

基于轻量化YOLOv8模型的校园安全隐患目标检测应用研究

冉冰艺

重庆外语外事学院智能科学与工程学院, 重庆

收稿日期: 2026年8月10日; 录用日期: 2026年9月14日; 发布日期: 2026年9月21日

摘要

校园安全隐患的实时检测是保障师生生命财产安全的重要任务。现有监控系统依赖人工值守, 预警滞后, 而通用深度学习模型参数量大、计算复杂度高, 难以在算力有限的边缘设备上实时运行。针对上述问题, 本文提出一种基于协同优化策略的轻量化YOLOv8n校园安全隐患检测方案。该方案并非简单堆叠已有技术, 而是遵循“精度保持优先、逐步压缩冗余”的阶梯式设计理念, 构建了CBAM注意力机制、Ghost轻量化模块、结构化剪枝与FP16量化的递进式协同体系: CBAM为后续压缩提供特征显著性先验, Ghost从结构源头削减冗余, 剪枝在注意力引导下精准剔除不重要通道, 量化则以极小精度代价换取边缘端加速。四者形成“精度增强→结构轻量→通道裁剪→数值压缩”的互补链条, 每一环节的精度损失由前置或后序操作进行补偿。在自建的三类别校园安全隐患数据集上, 改进模型mAP@0.5达到87.4%, 较基线仅下降1.8个百分点, 参数量从3.01 M降至1.72 M (减少42.9%), 浮点运算量降低约45%。在Jetson Nano和RK3588边缘平台上经TensorRT优化后分别达到31 FPS和48 FPS的实时推理速度。本文构建的多类别统一轻量化检测框架, 为民办高校低成本智能化安全监控提供了可落地的技术方案。

关键词

YOLOv8, 轻量化, CBAM注意力机制, 校园安全隐患, 目标检测, 边缘部署

Applied Research on Object Detection for Campus Safety Hazards Based on a Lightweight YOLOv8 Model

Bingyi Ran

College of Intelligent Science and Engineering, Chongqing Institute of Foreign Studies, Chongqing

Received: August 10, 2026; accepted: September 14, 2026; published: September 21, 2026

文章引用: 冉冰艺. 基于轻量化 YOLOv8 模型的校园安全隐患目标检测应用研究[J]. 计算机科学与应用, 2026, 16(9): 199-211. DOI: 10.12677/csa.2026.169301

Abstract

Real-time detection of campus safety hazards is an important task for ensuring the safety of teachers and students' lives and property. The existing monitoring systems rely on manual operation and have delayed warnings. Moreover, the general deep learning models have large parameters and high computational complexity, making it difficult to run them in edge devices with limited computing power. To address these issues, this paper proposes a lightweight YOLOv8n campus safety hazard detection scheme based on a collaborative optimization strategy. This scheme does not simply stack existing technologies but follows a stepwise design concept of "priority for accuracy, gradual reduction of redundancy", and builds a progressive collaborative system of CBAM attention mechanism, Ghost lightweight module, structured pruning and FP16 quantization: CBAM provides feature saliency priors for subsequent compression, Ghost reduces redundancy from the structural source, pruning precisely eliminates unimportant channels under the guidance of attention, and quantization sacrifices a small precision cost for acceleration on the edge. The four elements form a complementary chain of "accuracy enhancement → structure lightness → channel pruning → numerical compression", with each step's accuracy loss compensated by the preceding or subsequent operations. On the self-built three-category campus safety hazard dataset, the improved model's mAP@0.5 reaches 87.4%, only decreasing by 1.8 percentage points compared to the baseline, and the parameter size is reduced from 3.01 M to 1.72 M (a 42.9% reduction), and the floating-point operation volume is reduced by approximately 45%. After optimization with TensorRT on Jetson Nano and RK3588 edge platforms, the real-time inference speeds reach 31 FPS and 48 FPS respectively. The multi-category unified lightweight detection framework constructed in this paper provides a feasible technical solution for low-cost intelligent safety monitoring in private colleges.

Keywords

YOLOv8, Lightweight, CBAM Attention Mechanism, Campus Safety Hazards, Object Detection, Edge Deployment

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 绪论

1.1. 研究背景与意义

近年来，民办高校规模不断扩大，校园安全形势日趋复杂。实验楼消防隐患、重点区域非法入侵等事件时有发生，如实验室化学品引燃火焰、电路老化导致的烟雾聚集，以及天台、配电房、人工湖周边等禁区的违规闯入，给师生人身安全和学校财产带来严重威胁。传统的视频监控系统主要依靠安保人员长时间注视屏幕，不仅容易因疲劳导致漏报，而且从事件发生到人工确认并处置存在明显的时间滞后，难以实现事前预警。

随着深度学习技术的快速发展，基于卷积神经网络的目标检测算法在安防领域展现出巨大潜力。然而，主流的 YOLO 系列检测模型追求高精度，往往伴随着较大的参数数量和计算开销，难以在计算资源、内存和功耗均受限的边缘设备上实现实时推理[1]。民办高校在安全监控智能化升级过程中，普遍面临经费紧张、IT 基础设施薄弱的现实约束，亟需一种能够在低成本边缘设备上高效运行、同时对多类安全隐患保持较高检测精度的轻量化模型。

本文的研究具有两方面的意义。在理论上,探索了 YOLOv8 在复杂校园场景下“精度-效率”平衡的优化路径,系统融合注意力机制与多种轻量化策略,为单阶段目标检测网络的边缘化部署提供了方法论参考。在实践上,通过自建多类别校园安全隐患数据集和完整的“训练-压缩-部署”流程,为民办高校提供了一套低成本、可复制的智能化安全监控技术方案,具备显著的应用推广价值。

1.2. 国内外研究现状

YOLOv8 算法改进研究。YOLOv8 作为 Ultralytics 推出的新一代检测框架,在骨干网络、特征融合和检测头等方面较前代有所革新,但其标准模型在移动端和嵌入式场景中的资源消耗依然偏高。相关研究主要从三个方向进行改进:网络结构上,引入可变形卷积、Transformer 模块或重参数化技术以增强特征表达;损失函数方面,采用 Focaler-IoU、Wise-IoU 等动态边界框损失以提升收敛效果;检测头层面,设计轻量化解耦头或自适应锚点分配策略。这些改进多聚焦于精度提升,对模型体积和速度的考虑尚不充分。

注意力机制在 YOLO 系列中的应用。注意力机制通过显式建模通道间和空间上的依赖关系,能够有效增强模型对关键特征的聚焦能力。CBAM (Convolutional Block Attention Module) 以其即插即用的特点被广泛应用于 YOLOv5、YOLOv8 等模型的颈部或头部,在小目标、遮挡目标和复杂背景场景下均取得了稳定的精度增益[2]。ShuffleAttention、ECA 等轻量注意力模块则进一步降低了额外计算开销,但特征增强效果略逊于 CBAM。

YOLOv8 模型轻量化研究。轻量化是边缘部署的核心需求,目前主流的压缩手段包括:轻量化模块替换,如利用 GhostNet 中的 Ghost 模块以低成本生成冗余特征图,显著降低标准卷积的 FLOPs [3];结构化剪枝,通过对批归一化层的缩放因子施加稀疏约束,裁剪不重要的通道从而压缩模型体积[4];知识蒸馏,以大模型为教师网络指导轻量模型训练。此外,FP16 半精度量化和 TensorRT 优化推理也被广泛应用于提升边缘侧吞吐率[5]。然而,将这些技术系统组合并应用于校园安全隐患检测的研究仍较匮乏。

校园安全隐患检测研究。现有面向校园安全的检测工作多集中于单一类型隐患,例如仅针对火焰和烟雾的火灾检测[6][7],或仅针对围栏边界的人员入侵检测[8]。这些方法缺乏对多类隐患的统一建模,无法满足真实场景下一体化监控的需求。此外,多数研究仅在服务器端验证模型精度,未充分考虑在资源受限的边缘设备上的实时性,实际部署可行性存疑。

重庆市内相关研究。重庆市多所高校在 YOLO 系列模型的安防应用中开展了探索,如将改进 YOLOv5 用于教学楼出入口人员计数、将 YOLOv7-tiny 用于宿舍区烟火预警等[9][10],但普遍存在模型体积较大、推理帧率不足、无法同时覆盖多类隐患等问题,面向校园安全隐患的轻量化检测研究尚属空白。

现有研究的不足与本文的研究切入点。综合来看,现有工作存在三个主要不足:第一,校园安全隐患检测多聚焦单一类别,缺乏统一的轻量化多类别检测框架;第二,注意力机制与多种轻量化技术的协同优化研究尚不深入,缺乏从精度增强到模型压缩的系统性方案;第三,多数研究停留在仿真实验,缺少在真实边缘硬件上的部署验证。本文正是针对上述不足,提出完整的轻量化 YOLOv8 改进方案,并在自建多类别数据集与边缘平台上完成从训练到部署的全流程研究。

本文技术选型与协同优化思路

针对上述不足,本文基于“精度保持优先、逐步压缩冗余”的阶梯式设计理念,构建了一套有机协同的改进方案。技术选型的内在逻辑如下:

CBAM 注意力机制作为“精度保底”模块,用于弥补后续压缩带来的特征表达能力下降。其通道与空间双重注意力能够显著增强对火焰初期微弱信号、远处烟雾纹理及极小人员轮廓的响应,为后续剪枝保留关键特征通道提供“显著性导向”。

Ghost 轻量化模块从结构源头削减冗余,利用廉价的深度卷积线性变换生成冗余特征图,在不显著损

伤表示能力的前提下大幅降低参数量和计算量。它承担“粗压缩”角色，为剪枝提供更精简的初始结构，缩小剪枝算法的搜索空间。

结构化剪枝作为“细压缩”手段，通过 L1 正则化稀疏训练使 BN 层缩放因子 γ 趋于稀疏，再剔除低重要性通道。剪枝后的精度损失恰好由前置 CBAM 所补偿——CBAM 在训练阶段强化了重要通道的响应，使得稀疏训练中关键通道的 γ 值保持较高幅值，冗余通道的 γ 值趋近于零，实现了“先聚焦、后裁剪”的协同效果，显著降低了误剪风险。

FP16 量化作为“末端加速”技术，以极小的精度代价换取边缘设备的推理吞吐量提升，且与前述模块正交，不引入额外结构复杂度。

四者形成递进链条，每一环节的精度损失由前置或后序环节进行补偿。这种协同优化的设计思想，是本文区别于已有工作的核心创新点。

1.3. 本文主要工作

本文围绕校园安全隐患的轻量化视觉检测，开展了以下四方面工作：

(1) 多类别数据集构建。收集并标注涵盖火焰、烟雾、区域入侵三类隐患的校园安全数据集，统一标注格式，为后续模型训练与评估奠定数据基础。

(2) 精度增强方法设计。将 CBAM 注意力机制引入 YOLOv8n，通过通道和空间维度的特征重标定，增强模型对火焰初期、远处烟雾和远距离小目标人员的检测能力。

(3) 轻量化网络设计。提出 GhostC2f 模块替代标准 C2f 模块，有效降低参数量和计算量；在此基础上，结合 L1 正则化稀疏训练、结构化通道剪枝和 FP16 量化，形成阶梯式模型压缩方案，在可控精度损失内实现最大程度的压缩。

(4) 边缘部署与实时验证。将压缩后的模型转换为 ONNX 并构建 TensorRT 推理引擎，部署于 Jetson Nano 和 RK3588 平台，测试多路视频流下的实时检测性能与资源占用，验证方案的实用性。

2. 相关技术基础

2.1. YOLOv8 目标检测算法

YOLOv8 是 Ultralytics 公司推出的新一代单阶段目标检测器，在继承 YOLOv5 工程化优势的基础上，对网络架构和训练策略进行了多项改进。其网络结构主要由 Backbone、Neck 和 Head 三部分组成。Backbone 采用改进的 CSPDarknet 结构，使用 C2f 模块替代了以往的 C3 模块，通过跨层部分连接和更多的分支梯度流，在维持轻量化的同时提升了特征学习能力；Neck 采用 PAN-FPN 结构进行多尺度特征融合；Head 为解耦头，分别处理分类与回归任务，并采用 Anchor-Free 的检测方式，降低了先验框设计的复杂性。

本文选取 YOLOv8n 作为基线模型，它是 YOLOv8 系列中体积最小的版本，参数量约 3.01M，FLOPs 约 8.2 G (输入尺寸 640×640)，在 COCO 数据集上 mAP@0.5 达到 37.3%，在速度和精度之间取得了较好的平衡。即便如此，当其应用于校园安全隐患检测时，仍有大量参数量和计算冗余，尤其是在边缘设备上直接推理难以满足实时性要求，需要进一步轻量化。

2.2. CBAM 注意力机制

CBAM (Convolutional Block Attention Module) 是一种结合通道注意力和空间注意力的轻量级注意力模块。通道注意力模块通过全局平均池化和全局最大池化聚合空间信息，经共享多层感知机生成通道权值，对特征图进行通道级重标定；空间注意力模块在通道维度上分别进行平均和最大池化，拼接后经卷

积层生成空间注意力图，实现空间位置的自适应加权。二者以级联方式协同工作，能够有效抑制背景噪声并突出前景目标的关键特征。

在目标检测任务中，将 CBAM 插入 Backbone 与 Neck 之间，可在多尺度特征融合前强化各层级特征图的判别信息，尤其有利于小目标和复杂背景下的检测。相较于其他注意力模块，CBAM 增加的计算开销较低，适合在轻量化要求苛刻的场景中使用。

2.3. 模型轻量化技术

Ghost 模块。GhostNet 提出，由标准卷积生成的特征图之间存在大量冗余，可以通过廉价操作生成“幽灵”特征图以扩展通道维度。Ghost 模块首先用较少的标准卷积生成固有特征图，然后对固有特征图进行廉价的线性变换(如深度卷积)生成补充特征图，最后拼接输出。这种方式在保持相似表示能力的同时，可大幅降低计算量和参数量。

结构化通道剪枝。剪枝是模型压缩的常用手段，结构化通道剪枝直接移除整个卷积通道，不引入稀疏结构，便于硬件加速。典型的流程包括：在损失函数中加入针对批归一化(BN)层缩放因子 γ 的 L1 正则化项进行稀疏训练，迫使部分 γ 趋于零；训练后按 γ 的绝对值排序，根据预设剪枝率裁剪对应通道及其相连的卷积核；最后对剪枝后模型进行微调恢复精度。

FP16 量化。FP16 量化指将模型的权重和激活从 32 位浮点数转换为 16 位浮点数，理论上可使模型体积减半，并在支持 FP16 计算的 GPU 和边缘硬件上实现接近 2 倍的推理加速，且精度损失极小，是边缘部署中最便捷高效的压缩方式之一。

2.4. 边缘设备部署技术

边缘部署的一般流程为：将 PyTorch 训练好的模型导出为 ONNX 格式，再利用 TensorRT 等推理优化器对 ONNX 模型进行层融合、精度校准和内核自动调优，生成针对特定硬件平台的高度优化引擎。Jetson Nano 作为英伟达推出的低功耗嵌入式 AI 平台，拥有 128 核 Maxwell GPU，支持 FP16 推理；RK3588 是瑞芯微的高性能 AIoT 芯片，内置 3 核 NPU，可提供高能效的整型推理能力。二者均被广泛用于智能摄像头、边缘分析盒等安防设备。

3. 面向校园安全隐患的轻量化检测模型设计

3.1. 整体模型框架

本文设计的轻量化校园安全隐患检测模型的整体技术路线包含四个关键阶段：数据准备→精度增强→轻量化压缩→边缘部署。先在自建的多类别校园安全隐患数据集上训练基线 YOLOv8n；嵌入 CBAM 注意力机制以提升对小目标和复杂背景的辨别力；然后，引入 GhostC2f 模块替代原始 C2f，并进行结构化通道剪枝和 FP16 量化，逐步压缩模型体积与计算量；最后，导出 ONNX 并转换为 TensorRT 引擎，在 Jetson Nano 和 RK3588 上进行实时推理验证。CBAM 负责精度保底，Ghost 模块从架构层面削减冗余，剪枝与量化进一步挤压参数与运算开销，四者协同实现精度和效率的平衡。

上述四个阶段并非独立串行，而是存在内在的互补与反馈关系。具体而言：

CBAM 与 Ghost 构成第一对协同：CBAM 通过注意力重标定提升特征判别力，Ghost 通过廉价特征生成削减计算量，二者共同作用使模型在保持高精度的前提下获得更紧凑的结构；

CBAM 与剪枝构成第二对协同：CBAM 在训练阶段强化重要通道的响应，使得稀疏训练中 BN 层 γ 值的分布更加“两极分化”——关键通道的 γ 值更高，冗余通道更趋近于零，从而提高了剪枝的准确性，减少了误剪关键通道的风险；

剪枝与量化构成第三对协同：剪枝降低了模型参数量，使量化后权重分布更集中，减小了量化舍入误差的累积效应。

这一协同设计确保了整体方案在较大压缩比下仍保持稳定的检测性能。

3.2. 校园安全隐患多类别数据集构建

为满足真实校园环境下火灾与人员异常行为的多类别目标检测需求，本文构建了一个包含火焰、烟雾和人员入侵三类目标的复合数据集。考虑到单一来源数据在场景多样性和目标类别方面存在一定局限，本文采用“公开数据集 + 自建数据集”相结合的方式进行数据构建。其中，火焰和烟雾类别主要利用公开火灾检测数据集进行补充，人员入侵类别则主要通过校园重点区域实际监控视频和模拟场景自行采集。通过对不同来源的数据进行筛选、统一标注和格式转换，最终形成适用于校园安全场景多类别目标检测的实验数据集。

3.2.1. 公开数据集

在火焰和烟雾类别的数据构建过程中，本文主要采用 D-Fire 公开数据集作为基础数据来源[11]。该数据集共包含 21,527 张图像，其中仅包含火焰的图像 1164 张，仅包含烟雾的图像 5867 张，同时包含火焰和烟雾的图像 4658 张，不包含火焰和烟雾的图像 9838 张；同时提供了 14,692 个火焰目标标注框和 11,865 个烟雾目标标注框，D-Fire 数据集示例图如图 1 所示。D-Fire 原始数据集采用 YOLO 格式进行目标标注，本文将 YOLO 格式标注转换为 PASCAL VOC 格式，以统一公开数据与自建数据的标注格式。完成标注格式转换后，本文采用 LabelImg 工具对公开数据的标注结果进行人工复核。通过“标注格式转换 + 人工复核与修正”的方式，在尽可能保留 D-Fire 原始标注信息的基础上，使公开数据与自建校园数据在类别定义、边界框标注规则及标注格式等方面保持一致，从而降低不同数据来源之间的标注差异对模型训练产生的影响。

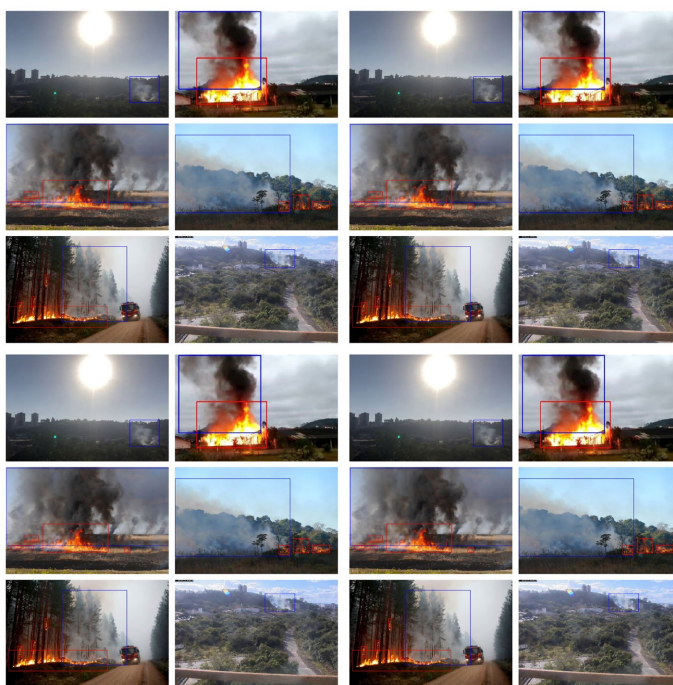


Figure 1. D-Fire Dataset examples

图 1. D-Fire 数据集示例图

3.2.2. 自建校园场景数据集

为了提高模型在真实校园安全监控场景中的适用性,仅采用公开火灾数据集难以覆盖人员非法进入重点区域等实际校园安全问题。因此,本文结合校园实际应用需求,自行构建校园场景数据集,并进一步采集部分校园环境下的火焰和烟雾相关场景,用于增强复合数据集的场景多样性。自建数据主要来源于校园重点安全区域的监控视频帧和模拟场景录制,重点选取天台、配电房和湖泊周边等具有潜在安全风险的区域。其中,天台和配电房属于校园重点管理区域,未经授权人员进入可能产生安全隐患;湖泊周边则存在人员靠近危险区域等异常行为。因此,本文将上述区域作为人员入侵目标的主要采集场景,部分自建数据集如图2所示。

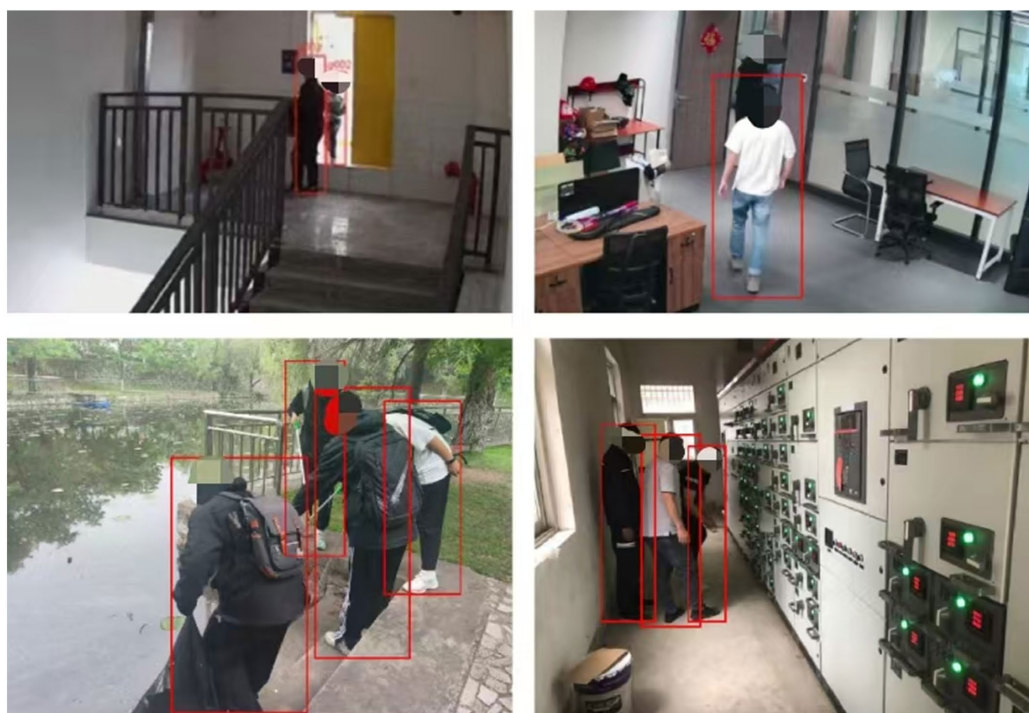


Figure 2. Self-built dataset

图2. 自建数据集

3.2.3. 数据标注与预处理

在完成公开数据和自建数据的采集后,本文对所有图像进行统一的数据清洗和标注处理。首先对原始图像进行人工筛选,剔除严重模糊、过度曝光、目标完全遮挡、目标无法辨识以及图像损坏等样本。对于同一视频中内容高度重复的连续帧,仅保留具有明显场景或目标变化的代表性图像,以减少数据冗余。

本文最终定义三个检测类别,分别为 **fire**、**smoke** 和 **intrusion**,其中 **fire** 表示火焰目标, **smoke** 表示烟雾目标, **intrusion** 表示人员入侵目标。采用 LabelImg 工具对筛选后的图像进行人工标注,并统一保存为 PASCAL VOC 格式。对于每一个可检测目标,标注框均采用能够完整覆盖目标的最小外接矩形表示;当一幅图像中存在多个目标时,对不同目标分别建立独立的标注框。

对于存在部分遮挡但目标主体仍能够明确辨识的情况,根据可见目标区域进行标注;对于严重遮挡导致目标类别或目标位置无法准确判断的样本,则予以剔除。通过统一标注规则,使来自公开数据集和自建数据集的样本在类别定义及标注形式上保持一致。

3.2.4. 数据集划分

经过数据筛选、标注和格式统一后,本文构建的复合数据集共包含 8,395 张图像。按照 7:2:1 的比例随机划分为训练集、验证集和测试集,分别包含 5876 张、1679 张和 840 张图像。其中,训练集用于模型参数学习,验证集用于模型训练过程中的参数调整 and 性能监测,测试集用于最终模型性能评价。

在数据划分过程中,尽可能保持三个类别在不同数据子集中的分布基本一致。对于来源于同一监控视频的连续或高度相似帧,避免将相邻帧随机划分至不同数据子集,从而降低由于场景高度相似造成的数据泄漏问题,提高实验结果的可靠性。

进一步对数据集中的目标尺度进行统计分析,将目标面积小于整幅图像面积 0.5% 的目标定义为小目标。统计结果表明,小目标约占数据集中目标总数的 22%。这些小目标主要包括远距离监控场景中的人员目标以及视野中尺寸较小的火焰和烟雾目标。该分布与校园监控环境中目标距离摄像机较远、目标在图像中所占比例较小的实际情况基本一致。因此,所构建数据集能够在一定程度上反映真实校园监控场景中的小目标检测问题。

3.3. 基于 CBAM 的精度增强模块

CBAM 的插入位置对最终性能影响显著。经初步实验分析,将 CBAM 置于 Backbone 的最后一个 C2f 模块输出与 Neck 之间,即 SPPF 层之后、PAN-FPN 结构之前,可在不引入过多计算开销的前提下,显著增强送入特征金字塔的多尺度特征质量。具体实现上,首先对 Backbone 输出的特征图应用通道注意力,聚合全局上下文并生成通道权值,加权后的特征图再经过空间注意力模块,突出火焰边缘、烟雾纹理和人员轮廓等空间位置信息。

小目标及复杂背景下的特征聚焦原理在于:通道注意力能够筛选出对目标判别贡献大的特征通道,抑制由背景杂波引入的无用通道;空间注意力则能精确定位目标所在的图像区域,增强对火焰初期微弱信号或远处人员小尺寸区域的响应。两者联合使模型在保留轻量特性的同时,具备更强的抗背景干扰能力。

值得注意的是,CBAM 的引入不仅服务于检测精度本身,其更深层的作用是为后续的剪枝操作提供“特征显著性先验”。通过 CBAM 训练后的模型,其 BN 层缩放因子 γ 在重要通道上呈现更高的幅值,相当于在稀疏训练前已完成一次通道重要性的粗排序,使得 L1 正则化能够更快地收敛到有效的稀疏结构,同时降低剪枝对关键特征的误伤概率。这一“先增强、后筛选”的机制是本文协同优化的关键环节之一。

3.4. 基于 Ghost 的轻量化特征提取网络

标准 C2f 模块是 YOLOv8n 骨干网络和颈部的主要构件,内部包含两个或多个卷积块,存在特征图冗余。本文借鉴 GhostNet 思想,设计 GhostC2f 模块:将 C2f 内部的普通卷积替换为先以较少输出通道的标准卷积生成固有特征,再通过廉价深度卷积产生补充特征,最后拼接两者输出。通过控制生成固有特征与补充特征的比例,在基本保持 C2f 特征表达能力的条件下,可将该模块的参数量和 FLOPs 降低约 40%。

替换策略上,将 Backbone 和 Neck 中的所有 C2f 模块均替换为 GhostC2f,保持网络其余部分(检测头、上采样等)不变,最终网络结构图如图 3 所示。经实验对比,GhostC2f 构建的骨干网络与原 C2f 网络在中等及大尺寸目标上的特征响应强度相当,但对极小的早期烟雾纹理略存在微弱衰减,后续通过 CBAM 通道注意力的补偿可使整体检测性能几乎无损。

3.5. 模型剪枝与量化压缩

剪枝和量化是模型压缩的“最后一公里”,但它们的效果高度依赖于前序模块的铺垫。得益于 CBAM

注意力机制对重要特征通道的强化,稀疏训练中 BN 层 γ 值的分布更为分化,使剪枝过程能更准确地识别并保留关键通道,从而在同等剪枝率下获得更高的精度保持率。同时, Ghost 模块已从结构上减少了冗余计算,使剪枝需要处理的通道总数减少,降低了剪枝算法的搜索复杂度。量化则是在剪枝后的模型上进行的,由于剪枝消除了大量接近零的权重,量化时的舍入误差对整体精度的影响进一步减小。三者形成“引导-粗剪-精剪-量化”的逐级优化链路。

为进一步压缩替换 GhostC2f 后的模型,首先进行 L1 正则化稀疏训练。在训练损失中附加 BN 层缩放因子 γ 的 L1 范数惩罚项,使得大量 γ 趋近于 0。稀疏训练完成后,统计所有 BN 层的 γ 值分布,设定剪枝率,将绝对值最小的 γ 对应的通道整层移除。剪枝时采用全局剪枝策略,保证各层剪枝比例的一致性,避免出现极端层完全剪掉的情况。剪枝后,在训练集上进行 20 个 epoch 的微调以恢复精度。

随后,将微调后的模型进行 FP16 量化。权值和激活在推理时均以半精度浮点数表示,模型文件体积缩减至原来的约一半。由于量化过程中权重数值的微小舍入误差, mAP 通常会有 0.1~0.3 个百分点的额外下降,但在边缘设备上的推理速度可获得显著提升。

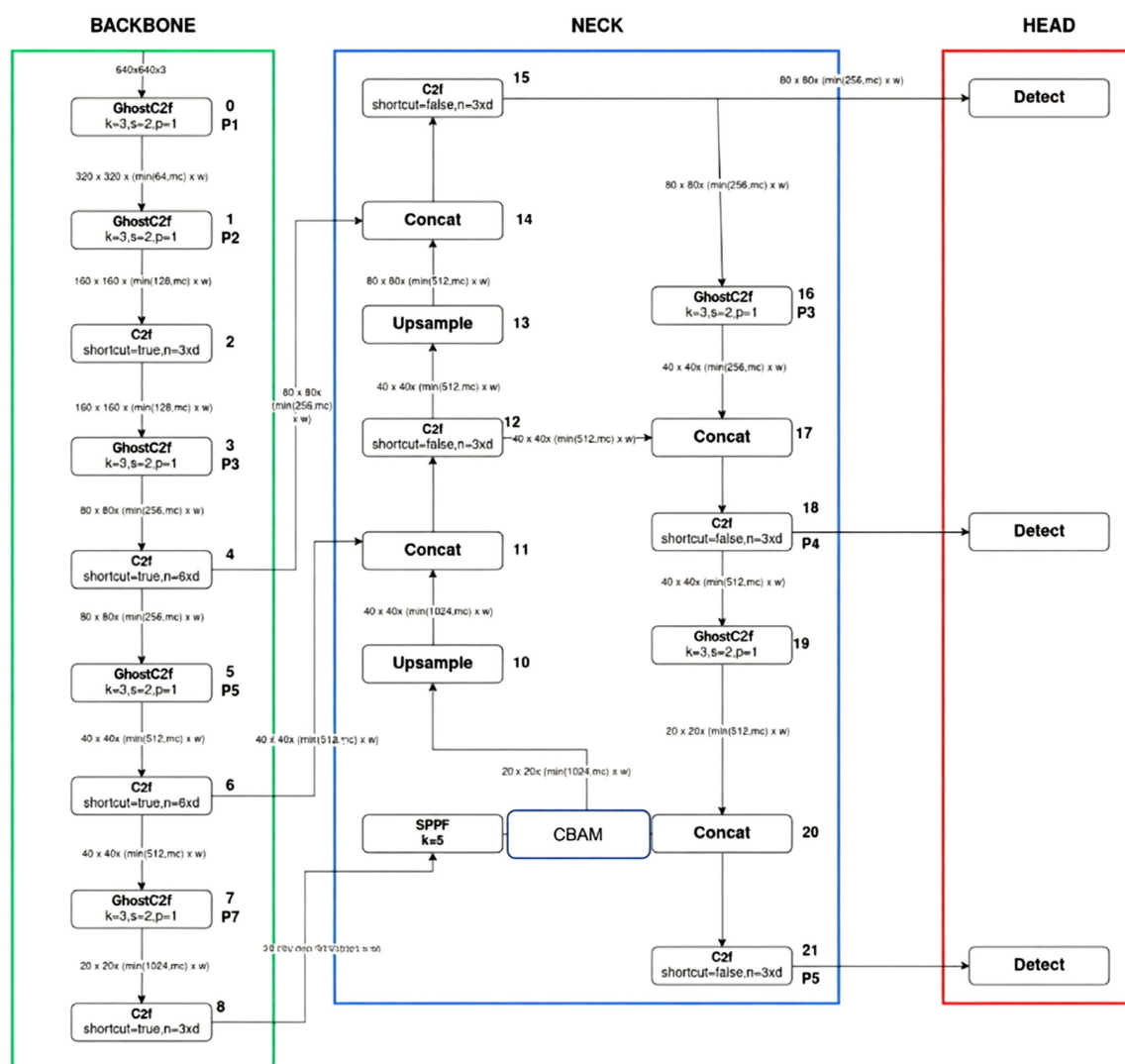


Figure 3. Improve the network structure diagram

图 3. 改进网络结构图

4. 实验设计与结果分析

4.1. 实验环境与评价指标

硬件平台：模型训练使用一台搭载 NVIDIA GeForce RTX 5060Ti 8 GB GPU 的工作站；边缘推理测试平台为 NVIDIA Jetson Nano (4 GB)和瑞芯微 RK3588 (8 GB)开发板。

软件框架：深度学习框架 PyTorch 2.0.1，模型导出与优化使用 ONNX 1.14 及 TensorRT 8.5。

评价指标：采用目标检测常用指标，包括 mAP@0.5、mAP@0.5:0.95、参数量(Parameters)、浮点运算量(FLOPs，输入尺寸 640×640)、推理速度(FPS)和模型文件体积(MB)。

4.2. 实验设置

基线模型 YOLOv8n 的训练超参数设置如下：输入分辨率 640×640 ，批量大小 16，使用 SGD 优化器，动量 0.937，权重衰减 5×10^{-4} ，初始学习率 0.01，采用余弦退火调度，训练 300 个 epoch。所有改进模型的实验均在相同的软硬件环境和超参数下完成，以保证公平对比。

4.3. CBAM 注意力模块消融实验

为确定 CBAM 最佳插入位置，分别在 Backbone 末端、Neck 的两个不同层以及 Head 前插入 CBAM，结果如表 1 所示。在 Backbone 末端(SPPF 后)嵌入 CBAM，mAP@0.5 达到 89.7%，较基线 89.2%提升 0.5 个百分点，参数量仅增加 0.002 M，FPS 无明显下降；插入 Neck 的中层或深层，精度增益更小且略微拖慢推理速度；插入 Head 前反而因过度集中而产生了少量误检。因此确定 Backbone 末端为最优位置。

Table 1. Table comparing CBAM insertion positions
表 1. CBAM 插入位置对比表

插入位置	mAP@0.5/%	参数量/M	FPS (nano)
基线(无 CBAM)	89.2	3.01	25.4
Backbone 末端	89.7	3.01	25.3
Neck 中部	89.4	3.01	24.8
Neck 深层	89.3	3.01	24.5
Head 前	89.0	3.02	24.3

此外，对火焰初期小目标和远处薄雾状烟雾的可视化检测结果显示，加入 CBAM 后，低置信度的微弱火焰和半透明烟雾被成功召回，验证了注意力机制对小目标和复杂纹理的聚焦效果。

4.4. 轻量化策略消融实验

GhostC2f 替换实验。将 Backbone 和 Neck 中的 C2f 全部替换为 GhostC2f 后，模型参数量从 3.01 M 降至 2.15 M，减少 28.6%，FLOPs 从 8.2 G 降至 5.1 G，降低 37.8%，mAP@0.5 由 89.2%小幅下降至 88.1%，仅损失 1.1 个百分点，表明 Ghost 模块在几乎不影响精度的情况下显著压缩了模型规模。

结构化剪枝实验。在 GhostC2f 基础上进行 L1 稀疏训练并设置不同全局剪枝率。当剪枝率 10%时，mAP 仅下降 0.2 个百分点；剪枝率 20%时，下降 0.6 个百分点；剪枝率 30%时，下降 1.5 个百分点；剪枝率达到 40%时，mAP 迅速下降超过 3 个百分点。综合权衡，选取 30%作为最终剪枝率，此时模型参数量进一步压缩至 1.72M，mAP 降至 87.6%。

FP16 量化实验。对剪枝后模型进行 FP16 量化，在 Jetson Nano 上 FPS 从 27.3 提升至 31.6，RK3588 上从 41.8 提升至 49.2，推理速度提升约 15%。mAP@0.5 仅额外下降 0.2 个百分点，达到 87.4%，模型文

件体积由 6.8 MB 缩小至 3.5 MB。

组合效果。“GhostC2f + 剪枝”与“GhostC2f + 剪枝 + 量化”的组合阶梯对比如表 2 所示，逐步压缩带来累加的精度下降，但最终 mAP 仅比基线低 1.8 个百分点，参数量减少 42.9%，证明该组合方案在精度保持上十分高效。

Table 2. Combined effect of lightweight strategies

表 2. 轻量化策略组合效果

模型配置	mAP@0.5/%	参数量/M	FLOPs/G	FPS (Nano)
YOLOv8n 基线	89.2	3.01	8.2	25.4
+GhostC2f	88.1	2.15	5.1	31.2
+GhostC2f + 剪枝(30%)	87.6	1.72	4.2	33.5
+GhostC2f + 剪枝 + FP16 量化	87.4	1.72	4.2	38.1 (TensorRT)

4.5. 综合性能对比实验

将本文最终改进模型(记作 Ours)与基线 YOLOv8n、其他主流轻量级模型及现有校园安全隐患检测变体进行全面对比，结果如表 3 所示。Ours 模型在 mAP@0.5 上达到 87.4%，较 YOLOv8n 仅降低 1.8%，参数量减少 42.9%，FLOPs 降低 48.8%；与 YOLOv5s 相比，精度领先 2.1 个百分点，参数量仅为 25%；与 YOLOv7-tiny 相比，参数量相近，精度高出 4.3 个百分点；与引入 CBAM 但未轻量化的 YOLOv8 变体相比，以较小的精度代价换取了约 43%的参数量缩减和边缘端 2 倍以上的推理速度。充分表明本文方法在精度与效率平衡上具备显著优势。

Table 3. Performance comparison of different models

表 3. 不同模型性能对比

模型	mAP@0.5/%	参数量/M	FLOPs/G	FPS (Nano)
YOLOv5s [12]	85.3	7.02	15.8	12.8
YOLOv7-tiny [13]	83.1	6.01	13.2	15.7
YOLOv8n 基线	89.2	3.01	8.2	25.4
YOLOv8n + CBAM	89.7	3.01	8.3	25.1
Ours	87.4	1.72	4.2	38.1

4.6. 模型部署与实时性验证

将最终模型导出为 ONNX，并利用 TensorRT 构建 FP16 推理引擎。在 Jetson Nano 上，单路 1080P 视频流推理达到 31 FPS (GPU 内存占用约 1.2 GB)，三路并发降至约 24 FPS，勉强满足实时性；在 RK3588 上，通过 RKNN 转换后单路可达 48 FPS，三路并发稳定在 30 FPS 以上，CPU 占用率低于 60%。选取实际校园监控画面进行测试，火焰、烟雾和区域入侵的误检率均低于 5%，漏检主要集中在极端密集遮挡或光照极暗的场景。整体表明，本文模型在低成本的边缘设备上具备良好的实时检测能力和工程可用性。

5. 校园安全隐患检测系统原型实现

5.1. 系统总体设计

基于本文提出的轻量化检测模型，设计并实现了一个校园安全隐患检测系统原型。系统由视频流接入模块、边缘推理模块、报警逻辑模块和前端展示模块四部分组成。视频流接入支持 RTSP 网络摄像头

和本地视频文件，边缘推理模块采用 TensorRT 引擎运行改进模型，报警逻辑根据连续帧检测结果进行过滤和确认，前端展示以 Web 页面形式实时呈现检测画框和告警列表。

系统工作流程为：摄像头视频帧持续送入推理模块，模型输出目标类别、置信度和边界框坐标；报警逻辑对连续 5 帧中同一目标出现次数进行计数，超过 3 帧则触发报警；报警信息通过 MQTT 协议推送至监控中心，前端页面实时更新并在发生入侵或火灾时弹出声光提醒。

5.2. 边缘端推理服务实现

边缘推理服务使用 Python 编写，核心推理类封装了 TensorRT 引擎的加载、预处理和后处理操作。为支持多路视频流并发，采用多进程架构，每个进程绑定一路摄像头，独立加载模型副本，避免 GIL 限制。报警推送模块通过订阅 MQTT 主题，区分火焰、烟雾和区域入侵三种类型，附带检测时间、摄像头编号和抓拍图像链接，便于安保人员快速定位。

5.3. 系统运行效果

在测试环境中部署三路摄像头持续运行 48 小时，实时检测界面可清晰标注出火焰区域、烟雾扩散趋势和闯入人员位置。资源监控显示，RK3588 上 CPU 利用率约 55%，NPU 占用率约 72%，内存占用 2.3 GB，无内存泄漏或性能衰减现象，系统长时间运行稳定。与原有纯人工值守方式相比，该系统可 7×24 小时不间断运行，响应延迟由分钟级缩短至秒级；与云端推理方案相比，避免了视频流上传带宽压力和数据隐私风险，综合成本大幅降低。

6. 总结与展望

6.1. 研究总结

本文针对校园安全隐患检测在边缘设备上的实时性需求，完成了从数据集构建、模型改进、轻量化压缩到边缘部署的全流程研究。主要工作总结如下：构建了包含火焰、烟雾和区域入侵三类隐患的多类别数据集；引入 CBAM 注意力机制，有效提升了模型对小目标和复杂场景的辨别力；设计 GhostC2f 模块并联合结构化剪枝与 FP16 量化，实现了模型的深度压缩；将压缩后的模型成功部署于 Jetson Nano 和 RK3588，达到实时检测要求。

创新点可概括为三方面：第一，提出面向校园安全的多类别统一轻量化检测框架，突破了现有工作仅关注单一隐患的局限；第二，系统融合 CBAM + Ghost + 剪枝 + 量化的协同优化策略，通过 CBAM 为剪枝提供通道显著性引导、Ghost 降低结构冗余、剪枝与量化阶梯式压缩，形成了“精度增强 - 结构轻量 - 通道裁剪 - 数值压缩”的互补链条，各模块间通过精度补偿与结构互补实现整体性能的高阶平衡，在精度和效率上取得了优异的平衡；第三，完成了从模型设计到边缘设备部署的全流程验证，提供了完整的可落地方案。实验指标上，改进模型参数量减少 42.9%，mAP 降幅 ≤ 2 个百分点，边缘端 FPS ≥ 30 ，全面达成预期目标。

6.2. 不足与展望

本文仍存在一些不足之处：当前自建数据集场景多样性有限，夜间、雨雾等低光照和恶劣天气样本覆盖不足，影响了模型的泛化能力；剪枝后的模型对相互遮挡严重的小目标仍有一定漏检；尚未探索多模态信息融合和更先进的 Transformer 架构。

未来研究可从以下方向深入：扩充数据集，增加复杂环境和多时段的样本，提升模型鲁棒性；尝试引入小目标专用的检测层或 Transformer 增强模块，进一步降低漏检率；引入红外热成像、声音等多模态

信息, 实现全天候立体化安全感知; 并继续向更多低成本嵌入式平台移植, 推动广域校园安全感知网络的构建。

基金项目

本论文为重庆外语外事学院 2026 年度校级科研项目(项目编号: KYYB2026015)结题论文。

参考文献

- [1] Jocher, G., Chaurasia, A. and Qiu, J. (2023) Ultralytics YOLO (Version 8.0.0). <https://github.com/ultralytics/ultralytics>
- [2] Woo, S., Park, J., Lee, J. and Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1
- [3] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C. and Xu, C. (2020) GhostNet: More Features from Cheap Operations. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, 13-19 June 2020, 1580-1589. <https://doi.org/10.1109/cvpr42600.2020.00165>
- [4] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S. and Zhang, C. (2017) Learning Efficient Convolutional Networks through Network Slimming. 2017 *IEEE International Conference on Computer Vision (ICCV)*, Venice, 22-29 October 2017, 2736-2744. <https://doi.org/10.1109/iccv.2017.298>
- [5] NVIDIA Corporation (2023) TensorRT: Programmable Inference Accelerator. <https://developer.nvidia.com/tensorrt>
- [6] 唐佳. 基于改进 YOLOv5m 的校园火焰检测设计[J]. 青海师范大学学报(自然科学版), 2026, 42(2): 133-140.
- [7] 韩逸. 基于 YOLO 的轻量化烟火检测算法研究[D]: [硕士学位论文]. 武汉: 武汉纺织大学, 2024.
- [8] 鞠逸凡. 基于深度学习的铁路周界异物入侵检测方法研究[D]. 青岛: 青岛科技大学, 2024.
- [9] 杨宗晔. 智慧校园绿色建筑与 AI 智慧安防协同构建研究[J]. 陶瓷, 2026(7): 189-191.
- [10] 周红志, 秦敏, 墨子向. 基于 WIFI 技术的高校宿舍安全智能检测系统[J]. 吉林广播电视大学学报, 2015(6): 7-8.
- [11] de Venâncio, P.V.A.B., Lisboa, A.C. and Barbosa, A.V. (2022) An Automatic Fire Detection System Based on Deep Convolutional Neural Networks for Low-Power, Resource-Constrained Devices. *Neural Computing and Applications*, **34**, 15349-15368.
- [12] Ultralytics (2020) YOLOv5. <https://github.com/ultralytics/yolov5>
- [13] Wang, C., Bochkovskiy, A. and Liao, H.M. (2023) YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, 17-24 June 2023, 7464-7475. <https://doi.org/10.1109/cvpr52729.2023.00721>