DOI:10.19651/j.cnki.emt.2517798

基于网络结构轻量化的道路监控检测模型*

来超凡[1,2] 花 强[1,2] 母静越[1,2] 张 博[3]
(1.河北大学数学与信息科学学院 保定 071000;2.河北省机器学习与计算智能重点实验室 保定 071000;
3.河北农业大学理学院 保定 071000)

摘 要:针对现有交通监控检测模型参数量大,计算复杂度高,在一些边缘设备上部署会受到硬件资源限制的问题,对 YOLOv8 模型的网络结构进行了针对性改进,提出一种基于网络结构轻量化的道路监控检测模型。首先在骨干网络部分:采用极简网络架构 VanillaNET 替代原本的主干网络的中间部分进行特征提取,以减少模型的参数量和整体的计算复杂度。接着将 FasterNet 的优势与 EMA 注意力机制相结合,应用到骨干网络的 C2f 模块,有效降低了内存访问量,并一定程度上提升了模型的检测能力。然后将 SPPCSPC 结合分组卷积,提出 G-SPPCSPC 模块,提升了模型对监控视角下不同大小尺度信息的提取能力。最后,在颈部网络:将轻量级注意力机制 MLCA 结合到 C2f 模块,目的是减少无关背景信息对于道路监控检测的干扰。实验结果表明,改进后的模型参数量降低了 53.3%,模型尺寸减小了 51.3%,计算复杂度下降了 48.1%,mAP/50%达到了 93.7%,FPS 达到了 280.5。模型在显著降低模型参数量和计算复杂度的同时,保持了较高的检测精度和速度,适用于边缘设备的部署,具有较高的实用价值。

关键词:道路监控检测;YOLOv8n;轻量化;注意力机制
中图分类号:TP391.7;U495;TN919.8 **文献标识码**:A **国家标准学科分类代码**:510.1050;510.50

Road surveillance detection model based on lightweight network architecture

Lai Chaofan[1,2] Hua Qiang[1,2] Mu Jingyue[1,2] Zhang Bo[3]
(1. College of Mathematics and Information Science, Hebei University, Baoding 071000, China; 2. Hebei Key Laboratory of Machine Learning and Computational Intelligence, Baoding 071000, China; 3. College of Science, Hebei Agricultural University, Baoding 071000, China)

Abstract: To address the issues of high parameter count and computational complexity in existing traffic surveillance detection models, which limit their deployment on edge devices due to hardware resource constraints, this study proposes a lightweight network architecture-based road surveillance detection model by specifically modifying the YOLOv8 model. In the backbone network, the minimalist architecture VanillaNET is introduced to replace the intermediate part of the original network for feature extraction, significantly reducing the model's parameter count and overall computational complexity. The advantages of FasterNet are combined with the EMA attention mechanism and applied to the C2f module in the backbone network, effectively reducing memory access and enhancing the model's detection capability. Additionally, the G-SPPCSPC module is proposed by integrating SPPCSPC with grouped convolutions, improving the model's ability to extract multi-scale information under varying surveillance perspectives. Finally, in the neck network, the lightweight attention mechanism MLCA is incorporated into the C2f module to reduce interference from irrelevant background information in road surveillance detection. Experimental results show that the improved model reduces the parameter count by 53.3%, model size by 51.3%, and computational complexity by 48.1%, while achieving a mAP50/% of 93.7% and an FPS of 280.5. The model maintains high detection accuracy and speed while significantly reducing parameter count and computational complexity, making it suitable for deployment on edge devices and demonstrating high practical value.

Keywords: road monitoring and detection; YOLOv8n; light weight; attention mechanism

0 引 言

道路监控检测技术通过将目标检测模型部署在道路监控设备上,能够有效提升交通安全性和管理效率。近年来,随着新能源汽车的普及,我国道路汽车数量急剧增加,导致交通拥堵问题日益严重[1]。因此,普及高效的道路监控检

收稿日期:2025-01-03
* 基金项目:河北省科技平台项目-创新能力提升计划(22567623H)项目资助

测设备显得尤为重要。然而,现有的目标检测模型往往因模型复杂、参数量大、计算复杂度高,不利于在硬件资源有限的边缘设备上部署。针对这一问题,对道路监控检测模型进行轻量化设计,以降低部署成本,具有重要的应用价值和研究意义。

目标检测算法主要分为多阶段和单阶段两大类。以基于区域的卷积神经网络(region-based convolutional neural network,R-CNN)系列为代表的多阶段算法通过候选区域生成和分类定位两个阶段,实现了较高的检测精度,但计算效率较低。相比之下,YOLO 系列算法作为单阶段算法的代表,通过一次前向传播即可获得检测结果,在检测速度上具有显著优势,更适合实时道路监控检测任务。例如,王宇宁等[2]基于 YOLOv1 实现了车辆实时检测,洪松等[3]基于 YOLOv3 提出了车辆行人检测算法。

近年来,国内外在目标检测模型轻量化方面取得了显著进展。主流的轻量级网络架构如 MobileNet、ShuffleNet 和 EfficientNet 系列,通过深度可分离卷积、通道混洗等技术有效降低了模型的计算复杂度和参数量。此外,注意力机制也被广泛应用于目标检测领域,通过增强重要特征通道或空间位置的权重来提升模型的检测精度和鲁棒性。例如,郭克友等[4]利用 MobileNetV3 替换 YOLOv4 的骨干网络实现轻量化,并引入卷积注意力模块(convolutional block attention module,CBAM)来提升昏暗场景下的车辆检测性能;范金豪等[5]基于 YOLOv4 提出了一种轻量化的 Yolo_ES 模型,通过 EfficientNet 和高效通道注意力机制(efficient channel attention,ECA)显著降低了参数量并提升了检测精度;史涛等[6]将可变形卷积和大核注意力机制融入 YOLOv8,提出了无人机车辆检测算法 YOLOv8-CX;顾杨海等[7]提出了一种基于多尺度特征融合与交互的目标检测算法,通过设计 C2f-CAST 模块和跨层级融合模块(spatial dual-feature fusion module,SDFM),有效解决了小目标漏检和复杂背景干扰问题,同时引入轻量级卷积进一步优化了模型复杂度。以上研究为模型的轻量化设计提供了重要参考,但在实际边缘设备部署中,仍然面临精度与效率的权衡、多尺度特征提取不足、注意力机制效率低下、硬件兼容性差以及泛化能力不足等问题。

YOLOv8 具有较高的检测精度和鲁棒性,但对于硬件资源有限的边缘设备,其计算复杂度和参数量仍然较高。因此,本文以降低设备部署成本为目标,基于 YOLOv8n 进行轻量化改进,提出了一种更加适用于道路监控检测的轻量化模型 YOLOv8_MLW。具体改进包括:

1)将 VanillaNET[8]作为骨干网络的主体部分进行高效的特征提取,大幅减少了模型的参数量和计算复杂度。

2)将 FasterNet[9]与高效多尺度注意力机制(efficient multi-scale attention,EMA)[10]结合应用到 C2f 模块中,在降低计算量和内存访问量的同时提高模型的检测能力。

3)将分组卷积[11]应用于空间金字塔池化跨阶段部分网络(spatial pyramid pooling cross stage partial networks,SPPCSPC)[12],保证轻量化同时,高效提取因视角远近造成的不同大小尺度的路况信息。

4)将轻量级的混合局部通道注意力机制(mixed local channel attention,MLCA)[13]与 C2f 模块结合,通过融合局部与全局特征以及通道与空间特征的信息,降低复杂背景信息对道路监控检测的干扰。

1 YOLOv8 模型

YOLOv8 基于缩放系数 X、L、M、S、N 由大到小提供了不同版本尺度的模型,除了可以被用来训练完成目标检测任务以外,还支持实例分割和图像分类。YOLOv8 网络使用了与 YOLOv5[14]相同的网络结构,同样由骨干网络(Backbone)、颈部网络(Neck)、检测头(Head)三部分组成,骨干网络负责对输入图像进行处理,提取特征图。在颈部网络,使用路径聚合网络-特征金字塔网络(path aggregation network with feature pyramid network,PA-FPN)[15]提取多尺度特征信息,对骨干网络提取的特征进一步增强。在检测头部分,YOLOv8 采用了解耦头结构,将分类和检测头分离,使得网络的训练和推理更加高效。

YOLOv8n 是 YOLOv8 系列中参数量最小的模型,其参数量仍达到了 300 万。对于一些硬件资源有限的道路监控检测设备,部署这样的模型仍然存在一定的限制。若可以设计一种模型在保证检测精度的同时,显著减少模型参数并降低计算复杂度,那么就可以使用较为廉价的硬件设备完成道路监控检测任务。这将有效降低部署道路监控检测设备的成本。

2 改进 YOLOv8 网络模型

针对道路监控检测设备的高效部署需求,本文对 YOLOv8 的网络结构进行了优化,提出了一种改进的目标检测网络模型 YOLOv8_MLW。如图 1 所示,改进模型首先通过两个 CBS 层对输入特征进行升维,随后利用结合了 FasterNet 和 EMA 注意力机制的(the C2f module combines FasterNet and EMA,FC2f_E)模块提取细粒度特征。接着,使用 1×1 的 CBS 层进行降维,并引入四层 VanillaBlock 模块,同时将整体通道结构调整为 96-192-384-768。此外,采用结合分组卷积的 G-SPPCSPC 模块替代原有的空间金字塔池化(spatial pyramid pooling fast,SPPF)模块。最后,在颈部网络中引入结合 MLCA 注意力机制的(C2f integrates MLCA,C2f_MA)模块。设计 FC2f_E 模块是为了通过结合 FasterNet 中的部分卷积(partial convolution,PConv)和 EMA 注意力机制,在优化内存访问效率的同时增强多尺度特征提取能力,从而实现计算效率与检测精度的平衡;引入 VanillaNET 旨在充分利用其极简高效的特性提取输入特征,在显著降低模型的参数量和计算复杂度的同时提升检测速度;G-SPPCSPC 模块是为了以最小的参数代价获取更丰

富的多尺度特征信息，增强模型对不同尺度目标的检测能力；C2f_MA 模块则是从轻量高效注意力机制的角度出发，利用 MLCA 协同局部与全局特征，降低复杂背景因素的干扰，从而提升模型的鲁棒性和检测精度。

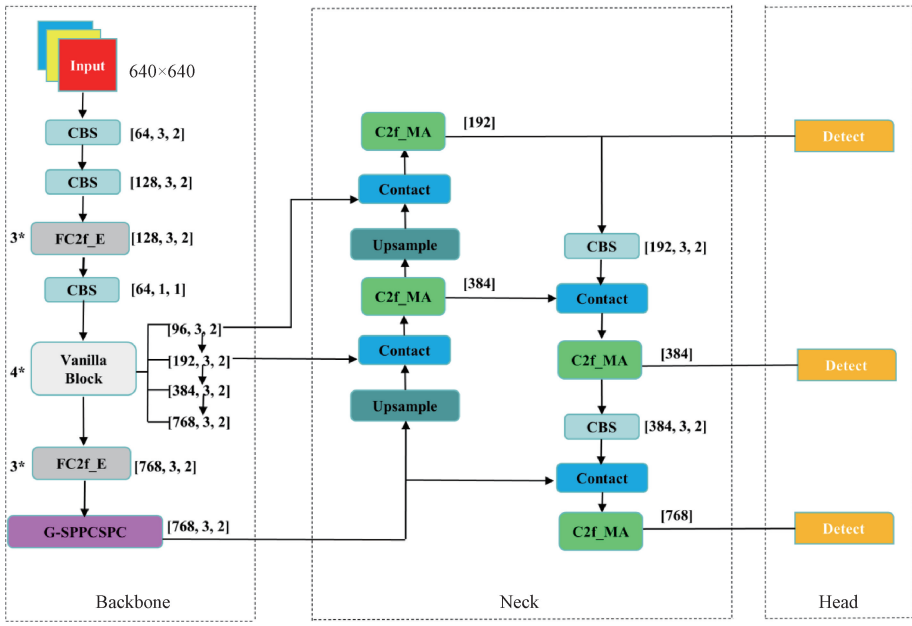


图 1　YOLOv8_MLW 网络结构

Fig. 1　YOLOv8_MLW network structure

2.1　基于 VanillaNET 结构的整体轻量化

YOLOv8 的骨干网络采用 Darknet-53 架构，这类复杂网络在多类别目标检测任务中表现出色。然而，在道路监控场景中，检测目标主要集中在行人和交通工具等少数类别，复杂网络不仅增加了模型部署的难度，还可能导致性能与参数过剩。针对这一问题，本文在骨干网络的中间部分引入 VanillaNET 以简化网络结构。VanillaNET 通过避免高深度、shortcut 连接和自注意力机制等复杂操作，在保持较高性能的同时，显著降低了模型的参数量和计算复杂度，其极简设计使其成为高效部署的理想选择。图 2 所示为以 6 层为例的 VanillaNET 网络架构，VanillaNET 由主干(Stem)、主体(Main body)和头部(Head)三部分组成。主干部分使用一个 4×4 卷积核(步长为 4)对输入图像的通道特征进行初步提取；主体部分通过步长为 2 的最大池化层调节特征图尺寸，并在Ⅰ、Ⅱ、Ⅲ阶段将通道数量扩充一倍，阶段Ⅳ则采用平均池化层，保持通道数量不变；头部部分通过全连接层输出分类结果。在每个 1×1 卷积操作后，依次应用激活函数和批量归一化，以最小的计算成本保持特征图的信息完整性，从而显著降低了模型的复杂度和计算成本，使其更适合在资源受限的边缘设备上高效部署。

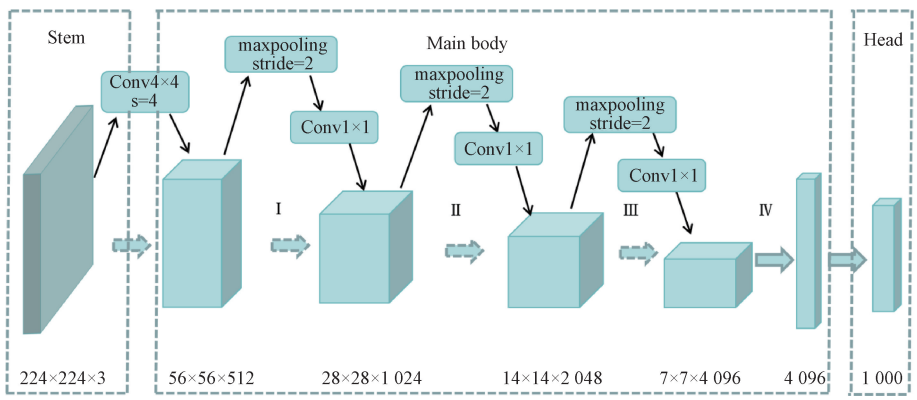


图 2　VanillaNet 网络架构(以 6 层为例)

Fig. 2　Vanillanet network architecture (Layer 6 as an example)

2.2　基于 FasterNet 与 EMA 注意力机制的 FC2f_E 模块

1) 面向内存访问的轻量化 C2f 模块设计

在实际的道路监控检测任务中，边缘设备通常需要处理高分辨率视频流或实时捕捉的连续图像帧，这会导致频

繁的内存访问,并对硬件设备的计算能力和带宽提出较高要求。传统的轻量化网络(如 MobileNet、GhostNet、ShuffleNet 等)虽然通过深度卷积[16]降低了计算量,但未能有效缓解内存访问压力。针对这一问题,本文在骨干网络的 C2f 模块中引入 FasterNet 进行优化,提出轻量化的(C2f integrates FasterNet, FC2f)模块,如图 3 所示,将 FasterNet 中的 FasterBlock 替换 C2f 中的 Bottleneck,显著减少了内存访问并降低了模型的计算量,同时提升了检测速度。FasterBlock 中的 PConv 是这一优化的关键。与普通卷积不同,PConv 仅对部分输入通道进行卷积操作,其余通道保持不变。在连续或周期性内存访问计算中,通常选择首部或尾部的连续通道作为特征图代表,从而在提升计算效率的同时,显著降低了内存访问量。

PConv 计算量(FLOPs)如式(1)所示。

$$FLOPs = h \times w \times k^2 \times c_p^2 \tag{1}$$

PConv 内存访问数量(MAC)如式(2)所示。

$$MAC = h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \tag{2}$$

其中,h 和 w 分别表示特征图的高度和宽度,k 代表卷积核的尺寸,c_p 表示 PConv 中进行普通卷积的特征通道数,通常仅仅为普通卷积的 1/4,因此基于上述公式可知在输入相同的情况下 PConv 的 $FLOP$s 和 MAC 要远远小于常规卷积。

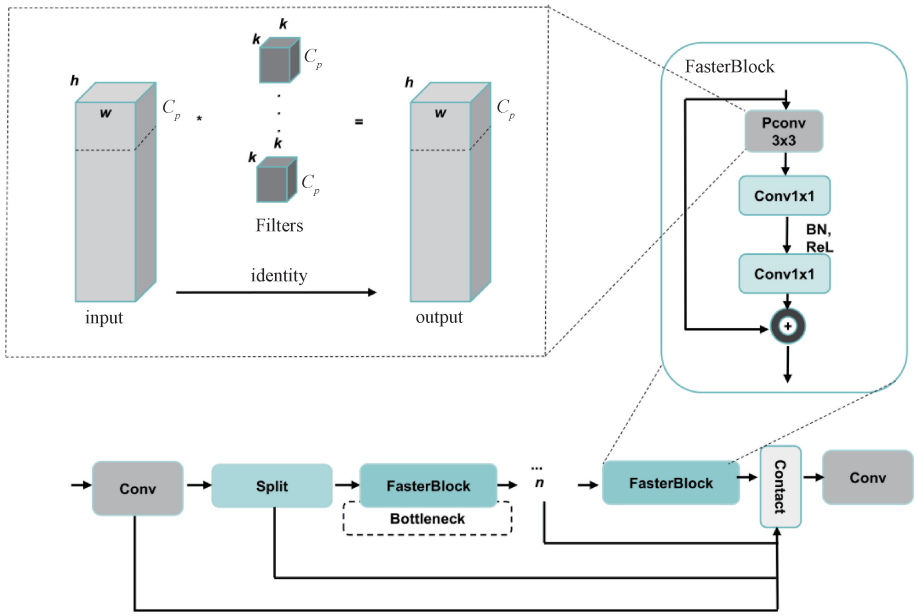


图 3 FC2f 模块结构

Fig. 3 Structure of FC2f module

2) 基于 EMA 注意力机制的精度补偿优化

尽管 PConv 有效降低了内存访问量与计算量,但仅对部分特征图进行卷积可能导致特征提取不够充分。为了在保证轻量化的同时提升检测精度,本文对 FC2f 模块进行二次优化,通过在 FasterBlock 的最后一个 1×1 卷积之后添加 EMA 注意力机制,提出 FC2f_E 模块。图 4 展示了 FasterBlock 中 EMA 注意力机制的集成结构。EMA 注意力机制通过跨空间学习方法,在不进行通道降维的情况下学习有效的通道描述,生成更好的像素级注意力,从而弥补 PConv 在特征提取上的不足。

与压缩与激励网络(squeeze-and-excitation networks, SE)[17]、坐标注意力(coordinate attention, CA)[18]、CBAM[19]等注意力机制相比,EMA 避免了大量池化操作带来的计算成本,更适合轻量化模型。其核心思想是在多尺度并行子网络中建立短期与长期依赖关系。具体而言,部分通道维度被重塑为批量维度,实现局部的跨通道交互,并将两个并行子网络的输出特征图通过跨空间学习方法融合。首先,输入特征图通道被分为 g 组,每组通过 3 个特征分支(x、y 和 3×3 卷积块)提取注意力权重。其中,x 和 y 分支使用 1×1 卷积和 1 D 全局平均池化沿两个空间方向对通道进行编码,而 3×3 卷积块用于提取多尺度特征表示。通过跨空间学习,EMA 增加了跨分支的特征交互,使模型提取的特征信息更加丰富,在保持低计算成本的同时实现更全面的特征聚合。

2.3 基于分组卷积的轻量化 G-SPPCSPC 模块

在道路监控目标检测任务中,目标与摄像头距离的变化导致目标尺度多样性,对检测性能产生显著影响。尽管 YOLOv8 骨干网络中的 SPPF 模块在空间金字塔池化(spatial pyramid pooling, SPP)[20]基础上进行了速度优化,但在处理复杂多变场景时,其检测精度仍有不足。为此,本文提出了一种轻量化的 G-SPPCSPC 模块,其核心创新在于将分组卷积与跨阶段空间金字塔池化相结合。

如图 5(a)所示,分组卷积通过将输入特征图沿通道维度均匀划分为多组,在各组内独立执行卷积运算后融合特

征。具体而言，分组卷积将输入特征图分为 g 组后，每组只需处理 $1/g$ 的的通道数，因此参数量仅为普通卷积的 $1/g$。分组数量 g 的选取直接影响模型的性能：当 g 值较小时，组间特征交互增强，模型复杂度提升，但特征提取能力更强；而随着 g 值增大，参数量和计算量进一步减少，但可能因组间信息交互减弱而影响特征表达能力。

如图 5 (b)所示，G-SPPCSPC 模块由两个关键组件构成：SPP 和跨阶段部分网络(cross stage partial networks，CSPC)。空间金字塔池化组件采用多尺度网格划分策略，对输入特征图进行不同粒度的最大池化操作，并将各尺度特征拼接，生成固定维度的特征向量，增强模型对目标尺度变化的鲁棒性。CSPC 组件通过将输入特征图划分为两个并行分支，分别进行卷积运算后拼接，在减少参数量的同时保持了模型的特征提取能力。该设计在降低计算复杂度的同时，有效提升了模型在复杂交通场景中的检测精度。

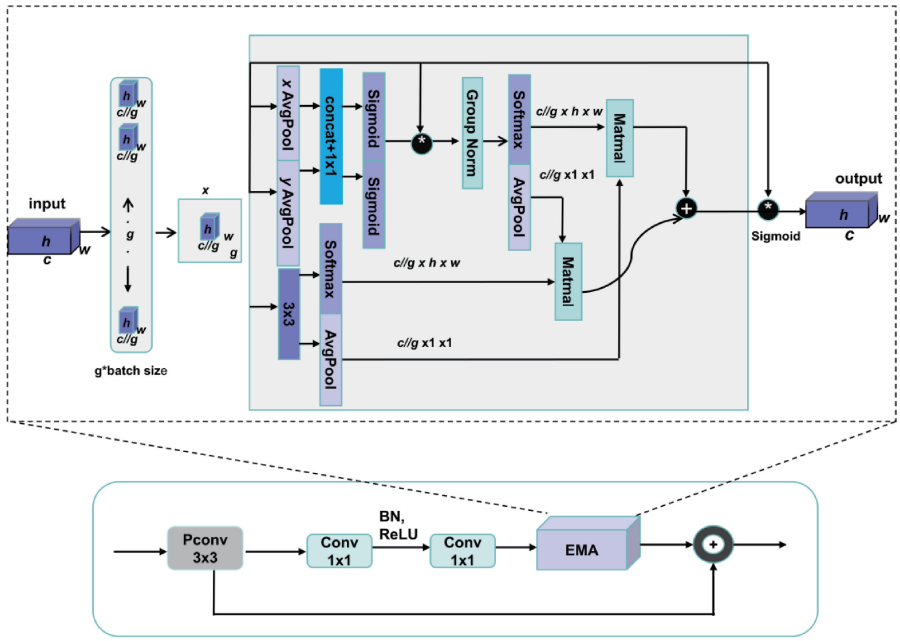
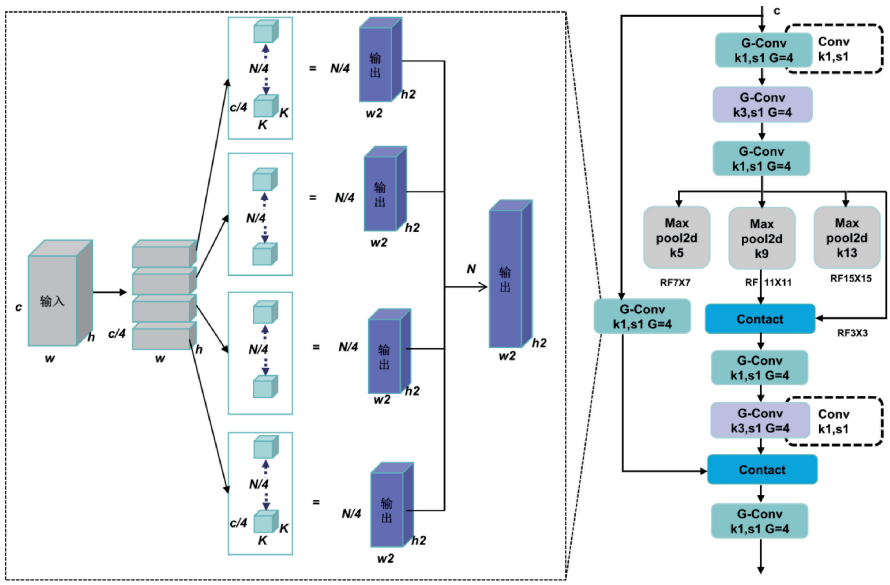


图 4　FasterNet 中 EMA 注意力机制的集成结构

Fig. 4　Integration structure of EMA attention mechanism in FasterNet



(a) 分组卷积过程

(a) Grouped convolution process

(b) G-SPPCSPC

图 5　G-SPPCSPC 结构

Fig. 5　G-SPPCSPC structure

2.4 基于轻量级混合局部通道注意力的 C2f_MA 模块

在道路监控检测任务中,复杂环境(如光照变化、背景杂乱)会显著影响检测性能。为在轻量化设计基础上增强模型对有效特征的感知能力并抑制背景干扰,本文在颈部网络的 C2f 模块中引入了一种轻量级混合局部通道注意力机制(MLCA),提出了 C2f_MA 模块,其结构如图 6 所示。MLCA 通过整合局部和全局的空间与通道注意力,在降低计算复杂度的同时提升关键特征捕捉能力。具体而言,MLCA 首先通过局部平均池化(local average pooling, LAP)和全局平均池化(global average pooling,GAP)分别提取局部细节和全局上下文信息,随后利用 1 D 卷积降低通道维度并保留空间结构,减少计算量。局部特征通过逐元素相乘与输入结合,增强有用特征;全局特征通过逐元素相加与局部特征融合,实现局部与全局互补。最后通过反池化(unpooling,UNAP)恢复特征图空间维度,确保信息完整性。该设计在保证性能的同时,最大限度地减少了参数量和计算开销。

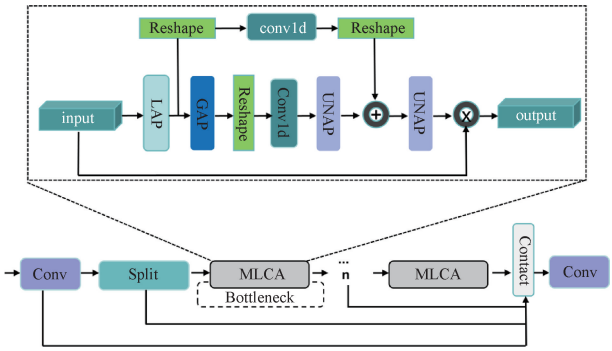


图 6 C2f_MA 结构

Fig. 6 Structure of C2f_MA

3 实验与结果

3.1 数据集

本文采用 Yusuf Berk Sarıdoğan 在 Roboflow 平台发布的公共交通安全检测数据集,包含自行车、摩托车、汽车、公交车和行人 5 个类别,共 9 675 张图像,划分为训练集(8 238 张)、验证集(854 张)和测试集(583 张)。该数据集涵盖多种道路监控场景,适用于复杂环境下的性能评估。

3.2 实验环境与参数设置

本实验在 Windows 11(64 位)系统下进行,硬件配置为 Intel ® Core™ i7-13700 H(20 核,2.4 GHz)和 NVIDIA GeForce RTX 4060 Laptop GPU(16 GB 显存)。软件环境包括 Python 3.8.19、PyTorch 2.3.1 及 CUDA 12.1。实验参数设置:输入图像尺寸为 640×640,批量大小为 32;损失函数中边界框损失权重为 7.5,类别损失权重为 0.5;优化器采用 SGD,初始学习率为 0.01,权重衰减率为 0.000 5;数据增强参数为:色相(0.015)、饱和度(0.7)和亮度(0.4)。模型训练总轮数为 200。

3.3 评价指标

为全面评估改进后模型的性能,本文采用平均精度(mAP50/%)、参数量(Params)、Giga 浮点运算次数(GFLOPs)、每秒帧数(FPS)和权重文件大小(MB)作为评价指标,并基于这些指标引入多目标优化(multi-objective optimization, MOO)[21] 方法进行综合评估。其中,mAP50/%综合考量各类别的准确率与召回率,反映模型的检测性能,其计算公式如式(3)所示,AP_i 表示第 i 个类别的平均精度。

$$mAP50 = \frac{1}{m}\sum_{i=1}^{m} AP_i \tag{3}$$

Params 衡量模型规模,值越大表示模型越复杂;GFLOPs 表示模型推理所需的计算量,值越小计算效率越高;FPS 反映模型推理速度,值越高实时性越好;权重文件大小影响模型内存占用,值越小部署需求越低。

由于上述指标的量纲和取值范围不同,直接比较难以全面反映模型性能。因此,本文采用多目标优化方法,对各指标进行归一化处理,消除量纲影响,权重分配基于以下原则:首先,mAP50/%作为核心性能指标,分配权重 0.3,以确保检测精度在评估中的主导地位;其次,Params、GFLOPs 和 FPS 分别分配权重 0.2,以平衡模型复杂度、计算效率和实时性需求;最后,权重文件大小分配权重 0.1,以兼顾模型部署的轻量化需求。该权重分配方案遵循多目标优化中的加权和法原则,并结合道路监控场景的实际需求,确保各指标间的合理权衡。加权得分(Score)的计算公式如式(4)所示,角标中的 max 和 min 分别代表同组模型对应指标的最大值和最小值。

$$\begin{aligned} Score = {} & 0.3\times\frac{mAP50/\%-mAP50/\%_{\min}}{mAP50/\%_{\max}-mAP50/\%_{\min}}+ \\ & 0.2\times\frac{FPS-FPS_{\min}}{FPS_{\max}-FPS_{\min}}+0.2\times\frac{Params_{\max}-Params}{Params_{\max}-Params_{\min}}+ \\ & 0.2\times\frac{GFLOPs_{\max}-GFLOPs}{GFLOPs_{\max}-GFLOPs_{\min}}+0.1\times\frac{MB_{\max}-MB}{MB_{\max}-MB_{\min}} \end{aligned} \tag{4}$$

表 1 不同分组数的 SPPCSPC 对比

Table 1 Comparison of SPPCSPC with different group numbers

分组数	Params /10^6	GFLOPs /10^9	mAP50/%	FPS	权重文件/MB	Score ×10
32	2.60	7.3	93.8	134.72	**5.20**	5.00
16	2.65	7.4	94.0	172.54	5.30	6.64
8	2.76	7.4	94.2	178.85	5.52	7.52
4	**2.99**	7.6	94.4	182.22	5.94	**7.88**
1	4.31	8.7	**94.5**	**210.02**	16.76	5.00

3.4 SPPCSPC 不同分组数量性能对比

分组卷积通过将输入特征图划分为若干组并独立卷

积，有效减少了模型参数量和计算复杂度。本文通过对比实验探究了不同分组数量对 SPPCSPC 模块性能的影响，结果如表 1 所示。实验表明，随着分组数量增加，模型参数量（Params）和计算量（GFLOPs）显著下降，但精度（mAP50/%）和推理速度（FPS）有所降低。例如，分组数量从 1 组增至 32 组时，参数量减少约 39.7%，GFLOPs 降低约 16.1%，而 mAP50/% 下降 0.7%，FPS 减少约 35.9%。为在轻量化和检测精度间取得平衡，本文采用多目标优化方法进行综合评估。结果表明，分组数量为 4 组的 SPPCSPC 在参数量（2.99×10^6）、计算量（7.6×10^9）、精度（94.4%）和推理速度（182.22 FPS）间实现了最佳权衡，加权得分为 7.88，显著高于其他分组。因此，本文最终选择将 SPPCSPC 的卷积核分为 4 组。

3.5　消融实验

本文通过消融实验验证了改进模块对模型性能的影响，实验结果如表 2 所示。实验表明，随着各模块的加入，模型在参数量（Params）、计算量（GFLOPs）、平均精度（mAP50/%）和推理速度（FPS）等方面均得到优化。VanillaNET 模块和 FC2f_E 模块显著降低了模型权重，并提升了 FPS，实现了轻量化与高速检测的平衡。G-SPPCSPC 模块与 C2f_MA 模块虽略微增加了参数量和计算复杂度，但有效缓解了轻量化带来的精度损失，相较于仅添加 VanillaNET 和 FC2f_E 模块，mAP50/% 提升了 1.7%。综合所有模块的 YOLOv8_MLW 模型，相较于初始模型，在检测精度上仅下降 0.4%，但参数量减少了 53.3%，模型尺寸减小了 51.3%，GFLOPs 降低了 48.1%，FPS 提升了 54.2%。该模型在轻量化的同时保持了较高的检测性能，更适合硬件资源有限的边缘设备。多目标优化加权得分表明，组合所有模块的模型性能最优，验证了各模块的优势互补性。

3.6　实际检测效果验证

为验证 YOLOv8_MLW 模型在实际道路监控检测任务中的效果，本文选取了测试集中不同分辨率和部署位置的代表性图片进行测试。如图 7 所示，改进后的 YOLOv8_MLW 在实现轻量化的同时，仍表现出优异的检测性能。在不同分辨率和复杂背景条件下，均能高效检测道路监控画面中的目标，证明了其较高的准确性和鲁棒性。

表 2　消融实验结果

Table 2　Fusion experiment results

Vanilla NET	FC2f_E	G-SPP CSPC	C2f_MA	Params /10^6	GFLOPs /10^9	mAP50 /%	FPS /fps	权重文件 /MB	Score ×10
×	×	×	×	3.01	8.1	94.1	181.90	5.98	2.62
√	×	×	×	1.36	4.1	93.1	226.91	**2.82**	6.81
×	√	×	×	2.31	6.5	93.6	191.51	4.66	4.69
×	×	√	×	2.99	7.6	**94.4**	182.22	5.94	3.25
×	×	×	√	2.85	7.7	**94.4**	189.46	5.68	3.60
√	√	×	×	**1.17**	**3.7**	92.0	243.21	2.48	6.25
√	√	√	×	1.44	4.1	92.9	245.25	2.99	6.78
√	√	√	√	1.40	4.2	93.7	**280.52**	2.91	**8.36**

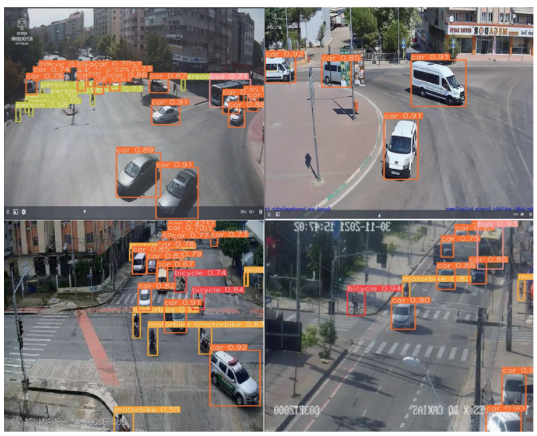


图 7　YOLOv8_MLW 在实际道路监控场景下的检测效果

Fig. 7　The detection performance of YOLOv8_MLW in real-world road surveillance scenarios

3.7　不同轻量化网络结构设计的对比实验

为验证 YOLOv8_MLW 网络结构设计的优越性，本文将 YOLOv8 的主干网络分别替换为当前主流的轻量化网络结构（MobileNetv3、ShuffleNet 和 EfficientVit），并与其进行性能对比。实验结果如表 3 所示。实验表明，YOLOv8_MLW 在参数量（1.40×10^6）、计算量（4.2×10^9）、推理速度（280.50 fps）和存储需求（2.91 MB）上均优于对比模型，同时保持了较高的平均精度（93.7%），仅次于 YOLOv8＋Efficientvit（94.3%）。此外，YOLOv8_MLW 在多目标优化得分（9.51）上显著优于其他设计，表明其在轻量化、检测精度和推理速度之间实现了最优平衡，凸显了其在资源受限的边缘设备上的部署优势。

表 3 不同轻量化网络结构设计的性能对比

Table 3 Performance comparison of different lightweight network architectures

网络结构	Params /10^6	mAP50 /%	FPS /fps	GFLOPs /10^9	权重文件 /MB	Score ×10
YOLOv8+MobileNetv3	2.30	92.5	196.51	4.3	4.67	6.01
YOLOv8+ShuffleNet	1.68	90.6	241.15	4.8	3.43	5.76
YOLOv8+Efficientvit	3.86	**94.3**	162.13	9.1	8.07	3.00
YOLOv8_MLW	**1.40**	93.7	**280.50**	**4.2**	**2.91**	**9.51**

3.8 不同模型之间的对比实验

为评估 YOLOv8_MLW 模型的性能,本文将其与当前广泛应用的轻量级单阶段目标检测模型进行对比,结果如表 4 所示。实验表明,YOLOv8_MLW 在多目标优化加权得分上表现最优,虽然在检测精度上未达到最高,但在参数量、计算量(GFLOPs)、推理速度(FPS)和权重文件大小方面均优于对比模型。具体而言,YOLOv8_MLW 的参数量为 1.40×10^6,计算量为 4.2×10^9,FPS 达到 280.5,权重文件大小仅为 2.91 MB。多目标优化得分(9.108)显示,YOLOv8_MLW 在轻量化、检测精度和推理速度之间实现了最佳平衡。其高效的计算性能、较低的存储需求和出色的实时性,使其非常适合在计算资源和存储空间有限的边缘设备上部署。综上所述,YOLOv8_MLW 为资源受限环境下的道路监控目标检测任务提供了一个理想的解决方案,具有显著的应用潜力和实践价值。

表 4 对比实验结果

Table 4 Comparative experimental results

模型	Params /10^6	mAP50 /%	FPS /fps	GFLOPs /10^9	权重文件 /MB	Score ×10
YOLOv3-tiny	8.67	90.1	57.6	12.90	16.67	2.09
YOLOv5n	1.76	93.2	66.8	**4.10**	3.74	6.80
YOLOx	5.03	92.0	87.5	15.24	38.70	3.49
YOLOv7-tiny	6.00	93.0	117.6	13.10	12.30	5.06
YOLOv8n	3.00	94.1	181.9	8.10	5.98	7.63
YOLOv8S	9.92	**95.2**	106.2	25.5	19.17	3.98
YOLOv10	2.70	94.3	231.7	8.20	5.51	8.27
YOLOv8_MLW	**1.40**	93.7	**280.5**	4.20	**2.91**	**9.11**

4 结 论

本文提出了一种基于网络结构轻量化的道路监控检测模型 YOLOv8_MLW,旨在为道路监控检测模型的高效部署提供可行的解决方案。通过在骨干网络中引入 VanillaNET 和结合 FasterNet 与 EMA 注意力机制的 C2f 模块,显著减少了模型参数量和内存访问量;基于分组卷积构建 G-SPPCSPC 模块,轻量且高效地增强了对多尺度目标的检测能力;并在颈部网络引入轻量级注意力机制 MLCA,有效抑制了背景干扰。实验结果表明,该模型在显著降低模型复杂度的同时,保持了较高的检测精度和实时性,可以有效降低边缘设备的部署成本,具有较高的应用价值。下一步的工作是在保持轻量化的同时,进一步提升模型检测精度。

参考文献

[1] 王宝乐,韩宝睿,董任,等. 基于交通态势数据的道路拥堵情况分析[J]. 物流科技,2024,47(15):80-84.
WANG B L, HAN B R, DONG R, et al. Analysis of road congestion based on traffic situation data[J]. Logistics Sci-Tech, 2024, 47(15): 80-84.

[2] 王宇宁,庞智恒,袁德明. 基于 YOLO 算法的车辆实时检测[J]. 武汉理工大学学报,2016,38(10):41-46.
WANG Y N, PANG ZH H, YUAN D M. Real-time vehicle detection based on YOLO algorithm[J]. Journal of Wuhan University of Technology, 2016, 38(10): 41-46.

[3] 洪松,高定国,三排才让. 基于 YOLO_v3 的车辆和行人检测方法[J]. 电脑知识与技术,2020,16(8):192-198.
HONG S, GAO D G, SANPAI C R. Vehicle and pedestrian detection method based on YOLO_v3[J]. Computer Knowledge and Technology, 2020, 16(8): 192-198.

[4] 郭克友,王苏东,李雪,等. 基于 Dim env-YOLO 算法的昏暗场景车辆多目标检测[J]. 计算机工程,2023,49(3):312-320.
GUO K Y, WANG S D, LI X, et al. Multi-target vehicle detection in dim scenes based on Dim env-YOLO algorithm[J]. Computer Engineering, 2023, 49(3): 312-320.

[5] 范金豪,崔立志. 基于 YOLO_ES 的垃圾分类目标检测模型[J]. 电子测量技术,2023,46(1):160-166.
FAN J H, CUI L ZH. Garbage classification target detection model based on YOLO_ES[J]. Electronic Measurement Technology, 2023, 46(1): 160-166.

[6] 史涛,崔杰,李松. 优化改进 YOLOv8 实现实时无人机车辆检测的算法[J]. 计算机工程与应用,2024,60(9):79-89.
SHI T, CUI J, LI S. Algorithm for real-time UAV vehicle detection based on optimized YOLOv8[J]. Computer Engineering and Applications, 2024, 60(9): 79-89.

[7] 顾杨海, 李富, 陈德基, 等. 基于多尺度特征融合与交互的路侧目标检测算法[J]. 电子测量技术, 2024, 47(23): 152-161.

GU Y H, LI F, CHEN D J, et al. Roadside target detection algorithm based on multi-scale feature fusion and interaction [J]. Electronic Measurement Technology, 2024, 47(23): 152-161.

[8] CHENG H T, WANG Y H, GUO J Y, et al. VanillaNet: The power of minimalism in deep learning [J]. ArXiv preprint arXiv:2305.12972, 2023.

[9] CHEN J R, KAO S H, HE H, et al. Run, don't walk: Chasing higher FLOPS for faster neural networks [C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023:12021-12031.

[10] OUYANG D L, HE S, ZHAN G ZH, et al. Efficient multi-scale attention module with cross-spatial learning[C]. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 2023: 1-5.

[11] IOANNOU Y, ROBERTSON D, CIPOLLA R, et al. Deep roots: Improving CNN efficiency with hierarchical filter groups[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 5977-5986.

[12] LI R J. Development of a micro-target detection algorithm based on refactored sppcspc and enhanced YOLOv7[C]. 2023 4th International Conference on Information Science and Education(ICISE-IE),2023:97-101.

[13] WAN D H, LU R SH, SHEN S Y, et al. Mixed local channel attention for object detection[J]. Engineering Applications of Artificial Intelligence, 2023, 123 (C): 106442.

[14] 邱天衡,王玲,王鹏,等. 基于改进 YOLOv5 的目标检测算法研究[J]. 计算机工程与应用,2022,58(13):63-73.
QIU T H, WANG L, WANG P, et al. Research on target detection algorithm based on improved YOLOv5 [J]. Computer Engineering and Applications, 2022, 58 (13): 63-73.

[15] HU J F, SUN J X, LIN Z H, et al. APANet: Auto-path aggregation for future instance segmentation prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3386-3403.

[16] CHOLLET F. Xception: Deep learning with depthwise separable convolutions [C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017: 1800-1807.

[17] LI H N, JIANG B ZH, MA ZH Y, et al. Dual-channel wind power forecasting model using squeeze and excitation network [C]. 2022 4th International Conference on Smart Power & Internet Energy Systems(SPIES), 2022:2098-2102.

[18] HOU Q B, ZHOU D Q, FENG J SH. Coordinate attention for efficient mobile network design[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2021: 13708-13717.

[19] SHYAMALA N, MAHABOOBBASHA S. Convolutional block attention module-based deep learning model for MRI brain tumor identification (ResNet-CBAM) [C]. 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), 2024: 1603-1608.

[20] HE K M, ZHANG X Y, REN SH Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37 (9): 1904-1916.

[21] LI J Y, ZHAN ZH H, LI Y, et al. Multiple tasks for multiple objectives: A new multiobjective optimization method via multitask optimization [J]. IEEE Transactions on Evolutionary Computation, 2025, 29 (1):172-186.

作者简介

来超凡,硕士研究生,主要研究方向为目标检测。
E-mail:laicf123@163.com

花强(通信作者),教授,主要研究方向为机器学习和智能计算。
E-mail:huaq@hbu.edu.cn

母静越,硕士研究生,主要研究方向为目标检测。
E-mail:594733638@qq.com

张博,讲师,主要研究方向为机器学习、深度学习。
E-mail:sxzhangbo@hebau.edu.cn