基于改进 YOLO11 的学生课堂行为识别研究

高 扬[1]
GAO Yang

摘 要 目前计算机视觉技术逐步应用于学生课堂行为识别场景，但学生课堂行为存在后排学生识别精度低、遮挡场景识别不准确等问题，文章基于 YOLO11 提出一种改进的学生课堂行为识别算法。通过在主干网络中引入感受野注意力卷积（RFAConv），扩大空间特征感受野，实现对全局信息的捕捉；在 Neck 部分中小目标检测层后加入多尺度膨胀注意力机制（MSDA），提升对遮挡场景中后排学生行为的识别准确率，同时提出自适应特征融合（ASFF）检测头，使用自适应特征融合机制对不同尺度的图像进行融合，提升检测头对图像的检测精度。在自建学生课堂行为数据集以及在公共数据集 SCB-Dataset3-S 上的 mAP50 值相较于基线 YOLO11n 模型分别提升 1.8% 与 2.1%。为学生课堂行为的深入分析提供了有效的途径与方法，对智慧课堂的发展具有重要意义。

关键词 YOLO；自适应特征融合；感受野注意力卷积；智慧课堂

doi：10.3969/j.issn.1672-9528.2025.08.002

0 引言

随着人工智能的不断进步，计算机视觉技术在教育领域内的应用范围正迅速扩大。在此背景下，智慧课堂的建设已成为推动教育发展的重要趋势 [1-2]。其中，学生课堂行为识别作为智慧教育的重要组成部分，受到了越来越多的关注。通过识别学生的课堂行为，教师可以及时了解学生的学习状态，调整教学策略，从而提高教学效率和质量，对于智慧教育的发展具有重要意义 [3]。

传统的课堂行为识别方法主要依赖人工观察或课后教师观看监控视频再分析等方法，这些方法因人工观察和分析无法提供实时反馈而导致效率低等问题 [4]。相较之下，使用深度学习方法对课堂行为识别效率更高，对于教育工作方面的应用具有重要意义。Lü 等人 [5] 使用 RF-SSD 模型检测课堂中学生的行为，通过替换 VGG 为 ResNet101 骨干网络与在骨干网络后引入 FPN 模块来提升对于不同特征图像识别的准确度；Zhu 等人 [6] 提出了 Csb-YOLO 模型，将 BiFPN 结构引入 YOLOv8 模型的同时设计了 C2f_SCConv，并采用剪枝优化模型结构，实现精度提升与参数量的降低；刘宏等人 [7] 提出基于视频时空特征的 YOLOv7-SlowFast 检测方法，通过结合 YOLOv7 与 SlowFast 对连续性课堂行为检测效果进行优化，加强对时间特征的感知，有效检测出学生课堂行为。

综上所述，在深度学习方面对课堂行为识别的研究已取得一定成果，但仍存在很多问题。大多数现有方法聚焦于课堂场景下特征的变化及摄像头角度的问题而忽略了在遮挡及密集场景下学生行为识别的准确率以及在不同角度监控下对后排学生行为的识别。为解决上述问题，本文提出一种基于改进 YOLO11 的学生课堂行为检测算法，用于学生课堂行为识别。本文主要贡献如下：

（1）在 YOLO11 的主干网络中引入感受野注意力卷积，使模型在忽略计算成本和参数量增加的同时，从空间特征转移到了感受野空间特征，提高了网络的性能，提升遮挡以及密集课堂场景下对学生行为的检测准确率。

（2）在 Neck 部分中小目标检测层后引入多尺度膨胀注意力机制，有效获得了上下文语义信息，提升对后排学生的识别准确率，从而解决对课堂场景下的后排学生目标的识别问题。

（3）提出自适应特征融合检测头，优化 YOLO11 检测头，提升不同尺度图像的处理能力，增强整个模型在课堂场景下的识别准确率。

1 YOLO11 算法

YOLO11 是 YOLO 系列最新一代的目标检测模型，继承并扩展了前几代的优势，并引入多项创新以提升检测精度和速度。根据网络深度与宽度可以分为 *n*、*s*、*m*、*l*、*x* 五种检

1. 太原师范学院计算机科学与技术学院 山西晋中 030619
[基金项目] 山西省科技战略研究专项重点项目（202304031401011）；山西省信创产业科教融汇、产教融合的模式与实施路径研究

测模型。

其网络架构的主要组成部分包括 Input（输入）、Backbone（骨干网络）、Neck（颈部网络）和 Head（头部网络），其中输入部分为对输入图像进行预处理和增强，以便为后续的特征提取和检测任务提供输入数据；骨干网络由 Conv、C3k2、SPPF 与 C2PSA 构成，Conv 为标准卷积层，负责提取输入图像的特征并传递给后续的网络层，C3k2 是在 C2f 模块的基础上做出了改进，通过设置 C3k 参数为 True 或者 False 来决定具体操作，当 C3k 为 False 时，C3k2 模块就是普通的 C2f 模块，当 C3k 为 True 时，C3k2 模块变成了 C3 模块，利用这种可切换的设计使得模型可以根据不同的需求和场景选择更合适的特征提取方式，增强了网络的检测性能；SPPF（空间金字塔池化）模块通过串联多个最大池化层实现了高效的多尺度特征提取；C2PSA 模块结合了 PSA 模块，通过多头注意力机制与前馈神经网络增强了特征提取能力；颈部网络由 FPN-PAN 结构组成，由一条自顶向下与自底向上的路径进行特征融合，高效且速度快；头部网络生成了最后的输出部分，包括目标的类别、边界框（Bounding Box）和置信度（Confidence）。

2 本文算法

为解决学生课堂场景中对于密集场景以及后排学生识别准确率低等问题，本文提出改进后的 YOLO11 学生课堂行为识别算法。在主干网络中引入感受野注意力卷积，依次替换主干网络中 0 层之后的所有标准卷积，强调感受野内各个特征的重要性，扩大空间特征感受野，使整体的检测精度提升；在 Neck 部分中小目标层后引入多尺度膨胀注意力机制，通过对不同自注意力头的多尺度膨胀率操作，满足了局部性与稀疏性，使得对后排学生的识别准确率提升；提出自适应特征融合检测头，在 YOLO11 检测头分类与预测之前引入自适应特征融合机制加强对不同尺度图像的特征融合。改进后的模型结构图如图 1 所示。

2.1 感受野注意力卷积

标准卷积运算凭借局部感受野的优势精准地捕捉输入数据中的空间局部特性。但这种卷积方式在特征抽取过程中难以处理不同位置间的信息差异，这导致了部分重要特征的遗漏，进而影响模型的整体性能，表现为漏检和误检的情况。

本文提出一种改进策略，在 YOLO11 主干网络中引入感受野注意力卷积（receptive field attention, RFA）[8]。在特征图获取之后执行平均池化（AvgPooL）操作以汇总感受野特征中的全局信息。然后采用 1×1 组卷积操作进行信息交互，并对整个过程应用 Softmax 函数进行归一化处理，从而生成注意力图 A_{rf}，强调感受野特征内各特征的重要性初始特征图经历一系列卷积操作后被转换成与注意力映射相匹配的大小和维度的感受野空间特征 F_{rf}。随后，通过将注意力映射的权重施加于感受野空间特征 F_{rf} 上，依据这些权重调整特征提取过程并对其进行尺度变换，得到最终输出结果 P。

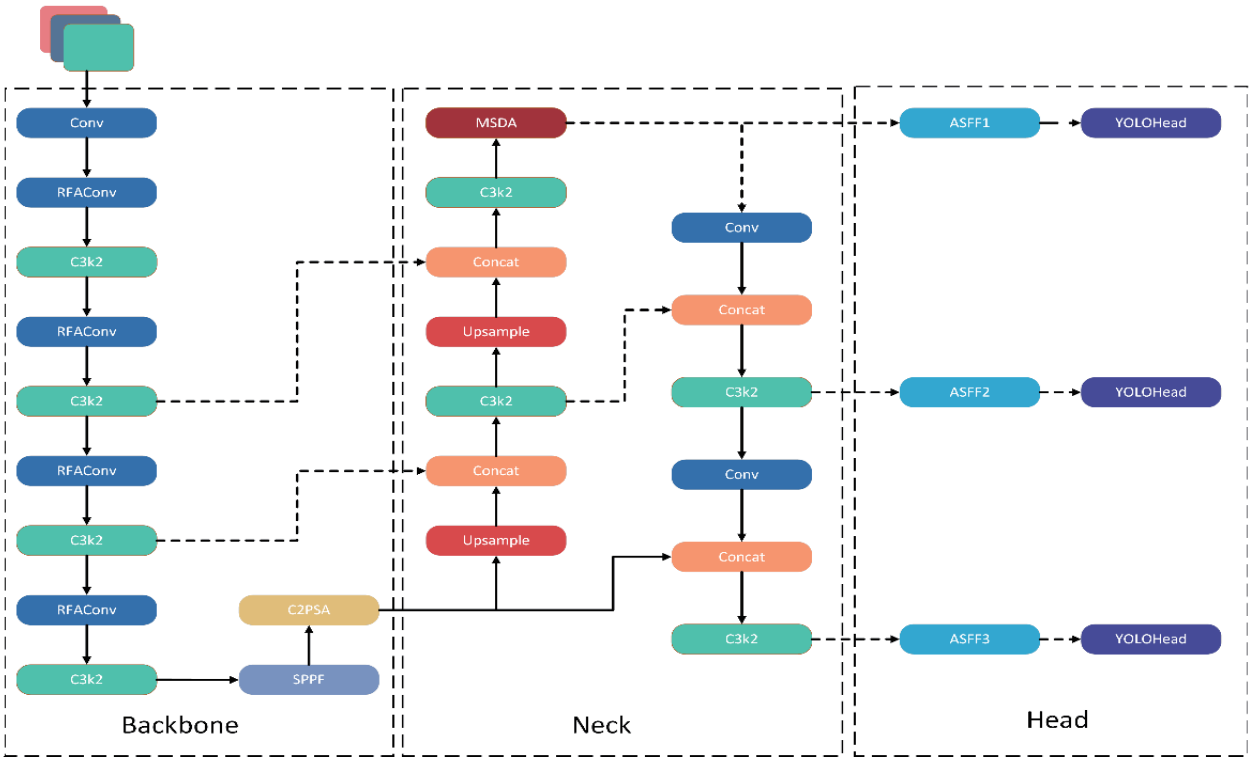

图 1 改进后的模型

感受野注意力卷积的计算方法为：

$$A_{rf} = \text{Softmax}(g^{1\times1}(\text{AvgPool}(X))) \quad (1)$$

$$F_{rf} = \text{ReLU}(\text{Norm}(g^{i\times i}(X))) \quad (2)$$

$$P = A_{rf} \times F_{rf} \quad (3)$$

2.2 多尺度膨胀注意力机制

本研究引入多尺度膨胀注意力机制（multi-scale dilated attention, MSDA）[9]，在 Neck 部分中小目标检测层后加入 MSDA 注意力机制，解决课堂场景中后排学生行为的识别不准确的问题。该注意力机制结构如图 2 所示。

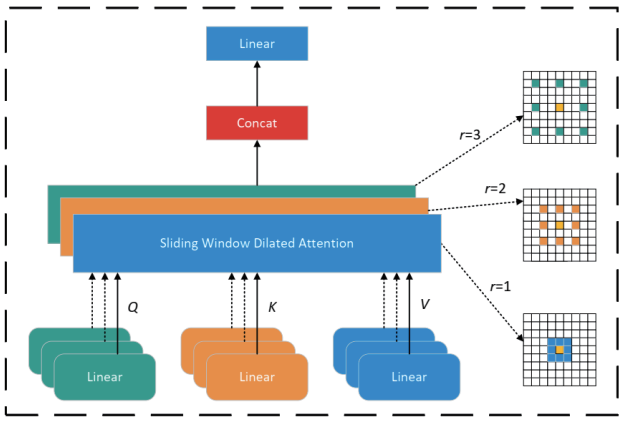

图 2 多尺度膨胀注意力机制

给定特征图 X，通过线性层分别获得对应的 Q、K、V，然后将对应特征图的通道划分为 n 个注意力头，并在不同的注意力头上按照不同的膨胀率执行多尺度的 SWDA。其中 SWDA 是一种以滑动窗口的方式对于中心特征图进行自注意力的操作，可以有效地获得上下文的语义信息，并满足对应的局部性与稀疏性。在执行完 SWDA 后，便通过在头部设置不同大小的内核与不同的膨胀率把不同头部的输出连接在一起，从而更准确地识别出学生行为。MSDA 的计算公式为：

$$h_i = \text{SWDA}(Q_i, K_i, V_i, r_i), \quad (4)$$

$$1 \le i \le n, X = \text{Linear}(\text{Concat}[h_1, ..., h_n]) \quad (5)$$

2.3 自适应特征融合检测头

本研究利用自适应特征融合（ASFF）[10] 机制提出了新的检测头 ASFFHead，提升在课堂遮挡场景下的识别精度，加强不同尺寸检测层的特征融合，具体结构如图 3 所示。

Level 0、Level 1、Level 2 为 Neck 部分后不同层的输出结果，把不同层级的特征 Level 0、Level 1、Level 2 调整到相同的分辨率后将不同层级的特征进行适应性融合，利用 ASFF 对不同尺度融合时的空间权重进行自适应调整的特性将输出后不同层的不同尺度的特征进行互相学习，由此提高检测的准确率，从而实现了不同层之间更好的融合效果，提升对于课堂场景下学生行为的检测精度。

3 实验与分析

3.1 数据集

为验证改进算法的有效性，本文使用自建学生课堂行为数据集与公开课堂行为数据集（SCB3-Dataset-S）[11] 进行验证。

本文自建一个学生课堂行为数据集，并在正面角度、45° 角度及其他部分角度的监控下收集公开教育平台下公开课的学生课堂数据，并选择出现次数最高的书写（Writing）、阅读（Reading）、听课（Listening）、站立（Standing）、举手（Raising Hand）、转身（Turning Around）、讨论（Discussing）7 种行为进行标注，并筛选出模糊、行为不明确等无效数据。数据集一共 4 500 张图片，按照 7:1:2 的比例随机划分为训练集、验证集和测试集。具体行为分类定义如表 1 所示。

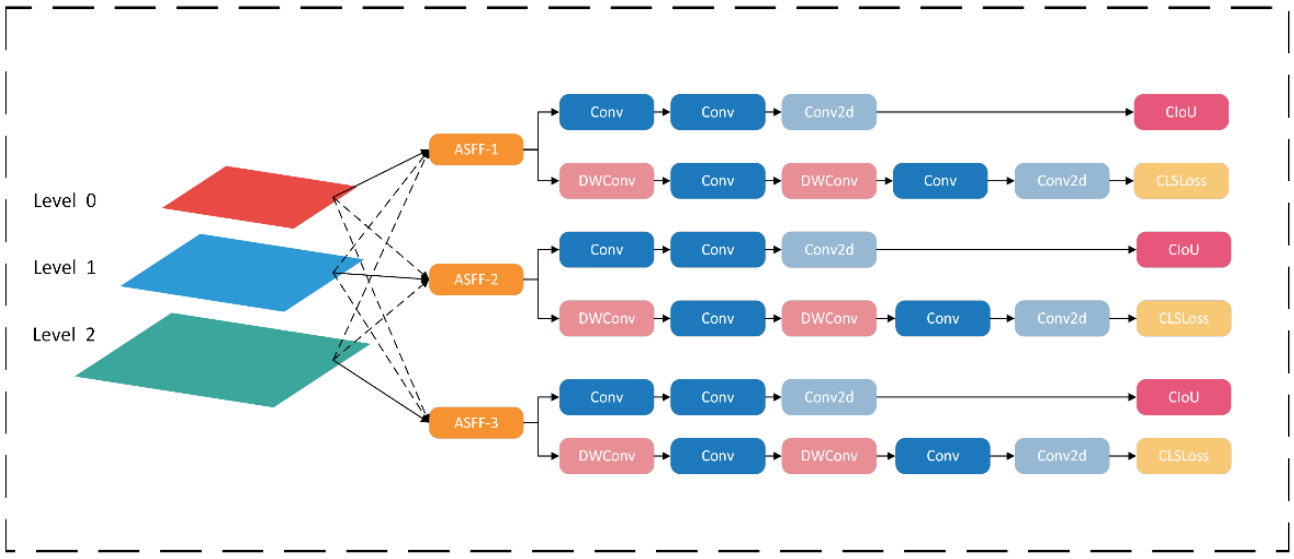

图 3 自适应特征融合检测头

表 1 行为定义

标签	行为	行为描述
Writing	书写	学生低头握笔
Reading	阅读	学生手拿课本看书
Listening	听课	学生抬头正视前方听讲
Standing	站立	学生双手自然下垂并站立
Raising Hand	举手	学生双手举起
Turning Around	转身	学生身体大幅倾斜或转头
Discussing	讨论	2 个及以上学生讨论

公开课堂行为数据集（SCB3-Dataset-S）数据集包含从幼儿园到高中的课堂行为数据，分类为举手（Raising Hand）、阅读（Reading）、书写（Writing），数据集一共 5 015 张图片，由于样本分类较少且样本数平衡，因此本文按照 8∶1∶1 的比例随机划分为训练集、验证集和测试集进行训练测试。

3.2 实验环境

本研究通过 PyTorch 框架来实现，实验环境如表 2 所示。

表 2 实验环境表

名称	版本
CPU	Intel(R)Xeon(R)Gold 6152 CPU @ 2.10 GHz
GPU	RTX 3090 24 GB
编程语言	Python3.8
操作系统	Ubuntu20.04
深度学习框架	PyTorch1.13.1+CUDA 11.7

其中在实验过程中参数设置如下：epochs 为 200，batch 为 16，优化器选择为 SGD，学习率为 0.01。

3.3 消融实验

为验证本文改进算法的有效性，本节设计了如下的消融实验。实验结果如表 3~4 所示。

在自建学生课堂行为数据集上的消融实验表明，在使用 RFAConv 时，平均精度提升 0.8%，在 Neck 的小目标层后加入 MSDA 后，平均精度提升 0.8%，在替换 YOLO11 的检测头为 ASFFHead 后平均精度提升 1.2%，在加入 RFAConv+MSDA 后平均精度提升 1.2%，在加入 RFAConv+ASFFHead 后平均精度提升 1.6%，在加入 MSDA+ASFFHead 后平均精度提升 1.3%，同时加入三者后平均精度提升 1.8%，说明改进的算法能够有效提升学生课堂行为识别的精度，在学生课堂行为检测方面具有重要意义。

在 SCB3-Dataset-S 数据集上的消融实验表明，在分别单独加入 RFAConv、MSDA、ASFFHead 的时候，mAP50 分别提升了 1.4%、0.5%、1.2%，在三者同时加入时总共提升了 2.1%，其余情况分别都提升 1.7%，表明在开源学生课堂行为数据集上同样有着优异的性能表现。

3.4 对比实验

为验证本文改进算法的性能，本文在自建数据集和 SCB-Dataset3-S 上分别进行对比实验，使用准确率、召回率、mAP50、mAP50-95、计算量 5 种指标进行评估，如表 5~6 所示。通过对在自建学生课堂行为数据集上的对比实验，在精度方面本文算法均优于其他模型，在 mAP50 指标上，比 YOLOv5s、YOLOv8n、YOLOv10n 分别提升 1.6%、1.5%、1.3%，然后在 GFLOPs 上也均优于其他模型，保持精度的同时，也具有很好的实时性。在 SCB-Datasets-S 数据集的对比上，本文算法在精度方面也优于其他模型，在 mAP50 指标上，比 YOLOv5s、YOLOv8n、YOLOv10n、YOLO11n 分别提升 2.1%、2.7%、3.4%、2.1%，在 GFLOPs 上同样也均优于其他模型，展现了优秀的性能。

表 3 自建学生课堂行为数据集的消融实验

模型	*P*	*R*	mAP50	mAP50-95	GFLOPs
YOLO11n	0.829	0.792	0.853	0.633	6.3
RFAConv	0.84	0.798	0.861	0.64	6.6
MSDA	0.838	0.792	0.861	0.638	6.5
ASFFHead	0.849	0.804	0.865	0.65	8.5
RFAConv + MSDA	0.829	0.808	0.865	0.645	6.9
RFAConv + ASFFHead	0.857	0.802	0.869	0.65	8.8
MSDA + ASFFHead	0.838	0.806	0.866	0.651	8.7
RFAConv + MSDA + ASFFHead	0.844	0.813	0.871	0.65	9.0

表 4 SCB3-Dataset-S 数据集的消融实验

模型	*P*	*R*	mAP50	mAP50-95	GFLOPs
YOLO11n	0.647	0.681	0.701	0.508	6.3
RFAConv	0.663	0.67	0.715	0.523	6.6
MSDA	0.64	0.701	0.706	0.513	6.5
ASFFHead	0.683	0.657	0.713	0.527	8.5
RFAConv + MSDA	0.681	0.666	0.718	0.528	6.8
RFAConv + ASFFHead	0.646	0.713	0.718	0.524	8.8
MSDA + ASFFHead	0.677	0.675	0.718	0.533	8.7
RFAConv + MSDA + ASFFHead	0.671	0.673	0.722	0.532	9.0

表 5 模型对比实验（自建数据集）

模型	*P*	*R*	mAP50	mAP50-95	GFLOPs
YOLOv5s	0.855	0.781	0.855	0.63	15.8
YOLOv8n	0.846	0.806	0.856	0.638	8.1
YOLOv10n	0.831	0.785	0.858	0.639	8.2
YOLO11n	0.829	0.792	0.853	0.633	6.3
本文算法	0.844	0.813	0.871	0.65	9.0

表 6 模型对比实验（SCB-Dataset3-S 数据集）

模型	*P*	*R*	mAP(0.5)	mAP(0.5:0.95)	GFLOPs
YOLOv5s	0.645	0.700	0.701	0.482	15.8
YOLOv8n	0.67	0.655	0.695	0.504	8.1
YOLOv10n	0.637	0.672	0.688	0.495	8.2
YOLO11n	0.647	0.681	0.701	0.508	6.3
本文算法	0.671	0.673	0.722	0.532	9.0

3.5 结果分析

为验证本文改进的模型在数据集上的效果，本文在自建学生课堂行为数据集上进行了检测结果的可视化展示，通过可视化结果可以看出在课堂场景下，本文算法对遮挡下学生课堂行为与后排学生的课堂行为识别的准确率均有一定提升。本文算法对具体不同场景的检测结果可视化如图 4 所示。

（a）基线模型　　（b）改进后的模型

图 4 可视化结果展示

4 结语

针对学生课堂行为识别在密集场景及对于后排学生识别准确率差的问题，本文基于 YOLO11 提出一种改进后的算法。实验结果表明该模型在学生课堂行为识别方面表现出明显优势，同时具备较强的泛化能力。为智慧教育的发展以及教师教学评估提供了切实可行的解决方案。未来研究应进一步拓展模型在不同课堂场景中的应用潜力，并结合更大规模的多模态数据集，探索更加高效的检测框架，为教育行业做出更大的贡献。

参考文献：

[1] 马瑞祯，徐娟．语言智能赋能国际中文智慧教育：现实境况与未来路向 [J]. 国际中文教育（中英文），2023, 8(2): 43-52.

[2] 李晓丹．大数据视角下的高校智慧课堂建设 [J]. 电脑知识与技术，2022,18(16):136-137.

[3] CHEN H W, ZHOU G H, JIANG H X. Student behavior detection in the classroom based on improved YOLOv8[J]. Sensors, 2023, 23(20): 8385.

[4] 贺子琴，黄文辉，肖嘉彦，等．基于 YOLOv5 的学生课堂行为分析系统设计 [J]. 电脑知识与技术，2023, 19(26):19-22.

[5] LÜ W C, HUANG M, ZHANG F Y, et al. Research on intelligent recognition algorithm of college students' classroom behavior based on improved SSD[C/OL]//2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI).Piscataway: IEEE, 2022[2025-01-03].https://ieeexplore.ieee.org/document/9807756. DOI:10.1109/CCAI55564.2022.9807756.

[6] ZHU W Q, YANG Z J. Csb-YOLO: a rapid and efficient real-time algorithm for classroom student behavior detection[J]. Journal of real-time image processing, 2024, 21(4): 140.

[7] 刘宏，张释文，孙程，等．基于视频时空特征的学生课堂行为检测 [J]. 软件导刊，2024, 23(8): 267-274.

[8] ZHANG X, LIU C, YANG D G, et al. RFAConv: innovating spatial attention and standard convolutional operation[EB/OL].(2023-04-06)[2024-05-01].https://github.com/Liuchen1997/RFAConv.

[9] JIAO J Y, TANG Y M, LIN K Y, et al. DilateFormer: multi-scale dilated Transformer for visual recognition [J].IEEE transactions on multimedia, 2023, 25: 8906-8919.

[10] LIU S T, HUANG D, WANG Y H. Learning spatial fusion for single-shot object detection[EB/OL]. (2019-11-25)[2025-02-14].https://doi.org/10.48550/arXiv.1911.09516.

[11] YANG F, WANG T. SCB-Dataset3: a benchmark for detecting student classroom behavior[EB/OL]. (2023-10-04)[2024-05-01]. https://arxiv.org/pdf/2310.02522v2.

【作者简介】

高扬（2000—），男，山西忻州人，硕士研究生，研究方向：图像处理、目标检测。

（收稿日期：2025-03-26 修回日期：2025-08-01）