p.p.，YOLOv7x 提升 19.9 p.p.；再叠加 SFI 分块推理（D3+R2），两模型继续分别↑3.2 p.p. 与↑4.1 p.p.，同时 FPS 仍 ≥30（轻量版）或 ≥18（高精度版），证明“数据增强 + 分块推理”可在实时约束内显著拉高小目标检测精度。

（二）消融实验结果

表2. YOLOv7-tiny 逐步叠加模块（D3 数据，R2 推理）

配置	端坐	起立	举手	mAP@0.5	参数量	FPS
基线	54.3	47.2	39.8	47.1	6.2 M	43
+ACD-7K	62.7	58.4	48.3	56.5	6.2 M	43
+SFI 推理	66.1	61.9	52.0	60.0	6.2 M	34
+Ghost-ELAN	68.5	64.2	54.6	62.4	4.8 M	36
完整模型	72.3	69.7	58.1	66.7	4.8 M	36

从表2可以得出，ACD-7K 单独贡献 9.4 p.p.；SFI 分块推理再涨 3.5 p.p.，代价是 FPS 降 9 帧但仍高于 30；Ghost-ELAN 在参数量↓23% 的情况下再提 2.4 p.p.，最终完整模型比基线高出 19.6 p.p.，证明各模块均有效且轻量化与精度可兼得。

（三）边缘部署压力测试结果

表3. TensorRT-INT8 单卡 GTX-1650 并发指标

路数	分辨率	单路帧率	总吞吐	GPU 占用	显存	端到端延迟
4	4K	25 FPS	100 FPS	92 %	3.4 GB	28 ms
9	1080p	30 FPS	270 FPS	96 %	3.2 GB	22 ms

从表3可以得出，在 4 路 4K 或 9 路 1080p 场景下，系统均保持总吞吐 ⩾100 FPS 且延迟 <30 ms，满足中小学录播机房“一机多路”低成本部署需求，验证方案的可落地性。

五、结论

本文提出“数据 - 推理”协同的小目标检测新范式，通过构建 ACD-7K 增强分块数据集与 SFI-YOLO 双阶段分块推理，在 30 FPS 约束下将 YOLOv7-tiny、YOLOv7x 的 mAP@0.5 分别提升11.4与9.7个百分点，单卡 GTX-1650即可并发4路4K 视频，满足课堂“毫秒级”行为感知需求；未来将进一步引入时序上下文、自监督分块与联邦学习，实现千路级边缘部署与隐私合规。

参考文献

[1] 王飞跃，王占宏，李未 . 小目标检测研究综述 [J]. 自动化学报 ,2021,47(1):1 - 14.

[2] 张瑞，王亮，王树新 . 基于 YOLOv4 的课堂学生行为检测方法研究 [J]. 现代教育技术 ,2021,31(5):89 - 95.

[3] 刘小虎，郭玉堂，何伟 . 基于增强分块数据集的课堂小目标行为检测方法 [J]. 电化教育研究 ,2023,44(8):102 - 108.

[4] 陈俊龙，杨静，王耀南 . 基于多尺度特征融合的小目标检测算法 [J]. 控制与决策 ,2020,35(6):1341 - 1348.

[5] 李宏亮，王田苗，王亮 . 基于滑动窗口与置信度融合的小目标检测优化方法 [J]. 计算机工程与应用 ,2022,58(10):123 - 129.

[6] 赵春江，李想，王儒敬 . 教育视频中学生行为自动识别研究综述 [J]. 中国电化教育 ,2020(9):75 - 82.

[7] 何凯明，任少卿，孙剑 . Focal Loss for Dense Object Detection[J]. IEEE Transactions onPatternAnalysisandMachineIntelligence,2018,40(2):2997 - 3007.

[8] 张宇，王亮，李波 . 基于 TensorRT 的 YOLO 模型边缘部署优化研究 [J]. 计算机工程 ,2021,47(12):256 - 262.

[9] 周志华 . 机器学习 [M]. 北京：清华大学出版社 ,2016:187 - 210.

