计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(1):200−210 [doi: 10.15888/j.cnki.csa.009749] [CSTR: 32024.14.csa.009749]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于轻量化 YOLOv8 模型的苹果快速识别①

聂忠强, 朱　明

(中国科学技术大学 信息科学技术学院 自动化系, 合肥 230026)
通信作者: 聂忠强, E-mail: niezhongqiang@mail.ustc.edu.cn

摘　要: 针对自然果园环境下苹果果实的识别, 本文提出了一种改进 YOLOv8n 模型的轻量化苹果检测算法. 首先, 通过使用 DSConv 和 FEM 特征提取模块的组合来替换主干网络中的部分常规卷积进行轻量化改进, 缩减卷积过程中的浮点数和计算量; 为了在轻量化过程中保持性能, 在特征处理的过程中, 引入结构化状态空间模型构建 CBAMamba 模块, 使用 Mamba 结构高效处理特征; 此后将检测头处的卷积替换为 RepConv, 并减小卷积层; 最后, 更改边界框损失函数为动态非单调聚焦机制 WIoU, 提高模型收敛速度, 进一步提升模型检测性能. 实验表明, 在公开数据集上, 本文提出的 YOLOv8 改进算法比原始 YOLOv8n 算法分别提升 1.6% 的 mAP@0.5 和 1.2% 的 mAP@0.5:0.95, 与此同时提升了 8.0% 的 FPS 并降低了 13.3% 的模型参数量, 轻量化的设计使之在机器人和嵌入式系统部署领域具有较强的实用性.

关键词: 苹果识别; 目标检测; YOLOv8; Mamba

引用格式: 聂忠强,朱明.基于轻量化 YOLOv8 模型的苹果快速识别.计算机系统应用,2025,34(1):200–210. http://www.c-s-a.org.cn/1003-3254/9749.html

Fast Apple Recognition Based on Lightweight YOLOv8 Model

NIE Zhong-Qiang, ZHU Ming

(Department of Automation, School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: This study proposes a lightweight apple detection algorithm based on an improved YOLOv8n model for apple fruit recognition in natural orchard environments. Firstly, the study uses a combination of DSConv and FEM feature extraction modules to replace some regular convolutions in the backbone network for lightweight improvements. In this way, the floating-point numbers and computational quantity during the convolution process can be reduced. To maintain performance during the lightweight process, a structured state space model is introduced to construct the CBAMamba module, which efficiently processes features through the Mamba structure, during the feature processing procedure. Subsequently, the convolutions at the detecting head are replaced with RepConv and the convolution layer is reduced. Finally, the bounding box loss function is changed to the dynamic non-monotonic focusing mechanism WIoU to accelerate model convergence and further enhance model detection performance. The experiments show that, on the public dataset, the improved YOLOv8 algorithm outperforms the original YOLOv8n algorithm by 1.6% in mAP@0.5 and 1.2% in mAP@0.5:0.95. Meanwhile, it also increases FPS by 8.0% and reduces model parameters by 13.3%. The lightweight design makes it highly practical in robotics and embedded system deployment fields.

Key words: apple recognition; object detection; YOLOv8; Mamba

① 基金项目: 科技创新特区计划 (20-163-14-LZ-001-004-01)
收稿时间: 2024-05-30; 修改时间: 2024-07-10; 采用时间: 2024-08-13; csa 在线出版时间: 2024-11-15
CNKI 网络首发时间: 2024-11-18

1 引言

苹果作为全球范围内广泛种植的水果, 不仅因其口感和营养价值而备受消费者喜爱, 也因其涉及庞大的市场和产业链而具有重要的经济意义[1]. 传统的苹果采摘方式主要依赖人工作业, 这不仅效率低下, 而且需要大量的劳动力投入, 特别是在收获季节, 劳动强度大且成本高[2]. 为了解决上述问题, 自动化采摘装备的研发和推广变得尤为重要. 这些装备能够显著减少对人力的依赖, 降低劳动成本, 同时提高采摘效率, 确保果实的品质和产量, 从而推动整个苹果产业的技术进步和经济增长[3]. 尽管自动化采摘技术不断进步, 但苹果果实的准确识别和检测仍是一个技术瓶颈. 机器人需要能够在复杂的自然环境中准确地定位和识别成熟的果实, 这一挑战限制了采摘机器人在实际果园中的应用. 果园的自然生长环境充满了变数, 如不同品种的果实大小、形状、颜色的差异, 以及枝叶遮挡、光照变化等因素, 这些都给识别任务带来了巨大的挑战[4,5].

传统的目标检测算法通常采用手工提取的特征, 这种方法在简单的控制环境中效果较好, 但在处理自然场景下动态和多变的图像时, 其性能往往受到限制, 无法提供稳定可靠的检测结果[6,7]. 随着技术的进步, 尤其是计算机性能的显著提升, 深度学习已成为图像识别领域的核心技术. 其中, 卷积神经网络因其出色的特征提取能力而在包括农业在内的多个领域展现出了卓越的性能. 这些算法通过学习大量数据中的层次特征, 能够有效地识别和分类视觉对象, 包括水果和其他农作物[8]. 例如, 通过对 Faster R-CNN 模型中的池化层和卷积层进行改进, Wan 等设计了一个适用于多种水果的检测方法, 并实现了 91% 的平均检测精度[9]. 在具体的应用场景中, 如复杂果园环境中的苹果识别, 传统的目标识别算法可能难以应对遮挡等问题. 为此, Fu 等提出了一种基于 Faster R-CNN 的方法[10], 该方法利用深度特征来过滤背景物体, 从而提高了苹果检测的准确性, 但其复杂的网络结构和对计算资源的高需求仍然是实现实时检测的障碍.

YOLO (you only look once) 系列, 因其实时性和准确性而在自然环境中的果实识别检测领域获得了广泛的应用[11,12]. Wu 等人采用了 YOLOv4 算法, 以实现对复杂环境图像中的无花果进行快速而准确的定位和识别[13]. 为了进一步提高模型在复杂果园环境中的鲁棒性和泛化性, 一些研究者尝试在检测模型中融入注意力机制. 例如, 龙燕等在 YOLOv7 的小目标检测层中加入了多头自注意力机制[14]. 这些改进提高了模型的检测精度. 尽管注意力机制能够提升模型的性能, 但它也带来了模型参数量和计算复杂度的增加. 这对于那些需要在资源受限的设备上运行的采摘机器人来说, 可能意味着更高的部署成本和实施难度.

在大部分机器人采摘系统中, 对于能够在资源有限的设备上高效运行的轻量化目标检测模型的需求日益增长. 为了解决这一需求, 研究者们专注于开发既快速又准确的网络模型, 以便实现在不牺牲过多精度的情况下提高模型的运行效率[15,16]. Wang 等[17]通过对 YOLOv5 模型进行结构优化, 用 MobileNetV3 替换原有的骨干网络, 并特别增加了针对小尺寸目标的检测层, 成功设计出了一种适用于番茄检测的轻量化模型. 这些轻量化的设计显著降低了模型的大小, 并提高了在嵌入式设备上的运行速度, 这对于实时应用至关重要. 然而, 轻量化过程往往需要在模型尺寸、计算效率与检测精度之间做出权衡. 在某些情况下, 尽管模型运行更高效, 但可能会损失一部分精度. 因此, 未来的工作需要继续探索如何平衡这 3 个方面, 确保在保持模型紧凑和快速的同时还能达到令人满意的检测精度.

现有的轻量化网络, 例如 MobileNet[18]、ShuffleNet[19,20]和 GhostNet[21], 采用了深度卷积或组卷积技术来有效提取空间特征. 这些方法在降低计算量的同时, 却可能会增加内存访问次数并降低碎片化计算的效率, 这是在设计轻量化网络时需要考虑的一个权衡点.

本文提出了一种创新的轻量化苹果检测方法, 该方法基于最新的 YOLOv8 算法进行改进. 首先, 为实现整体轻量化采用深度可分离卷积 (depthwise separable convolution) 代替原本的卷积核以降低模型参数量, 接着设计特征增强模块修饰主干网络, 使用不同尺度和数量的常规卷积和扩展卷积构造多分支结构提高网络的特征提取能力; 设计的 CBAMamba 模块引入结构化状态空间模型 (SSIM) 捕捉特征之间的依赖关系; 在输出端一方面使用 RepConv 实现模块轻量化, 后选用 WIoU (wise intersection over union) 损失函数替换原来的损失函数以提高模型的泛化能力和精确度. 实验结果证明, 该轻量化方法在提升了识别精度和处理速度的前提下, 也显著减小了模型的体积.

2 相关工作

2.1 YOLOv8 网络模型

作为一个尖端的目标检测模型, YOLOv8 不仅继承了 YOLO 系列的核心技术优势, 如速度快、准确率高等, 而且通过引入新的改进和特性, 进一步提升了模型的性能. 在模型的主干网络部分, C2f 模块取代了之前的 C3 模块. 此外, 通过去掉传统的上采样卷积步骤, 进一步简化了模型结构.

图 1 展示了标准 YOLOv8 网络的结构, 以功能作为标准可以视作由 3 个核心部分组成: 输入端 (input) 将负责接收经过增强处理的数据, 主干网络 (backbone) 负责从底层到高层逐步提取特征, 而颈部网络 (neck) 则专注于特征的深度融合和最终的目标检测任务. 输入端采用了 Mosaic 数据增强技术, 这有助于模型更好地泛化并处理各种条件下的图像, 例如不同光照、角度或遮挡情况. 主干网络基于强大的 DarkNet53 结构以实现更有效的特征提取. C2f 模块包含了多个残差瓶颈结构, 以及用于特征变换的卷积层, 这些设计有助于信息的深层传递和表达. 为了适应不同尺寸目标的检测, SPPF 模块利用多种内核大小的池化层来提取多尺度特征, 将它们融合来增强模型对于大小变化目标的适应性.

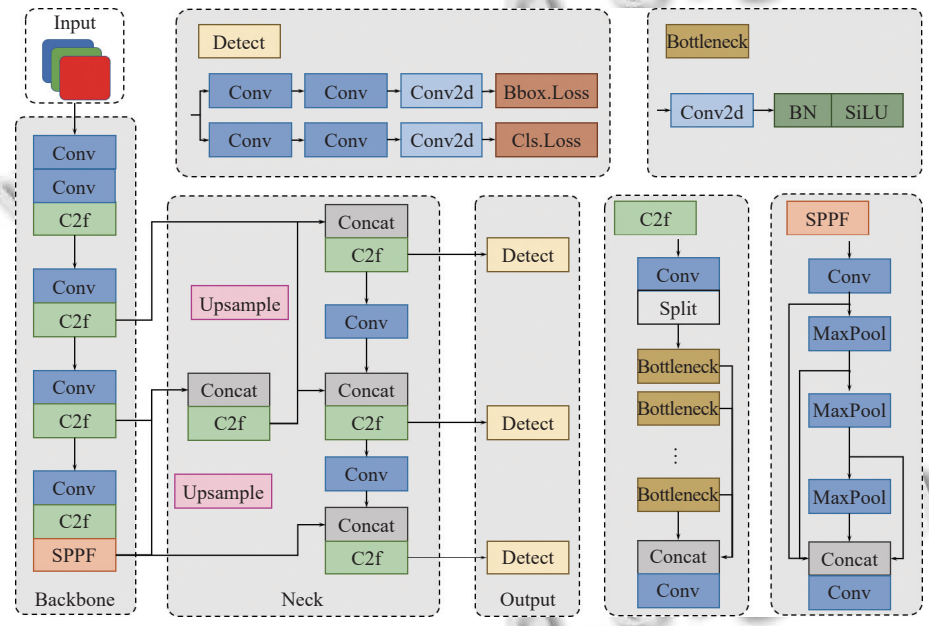


图 1 标准 YOLOv8 网络的结构

颈部网络的设计中, 路径聚合网络 (path aggregation network, PANet) 扮演了至关重要的角色, 它结合了自上而下的特征金字塔网络 (FPN) 以及自下而上的特征聚合 (PAN), 确保了丰富的上下文信息和精确的位置信息被有效地结合在一起. 最终, 预测头采用了 3 种不同尺寸的分支输出, 能够精细地进行目标的分类和定位. 这种结构不仅提高了模型对小目标的检测能力, 也使得模型能够更好地处理密集目标以及复杂背景下的目标检测问题.

2.2 Mamba 网络模型

基于 CNN 和 Transformer 的模型各有局限性. CNN 在捕获长距离信息上存在局部感受野限制, 导致在某些情况下难以有效捕获长距离信息. Transformer 在全局建模方面表现出色, 能够有效捕获长距离依赖关系, 但自注意力机制在处理图像尺寸较大时的复杂度较高. 而状态空间模型 (SSM) 通过建立远距离依赖关系并保持线性复杂度, 展现出在各种任务中的潜力. 在保持全局感受野的情况下, 状态空间模型通过 CSM 的设计替代了注意力机制, 将计算复杂度降低至线性. Mamba 的提出将特定输入参数化与可扩展的硬件优化计算方法结合起来, 在处理跨语言和基因组学的广泛序列方面实现了前所未有的效率和简单性.

S4ND 的出现标志着 SSM 块首次在视觉任务中得到应用, 成功地将视觉数据作为跨 1D、2D 和 3D 域的连续信号处理. 随后, 受 Mamba 模型成功的启发, Vmamba 和 Vim 扩展到通用视觉任务, 通过提出双向扫描和交

叉扫描机制来解决 SSM 中的方向灵敏度挑战, 也展示了视觉 Mamba 模型在解决复杂视觉挑战方面的适应性和有效性.

3 数据和方法

3.1 数据集构建

在构建苹果图像数据集时, 需要考量机器人实际从果树侧方或自下而上的视角采集图像的能力, 并确保为采摘机器人留出充裕的操作空间. 在这种条件下果实不一定能够无遮挡地显露出来, 这不利于相机捕捉到清晰的果实图像, 因此在数据集构建上, 除了清晰且主体明确的图片, 也需要有包含类似复杂场景以及不完整苹果果实的图片.

以此标准我们收集了公开数据集上的苹果图片共计 3 674 张, 在本文中选择的图像中, 部分图像能够反映复杂的环境条件, 例如不同光照情况, 不同的拍摄视角 (如图 2). 为了注释苹果图像, 使用 LabelImg 图像注释软件手动标记矩形区域. 注释文件以 XML 格式保存, 随后转换为 TXT 格式以使用 YOLOv8 算法.

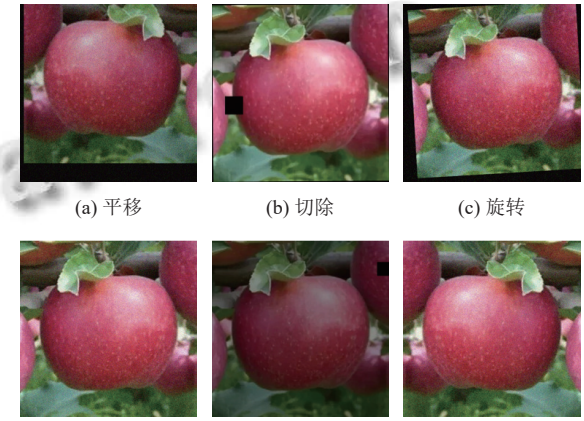
(a) 顺光 (b) 逆光 (c) 侧光

(d) 侧视 (e) 俯视 (f) 局部

图 2 数据集中的各类图片

3.1.1 数据增强算法

深度学习工作需要大量标记数据来提取特征和学习. 数据集中的噪声和错误会对模型的性能产生负面影响. 如果数据集中存在大量错误标记或不准确的数据, 则模型可能无法学习有效的模式. 当数据集有限时, 会出现过拟合, 导致网络过度关注图像中的噪声和干扰信息, 导致测试精度下降. 为此本文使用数据增强方法, 引入高斯噪声、模糊图像、随机旋转图像、切除小块、随机平移以及调整亮度等操作修改原始图像, 以增强模型的泛化能力和鲁棒性削弱过拟合的影响. 在这项研究中, 对 3 674 张分辨率为 640×640 带注释的图像进行了数据增强, 得到了 18 370 张增强图像, 将它们按比例 7:2:1 分为训练集、验证集和测试集, 图像数量分别为 12 840、3 685 和 1 845 张. 图 3 展示增强图片的示例, 每张图片进行不定次数的数据增强操作, 如图 3 中生成的图片 (e) 和 (f) 均包含两种数据增强操作.



(a) 平移 (b) 切除 (c) 旋转

(d) 随机噪声 (e) 降低亮度+切除 (f) 翻转+噪声

图 3 数据增强得到的苹果图片

3.2 主干网络改进

基于 YOLOv8n 基础模型, 本文针对问题的主要目标是实现模型的轻量化, 故选择使用深度可分离卷积替代原本的卷积块, 提高模型的训练效率和推理速度; 减少模型参数后通过设计 FEM 模块和 CBAMamba 模块来帮助网络提取更加精细的目标特征, 以克服轻量级神经网络在特征提取方面的弱点. 在检测头端将传统的卷积操作替换为重复卷积 (RepConv) 操作并减小检测头卷积层的尺寸, 减少检测过程中的浮点数运算次数, 进一步提升计算效率. 最后, 使用动态非单调聚焦机制 WIoU, 可以提高模型的收敛速度, 进一步提升模型的检测性能. 在图 4 中本文改进后的 YOLOv8n 模型的完整结构得以展示.

3.2.1 深度可分离卷积

相对于传统的卷积操作, 深度可分离卷积 (DSConv) 能够在保留较为优秀的特征提取能力的同时显著减少计算量. 它由深度卷积和逐点卷积两个步骤组成.首先进行深度卷积, 对输入特征图 X 的每个通道使用相应的卷积核独立地进行卷积, 生成相同通道数的中间特征图 Y, 可以用公式表示为:

$$Y_{i,j,k}=\sum_{m,n}X_{i+m,j+n,k}\cdot K_{m,n} \tag{1}$$

其中, i,j,k 分别是特征图的空间像素位置和通道数, m,n 是卷积核的空间位置. 事实上, 深度卷积只是单独对每个通道进行卷积运算, 但是没有设计跨通道的计算, 这样无法进行完善的特征提取, 而且它也不具备改变通道数的能力. 于是增加一个逐点卷积的操作交互通道之间的信息并调节输出特征图通道数量.

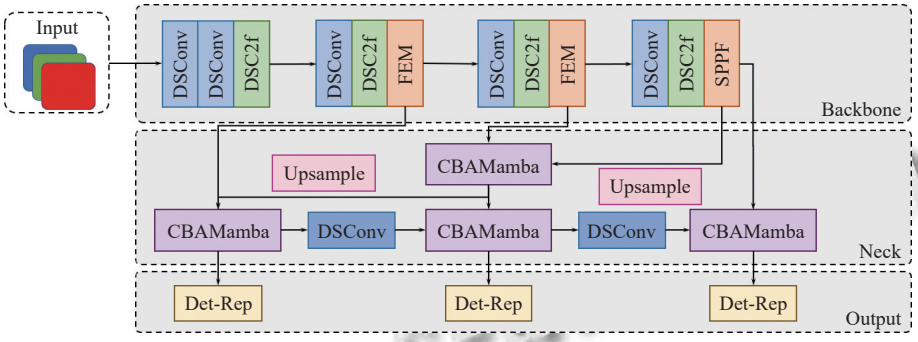


图 4 提出的改进 YOLOv8 网络框架

逐点卷积是将 1×1 卷积核应用于中间特征图的每个通道, 从而得到最终的输出特征图. 它补充了深度卷积忽略的通道之间的信息, 如式 (2):

$$Y_{i,j,k'}=\sum_{k}X_{i,j,k}\cdot K_{k,k'} \tag{2}$$

其中, k' 是逐点卷积之后的通道索引. 深度可分离卷积的主要优点是参数数量和计算复杂度的显著减少. 与普通卷积相比, 深度可分离卷积可以显著降低计算量成本, 同时保持相似的性能. 因此, 深度可分离卷积对于计算能力和内存有限的移动和边缘计算设备来说是首选. 在本文中, 将基础模型中的普通卷积全部进行替换为深度可分离卷积实现轻量化目标.

3.2.2 特征增强模块

特征图在神经网络中用于预测目标. 特征图中包含的信息量对预测输出有直接影响, 在设计追求轻量化的基础上, 如果特征图处理不充分, 会导致低层特征图中缺乏语义信息, 从而降低检测效果. 本文提出了一种特征增强模块 FEM, 它通过在多个分支上使用各种卷积和尺度来构建多分支结构以连接多通道特征图来水平扩展网络宽度, 由此提高了网络检测各类物体的适应性、灵敏度和感受野. 引入新模块帮助主干网络提取信息的原因是, 我们为了模型的轻量化对卷积操作进行了替换, 但前期的提取特征是网络功能实现的基础, 我们选择增加模块提升参数量来确保整体性能.

为了增强网络的特征提取能力, 主网络中包含有限元模块来提取全局特征, 与之前的卷积层协同工作以提高检测性能. 此外, 中间的两个分支结合了扩张的卷积层来扩大感受野并增加上下文信息, 从而提高特征的有效性. 图 5 显示了 FEM 的整体结构, 它由 4 个分支组成. 前 3 个分支执行 1×1 卷积运算来处理和调节特征图通道数以进行进一步处理. 第 4 个分支包含一个残差结构, 可在输出处生成等效图, 保留可有效检测小目标的高质量特征. 其余 3 个分支由级联的 3×3 传统卷积和扩张卷积组成, 通过各种尺度的卷积获取更细粒度的目标特征. FEM 的计算过程如下:

$$Y_1=f_{\text{dconv1}}^{3\times3}[f_{\text{conv}}^{1\times1}(X)] \tag{3}$$

$$Y_2=f_{\text{dconv3}}^{3\times3}\left\{f_{\text{conv}}^{3\times3}[f_{\text{conv}}^{1\times1}(X)]\right\} \tag{4}$$

$$Y_3=f_{\text{dconv5}}^{3\times3}\left\{f_{\text{conv}}^{3\times3}[f_{\text{conv}}^{1\times1}(X)]\right\} \tag{5}$$

$$Z=\text{Concat}(Y_1,Y_2,Y_3)\oplus X \tag{6}$$

其中, 核大小为 1×1 和 3×3 的常规卷积运算分别由 $f_{\text{conv}}^{1\times1}$ 和 $f_{\text{conv}}^{3\times3}$ 表示. 膨胀率为 1、3 和 5 的膨胀卷积的操作分别由 $f_{\text{dconv1}}^{3\times3}$、$f_{\text{dconv3}}^{3\times3}$ 和 $f_{\text{dconv5}}^{3\times3}$ 表示. Concat 表示特征图串联操作, 符号“$\oplus$”表示逐元素特征图求和操作. 此外, X 表示输入特征图, 而 Y_1、Y_2 和 Y_3 分别表示前 3 个分支在执行常规卷积和扩张卷积后获得的特征图. 最后, Z 表示增强后的特征图. 卷积操作的替换使得模型在小范围的特征提取上可能有性能下降, 如上的多尺度识别在扩充感受野的同时增强了小尺度邻近像素的关联. 通过特征增强模块处理主网络的低级特征

图. 能够有效提高在遮挡和重叠情况下的目标特征提取能力.

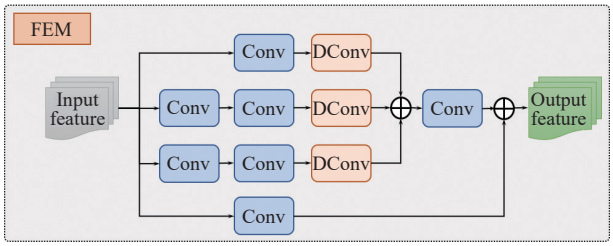


图 5 FEM 模块具体结构

3.2.3 CBAMamba 模块

使用注意力机制来为卷积网络模型提升精度是一种自然且合理的方法, 然而在关注轻量化的设计中, 注意力机制带来的大量计算负担对于本身就受限的机器人设备而言部署难度过大. 考虑到这一点, 本文引入了同样得以建立远距离依赖关系但能够保持线性复杂度的状态空间模型, 对特征图从通道和空间层面增强特征后, 由状态空间模型来选择性地关注和过滤, 这里的每一个信息提取机制均与主干网络中的卷积机制不同, 这样的结合能够有效增强模型性能.

具体实现上, 首先分解出通道注意力模块和空间注意力模块构造类似于深度可分离卷积的级联结构. 通道注意力模块对所在通道进行全局最大池化和平均池化, 提取出两个包含通道数的向量, 用一个共享全连接层学习每个通道的注意力权重进而判断出更重要的通道, 输出注意力权重与原始特征图相乘, 从而强调对识别任务有帮助的通道, 并抑制无关的通道. 空间注意力模块的操作类似, 加权的目标变更为每个空间位置的特征, 以此突出重要的图像区域. 将通道注意力模块和空间注意力模块的输出特征逐元素相乘, 得到最终的注意力增强特征. 这样的设计大幅度简化了模型提取自注意力关系所需的计算过程. 将得到的中间特征图输入 Mamba 模块, Mamba 模块代表一个复杂的神经网络模块, 包括线性投影、卷积、激活函数、自定义 S6 模块和残差连接. 该模块是 Mamba 模型的基本组成部分, 通过一系列转换处理输入序列, 以捕捉数据中的相关模式和特征. 这些不同网络层和操作函数的组合使它能够有效处理复杂的序列建模任务, 模块详细结构如图 6 所示.

3.2.4 检测头改进

RepConv 是一种基于重复卷积核的卷积操作, 它通过将输入特征图与多个不同的卷积核进行卷积, 并将结果相加来生成输出特征图. 具体的实现上, RepConv 通过重复使用相同的卷积核来对输入特征图进行卷积, 对每个输入位置只进行一次卷积操作, 从而减少了模型的参数数量和计算成本. 除此之外, 由于只使用一个卷积核进行卷积, 浮点数运算的次数得以大量减少. 这对于嵌入式设备或资源受限的环境特别重要, 考虑到这一点, 本文利用 RepConv 对检测头进行了改进. 结果如图 7 所示.

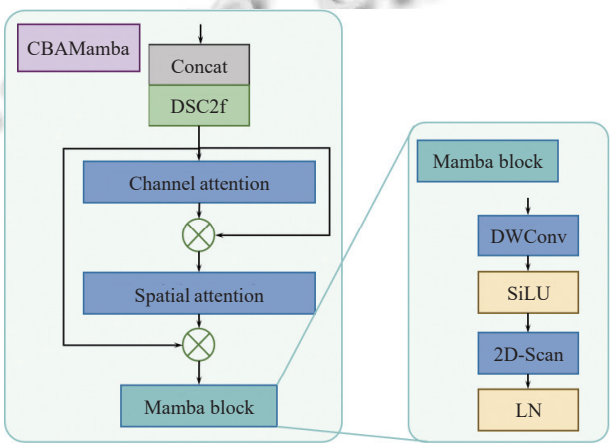


图 6 CBAMamba 模块具体结构

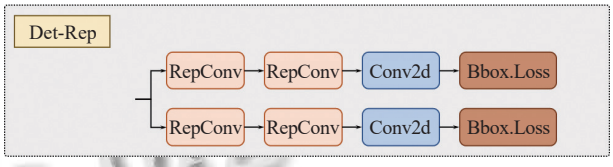


图 7 改进的 RepConv 检测头

3.2.5 损失函数改进

YOLOv8 网络采用 CIoU (complete intersection over union) Loss[22]建立预测边框坐标损失, 通过预测框和真实框的重叠面积、中心点距离和长宽比来计算损失, 其计算如式 (7) 所示:

$$L_{\mathrm{CIoU}} = 1 - IoU + \frac{\rho^2\left(B_{\mathrm{gt}}, B_{\mathrm{pred}}\right)}{C^2} + \alpha v \tag{7}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{8}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{\mathrm{pred}}}{h_{\mathrm{pred}}} - \arctan\frac{w_{\mathrm{gt}}}{h_{\mathrm{gt}}}\right)^2 \tag{9}$$

$$IoU = \frac{B_{\mathrm{gt}} \cap B_{\mathrm{pred}}}{B_{\mathrm{gt}} \cup B_{\mathrm{pred}}} \tag{10}$$

其中, B_{gt} 表示真实框, B_{pred} 表示预测框; $\rho^2\left(B_{gt},B_{pred}\right)$ 代表预测框和真实框中心点的欧氏距离; C 表示能够包含预测框和真实框的最小外接矩形的对角线长度, α 是平衡参数, v 为纵横比度量函数, 用于衡量高宽比一致性. 在目标检测领域, 准确地预测对象的边界框是至关重要的任务. 在使用的过程中发现, 尽管 CIoU 在很多情况下表现良好, 但它仍有一个缺陷: 当预测框与真实框的高宽比呈线性关系时, 其惩罚项会退化为 0, 这可能导致在这些情况下模型无法正确回归边界框. 如图 8, 粗框为真实框, 细框为预测框, 在比例一致的情况下 CIoU 损失函数发生失效.

(a) L_{CIoU}=0.76; L_{WIoU}=0.81 (b) L_{CIoU}=0.76; L_{WIoU}=0.72

图 8 两种不同预测情况下的损失函数计算结果

为了克服 CIoU 的这一限制并进一步优化模型的性能, 本文采用动态非单调聚焦机制 WIoU 进行替换. 这个损失函数采用了一种动态非单调聚焦机制, 它能够根据预测框的质量动态调整损失函数的焦点, 其损失函数如式 (11) 所示:

$$L_{WIoU}=rR_{WIoU}L_{IoU},\ R_{WIoU}\in[1,\mathrm{e}),L_{IoU}\in[0,1] \quad (11)$$

WIoU 的核心在于其两个主要组成部分: L_{WIoU} (距离聚焦机制) 和动态非单调聚焦系数 r. L_{WIoU} 负责放大普通质量预测框的损失, 而聚焦系数 r 则用于调节预测框的梯度增益, 从而在训练过程中减少低质量预测框产生的有害梯度. L_{WIoU} 定义见式 (12), r 定义如式 (13) 所示.

$$L_{WIoU}=\exp\left(\frac{\left(x-x_{gt}\right)^2+\left(y-y_{gt}\right)^2}{C^{2^*}}\right) \quad (12)$$

$$r=\frac{\beta}{\delta\alpha^{\beta-\delta}} \quad (13)$$

$$\beta=\frac{L_{IoU}^*}{\overline{L_{IoU}}}\in[0,+\infty) \quad (14)$$

其中, β 为离群度, $\overline{L_{IoU}}$ 表示动态滑动平均值, 在值大或值小时都分配较小的梯度增益, 降低对边界框回归的影响. α,δ 表示超参数. r 通过降低高质量样本对损失值贡献, 动态给予边界框梯度增益, 并在训练后期减少低质量预测框产生的有害梯度, 聚焦于普通质量的预测框, 提高模型定位能力. 通过这种方式, WIoU 成功地去除了 CIoU 中的纵横比惩罚项, 并在高质量和低质量预测框之间实现了更好的平衡. 这种改进不仅增强了模型对各种预测框的泛化能力, 还提升了模型的整体性能.

4 实验与分析

4.1 实验环境和参数设置

实验在 Ubuntu 22.04 系统环境下进行, 硬件设备为 ADM 5950X 处理器、一块 24 GB 显存的 NVIDIA GeForce RTX 3090 显卡和 32 GB 内存的计算机, 软件环境为 Python 3.8, 使用 PyTorch 1.11.0 框架. 不采用预先训练模型进行训练, 设置超参数如表 1 所示.

表 1 训练参数

训练参数	数值
初始学习率	0.01
优化器	AdamW
优化器动量	0.937
优化器权重衰减系数	0.000 5
训练批次数	32
迭代周期数	500

4.2 评价指标

为了验证模型的有效性, 选取了平均精度均值 (mean average precision, mAP)、参数量、计算量和推理速度对网络的性能进行了测试. 其中, 参数量和计算量分别用来衡量模型的空间复杂度和时间复杂度. 推理速度使用每秒检测图片数量进行衡量, 是单张图片预测时间的倒数. mAP 用以评价模型的准确性, mAP@0.5 表示 *IoU* 阈值设置为 0.5 时, 计算每一类的平均精度 AP, 后对所有类别 AP 取平均值. mAP@0.5:0.95 则是指 *IoU* 阈值从 0.5 开始到 0.95, 以 0.05 为步长逐一计算 mAP 再求取平均值.

4.3 改进方法效果对比

为探索本文提出的改进方法实际效果, 分别对 YOLOv8n 主干网络、Check 端和损失函数的选择进行实验, 对比各种改进方法的效果.

4.3.1 主干网络改进实验

为了验证主干网络替换 DSConv 并加入 FEM 模块对于 YOLOv8n 检测模型在苹果数据集上的提升效

果, 采用常用的 C2f_ScConv 对 YOLOv8 模型进行改进, 并在相同配置环境下对同一苹果数据集进行检测性能的对比试验. 试验如表 2 所示.

表 2 主干网络改进实验

模型	参数量 (M)	mAP@0.5	mAP@0.5:0.95	FPS
origin	3.0	0.883	0.682	87
C2f_ScConv	2.8	0.885	0.684	79
MnasNet	2.7	0.868	0.668	41
ShuffleNet	2.6	0.881	0.681	69
MobileNet	2.5	0.879	0.678	90
DS+FEM	2.7	0.889	0.699	89

分析对比检测结果可以得出, C2f_ScConv 模块的引入, 虽然提升了一定的检测性能, 但是在检测速度上有一定的减慢, 这点对于我们将其轻量化应用在端侧如采摘机器人上是不合理的. MnasNet 使用的弱连接搜索算法在多图像分类和目标检测任务上性能较好, 但训练和推理时间上存在的弊端不能轻易忽视. ShuffleNet 在轻量化和扩展性上均有优势, 但重排操作使得其处理小目标时性能不佳. MobileNet 实现了模型的高效性和轻量化, 但实验中也表明受到感受野和模型规模的限制, 它在处理复杂场景或近距离物体时面临困难. 而我们的方法则同时减少了参数量和提升了检测效果, 替换后的主干网络在参数量减少了 10% 的同时, 得到了比 YOLOv8n 模型分别高出 0.6% 的 mAP@0.5 和 1.7% 的 mAP@0.5:0.95.

4.3.2 Check 端改进实验

为了验证 RepConv 对于 YOLOv8n 检测模型在苹果数据集上的提升效果, 在相同配置环境下对同一苹果数据集进行检测性能的对比试验. 实验结果如表 3 所示, 更换 RepConv 测头, 同时减小检测头卷积层数, 均可提升效果. 仅减小检测头卷积层数相较于 YOLOv8 模型分别下降了 0.1% 的 mAP@0.5 和 0.3% 的 mAP@0.5:0.95, 但处理速度上升了 4.6%. 在更换 RepConv 测头, 同时减小检测头卷积层数的情况下相较于 YOLOv8 模型分别上升了 0.5% 的 mAP@0.5 和 0.1% 的 mAP@0.5:0.95, 但处理速度上升了 5.7%.

4.3.3 损失函数改进实验

改进前 YOLOv8 目标检测模型使用 CIoU 作为边界框回归损失函数, 在训练过程中拟合能力较强, 但由于预测框高宽比描述的是一个相对值, 计算时存在不确定性, 同时质量不确定的预测框对回归损失不利, 本实验将 WIoU 损失函数和常用的损失函数 SIoU、DIoU、GIoU、CIoU 进行实验对比. 由图 9 可知, 采用 WIoU 的方案时, 训练时模型的梯度下降速度在 5 个损失函数之间达到最快, 同时收敛达到稳定后的损失值相对于另外 4 种损失函数也有明显下降. 综合比对得到的数据, 采用 WIoU 作为损失函数能够帮助平均精度、帧率等评价指标均得到提升, 且能够使模型对边界框的回归更加稳定, 预测精度更高.

表 3 Check 端改进实验

模型	参数量 (M)	mAP@0.5	mAP@0.5:0.95	FPS
origin	3.0	0.883	0.682	87
light_conv	2.7	0.882	0.679	91
light_repconv	2.7	0.888	0.683	92

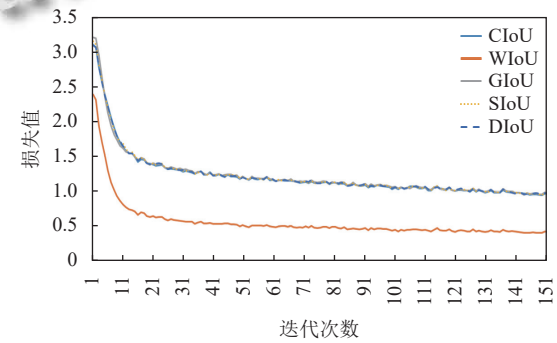


图 9 改进损失函数的对比

4.4 消融实验

为了评估改进算法的有效性, 本文设计了 6 组消融实验, 使用相同的设备和数据集进行训练和测试, 以确保可比性. 这 6 个实验包括我们提出的综合方法, 原始的 YOLOv8n 以及分别加入我们设计的各个模块后的网络模型, 实验结果如表 4 所示.

表 4 消融实验

模型	参数量 (M)	GFLOPs	mAP@0.5	mAP@0.5:0.95	FPS
YOLOv8n	3.0	8.2	0.883	0.682	87
YOLOv8n+DSConv	2.5	6.8	0.879	0.679	89
YOLOv8n+FEM	3.2	8.7	0.885	0.684	89
YOLOv8n+RepConv	2.7	7.9	0.880	0.676	91
YOLOv8n+CBAMamba	3.4	8.8	0.891	0.690	85
Ours	**2.6**	**7.6**	**0.899**	**0.694**	**94**

从表 4 可得出结论: 虽然 DSConv 可能会降低算法的检测精度和召回率, 但它可以显著降低模型的权重并提高推理速度, 有效促进后续模型的部署. 作为配合在主干网中引入 FEM 模块, 苹果的检测准确率模型分别提高了 2.46% 的 mAP@0.5 和 2.35% 的 mAP@0.5:

0.95. 将 RepConv 引入, 并缩小检测头卷积层数, 精度虽然分别下降了 0.3% 的 mAP@0.5 和 0.6%, 但 FPS 提高了 5.7%. 在再考量设计得到的 CBAMamba 层增强检测能力的效益, 苹果的检测准确率模型分别提高了 0.8% 的 mAP@0.5 和 1.2% 的 mAP@0.5:0.95. 最终改进的算法与 YOLOv8 相比较分别提高了 1.6% 的 mAP@0.5 和 1.2% 的 mAP@0.5:0.95, 同时 FPS 提高了 8.0%.

在本文中比较的 6 个网络模型中, 考量到应用需求的指标, 本文最终改进 YOLOv8n 得到的网络表现出最佳的整体检测性能. 与原始 YOLOv8n 网络相比, 改进后的网络同时在拥有最好的检测性能的同时, 参数量较小且检测速度最快, 这些都是在机器人或嵌入式系统上部署时所必要的.

4.5 不同模型的检测对比实验

为验证本文算法在苹果检测方面的优越性, 采用相同的数据集、实验条件, 将算法与主流的目标检测算法, 包括 Faster R-CNN[23]、SSD[24], 以 YOLO 系列模型 YOLOv5、YOLOv7[25]等网络模型进行测试对比.

从表 5 可得, 本文提出的算法对比 YOLOv5s、YOLOv7-tiny 和 YOLOv8n, mAP@0.5 分别提高了 9.7%、11.0% 和 1.6%, mAP@0.5:0.95 分别提高了 13.0%、17.6% 和 1.2%. 本文所提算法相比 YOLOv5s、YOLOv7-tiny 和 YOLOv8n, FPS 也分别提高了 27.1%、36.2% 和 8.0%. 最后本文改进的网络与 Faster R-CNN、SSD 网络相比, FPS 和平均精度则具有较大幅度提升. 由此可见, 改进后算法在显著提高检测精度的同时, 保持了较高的检测速率, 综合性能优于目前主流检测算法和改进后的其他算法. 本文提出的方法能在检测能力达到最优的同时满足实时检测要求. 相比之下, Faster R-CNN 和 SSD 都不能满足实时性要求, YOLOv4、YOLOv5 和 YOLOv7 的实时性相对较弱. 此外, 本文提出的改进算法参数量大小为 2.6M, 比 Faster R-CNN 和 SSD 模型都要小很多, 但在检测精度和帧率方面却表现更好.

同时, 我们进行了 YOLOv8 系列模型的多尺度实验, 实验结果如表 6 所示. 可以发现, 在原始的模型基础上增加尺度, 虽然能够一定程度上提升性能, 但是相比于高昂的参数量提升对于我们的任务来说并不合理. 而事实上, 我们的模型在使用最少参数量的基础上达到了最优性能, 这是因为任务本身并不复杂, 模型的尺度增加带来的优势并不明显, 而我们的方法在保证模型轻量化的基础上引入的设计关注于任务本身可能遇到的困难, 不仅满足了实时检测的要求, 而且提高了检测精度, 减小了模型, 具有较高的实用价值.

表 5 不同模型的检测性能对比

模型	mAP@0.5	mAP@0.5:0.95	FPS
Faster R-CNN	0.674	0.432	19
SSD	0.598	0.393	36
YOLOv5s	0.802	0.564	74
YOLOv7-tiny	0.789	0.518	69
YOLOv8n	0.883	0.682	87
Ours	**0.899**	**0.694**	**94**

表 6 多尺度模型的检测性能对比

模型	mAP@0.5	参数量 (M)	GFLOPs
YOLOv8n	0.883	3.0	8.2
YOLOv8s	0.892	11.1	28.4
YOLOv8m	0.895	25.8	78.7
Ours	**0.899**	**2.6**	**7.6**

4.6 可视化分析

为了验证改进后的检测模型对实际果园环境中苹果果实的检测效果及泛化性能, 在不同的光照条件下和多类别对象存在的情况, 以及树叶、树枝和果实重叠遮挡等等复杂场景下的苹果图像进行检测性能, 如图 10 所示.

此后我们将训练得到的网络应用到实际果园情景中, 结果如图 11 所示, 在实际场景中, 遮挡和光照条件都会干扰到采集的图像, 结果显示, 距离较远的果实存在不容易识别的情况, 但是近处的果实都正确进行了识别, 这符合实际情况的需要. 可以得到的结论是, 本文提出的方法能够有效地进行检测, 而且在各类场景的泛化性很强, 有多种使用场景的潜力.

5 结论与展望

本文提出了一种改进 YOLOv8 的苹果检测算法. 通过引入 DSConv 和 FEM 模块的组合对主干网络中的常规卷积进行轻量化改进, 缩减卷积过程中的浮点数和计算量; 其次加入 CBAMamba 模块引入状态空间模型对提取的特征进行整理; 此后将检测头中的 Conv 替换为 RepConv, 并减小检测头卷积层, 缩减检测过程中的浮点数和计算量; 最后, 更改边界框损失函数为动态非单调聚焦机制 WIoU, 提高模型收敛速度, 进一步提升模型检测性能. 这样的设计在维持性能的同时实

现了有效轻量化, 有利于后续在嵌入式系统和机器人上的部署. 本研究提出的改进框架能够有效泛化到其他目标检测任务, 下一步尝试设计网络使其能够在恶劣场景下的识别率得到提升.

不同光照条件的验证结果　　遮挡和多种类对象的验证结果

其他场景的验证结果

图 10　不同场景下的检测结果

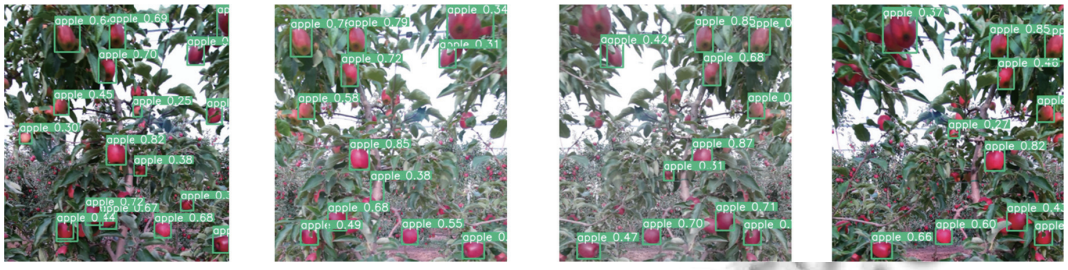


图 11　实际场景中的检测结果

参考文献

1 王丹丹, 何东健. 基于 R-FCN 深度卷积神经网络的机器人疏果前苹果目标的识别. 农业工程学报, 2019, 35(3): 156–163. [doi: 10.11975/j.issn.1002-6819.2019.03.020]

2 Chu PY, Li ZJ, Lammers K, *et al*. Deep learning-based apple detection using a suppression mask R-CNN. Pattern Recognition Letters, 2021, 147: 206–211. [doi: 10.1016/j.patrec.2021.04.022]

3 Baeten J, Donné K, Boedrij S, *et al*. Autonomous fruit picking machine: A robotic apple harvester. In: Laugier C, Siegwart R, eds. Field and Service Robotics. Berlin, Heidelberg: Springer, 2008. 531–539.

4 张毅, 郝骞. 苹果采收机器人结构设计与仿真分析. 机械设计与制造, 2022(2): 291–294, 299. [doi: 10.3969/j.issn.1001-3997.2022.02.062]

5 Zhuang JJ, Luo SM, Hou CJ, *et al*. Detection of orchard citrus fruits using a monocular machine vision-based method for automatic fruit picking applications. Computers and Electronics in Agriculture, 2018, 152: 64–73. [doi: 10.1016/j.compag.2018.07.004]

6 Lin GC, Tang YC, Zou XJ, *et al*. Color-, depth-, and shape-based 3D fruit detection. Precision Agriculture, 2020, 21(1): 1–17. [doi: 10.1007/s11119-019-09654-w]

7 Lin GC, Tang YC, Zou XJ, *et al*. In-field citrus detection and localisation based on RGB-D image analysis. Biosystems Engineering, 2019, 186: 34–44. [doi: 10.1016/j.biosystemseng.2019.06.019]

8 Tang YC, Chen MY, Wang CL, *et al*. Recognition and localization methods for vision-based fruit picking robots: A review. Frontiers in Plant Science, 2020, 11: 510. [doi: 10.3389/fpls.2020.00510]

9 Wan SH, Goudos S. Faster R-CNN for multi-class fruit

detection using a robotic vision system. Computer Networks, 2020, 168: 107036. [doi: 10.1016/j.comnet.2019.107036]

10 Fu LS, Majeed Y, Zhang X, *et al*. Faster R-CNN-based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. Biosystems Engineering, 2020, 197: 245–256. [doi: 10.1016/j.biosystemseng.2020.07.007]

11 孙丰刚, 王云露, 兰鹏, 等. 基于改进 YOLOv5s 和迁移学习的苹果果实病害识别方法. 农业工程学报, 2022, 38(11): 171–179. [doi: 10.11975/j.issn.1002-6819.2022.11.019]

12 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

13 Wu YJ, Yang Y, Wang XF, *et al*. Fig fruit recognition method based on YOLOv4 deep learning. Proceedings of the 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). Chiang Ma: IEEE, 2021. 303–306.

14 龙燕, 杨智优, 何梦菲. 基于改进 YOLOv7 的疏果期苹果目标检测方法. 农业工程学报, 2023, 39(14): 191–199. [doi: 10.11975/j.issn.1002-6819.202305069]

15 黄杰, 王相友, 吴海涛, 等. 基于轻量型卷积神经网络的马铃薯种薯芽眼检测算法. 农业工程学报, 2023, 39(9): 172–182. [doi: 10.11975/j.issn.1002-6819.202303035]

16 Bazame HC, Molin JP, Althoff D, *et al*. Detection of coffee fruits on tree branches using computer vision. Scientia Agricola, 2022, 80: e20220064.

17 Wang XF, Wu ZW, Jia M, *et al*. Lightweight SM-YOLOv5 tomato fruit detection algorithm for plant factory. Sensors, 2023, 23(6): 3336. [doi: 10.3390/s23063336]

18 Sinha D, El-Sharkawy M. Thin MobileNet: An enhanced MobileNet architecture. Proceedings of the 10th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). New York: IEEE, 2019. 280–285.

19 Ma NN, Zhang XY, Zheng HT, *et al*. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 122–138.

20 Zhang XY, Zhou XY, Lin MX, *et al*. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6848–6856.

21 Han K, Wang YH, Tian Q, *et al*. GhostNet: More features from cheap operations. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1577–1586.

22 Zheng ZH, Wang P, Liu W, *et al*. Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the 38 AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2020. 12993–13000.

23 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

24 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

25 Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 7464–7475.

(校对责编: 王欣欣)