R-YOLOv8s：一种用于教师课堂教学行为的检测方法

刘丽华[1] 李凤霞[1] 冯余佳[1] 张 伟[1]
LIU Lihua LI Fengxia FENG Yujia ZHANG Wei

摘 要 区别于传统的教学分析模式，人工智能技术为智能化教学分析评价提供了有力支撑。文章基于最新的人工智能技术提出一种新颖的R-YOLOv8s模型，融合ResT（resnet-like vision transformer）架构和极化注意力模型到YOLOv8s模型中，用于对智慧课堂下教师的教学行为进行检测识别。该模型利用Transformer建模图像的全局依赖关系信息，在图像特征提取阶段增强模型对输入图像的特征提取与表达能力，同时在特征采样与融合阶段中以C2PSA模块实现不同尺度特征图的细粒度融合与表征，实现高精度的教师课堂教学行为检测。通过实验表明，相较于基准YOLOv8s模型，R-YOLOv8s在自建教师课堂行为检测数据集上的Precision、Recall、mAP50和mAP50-95指标分别提升了6.0%、4.7%、3.8%和3.0%。

关键词 智慧课堂；教师教学行为；深度学习；目标检测；Transformer

doi：10.3969/j.issn.1672-9528.2025.09.043

0 引言

《近年来，人工智能技术与教育的创新融合成为热门研究领域之一。课堂作为教育的主要场所，教师课堂行为分析是教学分析评价中十分重要的组成部分，如背身板书、指向黑板或课件、与学生互动等行为。随着智慧课堂的普及，海量的课堂数据为将人工智能技术引入推进教育信息化数字化打下良好基础，研究者们逐步探索运用计算机视觉领域相关技术来优化传统的教学分析模式，从更多维度来挖掘智慧课堂中的学生与教师的行为数据，极大地节省了人力、物力、财力成本，进而有针对性地分析此类数据，达到自动化、实时化，也使得教学评价方法更为客观公正，有助于高效地帮助教师了解自身教学情况、及时反思与调整教学方法，促进教学质量的提升。

现有教师课堂行为识别以人体骨骼关键点检测[1]、目标检测[2]等技术为主，侧重于对教师的面部、头部、手势、行为动作等进行一系列分析。郑誉煌[3]基于HRNet评估教师鼻子、左手和右手的关键点信息，并在其基础上构建教学行为评价指标，旨在实现智能化分析和自动化评价；Pang等人[4]则采用Openpose人体姿势估计模型获取教师骨骼数据信息，达到对教师的非言语行为分类的目的；Peng等人[5]将检测算法YOLOv3引入课堂，对教师上课时的手势动作进行识别，将手势划分为象征手势、描述手势和情感手势三类，助力于后续关联教师的言语思维和认知等；江淑娜[6]实现对教师节拍手势识别的量化计算，自定义构建手势数据集，在真实课堂环境中检测成功率可达到95.6%。

但上述方法受限于数据集、早期算法臃肿等，使得其难以实际应用于智慧课堂等场景。因此，本文首先通过采集数据并构建得到一个教师行为检测数据集，其次基于最新的深度学习和目标检测技术，本文提出一种基于YOLOv8s模型的教师课堂教学行为检测算法R-YOLOv8s。

1 R-YOLOv8s模型

在目标检测任务中，YOLO（you only look once）系列算法凭借其出色的检测性能与极低的检测延时备受学术界与工业界的关注。YOLOv8[7]整体包含图像特征提取、特征采样与融合以及模型预测三大阶段。尽管YOLOv8相较于前几代算法无论是在检测精度还是在检测速度上均有不同幅度的提升，但将其应用于教师课堂行为检测任务时依旧存在不足，基于此，本文提出一种融合ResT架构和极化注意力模型的改进YOLOv8s模型。使用ResT（resnet-like vision transformer）架构模型[8]替换原始的CBS和C2f模块，利用Transformer模型建模图像的全局依赖关系信息，更好地提取图像的多尺度特征表达，完成模型对复杂图像的细粒度

1. 江苏开放大学数字化建设中心 江苏南京 210036
[基金项目] 2022年江苏高校思政课教育教学改革创新示范点项目“高校思政课教育教学数字化评价改革创新”（立项编号：12）；2023年度江苏省社科应用研究精品工程课题“江苏提高全民终身学习数字化公共服务均衡性与可及性对策研究”（23SYB-012）；2023年校级教学改革研究课题“基于完全学分制的开放教育数字化平台改革及创新研究”（23-QN-01）

特征描述。其次为了更好地融合多尺度特征图与突出表达目标特征信息，本文在特征采样与融合阶段中提出一种全新的 C2PSA 模块，基于卷积块、瓶颈层以及极化注意力模型设计得到一种梯度流更丰富且目标特征更突出的一种结构，能够较好地突出目标对象特征以及抑制无效的背景信息，实现不同尺度特征图的细粒度融合与增强，为后续的目标精准预测提供特征基础。图 1 为本文提出的模型整体结构框图。

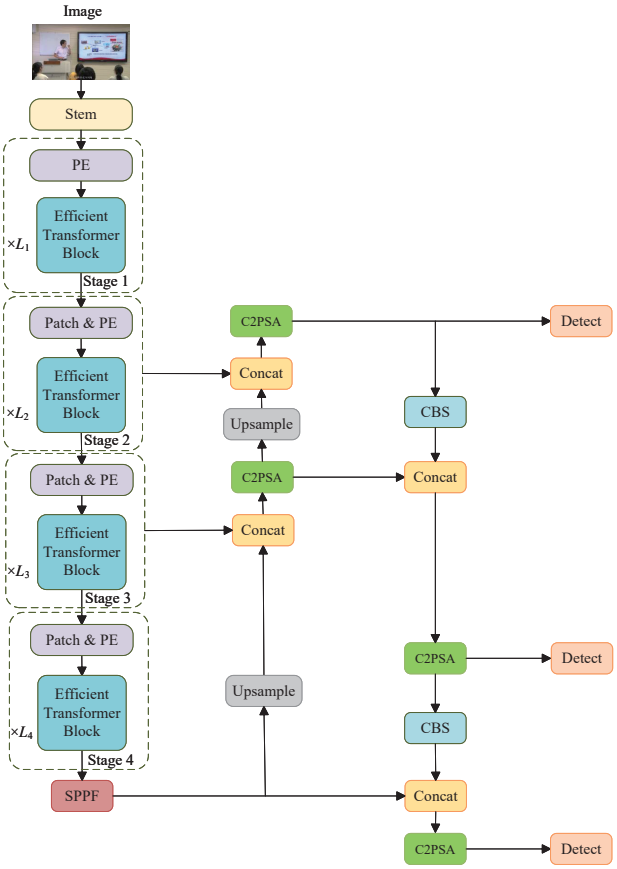


图 1 本文算法结构图

1.1 ResT 模型

在 ResT 模型 [8] 中，首先输入图像经过一个 STEM 模块来提取低维特征，再分别经过 4 个子阶段模块 Stage 来获取不同尺寸下的特征图。第一个 Stage 包含一个图像块嵌入层（Patch Embedding）以及若干个 Efficient Transformer Block；在后续的 3 个 Stage 中，分别包含一个 Patch Embedding、一个位置编码层（Position Encoding）以及若干个 Efficient Transformer Block。Efficient Transformer Block 整体结构如图 2 所示。在 Efficient Transformer Block 中，输入 Token 即 $\boldsymbol{X}$，首先经过一个线性层得到 query 即图中的 $\boldsymbol{Q}$，同时为了减少计算量，$\boldsymbol{X}$ 会由二维变换到三维，之后送入到一个深度可分离卷积和层归一化中压缩特征，再由三维特征图变换到二维特征图，同时分别送入两个独立的线性层来提取特征得到 key 和 value 即图中的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$；再通过式（1）计算注意力函数得到结果 $\boldsymbol{M}$，最后经过一个线性层变换维度外加一个残差连接得到最终的特征 $\boldsymbol{F}$。在进行多头注意力计算时，输入 Token 会平均分为 n 份同时计算，最后的结果即 n 份结果相连。

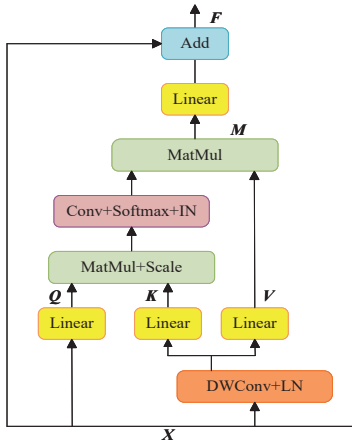


图 2 Efficient Transformer Block 结构图

相较于普通 Transformer 中的多头自注意力模块，Efficient Transformer Block 使用深度可分离卷积降低了输入 Token 的维度和分辨率，使得整个模块的计算复杂度大幅下降，另外在多头注意力计算阶段引入了层归一化操作，用于增强多头之间的信息交互。也正是由于这两大改进，使得 ResT 模型在显著降低计算量的同时表现出比 Transformer 模型更强的特征提取能力。

$$\mathrm{EMSA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{LN}\left(\mathrm{Softmax}\left(\mathrm{Conv}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\right)\right)\boldsymbol{V} \tag{1}$$

式中：EMSA 为 Efficient Transformer Block 的高效多头自注意力模块；Conv 为标准的 1×1 卷积操作，它用于建模不同 head 之间的交互。为了建模不同多头注意力模块之间的交互能力，EMSA 在注意力计算后外加一个 Softmax 和层归一化的操作来统计特征。

此外，在 Position Encoding 模块，ResT 模型提出一种简单高效的像素级注意力（pixel-wise attention, PA）来编码位置。具体而言，PA 采用一个卷积核为 3×3 的深度可分离卷积操作来获得像素级权重，再使用 Sigmoid 激活。使用 PA 获得的位置编码的整体步骤用公式表示为：

$$\boldsymbol{x}^{a} = \mathrm{PA}(\boldsymbol{x}) = \boldsymbol{x} \times \sigma(\mathrm{DWConv}(\boldsymbol{x})) \tag{2}$$

式中：$\boldsymbol{x}$ 为输入特征；$\boldsymbol{x}^{a}$ 为输出特征；PA 是像素级注意力；σ 为 Sigmoid 函数；DWConv 为卷积核 3×3 的深度可分离卷积。

本文方法使用 ResT 模型代替原始的卷积块和 C2f 结构，利用其高效的特征提取能力能更好地捕捉视频帧的细粒度信息，同时，对于本文研究的智慧课堂下的教师课堂行为检测任务，基于长时间依赖关系也能更好地学习教师行为动作的细粒度特征。

1.2 C2PSA 模块

本文聚焦关注空间注意力类方法，在初期该类方法通常需要计算区域之间的注意力权重，这会导致计算复杂度的显著增加，后续（polarized self-attention, PSA）[9] 通过引入极化因子来避免所有区域之间的权重计算，有效降低了该模块的整体计算量，提高模型的效率和准确性。图 3 为 PSA Block 的整体结构。

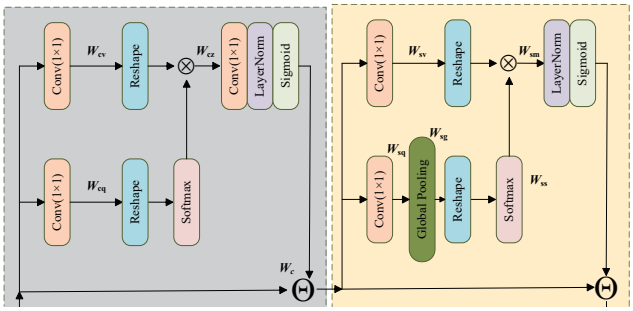


图 3 PSA Block 结构图

PSA Block 有两种独立的形式即并联和序列化，输入特征 $\boldsymbol{X}$ 分别送入仅通道自注意力模块（channel-only self-attention）和仅空间位置自注意力模块（spatial-only self-attention）得到关注后的两个输出特征图，最后将这两个输出特征图逐像素相加即可得到最后的融合结果特征图。在仅通道自注意力模块中，输入特征 $\boldsymbol{X}\in\boldsymbol{F}^{C\times H\times W}$（其中，$\boldsymbol{F}$ 为特征图，C、H 和 W 分别表示特征图的通道数、高和宽）分别经过两个独立的 1×1 卷积得到 $\boldsymbol{W}_{cv}\in\boldsymbol{F}^{C/2\times H\times W}$ 和 $\boldsymbol{W}_{cq}\in\boldsymbol{F}^{1\times H\times W}$，其中 $\boldsymbol{W}_{cv}$ 的通道维度减半，而 $\boldsymbol{W}_{cq}$ 的通道维度为 1 被完全压缩；后续 $\boldsymbol{W}_{cq}$ 在经过维度调整和 Softmax 操作后得到注意力权重，与维度变换后的 $\boldsymbol{W}_{cv}$ 进行矩阵相乘得到注意力加权后的特征 $\boldsymbol{W}_{cz}\in\boldsymbol{F}^{C/2\times 1\times 1}$；之后接着进行 1×1 卷积以及层归一化操作将特征图的通道维度由 $C/2$ 上升到 C，最后使用 Sigmoid 函数对结果归一化再与原始输入特征 $\boldsymbol{X}$ 相乘得到最后的输出 $\boldsymbol{W}_{c}\in\boldsymbol{F}^{C\times H\times W}$。在仅空间位置自注意力模块中，输入特征 $\boldsymbol{X}\in\boldsymbol{F}^{C\times H\times W}$ 同样先经过两个独立的 1×1 卷积得到 $\boldsymbol{W}_{sv}\in\boldsymbol{F}^{C/2\times H\times W}$ 和 $\boldsymbol{W}_{sq}\in\boldsymbol{F}^{C/2\times H\times W}$；后续全局平均池化操作被用于 $\boldsymbol{W}_{sq}$ 来对特征图的通道维度进行压缩得到 $\boldsymbol{W}_{sg}\in\boldsymbol{F}^{C/2\times 1\times 1}$，再进行维度变换将三维特征图压缩成二维特征图，使用 Softmax 对 $\boldsymbol{W}_{sg}$ 的信息进行加权得到 $\boldsymbol{W}_{ss}\in\boldsymbol{F}^{1\times C/2}$，之后 $\boldsymbol{W}_{sv}$ 也先进行维度变换再与 $\boldsymbol{W}_{ss}$ 进行矩阵相乘得到 $\boldsymbol{W}_{sm}\in\boldsymbol{F}^{1\times H\times W}$，最后再进行维度变换将二维特征转换到三维特征并进行 Sigmoid 归一化，与原始输入特征 $\boldsymbol{X}$ 相乘得到最后的输出 $\boldsymbol{W}_{s}\in\boldsymbol{F}^{C\times H\times W}$。输入特征图 $\boldsymbol{X}$ 分别经过仅通道自注意力模块和仅空间位置自注意力模块得到结果输出 $\boldsymbol{W}_{c}$ 和 $\boldsymbol{W}_{s}$ 之后，最后直接将这两个特征图的元素相加即可得到最终的视觉关注后的特征图 $\boldsymbol{W}_{z}$。PSA Block 在通道注意力计算时保证特征图的高通道数，在空间注意力计算时保证了特征图的高分辨率，这种方式能够最大程度上减少特征信息的损失和遗忘，使得该模型能最大限度地突出目标特征信息。同时 PSA Block 充分利用了自注意力结构的建模能力，实现了一种非常有效的长距离建模。

在特征融合阶段中，本文基于原始 YOLOv8s 的 C2f 模块与 PSA Block 提出一种改进的 C2PSA 模块，整体模块结构如图 4 所示。输入特征图 $\boldsymbol{F}_{in}$ 首先经过一个 CBS 模块来提取特征得到 $\boldsymbol{F}_{c1}$，之后 $\boldsymbol{F}_{c1}$ 通过 PSA Block 进一步突出表征目标特征得到 $\boldsymbol{F}_{c2}$，$\boldsymbol{F}_{c2}$ 送入到 Bottleneck 模块进一步缓化特征得到 $\boldsymbol{F}_{b1}$，接着将 $\boldsymbol{F}_{b1}$ 送入另一个 Bottleneck 模块得到输出特征图 $\boldsymbol{F}_{b2}$，最后将 $\boldsymbol{F}_{c2}$、$\boldsymbol{F}_{b1}$ 和 $\boldsymbol{F}_{b2}$ 拼接后送入一个卷积块来融合特征得到最终的输出特征图 $\boldsymbol{F}_{out}$。C2PSA 模块有效以视觉注意力的方式突出特征图中的关键目标特征信息，以细粒度的方式提取教师课堂行为特征，这为后续的正确分类识别提供可靠的特征基础。

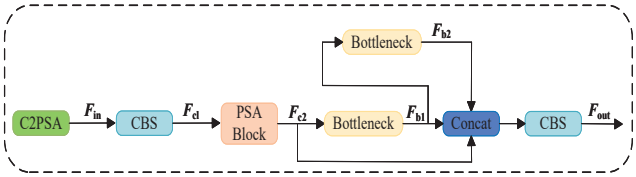


图 4 C2PSA 结构图

2 实验数据集

因现有相关教师行为的公开数据集较少，本文自行构建教师行为检测数据集，主要从全国高校思想政治理论课示范课堂展示中选取教学视频数据作为主要研究对象，将教师在智慧课堂下的行为分为 7 类，分别是板书、指引、环顾、单手势、开放性双手势、保守性双手势、无明显活动。本文本教师行为检测数据集包含 3495 张训练图片，891 张验证图片，1500 张测试图片。

3 实验分析

图 5 为本文提出的 R-YOLOv8s 模型在自建数据集的验证集上的指标结果图。

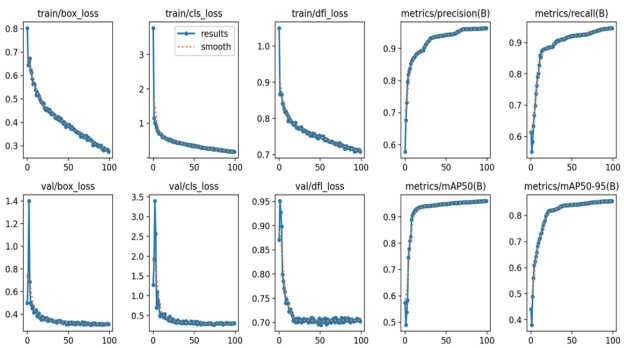
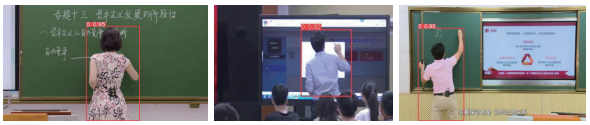


图 5 R-YOLOv8s 模型在自建数据集上的指标结果图

本文方法在验证集上的目标边框损失 box_loss、目标分类损失 cls_loss 以及 DFL 损失 dfl_loss 分别能收敛到 0.313、0.303 和 0.702。相比于原始的 YOLOv8s 模型，更强的特征提取网络带来更细粒度的目标特征提取，最终使得目标损失下降。在 Precision、Recall、mAP50 和 mAP50-95 指标上，本文方法分别达到 96.3%、94.6%、96.0% 和 85.6%。

4 可视化展示

图 6 直观地展示了本文提出的 R-YOLOv8s 模型在真实智慧课堂下对 7 个教师行为类别的实际检测效果。从图 6 中可以看出，对于指引和开放性双手势的行为检测准确率可达到 97%，这验证了本文方法对于教师行为识别具有良好的效果，能够满足正常的全过程教学的教师行为检测的任务需求。



（a）板书

（b）指引

（c）环顾

（d）单手势

（e）开放性双手势

（f）保守性双手势

（g）无明显活动

图 6 可视化检测结果

5 结语

针对教师课堂行为检测任务，本文提出一种改进的 R-YOLOv8s 模型，其有效地增强了模型对输入图像的特征提取能力，同时 C2PSA 模块能够有效地突出目标对象特征信息，且在自建教师课堂行为检测数据集上，R-YOLOv8s 在多个指标上相较于基准模型 YOLOv8s 有提升明显。

参考文献：

[1] 张宇，温光照，米思娅，等．基于深度学习的二维人体姿态估计综述 [J]. 软件学报，2022, 33(11): 4173−4191.

[2] 程旭，宋晨，史金钢，等．基于深度学习的通用目标检测研究综述 [J]. 电子学报，2021,49(7):1428-1438.

[3] 郑誉煌．一种基于姿势识别的教师教学行为评价方法 [J]. 软件工程，2021, 24 (4):6-9.

[4] PANG S Y, ZHANG A R, LAI S H, et al. Automatic recognition of teachers' nonverbal behavior based on dilated convolution[C]//2ICETC'22: Proceedings of the 14th International Conference on Education Technology and Computers. NewYork:ACM,2023: 429-435.

[5] PENG Z L, YANG Z D, XIAHOU J B, et al. Recognizing teachers' hand gestures for effective non-verbal interaction[J]. Applied sciences. 2022, 12(22): 11717.

[6] 江淑娜 .YOLOv5 实现网络环境下教师节拍手势智能识别 [J]. 福建电脑 ,2023,39(9):8-13.

[7] RAHIM A, YUAN F Q, BARABADY J. An ultralytics YOLOv8-based approach for road detection in snowy environments in the arctic region of norway[J]. Computers, materials & continua, 2025, 83(3):4411-4428.

[8] ZHANG Q L, YANG Y B. ResT: an efficient transformer for visual recognition[EB/OL]. (2010-10-14)[2024-05-25].https://doi.org/10.48550/arXiv.2105.13677.

[9] LIU H J , LIU F Q, FAN X Y ,et al.Polarized self-attention: towards high-quality pixel-wise regression[J].Neurocomputing, 2022,56:158-167.

【作者简介】

刘丽华（1998—），女，江苏泰州人，硕士，研究方向：计算机视觉、教育信息化。

李凤霞（1995—），女，山东菏泽人，硕士，研究方向：在线教育、教育信息化。

冯余佳（1992—），女，浙江衢州人，硕士，研究方向：在线教育、教育信息化。

张伟（1995—），男，江苏南京人，硕士，研究方向：在线教育、教育信息化。

（收稿日期：2025-02-19　修回日期：2025-07-28）