

DOI:10.19651/j.cnki.emt.2416512

基于轻量级算法的水上垃圾小目标检测研究^{*}徐尽达^{1,2,3} 陈慈发^{2,3,4} 张 上^{1,2,3}

(1.三峡大学水电工程智能视觉监测湖北省重点实验室 宜昌 443002; 2.三峡大学湖北省建筑质量检测装备工程技术研究中心 宜昌 443002; 3.三峡大学计算机与信息学院 宜昌 443002; 4.荆楚理工学院大数据研究中心 荆门 448000)

摘要: 针对水上漂浮垃圾检测中微小目标物体数量占比高、检测画面易受到水面波动和岸边环境反光等多重因素的干扰、检测模型庞大的参数量和计算量对终端的部署的设备性能要求高等问题,拟提出一种轻量化、高精度、实时性的检测模型 LS-YOLO。首先,该算法利用 HS-FPN 金字塔网络设计构造 YOLOv8 的 Neck 网络结构,构建的网络结构牺牲小部分精度,显著降低了模型的参数数量和计算复杂度。其次,通过引入 CAA 上下文锚定注意机制改进 HS-FPN,捕获远程上下文信息来回升检测精度。然后通过更换损失函数为具有动态聚焦机制的 Wise-IoUv3,大幅提升检测效果,增加模型鲁棒性。最后,使用 LAMP 剪枝技术对模型进行剪枝,减小模型的参数量和计算量。实验结果表明,改进后的 LS-YOLO 相比基准模型 mAP50 提升了 0.9%,召回率提升了 3.2%,参数量降至基准模型的 19.83%,计算量降至基线的 44.44%,模型大小降至基线的 22.22%。经过优化后的检测算法不仅显著提升了检测性能与特征提取的精准度,同时也便于在资源受限的硬件平台上的部署操作。

关键词: 水上垃圾检测;轻量化;小目标;损失函数;通道剪枝

中图分类号: TP391.41; TN40 **文献标识码:** A **国家标准学科分类代码:** 520.6040

Research on small object detection of waterborne debris based on
lightweight algorithmsXu Jinda^{1,2,3} Chen Cifa^{2,3,4} Zhang Shang^{1,2,3}

(1. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, China; 2. Hubei Province Engineering Technology Research Center for Construction Quality Testing Equipment, China Three Gorges University, Yichang 443002, China; 3. College of Computer and information Technology, China Three Gorges University, Yichang 443002, China; 4. Big Data Research Center, Jingchu University of Technology, Jingmen 448000, China)

Abstract: To address the high proportion of small target objects in waterborne debris detection, the interference caused by multiple factors such as water surface fluctuations and shoreline reflections, and the high demands on device performance due to the large number of parameters and computational load of detection models, we propose a lightweight, high-precision, real-time detection model, LS-YOLO. First, this algorithm uses the HS-FPN pyramid network design to construct the Neck network structure of YOLOv8. The constructed network structure sacrifices a small part of the accuracy and significantly reduces the number of parameters and computational complexity of the model. Secondly, HS-FPN is improved by introducing the CAA context-anchored attention mechanism to capture remote contextual information to improve detection accuracy. Then, by replacing the loss function with Wise-IoUv3, which features a dynamic focusing mechanism, the detection performance is significantly improved, increasing the robustness of the model. Finally, LAMP pruning technology is used to prune the model to reduce the number of parameters and calculations of the model. The experiment shows that the improved LS-YOLO has a 0.9% increase in mAP50 compared to the baseline model, a 3.2% increase in recall, a reduction in parameters to 19.83% of the baseline model, a reduction in computational cost to 44.44%, and a reduction in model size to 22.22%. The optimized detection algorithm not only significantly improves detection performance and feature extraction accuracy, but also facilitates deployment on resource-constrained hardware platforms.

Keywords: water garbage detection; lightweight; small target; loss function; channel pruning

0 引 言

伴随着经济的高质量发展,环境保护逐渐成为深入人

心的社会共识。内陆河流为农业灌溉、工业生产和水力发电提供关键水源,同时也是许多地区的饮用水源和运输通道,为经济的可持续发展贡献着清洁能源。然而,随着工业

收稿日期:2024-07-25

^{*} 基金项目:湖北省大学生创新创业训练计划(S202311075047)项目资助

化、城镇化进程的加快,内陆河流也面临着前所未有的挑战,水体污染、垃圾漂浮、生态退化等问题日益凸显,严重威胁到河流的健康与安全。因此,内陆河流的垃圾清理工作变得尤为迫切和重要^[1]。目前,内陆河的水上垃圾清理工作主要还是依赖于人力操作,这不仅效率低下,人力成本高昂,而且伴随着一定的安全风险。因此,探索一种更为高效、安全的自动化清理方式显得尤为重要。

随着科技的进步,基于深度学习的目标检测方法因其成本低廉、实时性强、准确度高等特点,在众多领域得到了广泛应用^[2]。特别是在水上垃圾检测领域,这一技术的应用潜力尤为巨大。然而,水上垃圾检测也面临诸多挑战。首先,检测的焦点往往集中在那些尺寸相对较小的目标上,小目标由于其和背景难以区分、包含的像素更少,不能像大目标一样获取足够的信息来对其进行精准定位^[3],这使得获取足够的目标信息变得尤为困难。更为复杂的是,水面特有的反光和倒影等环境因素进一步增加了检测的难度和复杂性,使得精确识别与定位垃圾变得更具挑战性。现阶段增强小目标检测可采取数据增强、多尺度策略、生成超分辨率细节信息等方法^[4]进行。戚玲珑等^[5]通过对 YOLOv7 添加 Acmix 注意力机制等方法改进,实现水上垃圾检测的精度提升,其平均检测精度度(mean average precision, mAP)达到 71.1%,但精度仍旧远达不到应用水准。岳旭生等^[6]增强数据集并增加小目标检测层等方法改进 YOLOv5s 有效提高了水上垃圾的检测精度。

在目标检测的实际应用中,由于需要将模型部署在性能受限的平台上,即便许多算法拥有较高的精确度,其庞大的模型规模、繁多的参数数量以及巨大的计算需求都成为了部署过程中的显著挑战。因此,网络结构的轻量化设计显得尤为重要。这种轻量化的核心理念在于,在不牺牲深度神经网络精度的前提下,通过优化模型结构来减少存储空间占用提升运行速度,从而实现模型的高效部署^[7-8]。

Lightdet^[9]、MobileNet^[10]、EfficientNet^[11]、GhostNet^[12]等众多模型为轻量化做出了卓越的贡献,YOLO 系列算法^[13-18]作为目标检测的佼佼者,提供了大中小多种模型,并被众多科研人研究出不同应用场景下的许多变体,但具体应用场景下仍然有着较大的提升空间。张新月等^[19]采用轻量化网络结构 Vanillanet 替换 YOLOv8n 的骨干特征提取网络、修改损失函数等手段将模型大小压缩至原来的一半左右,实现了较好的轻量化效果,但并没有解决伴随而来的检测效果的下降。杨洋等^[20]利用批量归一化(batch normalization, BN)层通道的网络剪枝算法改进 YOLOv5m,较大程度降低了模型的参数量,为部署在终端设备提供方便。张丽丽等^[21]提出了关联剪枝方法,在 SAR-Ship-Dataset 等数据集验证中,可轻松达到 70% 剪枝率,但平均检测精度下降也高达 2% 之多。

前人这些深度学习方法虽然可以有效检测小目标或降低模型大小,但现实场景中水上垃圾的检测精度和检测速

度仍受到环境、天气等因素的影响,在优化模型规模同时提升检测精度方向的研究仍有欠缺。YOLOv8 作为目标检测的佼佼者,至今仍然保持规律的维护更新,在 YOLOv8 系列模型中共包含 x/l/m/s/n 从大到小 5 种模型,为了更好地提升水上垃圾检测效果,优化模型参数量、计算量,降低终端设备部署难度,本文基于 YOLOv8 系列最小模型 YOLOv8n 模型进行改进优化,提出了精度更高,更轻量化的模型 LS-YOLO。本文创新的提出了 CAA-HS-FPN 模块,搭建了全新的网络结构、选取适配度最高的损失函数并进行通道剪枝进行轻量化,具体贡献如下:

1) 在 Chen 等^[22]提出的高级筛选功能金字塔网络(high-level screening-feature pyramid networks, HS-FPN)的思想启发下,构建轻巧的 CA-HS-FPN 通道注意力网络模块,重新搭建轻量级的 YOLOv8 的 Neck 网络结构,牺牲较少的精度以实现模型参数量的大幅下降。

2) 引入上下文锚定注意(context anchor attention, CAA)^[23]机制对 CA-HS-FPN 进行优化,以捕捉远距离上下文信息来提升检测精度。

3) 采用具有动态非单调的聚焦机制的损失函数 Wise-IoUv3^[24]替换默认的 CIoU,使模型能够更加注重数量最多的普通质量锚框,提升小目标检测能力。

4) 对改进模型采取层自适应幅度的剪枝(layer-adaptive sparsity for the magnitude-based pruning, LAMP)^[25]方法进行剪枝,对网络大幅瘦身助力模型轻量化部署。

1 LS-YOLO 水上垃圾检测算法

1.1 轻量化网络结构设计

针对水上垃圾目标检测中小目标检测精度不足以及模型参数量、计算量庞大难以在终端部署的问题,本节中对 YOLOv8n 网络进行优化设计,使其拥有更高的检测精度及更轻量的网络结构。

1) CA-HS-FPN 构建 Neck 网络

HS-FPN 是一种旨在高效应对多尺度目标检测挑战的网络结构。其核心组成部分:(1)特征选择模块借助通道注意力(channel attention, CA)和维度匹配机制,巧妙地对不同尺度的特征图进行精准筛选。通过实施全局平均池化和全局最大池化这两种池化操作,并结合权重计算策略,有效地提取了每个通道中的重要信息。(2)特征融合模块通过选择性特征融合机制将筛选后的低级特征和高级特征进行结合。在这一过程中,高级特征经过扩展后,通过双线性插值和转置卷积进行精确的尺度调整,确保与低级特征在尺度上保持一致。从而实现特征融合,这种融合策略极大地丰富了模型的特征表达能力,进一步提升了多尺度目标检测的准确性和效率。在水上垃圾检测中,小目标的位置和角度的微小变化都可能导致其尺度发生显著变动。这一特性恰好与 HS-FPN 检测模块相得益彰,使其在小目标检测方面展现出卓越的性能。

利用 HS-FPN 的思想,进行设计轻量的通道注意力网络结构 CA-HS-FPN(如图 1 所示),通过全局平均池化和全局最大池化来提取特征,通过 1×1 卷积和修正线性单元 ReLU 激活来调整通道数,并通过 Sigmoid 激活函数生成通道注意力权重。最终输出的特征图可以通过权重进行调制,从而增强重要特征,抑制不重要特征,提高模型的表征能力。

由于 CA-HS-FPN 模块具有更强的多尺度检测能力,对检测小目标更加匹配,但并未将其直接加入网络用于提升检测能力,而是采用该模块搭建全新的轻量化 Neck 网络结构,在保证高检测精度的前提下,减少消耗高计算量的 C2f 模块^[26]的使用,精简网络的参数量和计算量,重新搭建

新的轻量化网络结构 CA-HS-FPN_YOLO,如图 2 所示。

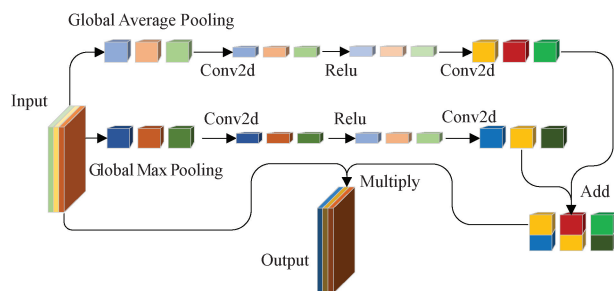


图 1 CA-HS-FPN 模块

Fig. 1 CA-HS-FPN module

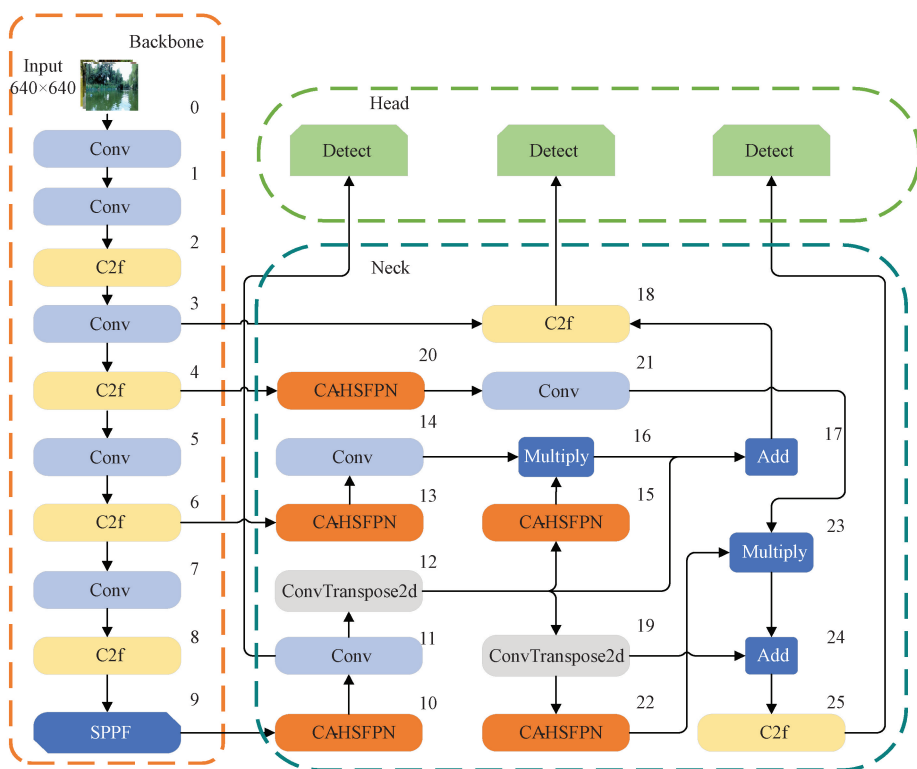


图 2 CA-HS-FPN_YOLO 网络结构

Fig. 2 CA-HS-FPN_YOLO network architecture

2) CAA 改进 CA-HS-FPN

通过精心设计的 CA-HS-FPN 架构,上文成功搭建了一个新型的轻量级网络结构。这一创新显著降低了模型的参数量和计算量,实现了更高的计算效率,但也不可避免地带来了一定程度的检测精度微小下降。为了有效回升优化检测精度,本文巧妙地引入了上下文锚定注意力 CAA 机制。

CAA 机制的核心在于其独特的全局平均池化和一维带状卷积的组合应用。这种组合策略使得网络能够精准地捕获图像中远距离像素之间的潜在关系,进而增强中心关键区域内的特征学习能力。CAA 机制旨在掌握图像中不同位置像素之间的上下文相互依赖性,特别是学习对检

测精度影响重大的远处像素信息,来强化中心区域的特征表示。

在具体实现上,CAA 应用全局平均池化操作对输入特征图进行空间压缩,随后通过 1×1 卷积来进一步提取和整合局部区域的特征信息。为了更有效地模拟和逼近传统大核深度卷积的功能,CAA 还采用了两个轻量级的深度条带卷积 H-Conv2d 和 W-Conv2d。这种条状卷积结构能够沿高和宽方向高效地捕获和整合像素信息,非常适用于识别和提取具有细长或特殊形状物体(如水瓶等垃圾)的特征,因此也更符合水上垃圾检测的要求。本文将 CAA 巧妙地融入并优化 CA-HS-FPN,从而有效缓解 CA-HS-FPN_YOLO 网络结构在检测任务中可能遭遇的精度

下滑问题,新的 CAA-HS-FPN 模块,不仅继承了原 CA-HS-FPN 的强大特征提取能力,还进一步增强了网络对关键特征区域的关注,实现了对检测精度的显著提升。改进后的模块替换上小节 Neck 网络中的 CA-HS-FPN,改进的 CAA-HS-FPN 模块如图 3 所示,该模块使得模型在精度回升方面取得了较好的效果。

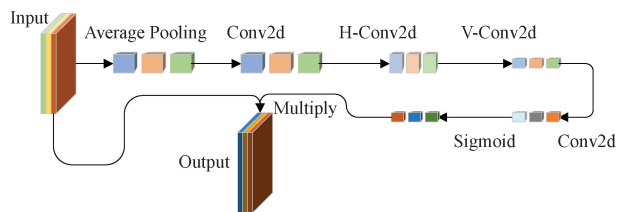


图 3 CAA-HS-FPN 模块

Fig. 3 CAA-HS-FPN module

1.2 损失函数

本文采用的 YOLOv8n 基线模型中的损失函数 (complete-IoU, CIoU) 如式(1)~(4)所示。

$$CIoU = IoU - \left(\frac{\rho^2(b, b_{gt})}{c^2} \right) + \alpha v \quad (1)$$

$$\alpha = \frac{v}{(1 - IoU) + v} \quad (2)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w_{gt}}{h_{gt}} - \arctan \frac{w}{h} \right)^2 \quad (3)$$

$$L_{CIoU} = 1 - CIoU \quad (4)$$

其中, w, h, w_{gt}, h_{gt} 分别代表预测框的宽、高和实际框的宽、高; IoU 代表预测边界框 b 和真实边界框 b_{gt} 的交并比; $\rho(b, b_{gt})$ 表示预测边界框 b 的中心点和真实边界框 b_{gt} 的中心点之间的欧几里得距离; c 表示预测边界框和真实边界框的最小外接矩形的对角线长度; α 是由式(2)计算的权重因子; v 表示预测边界框和真实边界框的宽高比一致性, L_{CIoU} 是最终得到的损失函数。

虽然 CIoU 引入纵横比机制已经实现了较好地检测效果,但它的主要局限在于它没有充分考虑到困难检测样本与容易检测样本之间的平衡问题,这导致在处理小目标检测任务时,其性能表现不尽如人意。而基于动态非单调聚焦机制的 Wise-IoUv3 损失函数弥补了这一缺陷, Wise-IoUv3 提出了动态非单调的聚焦机制,使用“离群度”替代 IoU 对锚框进行质量评估,采用了梯度增益分配策略,降低高质量锚框的竞争力的同时减小低质量目标产生的有害梯度以实现平衡低质量样本和高质量样本的权重,使得模型能够更专注数量最多的普通目标,提升锚框的检测能力。Wise-IoUv3 如式(5)~(8)所示。

$$L_{WIoUv3} = \left(\frac{\beta}{\delta \alpha^{\beta-\delta}} \right) L_{WIoUv1} \quad (5)$$

$$\beta = \frac{L_{IoU}^r}{L_{IoU}} \in [0, +\infty) \quad (6)$$

$$L_{WIoUv1} = R_{WIoU} (1 - IoU) \quad (7)$$

$$R_{WIoU} = \exp \left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{W_g^2 + H_g^2} \right) \quad (8)$$

其中, (x, y) 是预测框的中心坐标, (x_{gt}, y_{gt}) 是真实框的中心坐标; W_g, H_g 分别为同时包含预测框和真实框的最小外围框宽高。 β 为离群度,它表示的是一个目标 IoU 损失 L_{IoU}^r 与当前 IoU 损失 L_{IoU} 的比值; δ 和 α 是超参数,用于调节比例因子; R_{WIoU} 是距离损失; L_{WIoUv1} 是具有距离注意力的 1 代 Wise-IoU。

作者在 MS-COCO 数据集下对 δ 和 α 两个超参数进行实验分析,在 $\alpha=1.9$ 和 $\delta=3$ 时表现出最佳检测效果。但不同数据集中的目标对普通锚框有着不同的定义,对于水上垃圾数据集来说,小目标数目众多以及受环境影响较为严重,本文将做多组实验探索当前应用的最佳效果。

1.3 LAMP 通道剪枝

当前的深度神经网络在各个应用领域都能实现很好的检测效果,但大多检测模型伴随着检测精度的提升,其模型参数量、计算量、模型大小也越来越大,这显然对实际的应用部署产生了巨大的困难。因此产生了众多的轻量化手段,而通道剪枝技术则作为轻量化最有效的手段之一被广泛应用。神经网络通道剪枝是一种通过删除模型中的不重要权重来满足实际约束、减轻过度拟合、增强可解释性或加深对神经网络训练理解的技术。通常的层系数剪枝需要精心选择层间稀疏度才能产生良好的剪枝效果,本文选择基于层自适应幅度的剪枝 LAMP 剪枝技术对模型进行轻量化, LAMP 提出了一种新的全局剪枝重要度得分方法如式(9)所示。

$$Score(u; W) = \frac{(W[u])^2}{\sum_{v \geq u} (W[v])^2} \quad (9)$$

其中, $Score(u; W)$ 表示权重中第 u 个索引的 LAMP 得分,分数是第 u 个权重的平方除以第 u 个权重之后的所有权重的平方和,这个分数反映了第 u 个权重相对于其后所有权重的重要性; W 表示卷积核展开为一维向量的权值; $\sum_{v \geq u} (W[v])^2$ 表示从 $W[u]$ 之后的所有元素的平方和; u 和 v 表示索引值来对权值进行按序排列。

该剪枝方法通过计算每个连接的 LAMP 分数,并根据这些分数对连接进行全局修剪,从而达到所需的稀疏性。这一过程等同于在不同层之间自动选择稀疏性来执行剪枝,因此该方法不需调试超参数,只需设置剪枝计算量压缩率 (Speed_up) 即可实现优秀的轻量化效果。Speed_up 如式(10)所示。

$$Speed_up = \frac{Compute_{Bef}}{Compute_{Aft}} \quad (10)$$

其中, $Speed_up$ 是剪枝计算量压缩率, $Compute_{Bef}$ 表示剪枝前的计算量, $Compute_{Aft}$ 代表剪枝后的计算量。当计算量压缩达到预设值时,剪枝完成。剪枝完成后检测精度的下降通过微调训练来进行回升。

2 实验结果与分析

2.1 FloW-IMG 数据集

为了验证改进后算法的效果,本实验选取 FloW-IMG 数据集^[27]作为评估对象。该数据集是欧卡智舶公司制作的全球首个从无人船视角出发,针对真实内陆水域环境中漂浮物垃圾检测而构建的数据集。数据集内容丰富,包含了 2 000 张高清图,共计精确标注了 5 271 个目标对象,为算法性能的全面验证提供了可靠的数据支撑。其中,取复杂光影场景图像如图 4(a)所示,远距离目标场景如图 4(b)所示。

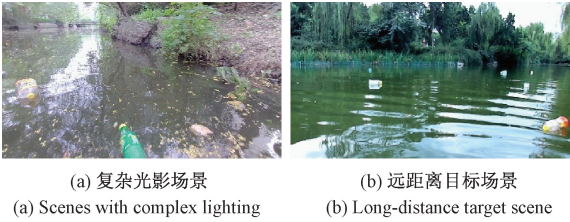


图 4 FloW-IMG 数据集不同场景图片

Fig. 4 FloW-IMG dataset images

通过 Python 脚本,将 2 000 张数据集标签文件和图像文件打乱随机分配,以 8 : 2 的比例划分为训练集和验证集,确保两个集合中的数据分布均匀,进而为模型训练和性能评估提供坚实的基础。

2.2 实验环境及参数设置

本实验环境具体配置如表 1 所示,实验设置了 500 个训练周期(epochs)来保证网络模型实现收敛,并设置模型在 50 轮训练没有提升(patience)的情况下提前终止训练;使用 2 个线程(workers)加载数据;每批次迭代数量(batch size)设置 32 张图像,以平衡训练速度和资源占用。

表 1 实验的环境配置

Table 1 Experimental environment configuration

实验环境	版本
操作系统	Ubuntu23.10
处理器(CPU)	Intel i5-12400f
显示卡(GPU)	Nvidia RTX 40600Ti
内存	16 GB
Python	3.10.13
Pytorch	1.11.0
CUDA	11.8

2.3 评价指标

本实验目的是水上垃圾目标检测算法优化,目的是保证精度上升的同时进行轻量化改进,为精准评估实验改进的有效性,设置以下多个评估指标,分别是:精度(precision,P)、回归率(recall,R)、平均精度均值 mAP50、检测速率(frames per second,FPS)、参数量、计算量和最佳

模型权重文件大小,计算公式如下:

$$P = \frac{TP}{TP + FP} \times 100\%$$
 (11)

$$R = \frac{TP}{TP + FN} \times 100\%$$
 (12)

$$mAP = \frac{\sum_{i=1}^N \int_0^1 P dR}{N} \times 100\%$$
 (13)

$$FPS = \frac{n \times batch}{\sum_{i=1}^n (T_{end} - T_{start})}$$
 (14)

其中,FP、TP、FN 分别表示目标被错误检测的数量、目标被正确检测的数量和没有检测到的目标数量;N 为目标的类别数量,mAP 是结合精度和召回率的综合性指标,它考虑了不同类别、不同阈值下模型的检测性能,mAP 能够评估整体检测能力的好坏,值越高意味着模型的检测性能更好;n 代表共检测的批次次数,batch 表示每批次检测的图片数量,(T_{end} - T_{start})表示每轮检测消耗的时间。本文采用的 FPS 检测方法为通过预先检测 200 轮进行预热,保证检测时的环境状态平稳,后进行 1 000 轮的正式推理来计算平均的 FPS 作为最终结果。

2.4 实验结果分析

1) Wise-IoU 实验

为探寻 Wise-IoU 在 FloW-IMG 数据集上的最佳表现,本文对损失函数的两个超参数 α 和 δ 的组合做大量实验进行探究。李岩超等^[28]在遥感场景中测试出 Wise-IoUv3 的最佳范围为 α ∈ [1.25, 1.75],δ ∈ [4, 5],同为小目标场景的水上垃圾数据集以该范围作参考,在此基础上继续扩大范围进行探究,测试出符合本实验场景的超参数。共 20 组实验结果如表 2 所示。实验以 mAP50 作为参考值,最终得出结论,当 α 取值 0.75,δ 取值 4.0 时,达到最大的 mAP50 为 0.91,相较于基础模型采用 CIoU 的 mAP 只有 0.891,检测能力提升明显。

表 2 Wise-IoU 超参数探究结果

Table 2 Exploration results of Wise-IoU hyperparameters

α	δ	mAP50	α	δ	mAP50
0.50	3.5	0.898	0.50	4.5	0.899
0.75	3.5	0.905	0.75	4.5	0.902
1.00	3.5	0.907	1.00	4.50	0.899
1.25	3.5	0.903	1.25	4.5	0.905
1.50	3.5	0.900	1.50	4.5	0.900
1.75	3.5	0.896	1.75	4.5	0.894
0.50	4.0	0.905	0.50	5.0	0.899
0.75	4.0	0.910	0.75	5.0	0.905
1.00	4.0	0.903	1.00	5.0	0.896
1.25	4.0	0.903	1.25	5.0	0.904
1.50	4.0	0.893	1.50	5.0	0.897
1.75	4.0	0.901	1.75	5.0	0.889
(CIoU)	—	0.889	—	—	—

实验结果验证了不同数据集应用 Wise-IoU 时需要设置不同的超参数,对于水上垃圾检测场景,超参数 $\alpha \in [0.5, 1.25], \delta \in [3.5, 4.5]$ 的范围内出现优质检测结果的概率更大,这一结果可以对 Wise-IoU 的类似场景应用提供参考。

2) 剪枝实验

LAMP 剪枝技术作为先进的层自适应剪纸能力,无需调试超参数即可实现自动化剪枝,可轻易地为训练好的网

络模型进行轻量化操作。但由于网络层之间的连接紧密程度不同,不同的剪枝程度对网络层之间的关联的损害程度也不同,剪枝不可避免地伴随精度的下降,虽然通过剪枝后的微调回升了精度,但回升水平也会有所差异。本文通过设置不同的 *Speed_up*,探究不同剪枝程度所能实现的效果,旨在通过剪枝进行轻量化的同时将精度的损失减轻到最低,实验数据均源自剪枝操作后的精细微调处理,具体结果如表 3 所示。

表 3 剪枝后微调实验结果
Table 3 Fine-tuning results after pruning

<i>Speed_up</i>	<i>FPS</i>	权重文件大小/MB	参数量	计算量(GFLOPs)	<i>P</i>	<i>R</i>	<i>mAP50</i>
改进网络结构(未剪枝)	336.9	4.5	2 089 811	7.3	0.884	0.845	0.901
1.7	340.1	1.8	763 492	4.3	0.872	0.850	0.891
2.0	354.9	1.4	596 136	3.6	0.860	0.864	0.898
2.3	359.2	1.2	484 901	3.2	0.876	0.828	0.894
2.7	347.9	1.0	388 469	2.7	0.879	0.830	0.886
3.0	352.3	0.9	335 511	2.4	0.859	0.843	0.893
3.3	351.8	0.9	303 180	2.2	0.881	0.831	0.890
3.7	347.8	0.7	256 191	1.9	0.866	0.823	0.872

随着剪枝率的提升,模型的参数量、计算量、权重文件的大小都在逐步降低,但也一定程度上伴随着 *P*、*R*、*mAP* 值的下降,本文以 *mAP50* 为主要参数并考量模型的轻量化程度,选择剪枝率 2 时作为最优模型。当 *Speed_up* 达到 2 时,相较于剪枝前模型 *mAP50* 下降幅度最低,*FPS* 有一定程度的提升,同时实现了较好的轻量化效果,权重

文件大小降低为之前的 31.1%,参数量降低至之前的 28.5%,计算量降低到之前的一半左右,实现了较好的均衡效果。此时各个通道剪枝的剪枝程度见图 5,图中上面突出部分为被剪掉的冗余通道,下面部分为剪枝并通过微调后模型的各层通道,剪枝效果明显,充分验证了 LAMP 通道剪枝的有效性。

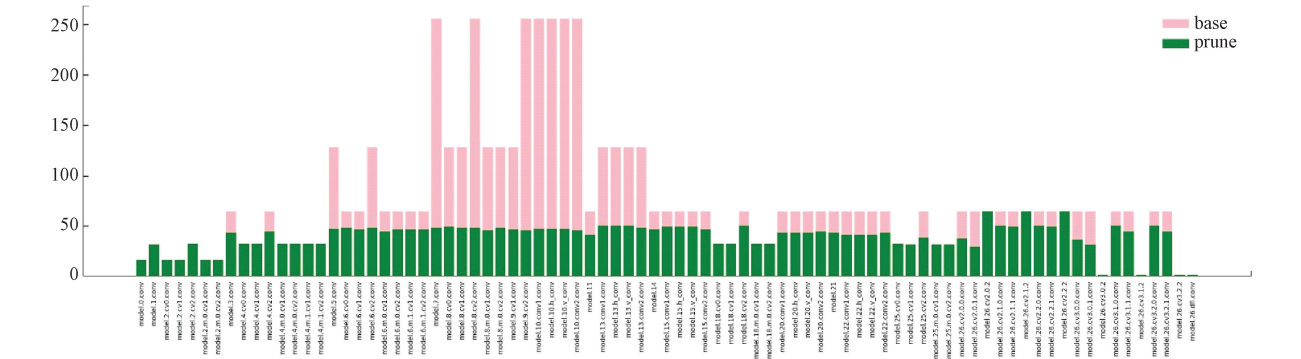


图 5 剪枝前后通道对比图
Fig. 5 Comparison of channels before and after pruning

3) 消融实验

为了全面评估算法检测效能及验证改进模块的有效性,本实施例在最轻量化模型 YOLOv8n 上进行消融实验,实验结果如表 4 所示,展示了各模块对检测性能的具体提升,被使用的方法以“√”进行标注。

实验结果表明了每个方法的有效性,具体而言,CA-HS-FPN 通过精心设计的轻型网络架构重构,在轻微牺牲检测精度的前提下,实现了模型计算量与参数量的显著削

减。随后,通过融入上下文锚定机制(CAA)对 CA-HS-FPN 进行深化优化,不仅有效弥补了检测性能的细微下滑,还在基准模型的基础上将 *mAP50* 提升了 0.2%。

为进一步优化模型在处理难易样本间的平衡问题,本文采用了 Wise-IoU 损失函数替代传统的 CIoU 损失,这一改进促使模型更加聚焦于数量占据主导的普通目标,从而实现了检测性能的飞跃式提升,相较于基础模型,*mAP50* 指标提升了 2.1%。

表 4 消融实验结果
Table 4 Ablation study results

模型	CA-HS-FPN	CAA	Wise-IoUv3 (0.75,4.0)	LAMP 剪枝(2.0)	FPS	权重文件 大小/MB	参数量	计算量 (GFLOPs)	P	R	mAP ₅₀
YOLOv8n					364.0	6.3	3 005 843	8.1	0.867	0.832	0.889
	✓				311.5	4.1	1 933 523	6.8	0.868	0.820	0.885
	✓	✓			371.3	4.4	2 089 811	7.3	0.854	0.840	0.893
			✓		369.2	6.3	3 005 843	8.1	0.892	0.853	0.910
				✓	348.9	1.9	840 704	4.0	0.846	0.828	0.875
	✓	✓	✓		336.9	4.5	2 089 811	7.3	0.884	0.845	0.901
LS-YOLO	✓	✓	✓	✓	354.9	1.4	596 136	3.6	0.860	0.864	0.898

最终引入层自适应剪枝技术 LAMP,对已完成训练的模型进行精细化轻量化处理。该技术在确保模型性能尽可能少受影响的同时,大幅降低了模型的计算复杂度。而最终模型 LS-YOLO 相较于基础模型进行剪枝 mAP_{50} 提升了 2.3%,进一步验证了网络改进方法的有效性。

最终,LS-YOLO 成功实现了模型的轻量化改进,不仅大幅度削减了其计算需求,同时保持了卓越的检测能力,

从而诞生了一个既轻量又具备强大检测性能 的模型。

4) 可视化分析

为了充分验证本文所提出算法在实际应用场景下检测任务的有效性,本文特别选取了 FloW-Img 数据集中具有挑战性的场景进行测试,这些场景包括但不限于目标距离较远、光照强度显著变化以及图像模糊或失帧等复杂条件。对比结果如图 6 所示。

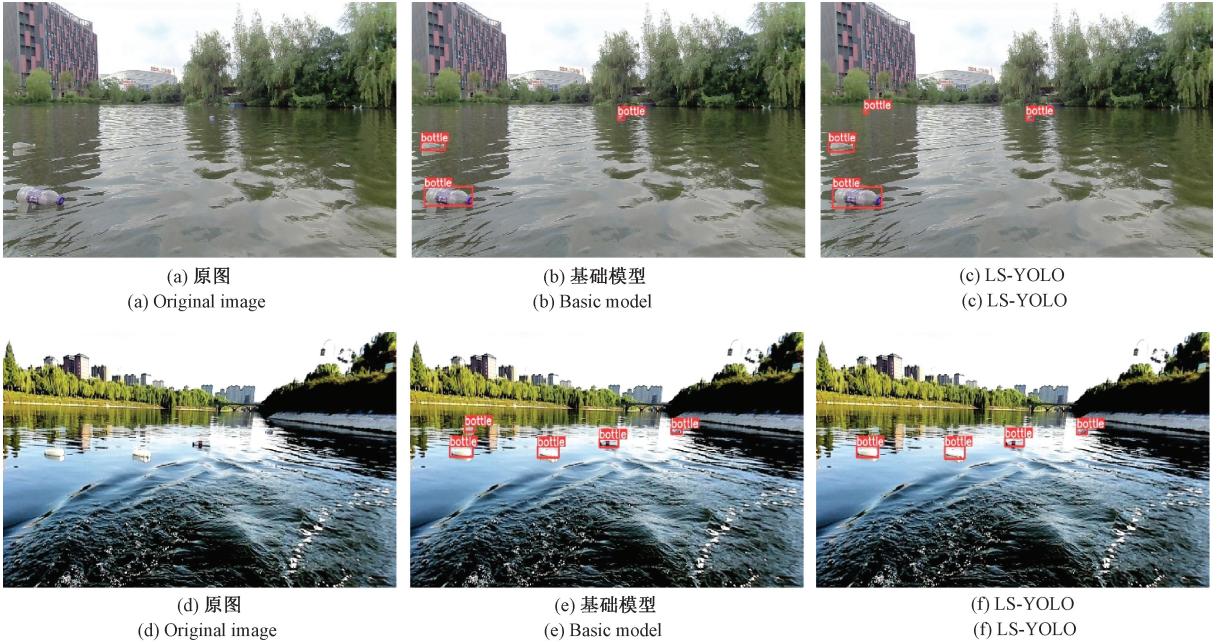


图 6 检测结果可视化
Fig. 6 Visualization of detection results

检测结果可视化如图 6 所示,图 6(a)的识别结果为图 6(b),基础模型未能检测到位于左上方的微小目标,显示出其在细节捕捉上的不足。相比之下,LS-YOLO 模型的检测结果(图 6(c))成功识别并标记了该微小目标,体现了改进模型具有更强的检测精度与敏感性。此外,在另一场景识别(图 6(d))中,当基础模型错误地将图 6(e)左上角的光照反射误判为目标时,改进后的 LS-YOLO 模型识别(图 6(f))展现出了卓越的区分能力,正确识别了真实目

标,忽略了非目标的光照干扰,进一步验证了其在复杂环境下的识别准确性与鲁棒性。实验结果表明,在基础模型出现的漏检、错检问题得到较大改善,模型的鲁棒性实现了有效增强。

5) 对比实验

为直观查看 LS-YOLO 的优越性,本文设置了对比实验,与一众经典轻量化模型进行对比。对比指标包含检测速度、轻量化程度和检测精度等多个项目,实验结果如表 5 所示。

表 5 对比实验结果

Table 5 Comparative experimental results

模型	FPS	权重文件 大小/MB	参数量	计算量 (GFLOPs)	P	R	mAP50
基础模型(YOLO8n)	364.00	6.3	300 5843	8.1	0.867	0.832	0.889
YOLOv6	397.20	8.7	4 233 843	11.8	0.852	0.830	0.877
YOLOv7-tiny	384.62	12.3	6 006 646	13.0	0.864	0.809	0.863
YOLOv10n	277.20	5.8	2 694 806	8.2	0.886	0.831	0.893
戚玲珑改进 YOLOv7 ^[5]	×	×	×	×	×	×	0.711
李安达改进 YOLOv7 ^[29]	43.00	×	×	×	×	×	0.893
张瑶改进 YOLOv8-WSSOD ^[30]	86.00	×	×	12.2	×	×	0.824
(our)LS-YOLO	354.90	1.4	596 136	3.6	0.860	0.864	0.898

注：“×”表示原文中未给出数据

实验结果中,本文引用了戚玲珑与李安达分别改进的 YOLOv7 模型,和张瑶等改进的 YOLOv8n 作为参考,尽管这 3 项实验基于相同的检测数据集展开,但由于实验环境的差异及结果呈现的不充分性,本文仅将其视为辅助性对比。在全面对比中,LS-YOLO 模型展现出了显著的优势;尽管在检测速度 FPS 这一单项指标上略显不足,但在轻量化设计与检测精度两大核心方面,LS-YOLO 均实现了对其他模型的全面超越。

LS-YOLO 模型因其卓越的轻量化特性与高精度检测能力,更适用于水上垃圾检测的应用场景,不仅能够有效提升检测效率,还极大地便利了在各种终端设备上的部署与运行,为实际应用带来了更多便利与可能性。

3 结 论

本文聚焦于模型轻量化与高精度检测的双重目标,创新性地提出了一种针对水上垃圾检测的轻量化 YOLOv8n 改进算法 LS-YOLO。与当前主流经典算法相比,LS-YOLO 展现出显著优势。该极致轻量化模型完美平衡了轻量化设计与高性能需求,这一特性为其在终端计算设备中的部署与应用奠定了坚实的基础。下一步研究方向是将算法部署在终端设备中并设计目标跟踪及路径规划算法,为水上垃圾自动化清理的应用作出贡献。

参考文献

[1] 王利. 我国长江垃圾倾倒防范机制的不足及完善策略——以“长江口倾倒垃圾案”为视角[J]. 平顶山学院学报,2018,33(4):64-68.

WANG L. The deficient prevention mechanism for waste dump into Yangtze River and the strategy to improve: From a viewpoint of the case of waste dump into the Yangtze River estuary [J]. Journal of Pingdingshan University, 2018, 33(4): 64-68.

[2] 李柯泉,陈燕,刘佳晨,等. 基于深度学习的目标检测算

法综述[J]. 计算机工程,2022,48(7):1-12.

LI K Q, CHEN Y, LIU J CH, et al. A survey of object detection algorithms based on deep learning[J]. Computer Engineering, 2022, 48(7): 1-12.

[3] TONG K, WU Y Q, ZHOU F. Recent advances in small object detection based on deep learning: A review [J]. Image and Vision Computing, 2020, 97: 103910.

[4] 曹健,陈怡梅,李海生,等. 基于深度学习的道路小目标检测综述[J]. 计算机工程,2023,49(10):1-12.

CAO J, CHEN Y M, LI H SH, et al. Survey of small target detection on roads based on deep learning [J]. Computer Engineering, 2023, 49(10): 1-12.

[5] 戚玲珑,高建瓴. 基于改进 YOLOv7 的小目标检测[J]. 计算机工程,2023,49(1):41-48.

QI L L, GAO J L. Small object detection based on improved YOLOv7[J]. Computer Engineering, 2023, 49(1): 41-48.

[6] 岳旭生,李军,王耀弘,等. 基于改进 YOLOv5s 的水面漂浮小目标检测算法[J]. 中国舰船研究,2024:1-9. DOI:10.19693/j.issn.1673-3185.03689.

YUE X SH, LI J, WANG Y H, et al. A small floating target detection algorithm on the water surfacebased on the improved YOLOv5s[J]. Chinese Journal of Ship Research, 2024: 1-9. DOI:10.19693/j.issn.1673-3185.03689.

[7] 王军,冯孙铖,程勇. 深度学习的轻量化神经网络结构研究综述[J]. 计算机工程,2021,47(8):1-13.

WANG J, FENG S CH, CHENG Y. Survey of research on lightweight neural network structures for

- deep learning [J]. Computer Engineering, 2021, 47(8): 1-13.
- [8] 赵倩,杨一聪.多重金字塔的轻量化遥感车辆小目标检测算法[J].电子测量技术,2023,46(13):88-94.
- ZHAO Q, YANG Y C. Small object detection algorithm for lightweight remote sensing vehicles with multiple pyramids [J]. Electronic Measurement Technology, 2023, 46(13): 88-94.
- [9] TANG Q K, LI J, SHI ZH P, et al. Lightdet: A lightweight and accurate object detection network[C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 2243-2247.
- [10] HOWARD A G, ZHU M L, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [J]. ArXiv preprint arXiv:1704.04861, 2017.
- [11] TAN M X, LE Q V. Efficientnet: Rethinking model scaling for convolutional neural networks [C]. International Conference on Machine Learning, 2019: 6105-6114.
- [12] HAN K, WANG Y H, TIAN Q, et al. Ghostnet: More features from cheap operations[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:1577-1586.
- [13] 毛少华,王文东.基于深度学习的 YOLO 系列物体检测算法研究综述[J].延安大学学报(自然科学版), 2024,43(2):88-95.
- MAO SH H, WANG W D. Research review of YOLO series object detection algorithms based on deep learning [J]. Journal of Yan'an University (Natural Science Edition), 2024, 43(2): 88-95.
- [14] 叶彦斐,程立,侯翔一.基于改进 YOLOv5 的钢轨内部伤损 B 显图像识别与分类[J].国外电子测量技术, 2023,42(12):70-76.
- YE Y F, CHENG L, HOU X Y. Recognition and classification of internal defects in railway tracks based on improved YOLOv5 [J]. Foreign Electronic Measurement Technology, 2023, 42(12): 70-76.
- [15] 方涛涛,王池社,王洁,等.基于 YOLOv8n 的探地雷达图像管线定位方法[J].国外电子测量技术,2023, 42(11):170-177.
- FANG T T, WANG CH SH, WANG J, et al. Ground penetrating radar image pipeline location based on YOLOv8n [J]. Foreign Electronic Measurement Technology, 2023, 42(11): 170-177.
- [16] WANG C Y, BOCHKOVSKIY A, LIAO H Y M, et al. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023:7464-7475.
- [17] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: Learning what you want to learn using programmable gradient information [J]. ArXiv preprint arXiv: 2402.13616, 2024.
- [18] WANG AO, CHEN H, LIU L, et al. YOLOv10: Real-time end-to-end object detection [J]. ArXiv preprint arXiv:2405.14458, 2024.
- [19] 张新月,胡广锐,李浦航,等.基于改进 YOLOv8n 的轻量化红花识别方法[J].农业工程学报,2024,40(13): 163-170.
- ZHANG X Y, HU G R, LI P H, et al. Recognizing safflower using improved lightweight YOLOv8n [J]. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(13): 163-170.
- [20] 杨洋,宋品德,杨思念,等.面向边缘计算平台及遥感影像的实时检测算法[J].电子测量技术,2024,47(2): 150-159.
- YANG Y, SONG P D, YANG S N, et al. Real-time detection algorithm for edge computing platforms and remote sensing imagery [J]. Electronic Measurement Technology, 2024, 47(2): 150-159.
- [21] 张丽丽,王贤俊,屈乐乐,等.SAR 图像舰船检测的神经网络关联剪枝方法[J].雷达科学与技术,2024, 22(3):284-290,299.
- ZHANG L L, WANG X J, QU L L, et al. Neural network associative pruning for ship detection in SAR images [J]. Radar Science and Technology, 2024, 22(3): 284-290,299.
- [22] CHEN Y F, ZHANG CH Y, CHEN B, et al. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases[J]. Computers in Biology and Medicine, 2024, 170: 107917.
- [23] CAI X H, LAI Q X, WANG Y W, et al. Poly kernel inception network for remote sensing detection [C].

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 27706-27716.

[24] TONG Z J, CHEN Y H, XU Z W, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[J]. ArXiv preprint arXiv:2301.10051,2023.

[25] LEE J, PARK S, MO S, et al. Layer-adaptive sparsity for the magnitude-based pruning[J]. ArXiv preprint arXiv:2010.07611, 2020.

[26] 刘丽,张硕,白宇昂,等.改进 YOLOv8 的轻量级军事飞机检测算法[J].计算机工程与应用, 2024, 60(18): 114-125.

LIU L, ZHANG SH, BAI Y ANG, et al. Improved lightweight military aircraft detection algorithm of YOLOv8 [J]. Computer Engineering and Applications, 2024, 60(18): 114-125.

[27] CHENG Y W, ZHU J N, JIANG M X, et al. FloW: A dataset and benchmark for floating waste detection in inland waters [C]. IEEE/CVF International Conference on Computer Vision, 2021: 10953-10962.

[28] 李岩超,史卫亚,冯灿.面向无人机航拍小目标检测的轻量级 YOLOv8 检测算法[J].计算机工程与应用, 2024,60(17):167-178.

LI Y CH, SHI W Y, FENG C. Lightweight YOLOv8 detection algorithm for small object detection in UAV aerial photography [J]. Computer Engineering and Applications, 2024,60(17):167-178.

[29] 李安达,吴瑞明,李旭东.改进 YOLOv7 的小目标检测算法研究[J].计算机工程与应用, 2024, 60(1): 122-134.

LI AN D, WU R M, LI X D. Research on improving YOLOv7 ’ s small target detection algorithm [J]. Computer Engineering and Applications, 2024, 60(1): 122-134.

[30] 张瑶,陈姚节.改进 YOLOv8 的水面小目标检测算法[J].计算机系统应用,2024,33(4):152-161.

ZHANG Y, CHEN Y J. Improved YOLOv8 algorithm for small object detection on water surface [J]. Computer Systems & Applications, 2024, 33(4): 152-161.

作者简介

徐尽达,硕士研究生,主要研究方向为人工智能、计算机视觉等。

E-mail:15512369018@163.com

陈慈发(通信作者),二级教授、研究员、硕士生导师,主要研究方向为嵌入式系统、物联网、计算机测控系统。

E-mail:chcf0415@126.com

张上,博士、副教授、硕士生导师,主要研究方向为物联网技术、计算机应用技术、图像处理。

E-mail:zhangshang@ctgu.edu.cn