DOI: 10.14083/j.issn.1001-4942.2025.10.019

面向边缘设备的改进 YOLOv7-tiny 线虫检测模型

李耀东[1],侯文进[1],侯华鑫[1],王秀丽[1],王东[2],曲建平[3],周波[3,4],刘璋[3]

(1. 山东农业大学信息科学与工程学院,山东 泰安　271018;2. 西北农林科技大学农学院,陕西 杨凌　712100;
3. 山东农业大学生命科学学院,山东 泰安　271018;4. 山东未来生物科技有限公司,山东 泰安　271000)

摘要:线虫是一种广泛用于生物学研究的模式生物,本研究针对在线虫活性筛选阶段存在个体目标小、易被遮挡以及现有线虫检测模型轻量化性能差、不易在边缘设备部署等问题,提出了一种面向边缘设备的改进 YOLOv7-tiny 线虫检测模型。采用 MobileOne 网络作为骨干网络,提高模型计算效率;引入广义特征金字塔网络(GFPN)改进 Neck 层,实现“跳层”与“跨尺度”的自适应融合,从而提供更丰富的图像特征信息;在 Neck 层引入双层路由注意力机制(BRA),加强对遮挡目标的特征提取能力;在 Head 层增加第四检测头,提高对小目标的检测能力;利用感知量化方法对模型进行 INT8 量化处理,并对激活值部分采用非对称量化策略,以降低计算量并实现模型轻量化。将改进后的模型部署在边缘设备 Jetson Nano 上进行测试,结果表明,改进模型与原模型相比平均精度均值(mAP@0.5)提高了 2.7 个百分点,计算量(GFLOPs)压缩了 67.71%,检测帧率(FPS)提高了 23.01%。可见,改进后的模型精度有明显提升,可在边缘设备上实现快速、精准检测线虫目标。

关键词:边缘设备;线虫检测;YOLOv7-tiny;轻量化

中图分类号:S126:S154.386;TP391.41　　**文献标识号**:A　　**文章编号**:1001-4942(2025)10-0149-09

Improved YOLOv7-tiny Nematode Detection Model for Edge Devices

Li Yaodong[1], Hou Wenjin[1], Hou Huaxin[1], Wang Xiuli[1], Wang Dong[2], Qu Jianping[3], Zhou Bo[3,4], Liu Zhang[3]

(1. *College of Information Science and Engineering, Shandong Agricultural University, Taian* 271018, *China*;
2. *College of Agriculture, Northwest A&F University, Yangling* 712100, *China*;
3. *College of Life Sciences, Shandong Agricultural University, Taian* 271018, *China*;
4. *Shandong Future Biotechnology Co., Ltd., Taian* 271000, *China*)

Abstract　Nematodes are widely used as model organisms in biological researches. To address the challenges during nematode activity screening stage, such as the small size of individual nematode target, easy to be obscured, and the poor lightweight performance and difficult to deploy on edge devices of existing nematode detection models, we proposed an improved YOLOv7-tiny nematode detection model tailored for edge devices. The MobileOne network was employed as the backbone network to boost the model's computational efficiency. The Generalized Feature Pyramid Network (GFPN) was incorporated to refine the Neck layer to enable adaptive fusion of "skip-layer" and "cross-scale" approaches, thereby enriching the representation of image features. Additionally, a dual-layer routing attention mechanism (BRA) was introduced into the Neck layer to enhance the feature extraction capability for obscured targets. The fourth detection head was added into the Head layer to enhance the detection capability for small targets. The INT8 quantization processing was adopted for the model using the perceptual quantization method, with an asymmetric quantization strategy applied to the

收稿日期:2024-06-16
基金项目:山东省重大科技创新工程项目(2019JZZY010716);山东农业大学横向科研项目(233024);山东省大学生创新创业训练计划项目(S202310434217)
作者简介:李耀东(2002—),男,山东济宁人,本科生,研究方向为农业信息化。E-mail:1746544789@qq.com
通信作者:王秀丽(1981—),女,山东泰安人,硕士,讲师,研究方向为农业信息化。E-mail:wxlmail@sdau.edu.cn

中国知网　https://www.cnki.net

activation values to further reduce computational load and achieve model lightweighting. The improved model was deployed and tested on the edge device Jetson Nano. The experimental results indicated that compared to the original model, the improved model showed an increase in mean average precision (mAP@0.5) by 2.7 percentage points, a reduction in computational demand (GFLOPs) by 67.71%, and an increase in detection frame rate (FPS) by 23.01%. These results demonstrated that the accuracy of the improved model was significantly enhanced, and it could be enabled rapid and precise detection of nematode targets on edge devices.

Keywords Edge device; Nematode detection; YOLOv7-tiny; Lightweight

线虫是一种广泛用于生物学研究的模式生物[1],具有繁殖快、基因序列已知等特点,在生化制药、毒理研究等领域发挥着重要作用,而线虫活性的快速检测则有利于推动线虫生物学研究的进程。传统的线虫检测方法仍停留在单显微镜人工肉眼观察阶段,该方法费力、耗时,而且存在主观性强和误差大等缺点[2]。近年来,随着人工智能的发展,深度学习方法提高了图像识别的精度和效率,克服了人工提取特征的众多弊端[3]。学者们对于深度学习方法在农业生物检测方面的应用已有大量研究。黄丽明等[4]利用改进 YOLOv4 模型识别遥感图像上因松材线虫病导致的异色木,识别平均精度(AP)为 80.85%。Liu 等[5]利用加入特征金字塔的 YOLOv3 模型对番茄病害进行识别,平均检测精度达到 92.39%。刘金涛等[6]基于 YOLOv5s 模型,通过引入 CARAFE 上采样模块和 GAM 注意力机制进行改进,实现了对咖啡叶病虫害的检测,平均精度均值(mAP)达 91.4%。赵鹏飞等[7]以甜椒为研究对象,基于 YOLOv7-Tiny 模型并在骨干网络添加 DBB 模块进行改进,mAP 达 93.95%。以上农业生物检测模型,虽然在检测精度上较传统模型有所提高,但对硬件算力过于依赖,不易在边缘移动设备部署,限制了其实际应用与推广。

边缘设备是分布在网络边缘或靠近数据源,用于分析处理数据的计算设备。边缘设备的主要特点是能够在本地执行计算、存储和网络通信功能,以便更快地响应用户请求或处理数据,同时可以减轻中心服务器的负载。边缘设备的应用场合涵盖了各个领域的实时数据处理、安全与隐私保护和服务优化等需求[8]。随着线虫研究的不断发展,筛选线虫的数量会持续增长,将数据处理任务分布式部署至显微镜设备侧,对于降低数据中心的处理压力、减少能源消耗、保障数据安全、实现实时检测等至关重要。

受线虫背景环境和个体表征因素影响,线虫目标检测任务需综合考虑检测精度与速度,而 one-stage 网络能够在检测中同时输出类别的锚框与概率,因此更加适合该任务场景。在现有的多种 one-stage 检测框架中,YOLOv7-tiny 能以轻量化结构实现更高的检测精度与速度,适合部署在资源受限的边缘设备上[9]。因此,为了实现检测精度与轻量化的平衡,本研究基于 YOLOv7-tiny 模型进行优化,并将改进后的模型经量化处理后部署在 Jetson Nano 边缘设备进行测试,检验改进模型在边缘端完成线虫检测任务的能力,以期为满足不同线虫生物学研究需求提供技术支持。

1 线虫图像数据集构建

1.1 图像数据采集

数据采集地为山东农业大学作物生物学国家重点实验室,采用光学显微镜、萨伽 500 万像素电子目镜设备,拍摄含有活体线虫的溶液液滴,为保持线虫的活性,每滴含有线虫的溶液液滴拍摄 2 min。由于活虫蜷曲扭动,拍摄的图像易出现线虫间遮挡现象,图 1 中分别为同等环境下非遮挡、遮挡的线虫样本图像。基于体态学判断线虫死活,如图 2 所示,虫体僵直不动、呈“J”状为死虫,虫体形态不定、扭动蜷曲为活虫[10]。

1.2 数据预处理

本研究使用标注软件 LabelImg,按照活虫(live)和死虫(dead)两个类别对数据集进行标注。原始图像为 2 402 张,通过添加不同类型的噪声、图像属性调整(调整图像的亮度、对比度和

中国知网 https://www.cnki.net

饱和度）、中心裁剪、水平和垂直随机翻转、按比例缩放和平移以及仿射变换的方法扩充至 14 550 张，分辨率维持原尺寸 1 280×960 像素。随后将数据集按照 8∶1∶1 划分为训练集、测试集、验证集，数量分别为 11 640、1 455、1 455 张。

图 1　非遮挡和遮挡线虫图像示例

图 2　死虫和活虫图像示例

2　改进 YOLOv7-tiny 目标检测算法

2.1　YOLOv7-tiny 模型

YOLOv7-tiny 模型是 YOLOv7 模型的精简版，属于轻量型检测模型，参数量仅 600 万左右，相当于 YOLOv7 参数量的 1/6，这使其检测速度较 YOLOv7 有了很大提升[11]。YOLOv7-tiny 的网络结构与以往版本的 YOLO 结构相似，如图 3 所示，主要包含 4 部分，分别为输入端（Input）、骨干网络（Backbone）、颈部（Neck）以及检测头（Head）。在输入端，主要对输入的图像进行数据增强、通道顺序调整和自适应图片缩放等预处理，然后将其输入至骨干网络进行图像特征提取。在骨干网络，使用简化后的 ELAN（Efficient Layer Aggregation Networks，高效层聚合网络）替代扩展的 ELAN（E-ELAN），仅采用池化进行下采样，保留优化后的 SPP（空间金字塔池化）结构，为 Neck 层提供更丰富的特征图。在颈部，采用 PANet（路径聚合网络）结构将来自不同层级的特征图融合，以提升目标检测性能。在检测头部分，YOLOv7-tiny 采用 Idetect 检测头，以隐式表示策略来预测结果[12]。相比于 YOLOv7，YOLOv7-tiny 牺牲了一定的精度，但在轻量化方面具有明显优势，适用于部署在边缘移动设备。

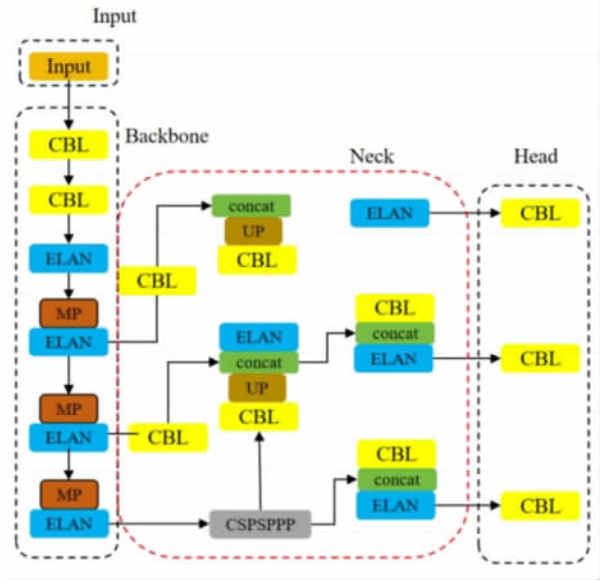

图 3　YOLOv7-tiny 网络结构

2.2　融入 MobileOne 轻量化骨干网络

为了便于线虫检测模型部署在边缘移动设备，在保持模型精度的情况下降低设备延迟，本研究将原 YOLOv7-tiny 的骨干网络替换成轻量级的 MobileOne 网络[13]。MobileOne 是一款由 Apple 公司提出的面向移动设备的新型轻量化骨干网络，旨在解决大量计算导致的预测延迟和精度下降等问题。其变体在移动设备上的推理时间低于 1 ms，推理时间和准确度相较于目前主流的轻量化神经网络 MobileNetv3 和 ShuffleNetv2 有一定优势[14]。MobileOne 网络可应用于图像分类、语义分割和目标检测任务，在保持模型精度的同时显著降低延迟，适合边缘端设备部署。

MobileOne 网络作为轻量化骨干网络，由多个 MobileOne Block 模块构成。这些 MobileOne Block 模块借鉴了极简架构 RepVGG 的结构重参数化思想[15]，由深度可分离卷积模块和点卷积模块组成。MobileOne Block 结构如图 4 所示，左侧为训练状态，右侧为推理状态。训练状态为多分支结构，其中深度可分离卷积模块有 3 个分支，分别为 1×1 卷积、3×3 卷积和一个 BN 层；点卷积模块有两个分支，分别为 1×1 卷积和一个 BN 层。推理状态为重参数化后的线性结构，深度可分离

中国知网　https://www.cnki.net

卷积模块和点卷积模块分别只有一个3×3深度可分离卷积和一个1×1点卷积，没有其他额外分支。在模型训练状态，通过增加模型复杂度提取更多有效的语义特征。在模型推理状态，将多分支结构变为单分支结构，不仅降低了计算量，还减少了concat等无参操作的使用，从而降低操作内存，使其能够在计算资源受限的边缘设备中高效运行。

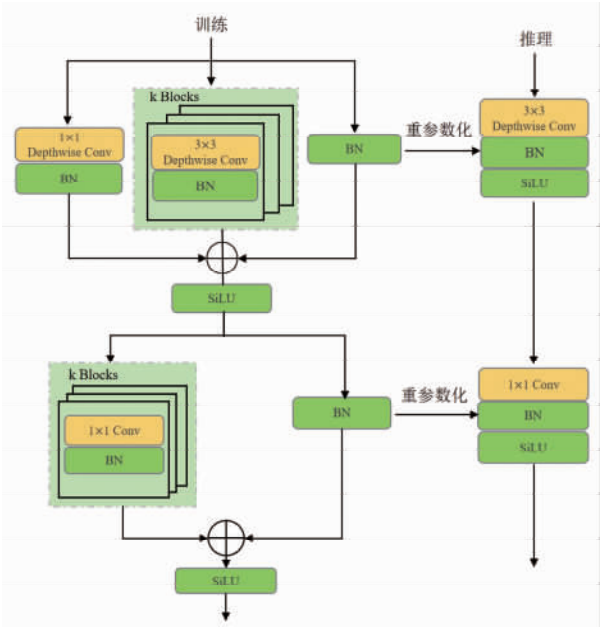

图4 MobileOne Block结构

2.3 引入GFPN特征融合网络

GFPN(Generalized-FPN)[16]是一种新颖的类“长颈鹿”的GiraffeDet架构，其基本结构如图5所示。传统的特征融合网络FPN、BiFPN等只关注于特征融合，忽略网络内部连接，而GFPN则有效地解决了这个问题，设计出一种新的路径聚合策略——跳层与跨尺度连接[17]。

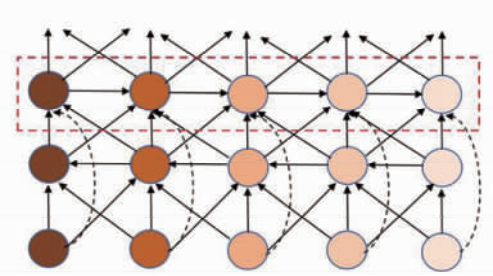

图5 GFPN基本结构

跳层连接提供了更为有效的信息传递和特征复用机制，并且可促使网络结构实现更深层次的拓展。图6为两种跳层连接的示意图。借鉴DenseNet(密集连接卷积网络)的思想，在每一层中，GFPN设计了Dense跳层连接方式，以促进更多的特征复用。但是这种方法给网络带来了巨大的负担，随着网络的深入，参数量将会增大，导致梯度消失。因此，GFPN又提出了一种名为$\log_2 n$的跳层连接方法。相对于Dense跳层连接方式，$\log_2 n$的跳层连接方式复杂度更低，在反向传播过程中，层与层之间的距离也会减少。

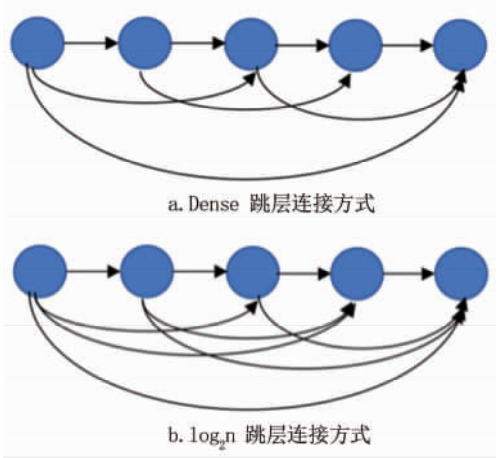

图6 GFPN的两种跳层连接方式示意图

跨尺度连接有助于提供更丰富的特征信息，能够克服目标尺度变化的问题。如图7所示，每个节点不仅接收上一个节点的输入，还同时接收斜上方和斜下方节点的输入。为了避免在特征融合过程中使用sum方法导致信息丢失的问题，采用concat方法进行特征融合。同时，采用双线性插值(Bi-Linear)与最大值池化(Max-Pooling)方法，实现特征的上采样和下采样。

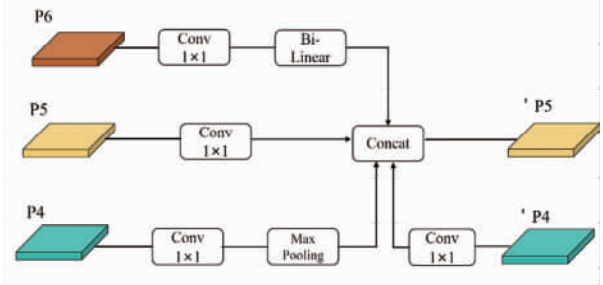

图7 跨尺度连接

通过以上论述可知，GFPN模块通过跳层和跨尺度连接使得高层语义信息能够更充分地与低层空间信息进行交互，具备更出色的图像处理能力。但是通过上采样和下采样实现的特征融合

中国知网 https://www.cnki.net

中，采样操作是较为粗略的下采样过程，因此会造成特征信息的损失。

2.4　引入 BRA 注意力机制

由于在线虫检测任务中存在线虫目标小、密集重叠导致目标被遮挡等挑战，同时为进一步减少 GFPN 结构造成的特征信息损失，本研究引入 BRA(Bi-Level Routing Attention) 注意力机制，使模型能够将注意力聚焦在重要位置，以获取物体的关键信息。BRA 注意力机制的主要优势在于其动态稀疏的特性[18]，它引入了两个层级的注意力机制来促进任务之间的关联和信息交互。首先，在粗糙区域级别上，它能过滤掉大部分不相关的信息交互，只保留少量路由区域，增强了有效信息之间的交互，更精准地关注重要目标。其次，在路由区域中，采用了细粒度的 token-to-token 注意力，通过相关联任务之间的深度交互，获取更多有效特征信息。相较于传统的注意力机制，BRA 能够更灵活地根据输入图像内容动态调整注意力的分布，从而更好地适应不同尺度和相互遮挡的目标，提高了对小目标特征的准确捕捉能力[19]。

BRA 注意力机制的作用过程如图 8 所示，其中，W、H、C 分别表示输入特征图的宽度、长度和通道数，S 为划分的每个区域的长宽，Q、K、V 分别表示查询、键(key) 和值(value) 向量。BRA 首先将一张输入特征图划分为 S×S 个不重叠区域，每个区域包含 HW/S^2 个特征向量。由线性映射可得 $Q,K,V \in R^{S^2\times HW/S^2\times C}$，线性预测如公式(1) :

$$Q=X^rW^q,\ K=X^rW^k,\ V=X^rW^v。\quad (1)$$

其中，$W^q,W^k,W^v \in R^{C\times C}$，分别为 Q、K、V 的投影权重。然后构造一个有向图确定不同键值对应的参与关系，最后采用细粒度的 token-to-token 注意力操作，计算公式如下:

$$O=\mathrm{Attention}(Q,K^g,V^g)+\mathrm{LCE}(V)\ 。\quad (2)$$

其中，K^g 和 V^g 是聚合后 key 和 value 的 tensor，函数 LCE(·) 使用深度卷积参数化。BRA 通过动态稀疏性操作舍去最不相关区域的计算，使模型关注重要特征，自适应地分配注意力权重，提升学习效果和泛化性能。

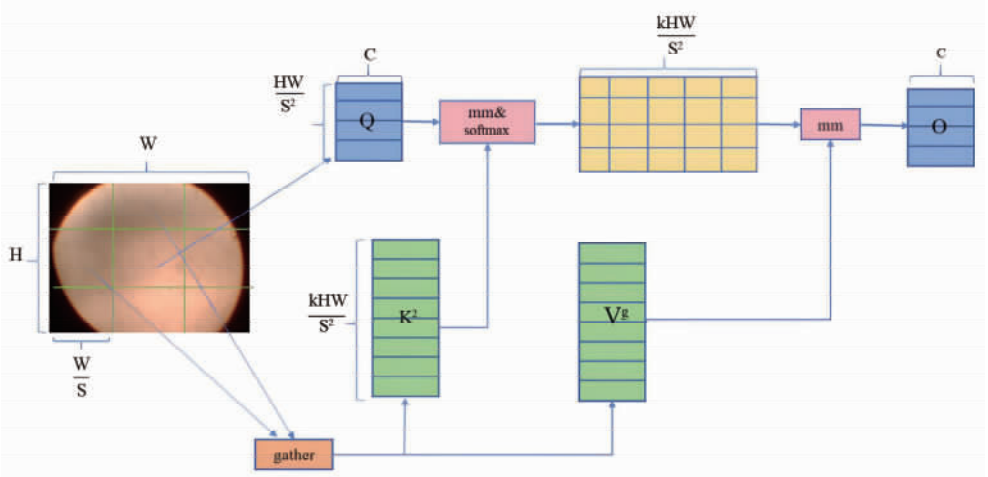

图 8　BRA 注意力机制作用过程

2.5　增加第四检测头

原始的 YOLOv7-tiny 模型只有 3 个检测头，其高度和宽度分别为 20×20、40×40 和 80×80，导致模型对于尺度更大物体的检测精度不理想，不能满足线虫检测场景的需求。因此，本研究在 Head 部分 80×80 检测头的旁边额外引入一个 160×160 检测头，以提高模型对不同尺度物体的检测能力。添加的第四检测头融合了输入图像中第一个 C2f 模块的浅层信息，不仅丰富了锚框的尺度，而且提升了检测精度[20]。

综上，本研究在 YOLOv7-tiny 模型的基础上，采用轻量化 MobileOne 网络替换 Backbone 层的原始骨干网络；引入 GFPN 网络改进 Neck 层，以提供更丰富的特征信息；在 Neck 层引入 BRA 注意力机制，以使模型能够将注意力聚焦在重要位置，减少特征信息的损失；最后增加第四检测头，提高对不同尺度目标的检测能力。基于改进 YOLOv7-tiny 的线虫检测模型网络结构如图 9 所示。

中国知网　https://www.cnki.net

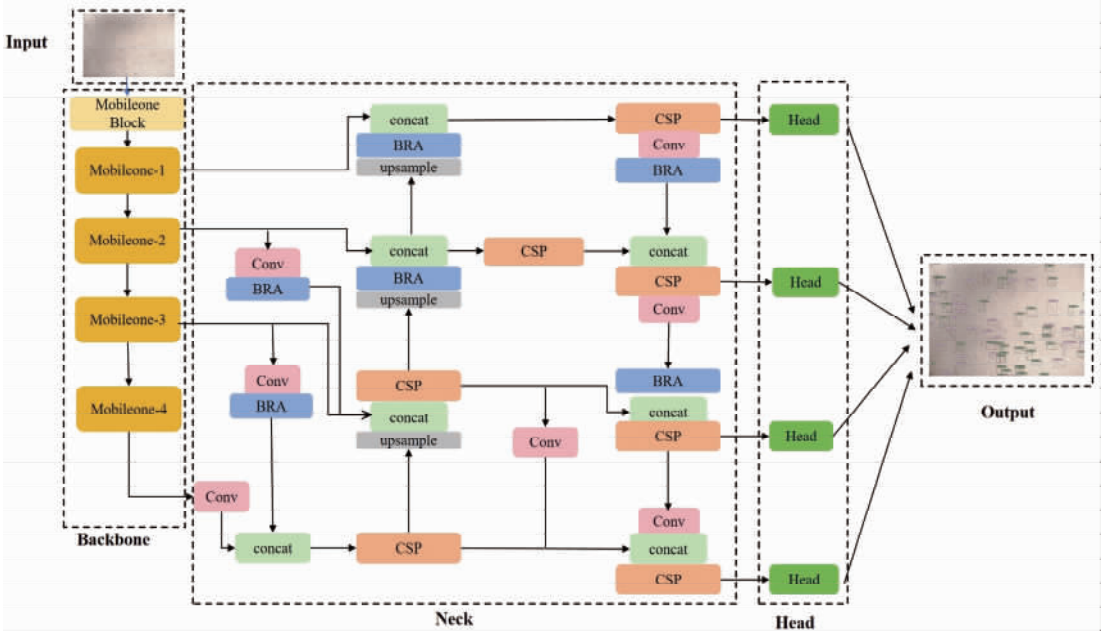

图 9 改进 YOLOV7-tiny 网络结构

3 模型感知量化

随着深度学习研究的进展，神经网络已广泛应用于多个领域。然而，随着模型检测性能的不断提升，模型尺寸和计算量也随之增加。为了降低模型计算强度、参数量大小和内存消耗，研究人员开发了模型量化技术，将浮点计算转换为低比特定点计算[21]。

本研究采用 PyTorch 框架下的感知量化方法将模型权重变为 INT8 类型。INT8 量化主要采用缩放因子和偏移量将神经网络中的权值和输入从 32 位或 64 位浮点数转换成 8 位定点数。

量化过程中的缩放因子(S)计算公式如下:

$$S=\frac{X_{max}-X_{min}}{Q_{max}-Q_{min}} \quad 。 \quad (3)$$

其中，X 为原始浮点数，Q 为量化之后的整型数。

量化过程中的偏移量计算如公式(4)所示:

$$Z=\text{round}(Q_{max}-Q_{min}) \quad 。 \quad (4)$$

其中，Z 对应量化过程中的零点。

最终的量化参数计算如公式(5)所示:

$$Q=\text{round}(X/S+Z) \quad 。 \quad (5)$$

由以上公式可知，模型可根据精度损失自动选择最优的 S 和 Z 进行量化。但实际上 YOLO 网络由于激活不均匀导致存在异常值，使性能下降。为削弱异常值对结果的不利影响，本研究在原模型 INT8 量化的基础上对激活值部分采用非对称量化策略，该策略通过分析激活值的分布特性，动态调整量化比例和零点，以降低量化引入的精度损失。首先，在校准阶段，使用验证集的一部分数据作为校准数据集，获取激活值的实际分布特性；根据分布特性，非对称策略将动态设定出上限截断阈值，而下限截断阈值则预设为一个经验值；然后通过迭代优化的方式找到量化误差最小的理想阈值范围。这不仅减少了校准时间，而且增强了激活值量化的准确性，因此，在保持高预测精度的同时，显著降低了计算和存储成本，从而使在计算能力有限的边缘设备上实现实时、高效的目标检测算法部署成为可行之举。

4 实验与结果分析

4.1 实验配置及评价指标

4.1.1 *实验环境及参数设置* 本研究采用 Ubuntu 操作系统，GPU 配置为 NVIDIA GeForce RTX 4090，软件配置为 CUDA12.0；使用 Python 3.8.1 和 PyTorch 1.11。损失函数采用 YOLOv7-tiny 官方版本的 CIoU。初始学习率设置为 0.01，并使用 SGD 优化器进行训练，Batch size 设置为 64，训练次数 Epoch 设置为 1 000。

4.1.2 *评价指标* 为了准确客观地评估网络模型的检测性能，主要选取的评价指标为平均精度

中国知网 https://www.cnki.net

均值(mAP@0.5) 、计算量(GFLOPs) 和检测帧率(FPS) 。

4.2　消融实验结果分析

为验证本研究对 YOLOv7-tiny 改进的有效性，在实验环境及参数配置相同的前提下使用同一线虫数据集进行消融实验，分析各种改进方法对模型检测性能的影响，实验结果如表 1 所示。模型 1 将原 YOLOv7-tiny 的骨干网络替换为 MobileOne 模块以实现网络的高效计算，结果显示该策略虽然使得 mAP@0.5 值降低了 0.94 个百分点，但高效率网络具有更强的实用价值。具体地，我们将神经网络中访存消耗与计算并行度一起考虑在内，如改善 concat 等无参操作带来的显著的访存消耗。模型 2、3、4 是在原模型的基础上，分别增加 GFPN、BRA、第四检测头，结果表明，其 mAP@0.5值均有所提升，尤其增加第四检测头的提升较大。模型 5 和模型 6 融合其中 3 种改进策略，mAP@0.5 值较原模型分别提升了 1.46、1.84 个百分点。而本研究提出的改进 YOLOv7-tiny 模型融合了 MobileOne、GFPN、BRA 和第四检测头四种改进策略，mAP@0.5 值较原模型提升了 2.06 个百分点，提升效果最佳。

表 1　消融实验结果

模型	MobileOne	GFPN	BRA	第四检测头	mAP@0.5/%
YOLOv7-tiny	—	—	—	—	91.18
1	√	—	—	—	90.24
2	—	√	—	—	91.67
3	—	—	√	—	91.48
4	—	—	—	√	91.70
5	√	—	√	√	92.64
6	—	√	√	√	93.02
改进 YOLOv7-tiny	√	√	√	√	93.24

注“√”表示实验中使用该方法改进模型，“—”则为不使用。

4.3　对比实验结果分析

4.3.1 使用不同量化方法的对比实验　为深入研究量化方法对模型性能的影响，本研究从线虫数据集中随机抽取 2 500 张图像样本作为校准数据集，用于指导和优化模型参数设定。同时选取一组线虫图像数据构建独立的验证集(与校准数据集互不重叠) ，进行实证检验和全面的性能评测。

表 2 为不同量化类型及量化比特数对模型性能影响的实验结果，考虑了仅量化权重、仅量化激活值以及同时量化权重和激活值。需要说明的是，该实验属于前置探索实验，旨在分析不同量化类型和比特数对模型性能的影响，为后续本研究量化方案设计提供依据。结果显示，相较于仅量化激活值，仅针对权重的量化操作通常会带来明显的性能衰减现象，且量化程度越深，即所使用的量化比特数越少，由此引发的精度损失也越明显。所量化的权重和激活值都为 4 bit 时，模型精度损失最大，此时模型的 mAP@0.5 为 65.4%。所量化的权重为 6 bit 且激活值不进行量化时，模型精度损失最小，此时模型的 mAP@0.5 为 89.4%。在 YOLO 框架下的神经网络中，权重参数实质上承载了网络习得的核心知识内容，因此权重的精度对于模型的整体性能具有决定性作用。与此相对，激活值作为输入数据在网络内部流转过程中的中间表达形式，对量化误差具有一定的容错能力，即便进行量化处理，也不至于对模型性能产生同等程度的影响。

表 2　不同量化类型对模型性能影响的比较

模型	比特数	量化类型	mAP@0.5/%
YOLOv7-tiny	实值(浮点型)		93.2
1	6-32	仅量化权重	89.4
2	32-6	仅量化激活值	89.2
3	6-6	同时量化权重和激活值	88.7
4	4-32	仅量化权重	68.9
5	32-4	仅量化激活值	83.1
6	4-4	同时量化权重和激活值	65.4

表 3 给出了未经量化的完整精度模型与运用 MinMax、Percentile 方法实现 8 bit 和 4 bit 量化模型的对比结果。实验中，本研究对权重部分实施对称通道量化技术，为了确保对比公正，统一采用 MinMax 量化方法，而对于激活部分，则采取非对称分层量化策略。鉴于输入层和输出层对模型精度的影响尤为关键，为维持模型总体性能的稳定性和精确度，本研究不对这两层进行量化处理，保留其原有高精度状态。通过比较不同量化方法和检测算法在计算效率及存储资源占用方面的表现，可以看出，本研究方法显著提升了运算速度，并减少了线虫检测模型所需存储空间。在检测精度方面，当本研究方法将模型量化至 8 bit 时，其

中国知网　https://www.cnki.net

mAP@0.5与未经量化的全精度模型持平。然而，对于参数量极大的基础模型，在转化为8 bit量化模型时，尽管整体上精度损失较小，但仍可能出现精度滑落。当模型进一步量化至4 bit时，由于4 bit整数表达能力的局限性，会导致精度明显下降。同时，采用MinMax量化方案的模型相较于未量化的实值模型在精度上略有下降，但模型大小和计算量则明显减少；而基于百分位数的Percentile量化方法并未显示出优势提升。综合考虑精度、计算量和模型大小，本研究最终采用的量化方案为模型3（比特数为8-8 bit），其中权重采用对称量化，激活采用非对称分层量化，在保证模型性能的同时实现了更为显著的优化效果。

表3 不同量化方法的比较实验结果

模型	方法	比特数	参数量(M)	计算量/GFLOPs	mAP@0.5/%
Base	实值(浮点型)	32-32	57.6	16.50	88.4
1	MinMax	8-8	14.4	4.23	88.2
2	Percentile	8-8	14.4	4.23	87.9
3	本研究方法	8-8	14.4	4.23	88.4
4	Percentile	4-4	7.7	2.16	58.4
5	本研究方法	4-4	7.7	2.16	67.3

4.3.2 边缘设备部署实验 本研究所开发的模型将主要聚焦于杀线虫菌剂活性筛选的现实应用场景，因此，为模拟现实环境中的线虫检测，将本研究提出模型以及Faster-RCNN、YOLOv3-tiny、YOLOv5n、YOLOv6n、YOLOv7-tiny、YOLOv8n分别部署在Jetson Nano边缘设备中（图10），在相同的实验环境及配置下对同一线虫数据集进行检测，结果见表4。可知，本研究提出的改进YOLOv7-tiny模型的mAP@0.5高达93.0%，优于其他6种检测模型，较原始模型提高2.7个百分点，比新版本的YOLOv8n也高出3.3个百分点。在轻量化性能方面，改进YOLOv7-tiny的计算量仅为4.23 GFLOPs，明显低于其他6种模型，较原始模型减少67.71%，改进效果显著。此外，本研究提出模型的检测帧率为9.57 FPS，较原始模型提高了23.01%，虽低于YOLOv5n、YOLOv6n和YOLOv8n，但在检测精度和计算量方面远高于它们。综合来看，本研究提出的改进模型在检测精度和轻量化方面表现出色，能够满足线虫实际检测的要求，相对于其他6种主流的目标检测模型，具有最佳的检测精度和最低的计算量，实现了检测精度与轻量化之间的平衡。

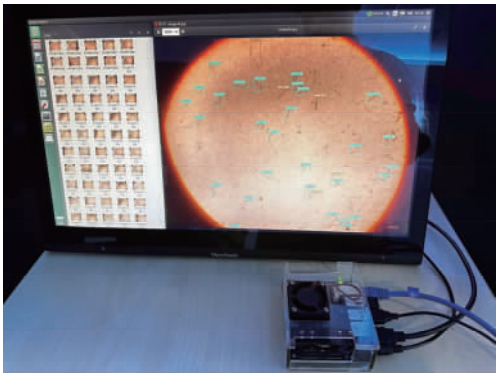

图10 Jetson Nano部署测试图

表4 边缘设备的模型部署对比实验结果

模型	mAP@0.5/%	mAP@0.5:0.95/%	延迟/ms	检测帧率/FPS	参数量(M)	计算量/GFLOPs
Faster-RCNN	88.9	61.2	1 052.6	0.95	60.3	66.50
YOLOv3-tiny	84.8	51.6	139.7	7.19	16.0	5.58
YOLOv5n	85.3	54.4	66.7	12.78	6.0	4.50
YOLOv6n	89.3	60.8	89.5	11.79	11.0	11.10
YOLOv7-tiny	90.3	61.3	128.6	7.78	15.0	13.10
YOLOv8n	89.7	61.2	97.9	10.21	9.0	8.10
改进YOLOv7-tiny	93.0	61.8	104.5	9.57	14.4	4.23

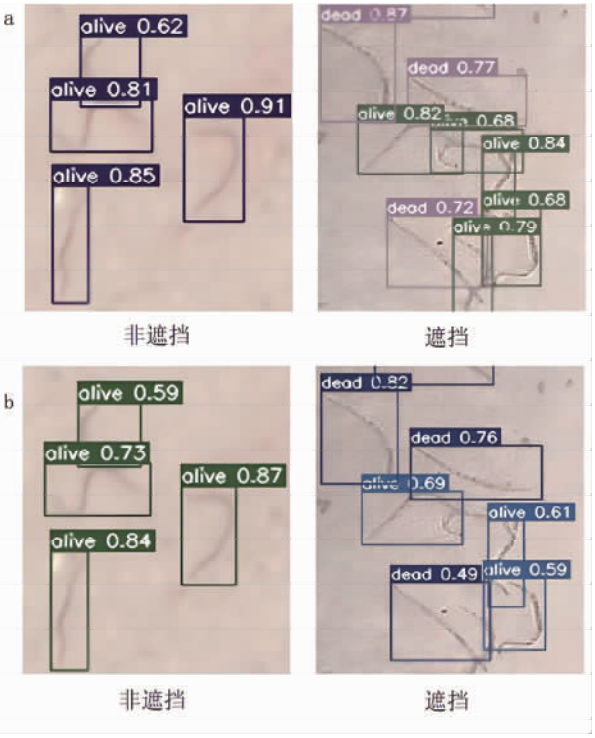

图11 改进YOLOv7-tiny(a)与原始YOLOv7-tiny的线虫检测效果对比

中国知网 https://www.cnki.net

4.4　模型对线虫的识别效果验证

对比改进 YOLOv7-tiny 与原始 YOLOv7-tiny 对线虫的识别效果(图 11)可以看出,YOLOv7-tiny 模型在遮挡环境下会出现漏判和精度低的现象,而改进 YOLOv7-tiny 模型成功实现了重叠遮挡线虫的识别,鲁棒性更好。

5　结论

本研究提出一种面向边缘设备的改进 YOLOv7-tiny 线虫检测模型。在 Backbone 层,利用 MobileOne 网络替换 YOLOv7-tiny 原始骨干网络,构建轻量化网络模型;在 Neck 层,引入广义特征金字塔网络(GFPN),增强高低层信息交互;同时融入双层路由注意力机制(BRA),使模型将注意力聚焦于重要位置,增强遮挡目标特征提取能力;在 Head 层,增加第四个检测头,提高小目标的检测能力;最后,对模型进行 INT8 量化处理,并部署在 Jetson Nano 边缘计算设备上进行测试。实验结果证明,本研究提出的改进 YOLOv7-tiny 模型 mAP@0.5 为 93.0%(其中活虫为 93.8%,死虫为 92.5%),计算量仅为 4.23 GFLOPs,检测帧率为 9.57 FPS,相比原始 YOLOv7-tiny 模型,mAP@0.5提升了 2.7 个百分点,计算量压缩了 67.71%,检测帧率提高了 23.01%,在模型准确性和实时性方面具有较大优势,可为线虫目标检测的边缘计算模式应用奠定基础。

参　考　文　献:

[1] 崔梦侨,杨俊霞,王松山. 模式生物秀丽隐杆线虫在吸入麻醉药实验教学中的应用[J]. 中国继续医学教育,2023,15(22):148-153.

[2] 陆健强,梁效,余超然,等. 基于坐标注意力机制与高效边界框回归损失的线虫快速识别[J]. 农业工程学报,2022,38(22):123-132.

[3] 任妮,鲍彤,沈耕宇,等. 基于深度学习的细粒度命名实体识别研究:以番茄病虫害为例[J]. 情报科学,2021,39(11):96-102.

[4] 黄丽明,王懿祥,徐琪,等. 采用 YOLO 算法和无人机影像的松材线虫病异常变色木识别[J]. 农业工程学报,2021,37(14):197-203.

[5] Liu J,Wang X W. Tomato diseases and pests detection based on improved Yolo V3 convolutional neural network[J]. Frontiers in Plant Science,2020,11:898.

[6] 刘金涛,李双,李佳骏,等. 基于改进 YOLOv5s 的咖啡叶病虫害识别方法[J]. 山东农业大学学报(自然科学版),2023,54(5):691-699.

[7] 赵鹏飞,钱孟波,周凯琪,等. 改进 YOLOv7-Tiny 农田环境下甜椒果实检测[J]. 计算机工程与应用,2023,59(15):329-340.

[8] 关欣,李璐,罗松. 面向物联网的边缘计算研究[J]. 信息通信技术与政策,2018(7):53-56.

[9] 于放. 基于 YOLOv7-tiny 的目标检测脱敏算法研究[D]. 大连:大连交通大学,2023.

[10] 万树青,周青春,庄农波,等. 杀线虫剂活性测定中线虫死活鉴别的染色方法[J]. 农药,1993,32(1):18-19.

[11] 孙迟,刘晓文. 基于 YOLOv7-tiny 改进的矿工安全帽检测[J]. 中国科技论文,2023,18(11):1250-1256,1274.

[12] 刘修政,王波. 改进 YOLOv7-tiny 的轻量化绝缘子缺陷检测算法[J]. 无线电工程,2024,54(10):2305-2314.

[13] Vasu P K A,Gabriel J,Zhu J,et al. MobileOne: an improved one millisecond mobile backbone[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2023:7907-7917.

[14] 靳红杰,马顾彧,唐梦圆,等. 复杂环境下黄花菜识别的 YOLOv7-MOCA 模型[J]. 农业工程学报,2023,39(15):181-188.

[15] 李婷,孙渊. 基于改进轻量型 YOLOv5 的太阳能电池板缺陷检测[J]. 组合机床与自动化加工技术,2023(11):95-99,106.

[16] Jiang Y Q,Tan Z Y,Wang J Y,et al. Giraffedet: a heavy-neck paradigm for object detection[J/OL]. arXiv: 2202.04256 [cs.CV],2022. https://doi.org/10.48550/arXiv.2202.04256.

[17] 梅礼坤,陈智利. YOLO-Plane:一种基于改进 YOLOv5 的飞机检测算法[J]. 激光杂志,2024,45(5):69-78.

[18] 田鹏,毛力. 改进 YOLOv8 的道路交通标志目标检测算法[J]. 计算机工程与应用,2024,60(8):202-212.

[19] 吴明杰,云利军,陈载清,等. 改进 YOLOv5s 的无人机视角下小目标检测算法[J]. 计算机工程与应用,2024,60(2):191-199.

[20] Kang M,Ting C M,Ting F F,et al. BGF-YOLO: enhanced YOLOv8 with multiscale attentional feature fusion for brain tumor detection[J/OL]. arXiv: 2309.12585 [cs.CV],2023. https://doi.org/10.48550/arXiv.2309.12585.

[21] 胡艺馨,张逸杰,方健,等. 面向目标检测任务的轻量化网络模型设计[J]. 计算机工程与设计,2023,44(2):548-555.

中国知网　https://www.cnki.net