

基于轻量化 YOLO v8s - GD 的自然环境下 百香果快速检测模型

罗志聪^{1,2} 何陈涛¹ 陈登捷¹ 李鹏博¹ 孙奇燕³

(1. 福建农林大学机电工程学院, 福州 350002; 2. 福建省农业信息感知技术重点实验室, 福州 350002;
3. 福建农林大学计算机与信息学院, 福州 350002)

摘要: 为了提高百香果检测精度, 并将深度学习模型部署在移动平台上, 实现快速实时推理, 本文提出一种基于改进 YOLO v8s 的轻量化百香果检测模型 (YOLO v8s - GD)。使用聚集和分发机制 (GD) 替换颈部特征融合网络, 提高模型对百香果图像特征信息跨层融合能力和模型泛化能力; 通过基于层自适应幅度的剪枝 (LAMP) 修剪模型, 损失一定精度换取减小模型体积, 减少模型参数量, 以实现在嵌入式设备上快速检测; 运用知识蒸馏学习策略弥补因剪枝而损失的检测精度, 提高模型检测性能。实验结果表明, 对于自然环境下采集的百香果数据集, 改进后模型参数量和内存占用量相比原 YOLO v8s 基线模型分别降低 63.88% 和 62.10%, 精确率 (Precision) 和平均精度 (AP) 相较于原模型分别提高 0.9、2.3 个百分点, 优于其他对比模型。在 Jetson Nano 和 Jetson Tx2 嵌入式设备上实时检测帧率 (FPS) 分别为 5.78、19.38 f/s, 为原模型的 1.93、1.24 倍。因此, 本文提出的改进后模型能够有效检测复杂环境下百香果目标, 为实际场景中百香果自动采摘等移动端检测设备部署和应用提供理论和技术支持。

关键词: 百香果; YOLO v8s; 轻量化; 检测模型; 聚集和分发机制

中图分类号: TP391.4; S24

文献标识码: A

文章编号: 1000-1298(2024)08-0291-10

OSID:



Passion Fruit Rapid Detection Model Based on Lightweight YOLO v8s - GD

LUO Zhicong^{1,2} HE Chentao¹ CHEN Dengjie¹ LI Pengbo¹ SUN Qiyang³

(1. College of Mechanical and Electronic Engineering, Fujian Agriculture and Forestry University, Fuzhou 350002, China
2. Fujian Key Laboratory of Agricultural Information Sensing Technology, Fuzhou 350002, China
3. College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China)

Abstract: In order to improve the accuracy of passion fruit detection and deploy the deep learning model on mobile platforms for rapid real-time inference, a lightweight passion fruit detection model was proposed based on an improved YOLO v8s. The model replaced the neck feature fusion network with a Gather-and-distribute mechanism (GD) to enhance cross-layer feature fusion and generalization capabilities for passion fruit images. Additionally, the model was pruned by using layer-adaptive sparsity for the magnitude-based pruning (LAMP), which traded off some accuracy to reduce model size and parameter count, facilitating rapid detection on embedded devices. Knowledge distillation was employed to compensate for the accuracy loss caused by pruning, further enhancing detection performance. Experimental results showed that for a passion fruit dataset collected in natural environments, the improved model reduced parameter count and memory usage by 63.88% and 62.10%, respectively, compared with the original YOLO v8s baseline model. The precision and average precision (AP) of the improved model were increased by 0.9 percentage points and 2.3 percentage points, respectively, outperforming other comparative models. Real-time detection frame rates (FPS) on Jetson Nano and Jetson Tx2 embedded devices were 5.78 f/s and 19.38 f/s, respectively, which were 1.93 times and 1.24 times higher than that of the original model. Therefore, the proposed improved model effectively detects passion fruit in complex environments, providing theoretical and technical support for the deployment and application of mobile detection devices in scenarios such as automatic passion fruit harvesting.

Key words: passion fruit; YOLO v8s; lightweight; detection model; gather-and-distribute mechanism

收稿日期: 2024-05-06 修回日期: 2024-06-12

基金项目: 福建农林大学自主创新平台项目 (2023FX0002)

作者简介: 罗志聪 (1982—), 男, 副教授, 主要从事生物电子集成电路设计和智能传感检测系统研究, E-mail: zel@fafu.edu.cn

通信作者: 孙奇燕 (1981—), 女, 讲师, 博士, 主要从事人工智能、图像识别和边缘计算研究, E-mail: sunqiyang99168@163.com

0 引言

目前百香果采摘方式仍以人工为主,消耗大量劳动成本。农业机器人采摘对于解放劳动力以及引领水果产业走向精准化和智能化模式具有重大意义^[1]。近年来,利用图像技术对果实检测已成为研究热点^[2]。

传统的基于机器学习的图像处理和识别方法主要基于手动设计特征和分类器组合,主要对水果颜色、形状等特征进行研究。BANSAL 等^[3]通过计算水果和叶片的快速傅里叶变换泄漏值,比较水果和其他物体的泄漏百分比,得到一个阈值,开发了一种从室外彩色图像中检测未成熟的绿色柑橘类水果技术。PENG 等^[4]在提取特征后采用 SVM 分类器进行分类识别,实现多种水果识别,并提高了采摘机器人适用性。然而,传统的图像识别方法存在鲁棒性差和处理大规模数据难等问题。

相较于机器学习方法,深度学习网络可快速有效提取目标特征,有效解决数据获取难、位置不准确和采摘效率低等问题。研究发现,针对不同目标有针对性地改进模型可以得到更高的检测精度。TU 等^[5]提出一种多尺度的 MS - F RCNN 网络模型,在 Faster R - CNN 的基础上添加多尺度特征提取器,再分别训练 RGB 检测器和深度检测器,提升了对百香果小目标的检测能力。WU 等^[6]通过改进 YOLO v3 卷积网络,在其特征提取模块中结合 DenseNet 网络,将多尺度预测化为单尺度预测,实现了在自然环境中快速识别和定位不同成熟度的百香果。YANG 等^[7]在原 YOLO v5s 基础模型的主干网络和颈部网络之间加入 BCAM,能更好地构建深层次位置特征关系,能够有效提升召回率,但精确度却只有 89.15%。GAI 等^[8]通过改进 YOLO v4 骨干网络,结合 DenseNet 将模型先验框的层间密度改为适合樱桃果实形状的圆形标记框,增强了模型特征提取和检测能力。AN 等^[9]基于 YOLOX 模型,用自主设计的特征提取模块 C3HB 替换原有的 CSP 模块,并加入归一化注意力模块 NAM,设计了一种草莓生长检测算法,能有效检测草莓果实生长状态。LI 等^[10]使用 Swin - Transformer 结合 ConvNext 模块对 YOLO v7 的骨干网络进行优化,能够快速有效检测籽棉中的异物纤维。以上研究虽提高了模型检测精度,但参数量大和计算成本高,增加了嵌入式平台等移动端设备部署难度。

为解决模型参数冗余和计算复杂度高等问题,研究人员开始关注轻量化网络设计^[11]、模型剪枝^[12]、知识蒸馏^[13]等压缩模型方法。轻量化网络

设计方法是通过调整网络内部结构,设计较少的模型参数规模和更快速的卷积网络,如 MobileNet^[14]、shuffleNet^[15]、GhostNet^[16]等。SHEN 等^[17]使用通道剪枝算法对 YOLO v5s 聚类检测模型进行压缩,模型参数量和内存占用量分别减少 79% 和 76%,设计出一种轻量级的葡萄簇检测模型。LYU 等^[18]采用多教师预激活特征蒸馏框架,使用相对复杂的 YOLO v4 和 YOLO v5l 作为教师模型,相对简单的 YOLO v4 - Tiny 作为学生模型,该方法高效地提取特征,大幅提高检测精度。

本文基于改进 YOLO v8s 模型,构建复杂环境下百香果数据集,构建一个百香果检测模型,以期确保在野外复杂环境中具有较高泛化能力的同时,能够在嵌入式平台上达到实时检测要求。

1 数据集建立

图像采集地点为福建省福州市闽侯县君之誉百香果基地,图像采集设备选用尼康数码相机,采集图像时,相机与果实距离保持在 80 ~ 100 cm,图像分辨率为 1 920 像素 × 1 080 像素,保存格式为 JPEG。所有数据图像均采集于自然环境中,采用顺光、逆光、俯拍和仰拍等多种拍摄方式,涵盖果实重叠、叶片遮挡、单个果实和多个果实等多种情况,共采集 1 500 幅图像。采集图像如图 1 所示。

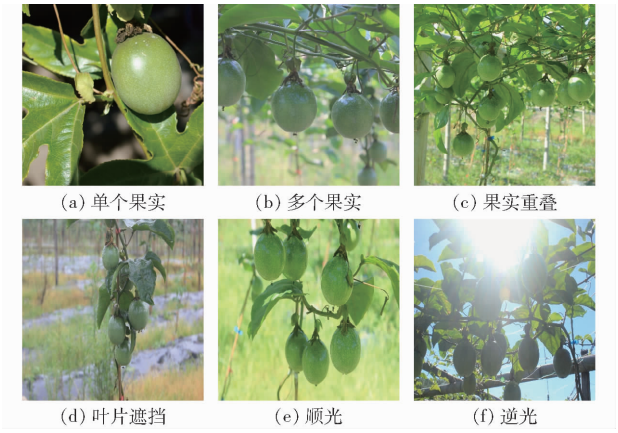


图 1 部分百香果图像

Fig. 1 Partial passion fruit images

鉴于单一百香果数据集较少,为了降低因样本不足而导致的模型过拟合现象,对原始数据进行旋转(90°、180°、270°)、添加噪声、镜像(水平镜像、垂直镜像)以及改变亮度等图像增强方式扩充数据集,强调百香果图像整体或局部特征,扩大不同物体特征之间的差别,有利于提高深层次网络训练学习能力,从而提高训练模型鲁棒性和泛化能力。增强后部分图像如图 2 所示。经过上述的离线数据扩增,最终获得百香果图像 3 000 幅。对

以上图像使用软件 Labellmg 手动标记百香果目标并绘制边界框,生成 xml 格式文件。将制作完成的数据集按照 8:1:1 随机分为训练集、验证集和测试集,最终得到训练集 2 400 幅、验证集 300 幅和测试集 300 幅。

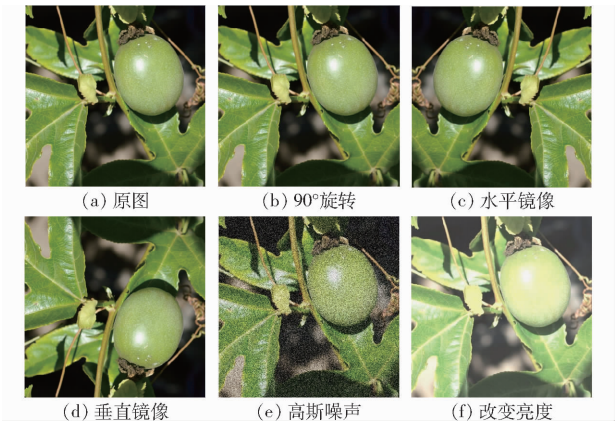


图 2 数据增强
Fig. 2 Data augmentation

2 检测模型改进

2.1 YOLO v8 模型

YOLO v8 目标检测算法于 2023 年由 Ultralytics 公司发布,具有高精度和快速推理速度,是现今性能

最好的目标检测模型之一^[19]。该模型可以分为输入端 (Input)、骨干网络 (Backbone)、颈部网络 (Neck) 和头部网络 (Head) 4 部分,主要结构如图 3 所示。输入端使用 Mosaic 数据增强方式对百香果图像随机缩放、裁剪、拼接后作为训练数据,此外,引入 YOLOX^[20]中提出的最后 10 个 epoch 关闭 Mosaic 的操作,该方式不仅丰富了数据集而且增强网络模型鲁棒性。主干网络主要负责提取图像特征,可分为 3 部分:标准卷积 (Conv) 模块、C2f 模块、SPPF 模块。Conv 模块包括卷积层、批量归一化层和激活函数层。C2f 模块主要是参考 C3 模块和 ELAN^[21]的思想进行设计,在一个维度上将输入特征图分成两部分,在减少计算量的同时丰富梯度流信息,有助于提高模型非线性表示能力。SPPF 模块连接主干网络和颈部网络,通过不同的卷积核提取百香果特征后再融合不同尺度的特征信息。颈部网络使用 PAN - FPN 结构,FPN 结构从上往下采样提取语义特征,提高模型多尺度检测精度,PAN 结构从下往上采样提取丰富的位置信息,并与高分辨率特征图拼接,增强目标定位能力。检测网络以目标概率类别、置信度分数以及目标框位置作为最终输出结果。

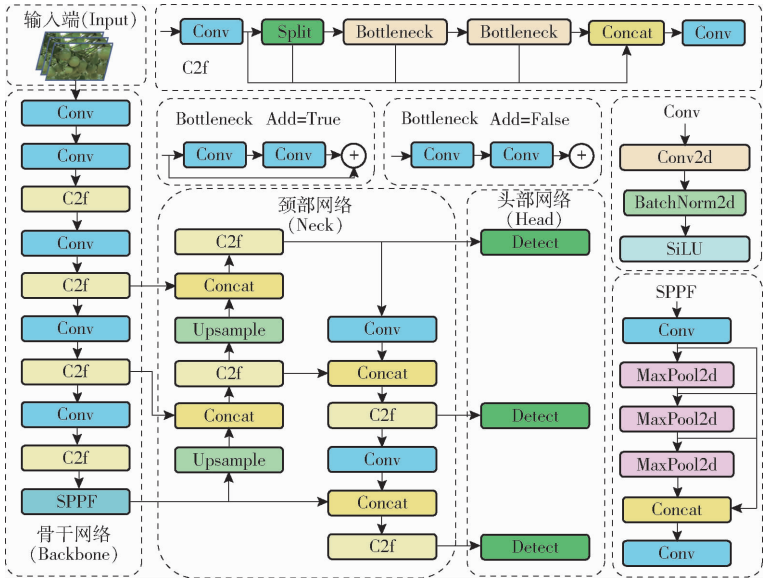


图 3 YOLO v8 网络结构
Fig. 3 YOLO v8 network structure

YOLO v8 算法具有出色的灵活性,研究人员可根据网络深度和宽度选定 YOLO v8 的不同模型。本研究使用自制百香果数据集测试 4 个版本的检测性能,结果如表 1 所示。为节约嵌入式设备内存和降低计算成本,选择 YOLO v8s 作为基准模型。

2.2 YOLO v8 模型改进

为了增加模型信息交换效率,提高其检测精度,在 YOLO v8s 模型基础上进行改进。图 4 为改进后

表 1 不同 YOLO v8 模型性能对比

Tab. 1 Comparison of performance of different YOLO v8 models

模型	精确率/ %	召回率/ %	平均精度/ %	模型内存占 用量/MB
YOLO v8s	93. 80	91. 80	94. 10	21. 40
YOLO v8m	93. 50	93. 00	95. 20	49. 60
YOLO v8l	95. 10	92. 50	95. 80	83. 50
YOLO v8x	94. 50	92. 10	95. 70	130. 00

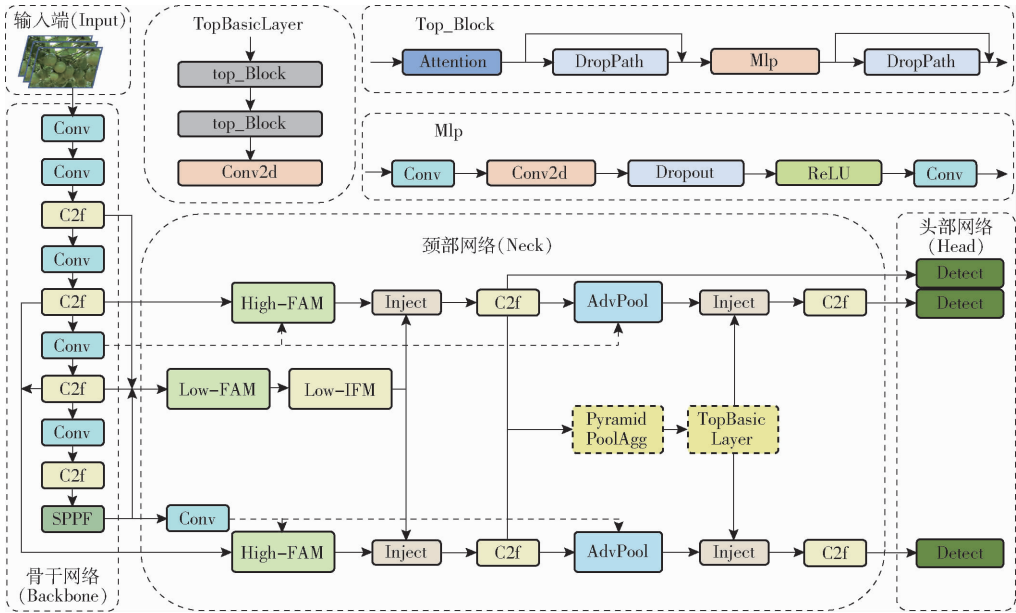


图 4 改进后 YOLO v8s - GD 模型结构图

Fig. 4 Improved YOLO v8s - GD model structure

模型结构图。

2.2.1 颈部网络重新构造

在深度学习网络中,改进模型鲁棒性和泛化能力不仅取决于修改主干网络,修改颈部网络也能产生类似效果。在 YOLO 系列模型中通常使用特征金字塔(FPN)及其变体来融合多层次特征,但这种融合方法有一个显著缺陷:当需要跨层融合信息时,传统 FPN 结构无法无损地传递信息,在一定程度上阻碍了信息融合。本研究选取聚集和分发机制^[22](Gather-and-distribute, GD)替换 YOLO v8 颈部网络, GD 机制通过全局融合多层次特征信息注入更高层,实现 YOLO v8 中高效的信息交换,能够显著增强颈部网络信息融合能力,提高模型在不同百香果尺寸下的性能。颈部网络选取的 GD 机制结构如图 5 所示。GD 机制包括浅层聚集和分发(Low -

GD)和深层聚集和分发(High - GD)两部分,分别通过卷积和自注意力机制提取和融合特征。Low - GD 主要用于融合模型浅层的特征信息,图中 B2、B3、B4 和 B5 为主干网络提取到的特征,经过 FAM 模块将特征图尺寸统一成与 B4 特征图相同的大小,再经过 IFM 模块将其按照通道数拆分成两部分,以适应不同的模型尺寸,最后经过 Inject 模块将全局特征信息注入到不同层次输出。High - GD 主要负责深层特征信息,其分辨率为 Low - GD 的 1/2,在结构上与 Low - GD 类似,但输入信息是由 Low - GD 生成的特征 P3、P4 和 P5。通过上述方案修改,低层次和高层次特征信息被有效融合。颈部网络同时具有高层次语义特征和低层次定位特征,与此同时,网络增加了大目标检测敏感度,提高了中大目标检测能力。

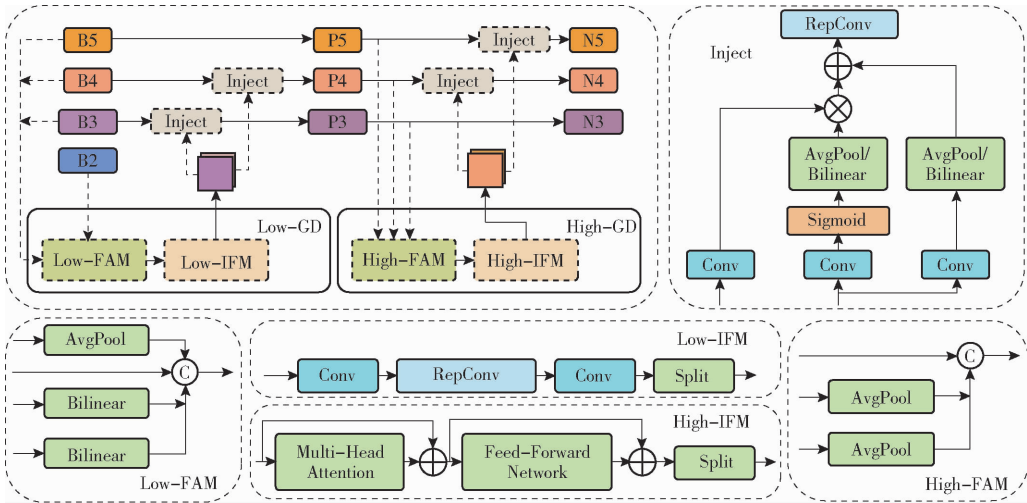


图 5 GD 机制结构图

Fig. 5 Architecture of GD mechanism

2.2.2 模型剪枝

由于在有限的存储空间和计算资源的嵌入式设备上部署神经网络相当困难,因此需要对模型计算量和参数量进一步压缩^[23]。神经网络模型从卷积层到全连接层存在过多冗余参数,神经网络剪枝即将网络结构中的冗余部分去除,这些部分移除后不会影响模型表达能力。采用基于层自适应稀疏度的幅度剪枝^[24] (Layer-adaptive sparsity for the magnitude-based pruning, LAMP) 对模型进行修剪。在剪枝过程中,判断是否去除某个通道结构的依据是这个通道的评分,在每一个连接中都有一个权重,目前评价标准是比较权重绝对值,而 LAMP 评分是权重平方,评分较小的权重被认为是冗余部分,即需修剪部分。LAMP 是一种全局剪枝,设计思路如下:

神经网络权重张量有二维矩阵,也有四维,为给出 LAMP 分数的统一定义,将所有权重张量都展开为一维向量,对于每个展开的向量根据给定索引排序, μ, ν 表示权重索引, $W[\mu]$ 和 $W[\nu]$ 表示索引 μ, ν 映射权重,假设权重根据其索引大小排序,即当 $\mu < \nu$ 时, $W[\mu] \leq W[\nu]$ 成立,即 LAMP 评分计算式为

$$score(\mu; W) := \frac{(W[\mu])^2}{\sum_{\nu \geq \mu} (W[\nu])^2} \quad (1)$$

$(W[\mu])^2$ 表示目标连接的权重平方, $\sum_{\nu \geq \mu} (W[\nu])^2$ 表示在同一层中索引大于 μ 的连接权重平方和。即分母表示当前同一层连接当中所有剩余连接的权重平方和,其他索引小于 μ 的连接已经被剪除。根据式(1)可知,权重平方越大,对应的 LAMP 评分越高,因此评分也反映连接重要程度。评分较低的权重项被认为是相对不重要的部分被剪除。在 LAMP 剪枝中设置好目标剪枝比率后,会对具有较小 LAMP 分数的连接进行全局剪枝,直到满足所需的全局稀疏性约束。而且 LAMP 评分每一层都存在一个分数为 1 的连接,意味着每一层中至少会保留一个存活连接,有效避免层崩溃。通过 LAMP 算法的应用,能够有效减少模型参数量,使其更适合在资源受限的嵌入式设备上部署。

2.2.3 知识蒸馏

知识蒸馏是进一步提高模型检测精度的有效方法^[25]。尽管目前蒸馏在 YOLO 系列网络改进中尚未被广泛使用,特别是对于单一目标的小模型。针对剪枝后模型,最终使用蒸馏技术在本模型上实现性能提升。首先,合理地选用教师模型。为了确保学生和教师模型在提取特征图的尺度一致,根据百香果数据集训练结果,选用改进后的 YOLO v8-GD

模型作为教师模型对剪枝后模型进行蒸馏。使用一种适用于密集目标检测的跨任务一致协议的蒸馏方法^[26],该蒸馏方法通过模仿教师检测器的 logit 输出来传递知识,主要由 2 个蒸馏损失值构成:二分类蒸馏损失值 L_{cls}^{dis} 和基于 IoU 的定位蒸馏损失值 L_{loc}^{dis} ,通过这两个损失值来分别传递分类知识和定位知识,具体流程如图 6 所示。

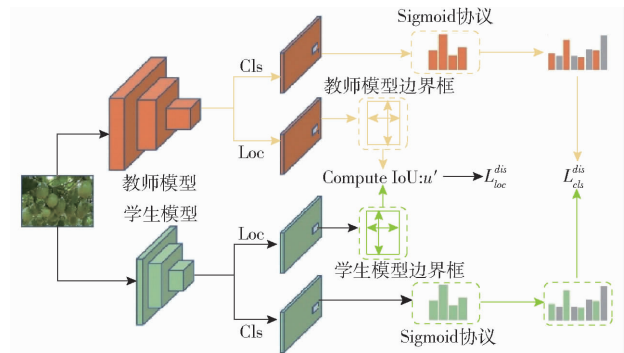


图 6 知识蒸馏流程图

Fig. 6 Knowledge distillation flowchart

数据集输入到蒸馏过程中有两个分支:①将百香果数据输入到 YOLO v8-GD 教师模型中。②将百香果数据送入到剪枝后学生模型中,经过预测后分别得到教师模型和学生模型的二元分类图。通过计算每个二元分类图的二进制交叉损失值来提取分类知识,再根据每个分类图的重要性赋予权重,得到二分类蒸馏损失值 L_{cls}^{dis} ,从而学生模型能够继承教师模型的正确预测能力。二分类蒸馏损失值计算式为

$$L_{cls}^{dis}(x) = \sum_{i=1}^n \sum_{j=1}^k w_{i,j} L_{BCE}(p'_{i,j}, p'_{i,j}) \quad (2)$$

其中

$$L_{BCE}(p'_{i,j}, p'_{i,j}) = -[(1 - p'_{i,j}) \lg(1 - p'_{i,j}) + p'_{i,j} \lg p'_{i,j}] \quad (3)$$

式中 $w_{i,j}$ ——样本 x 重要性权重

L_{BCE} ——二进制交叉熵损失值

$p'_{i,j}, p'_{i,j}$ ——二元分类分数第 i 个位置和第 j 类

在得到二元分类图的同时会得到教师模型和学生模型定位图,对于输入的百香果样本,利用锚点位置和定位预测得到每个百香果位置边界框,通过计算教师模型边界框和学生模型边界框之间的 IoU,引入损失加权策略,得到基于 IoU 的定位蒸馏损失值 L_{loc}^{dis} ,将定位知识从教师模型转移到学生模型。定位蒸馏损失值计算公式为

$$L_{loc}^{dis}(x) = \sum_{i=1}^n \max(w, j) (1 - u'_i) \quad (4)$$

式中 u'_i ——教师模型边界框和学生模型边界框之间的 IoU

本文使用的知识蒸馏算法总损失函数由二分类蒸馏损失值和基于 IoU 的定位蒸馏损失值组成,以提高分类和定位任务的能力。综合蒸馏损失值计算式为

$$L_{total}^{dis}(x) = a_1 L_{cls}^{dis}(x) + a_2 L_{loc}^{dis}(x) \tag{5}$$

式中 a_1 ——分类蒸馏损失值权重
 a_2 ——局部蒸馏损失值权重

2.3 实验环境

为了验证文中所改进算法在不同实验环境下的有效性,搭建 3 种不同的实验平台,包括计算机开发平台以及两种嵌入式测试平台。在计算机开发平台上选用 Windows 10 x64 操作系统, CPU 为 Intel® Core™ i7 - 10700F CPU 2.90 GHz, GPU 为 NVIDIA GeForce RTX 3070 8 G, 运行内存为 32 GB RAM。

嵌入式测试平台选用英伟达公司台的 Jetson Nano 和 Jetson Tx2 设备,主要用于模型推理和测试。实验环境选用 Ubuntu 18.04 系统,系统配置为 Jetpack 4.5、CUDA 10.2、cuDNN 8.0。编程语言为 Python 3.6,深度学习框架采用 Pytorch 1.8.1、Torchvision 0.9.1。测试图如图 7 所示。

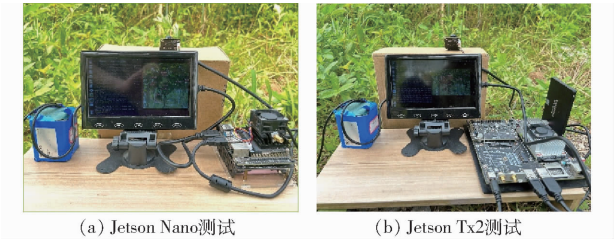


图 7 嵌入式设备测试图

Fig.7 Embedded device test chart

2.4 模型评价指标

为了评估改进后模型检测性能,采用精确率(P)、召回率(R)、平均精度(AP)、参数量(Param)、每秒浮点运算次数(FLOPs)、模型内存占用量、帧率

(FPS)作为评价指标。

在计算机开发平台训练具体参数如下:输入图像为 640 像素 × 640 像素,批量大小、初始学习率和初始动量分别设置为 16、0.001 和 0.937。测试时设置置信度为 0.5,交并比(IoU)为 0.7。优化器设置为 Adam。将训练迭代次数设置为 70,以获得最佳模型,训练 70 次后,测试集精确率、召回率和平均精度变化曲线如图 8 所示。网络训练前 10 次内,所有指标迅速上升,然后进入稳定收敛阶段。60 次后,平均精度平缓上升,曲线收敛,表明模型训练效果优。

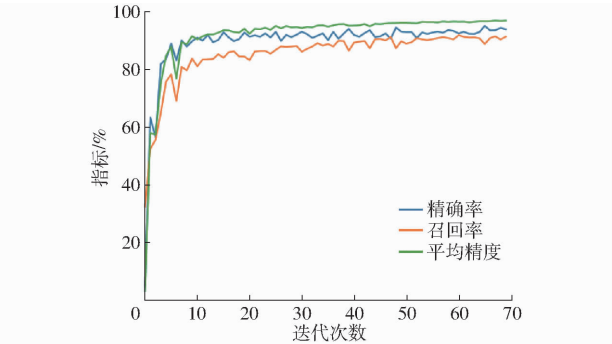


图 8 训练模型参数指标变化曲线

Fig.8 Trend of parameters of training model

3 结果与分析

3.1 不同颈部网络对比

为了选取一个特征融合效果最好的网络作为 YOLO v8s 模型的颈部网络,比较不同特征融合层和配置对颈部网络性能的影响,选取 5 种类型的颈部网络,即加权双向特征金字塔网络(BIFPN)^[27]、尺度序列特征融合网络(ASF)^[28]、通用特征金字塔网络(GFPN)^[29]、高级筛选特征金字塔网络(HSFPN)^[30]、GD 机制。为了避免偏差,保持主干网络和检测头不变,实验结果如表 2 所示。

表 2 不同颈部网络性能对比

Tab.2 Comparison of different neck networks

模型	精确率/%	召回率/%	平均精度/%	浮点计算次数/FLOPs	参数量	模型内存占用量/MB
YOLO v8 - BIFPN	94.60	91.30	95.00	2.500×10^{10}	7.360×10^6	14.30
YOLO v8 - ASF	93.70	92.70	95.60	3.010×10^{10}	1.128×10^7	21.70
YOLO v8 - GFPN	95.00	91.90	95.40	2.920×10^{10}	1.211×10^7	23.60
YOLO v8 - HSFPN	93.70	92.10	94.60	2.390×10^{10}	7.130×10^6	13.80
YOLO v8 - GD	94.90	91.60	95.90	2.990×10^{10}	1.361×10^7	26.40

相较于原始 YOLO v8s 模型,改进后模型(YOLO v8s - GD)精确率和平均精度(AP)分别提升 1.1、1.8 个百分点,表明有利于复杂环境下百香果特征提取,但是参数量和模型内存占用量分别提升 2.48×10^6 、5 MB,降低了模型检测速度,不利于在嵌入式设备上部署。相比于 YOLO v8 - BIFPN、

YOLO v8 - ASF、YOLO v8 - GFPN、YOLO v8 - HSFPN 模型平均精度提高 0.9、0.3、0.5、1.3 个百分点,说明 GD 机制在特征融合方面的有效性,保障了在后续轻量化过程中模型精度不会受到太大的影响。所提出的 5 种网络代替颈部后的结果和原网络结果有差距,表明网络有效性受到其参数量和特殊网络

结构的影响。经过替换颈部网络提升了模型平均精度。

3.2 模型剪枝实验

在模型修剪时,发现随着压缩比率的增加,模型损失精度也随之增加。表 3 为不同压缩比率下模型性能指标。当压缩比率小于 3 时,精确率、召回率和平均精度影响较小,浮点计算次数、参数量和模型内存占用量大幅减小;在压缩比率大于 3 后,召回率和平均精度大幅降低,但参数量和模型内存占用量变化不大,说明在压缩比率大于 3 后,修剪通道已经影响到模型性能,所以本研究取压缩比率为 3 的剪枝策略对模型进行修剪。

图 9 为 LAMP 剪枝策略下,按照修剪程度排序

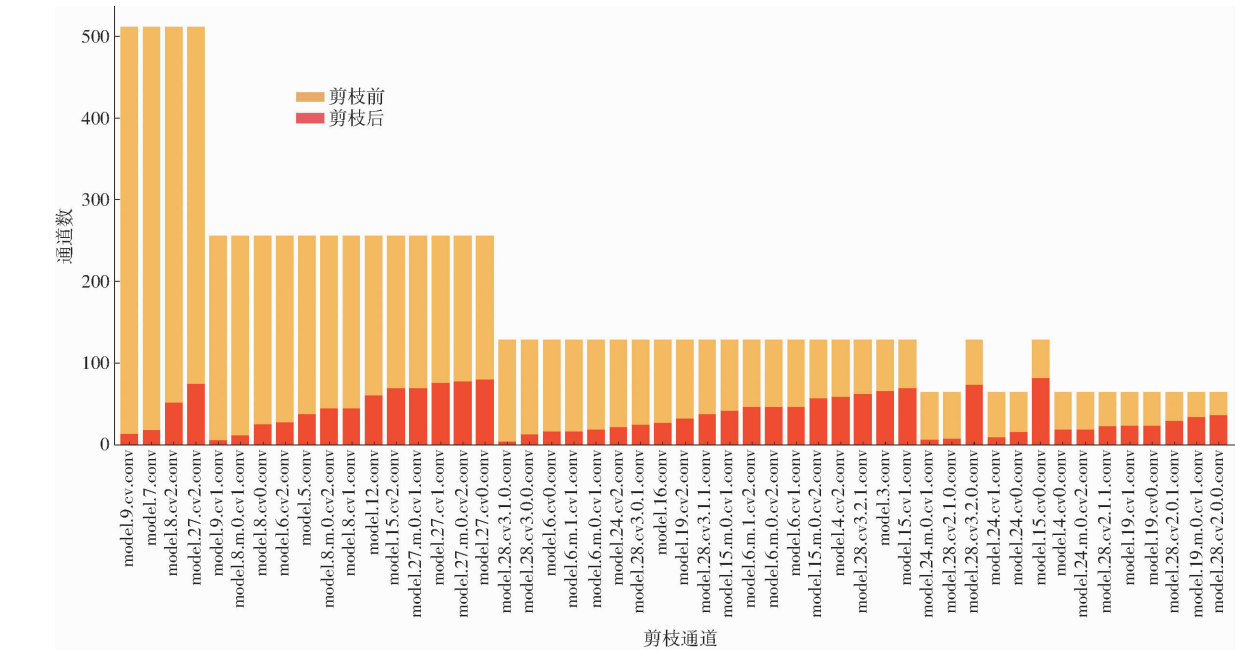


图 9 各通道修剪前后对比

Fig. 9 Comparison of each channel before and after pruning

3.3 知识蒸馏实验

在蒸馏实验中发现蒸馏权重系数对蒸馏效果有很大的影响。因此在学生模型和教师模型 (YOLO v8s-GD) 基础上探讨不同蒸馏权重系数对知识蒸馏的影响。结果如图 10 所示。从图 10 可以看出,当蒸馏权重系数为 2.0 时,蒸馏后模型具有较高精确率和平均精度,分别为 94.70% 和 96.40%。随着蒸馏权重系数的不同,精确率和召回率曲线出现上下波动,说明不同权重系数会导致模型关注百香果特征信息不同。而平均精度均值一直趋于较高的水平,说明蒸馏后模型对百香果检测性能良好。

3.4 消融实验

消融实验旨在分析每一种改进存在的价值。以自制数据集为基础进行了消融实验,对改进后模型进行测试,实验结果如表 4 所示。从表 4 可看出,在

的各通道的修剪前后情况对比,可以看出在替换颈部网络的同时,也带来了大量冗余参数。

表 3 不同压缩比率对改进模型性能的影响

Tab. 3 Impact of different compression ratios on improved model

压缩比率	精确率/%	召回率/%	平均精度/%	浮点计算次数/ FLOPs	参数量	模型内存 占用量/ MB
2.0	94.10	91.20	95.60	1.480×10^{10}	5.91×10^6	11.7
2.5	93.50	90.70	95.20	1.180×10^{10}	4.64×10^6	9.29
3.0	93.70	90.30	94.70	9.80×10^9	4.02×10^6	8.11
3.5	93.50	88.20	91.20	8.40×10^9	3.77×10^6	7.62
4.0	92.70	87.90	89.50	7.30×10^9	3.59×10^6	7.28

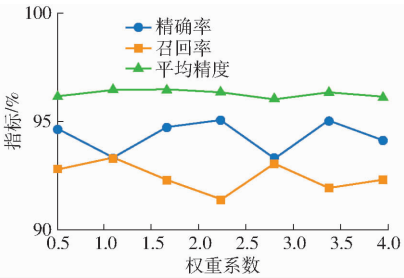


图 10 不同权重系数下模型性能

Fig. 10 Model performance with different weight coefficients

YOLO v8s 基线网络上进行更换颈部网络、轻量化剪枝、知识蒸馏增强等操作,改进后网络精度可满足检测要求。使用 GD 机制替换颈部网络更有利于百香果特征融合,模型精度相较于基线网络提升 1.8 个百分点。然而,模型参数量和模型内存占用量也随之增加,造成上述原因是改进模型提取更多的百香

果参数,不利于嵌入式模型部署。为了弥补上述缺陷,利用剪枝算法对模型冗余参数进行修剪,可以看出,模型参数量和模型内存占用量都有大幅度降低,但由于参数量减少引起精确率、召回率和模型精度有所下降。使用知识蒸馏增强的方式,教师网络给

学生网络传递丰富的百香果特征信息,大幅度提升了剪枝后模型精度。与基线网络相比,经过上述改进后模型精确率、召回率和平均精度分别提高 0.9、0.5、2.3 个百分点,模型参数量和模型内存占用量分别减小 63.88% 和 62.10%。

表 4 消融试验结果
Tab.4 Ablation experiment

基线网络	颈部网络	剪枝	知识蒸馏	精确率/%	召回率/%	平均精度/%	参数量	模型内存占用量/MB
YOLO v8s	×	×	×	93.80	91.80	94.10	1.113×10^7	21.40
	√	×	×	94.90	91.60	95.90	1.361×10^7	26.40
	√	√	×	93.70	90.30	94.70	4.020×10^6	8.11
	√	√	√	94.70	92.30	96.40	4.020×10^6	8.11

注: × 表示不使用该改进方法;√ 表示使用该改进方法。

3.5 与不同检测模型对比

为了比较改进后模型与目前主流目标检测模型性能,以浮点计算次数、参数量、帧率、精确率、召回率、平均精度和模型内存占用量为指标,将本文模型与模型 Faster R - CNN、YOLO v5s、YOLO v6s、YOLO v7、YOLO v8s、YOLO v9 和 Gold - YOLO - S 在同一自制数据集上进行对比,实验结果如表 5 所示。由表 5 可看出,Faster R - CNN 作为二阶段目标检测算法,其参数量和浮点计算次数都较大,且模型内存占用量更是比一阶段目标检测算法大,因此认为二阶段目标检测算法不太满足本数据集轻量化的实时检测需求。本研究改进后模型平均精度为 96.4%,相较于原始 YOLO v8s 提高 2.3 个百分点,相较于 Faster R - CNN、YOLO v5s、YOLO v6s、YOLO v7、YOLO v9 和 Gold - YOLO - S 分别提高 8.5、4.6、8.7、1.5、2.6、6.1 个百分点。本文模型平均精度最

优,说明对百香果检测能力强。改进后模型内存占用量和参数量分别为 8.11 MB、和 4.02×10^6 ,约为原始 YOLO v8s 模型的 1/3。从实时检测速度来看,本文模型在计算机的检测速度均优于其他 7 类模型,在 Jetson Nano 和 Jetson Tx2 嵌入式设备上的实时检测速率分别为 5.78、19.38 f/s,约为 YOLO v5s 的 1.61、1.23 倍,约为 YOLO v8s 的 1.93、1.24 倍,基本满足实时检测的要求。浮点计算次数相比于 Faster R - CNN 减少 95.91%,相比于 YOLO v5s 和 YOLO v6s 减少 39.88% 和 78.30%,相比于 YOLO v7 和 YOLO v8s 减少 90.68% 和 65.50%,相比于 YOLO v9 和 Gold - YOLO - S 减少 91.66% 和 78.71%。通过以上分析,本文模型在多个指标方面均展现出优越性,具有良好的综合性能,在移动终端上具有更好的实时性,适用于轻量级实时目标检测场景。

表 5 不同网络检测性能对比
Tab.5 Comparison on detection performance of different networks

模型	浮点计算次数/ FLOPs	参数量	Nano 帧率/ (f·s ⁻¹)	Tx2 帧率/ (f·s ⁻¹)	计算机帧率/ (f·s ⁻¹)	精确率/ %	召回率/ %	平均精度/ %	模型内存占 用量/MB
Faster R - CNN	2.394×10^{11}	4.14×10^7	0.33	1.79	5.05	88.50	77.50	87.90	321.00
YOLO v5s	1.630×10^{10}	7.05×10^6	3.60	15.72	71.42	93.90	91.90	91.80	13.70
YOLO v6s	4.517×10^{10}	1.85×10^7	2.85	13.60	71.60	91.80	87.40	87.70	38.70
YOLO v7	1.051×10^{11}	3.72×10^7	1.05	4.41	52.63	92.30	90.20	94.90	71.30
YOLO v8s	2.840×10^{10}	1.11×10^7	3.00	15.58	83.33	93.80	91.80	94.10	21.40
YOLO v9	1.175×10^{11}	3.14×10^7	0.91	4.14	49.16	94.70	89.10	93.80	60.30
Gold - YOLO - S	4.603×10^{10}	2.15×10^7	2.55	12.60	68.40	93.80	88.60	90.30	44.60
本文模型	9.800×10^9	4.02×10^6	5.78	19.38	97.20	94.70	92.30	96.40	8.11

3.6 改进前后识别效果对比

根据模型对比验证结果,选用 YOLO v5s、YOLO v8s 以及本文模型在百香果的检测中进行可视化对比,随机选择光照影响、遮挡和多目标情况下进行测试。选取置信度阈值为 0.5,IoU 阈值为 0.7,其检

测结果如图 11 所示。图中红色矩形框为预测目标框,黑色矩形框为漏检目标。根据 A 组的对比实验结果可以看出,当受光照影响较大时,3 种模型均出现漏检现象,其中 YOLO v5s 和 YOLO v8s 漏检 2 个百香果对象,而本文模型仅漏检 1 个百香果对象;根

据 B 组实验可以看出,遮挡较为严重时,本文模型检测效果优异,而 YOLO v5s 和 YOLO v8s 分别漏检 3 个和 2 个百香果对象;根据 C 组实验,在多目标检测时,可以看出本文模型可以检测到另外两个模型检测不到的小目标。从置信度方面看,本文模型置信度较高,说明改进后模型检测效果优。本文模型具有良好的鲁棒性和泛化能力。

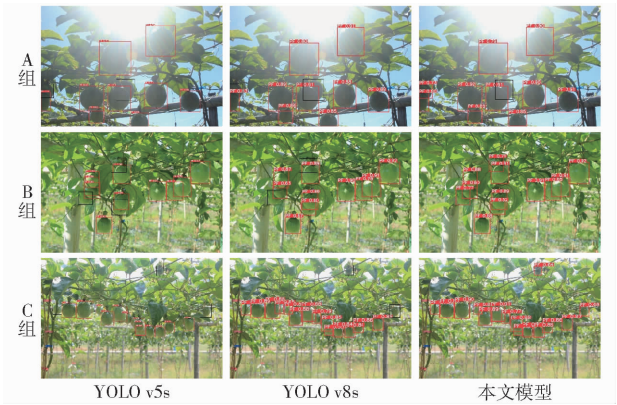


图 11 3 种模型检测效果对比

Fig. 11 Comparison of detection effects of three models

4 结论

(1) 针对复杂自然环境下轻量化百香果目标快速检测问题,在 YOLO v8s 的基础上,替换模型中的颈部网络为聚集和分发机制,提升模型特征融合能

力;采用剪枝策略对改进后模型修剪,减少模型浮点计算次数和模型参数量,同时减小模型内存占用量;使用知识蒸馏学习策略提高网络模型的检测精度。最终实现了具有低计算量、低时延性和高精度百香果快速检测模型。

(2) 实验结果表明,改进模型精确率、召回率和平均精度分别为 94.70%、92.30% 和 96.40%。与 Faster R - CNN、YOLO v5s、YOLO v6s、YOLO v7 和 YOLO v9 模型相比,在百香果数据集上该模型具有最高的平均精度和最佳的综合性能。相比较于原 YOLO v8s 基线网络模型,精确度、召回率和平均精度分别提高 0.9、0.5、2.3 个百分点,模型参数量、浮点计算次数和模型内存占用量分别下降 63.88%、65.50% 和 62.10%。并且在计算机和两种嵌入式设备上检测帧率分别提高 13.87、2.78、3.8 f/s。改进后模型平均精度等均优于对比模型,且参数量和模型内存占用量均最低。

(3) 结果表明,改进后轻量化模型检测效果优于 YOLO v5s 和 YOLO v8s,对于百香果识别效果更好,能够快速精准检测出复杂环境下和多目标场景的百香果,兼顾精度同时满足实时性需求,说明改进模型综合性能最优,具有较好的鲁棒性,可为百香果采摘识别提供支持。

参 考 文 献

[1] 鲍秀兰,马志涛,马萧杰,等. 丘陵果园自然环境下柑橘采摘机器人设计与试验[J]. 农业机械学报,2024,55(4):124 - 135.
BAO Xiulan, MA Zhitao, MA Xiaojie, et al. Design and experiment of citrus picking robot in hilly orchard natural environment [J]. Transactions of the Chinese Society for Agricultural Machinery,2024,55(4):124 - 135. (in Chinese)

[2] 张领先,景嘉平,李淑菲,等. 基于图像自动标注与改进 YOLO v5 的番茄病害识别系统[J]. 农业机械学报,2023,54(11): 198 - 207.
ZHANG Lingxian, JING Jiaping, LI Shufei, et al. Tomato disease recognition system based on image automatic labeling and improved YOLO v5[J]. Transactions of the Chinese Society for Agricultural Machinery,2023,54(11):198 - 207. (in Chinese)

[3] BANSAL R, LEE W S, SATISH S. Green citrus detection using fast Fourier transform (FFT) leakage [J]. Precision Agriculture, 2013, 14(1): 59 - 70.

[4] PENG H, SHAO Y, CHEN K, et al. Research on multi-class fruits recognition based on machine vision and SVM[J]. IFAC-PapersOnLine, 2018, 51(17): 817 - 821.

[5] TU S, PANG J, LIU H, et al. Passion fruit detection and counting based on multiple scale Faster R - CNN using RGB - D images[J]. Precision Agriculture, 2020, 21(5): 1072 - 1091.

[6] WU X, TANG R. Fast detection of passion fruit with multi-class based on YOLO v3 [C] // Proceedings of 2020 Chinese Intelligent Systems Conference: Volume II. Springer Singapore, 2021: 818 - 825.

[7] YANG R, HU Y, YAO Y, et al. Fruit target detection based on BCo - YOLO v5 model[J]. Mobile Information Systems, Hindawi, 2022, 2022: e8457173.

[8] GAI R, CHEN N, YUAN H. A detection algorithm for cherry fruits based on the improved YOLO - v4 model[J]. Neural Computing and Applications, 2023, 35(19): 13895 - 13906.

[9] AN Q, WANG K, LI Z, et al. Real-time monitoring method of strawberry fruit growth state based on YOLO improved model [J]. IEEE Access, 2022, 10: 124363 - 124372.

[10] LI Q, MA W, LI H, et al. Cotton-YOLO: improved YOLO v7 for rapid detection of foreign fibers in seed cotton[J]. Computers and Electronics in Agriculture, 2024, 219:108752.

[11] 张震,周俊,江自真,等. 基于改进 YOLO v7 轻量化模型的自然果园环境下苹果识别方法[J]. 农业机械学报,2024,

- 55(3):231–242.
- ZHANG Zhen, ZHOU Jun, JIANG Zizhen, et al. Lightweight apple recognition method in natural orchard environment based on improved YOLO v7 model[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(3):231–242. (in Chinese)
- [12] MENG F, CHENG H, LI L, et al. Pruning filter in filter[J]. Advances in Neural Information Processing Systems, 2020, 33: 17629–17640.
- [13] 黄海松, 陈星燃, 韩正功, 等. 基于多尺度注意力机制和知识蒸馏的茶叶嫩芽分级方法[J]. 农业机械学报, 2022, 53(9): 399–407, 458.
- HUANG Haisong, CHEN Xingran, HAN Zhenggong, et al. Tea buds grading method based on multiscale attention mechanism and knowledge distillation[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(9):399–407, 458. (in Chinese)
- [14] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 1314–1324.
- [15] MA N, ZHAN X, ZHENG H T, et al. Shufflenet v2: practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 116–131.
- [16] TANG Y, HAN K, GUO J, et al. GhostNetv2: enhance cheap operation with long-range attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 9969–9982.
- [17] SHEN L, SU J, HE R, et al. Real-time tracking and counting of grape clusters in the field based on channel pruning with YOLO v5s[J]. Computers and Electronics in Agriculture, 2023, 206: 107662.
- [18] LYU S, ZHAO Y, LIU X, et al. Detection of male and female litchi flowers using YOLO–HPFD multi-teacher feature distillation and FPGA-embedded platform[J]. Agronomy, 2023, 13(4): 987.
- [19] 梁秀英, 贾学镇, 何磊, 等. 基于 YOLO v8n-seg 和改进 Strongsort 的多目标小鼠跟踪方法[J]. 农业机械学报, 2024, 55(2):295–305, 345.
- LIANG Xiuying, JIA Xuezheng, HE Lei, et al. Multi-object mice tracking based on YOLO v8n-seg and improved Strongsort [J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(2):295–305, 345. (in Chinese)
- [20] GE Z, LIU S, WANG F, et al. YOLOX: exceeding YOLO series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.
- [21] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLO v7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 7464–7475.
- [22] WANG C, HE W, NIE Y, et al. Gold-YOLO: efficient object detector via gather-and-distribute mechanism[J]. Advances in Neural Information Processing Systems, 2024, 36: 51094–51112.
- [23] GUI Z, CHEN J, LI Y, et al. A lightweight tea bud detection model based on YOLO v5[J]. Computers and Electronics in Agriculture, 2023, 205: 107636.
- [24] LEE J, PARK S, MO S, et al. Layer-adaptive sparsity for the magnitude-based pruning[J]. arXiv preprint arXiv:2010.07611, 2020.
- [25] GOU J, YU B, MAYBANK S J. et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129:1789–1819.
- [26] YANG L, ZHOU X, LI X, et al. Bridging cross-task protocol inconsistency for distillation in dense object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 17175–17184.
- [27] TAN M, PANG R, LE Q V. Efficientdet: scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781–10790.
- [28] KANG M, TING C M, TING F F, et al. ASF–YOLO: a novel YOLO model with attentional scale sequence fusion for cell instance segmentation[J]. arXiv preprint arXiv:2312.06458, 2023.
- [29] XU X, JIANG Y, CHEN W, et al. Damo-YOLO: a report on real-time object detection design[J]. arXiv preprint arXiv: 2211.15444, 2022.
- [30] CHEN Y, ZHANG C, 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.