基于多模态图像的自然环境下油茶果识别

周宏平，金寿祥，周 磊，郭自良，孙梦梦
（南京林业大学机械电子工程学院，南京 210037）

摘 要：针对自然条件下油茶果生长条件复杂，存在大量遮挡、重叠的问题，提出了一种基于 RGB-D（red green blue-depth）多模态图像的双主干网络模型 YOLO-DBM（YOLO-dual backbone model），用来进行油茶果的识别定位。首先，在 YOLOv5s 模型主干网络 CSP-Darknet53 的基础上设计了一种轻量化的特征提取网络。其次，使用两个轻量化的特征提取网络分别提取彩色和深度特征，接着使用基于注意力机制的特征融合模块将彩色特征与深度特征进行分级融合，再将融合后的特征层送入特征金字塔网络（feature pyramid network，FPN），最后进行预测。试验结果表明，使用 RGB-D 图像的 YOLO-DBM 模型在测试集上的精确率 *P*、召回率 *R* 和平均精度 AP 分别为 94.8%、94.6% 和 98.4%，单幅图像平均检测耗时 0.016 s。对比 YOLOv3、YOLOv5s 和 YOLO-IR（YOLO-InceptionRes）模型，平均精度 AP 分别提升 2.9、0.1 和 0.3 个百分点，而模型大小仅为 6.21MB，只有 YOLOv5s 大小的 46%。另外，使用注意力融合机制的 YOLO-DBM 模型与只使用拼接融合的 YOLO-DBM 相比，精确率 *P*、召回率 *R* 和平均精度 AP 分别提高了 0.2、1.6 和 0.1 个百分点，进一步验证该研究所提方法的可靠性与有效性，研究结果可为油茶果自动采收机的研制提供参考。

关键词：图像识别；深度学习；模型；油茶果；多模态；多尺度融合
doi：10.11975/j.issn.1002-6819.202303054
中图分类号：S24；TP391.4 **文献标志码**：A **文章编号**：1002-6819(2023)10-0175-08

周宏平，金寿祥，周磊，等. 基于多模态图像的自然环境下油茶果识别[J]. 农业工程学报，2023，39(10)：175-182. doi: 10.11975/j.issn.1002-6819.202303054 http://www.tcsae.org
ZHOU Hongping, JIN Shouxiang, ZHOU Lei, et al. Recognition of camellia oleifera fruits in natural environment using multi-modal images[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2023, 39(10): 175-182. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202303054 http://www.tcsae.org

0 引 言

油茶是中国重要的木本油料作物，主要分布于中国南方低山丘陵和山区，是中国栽培面积最大、分布最广的经济树种之一[1-3]。油茶是一种花果同期作物，导致采收难度较大，振动式和梳齿式采收的方法会导致部分茶花脱落，导致来年产量下降[4-5]。摇枝式采收是目前较为合适的一种油茶果果实采收方法，其具有振动式采收速度快的优点，并且对茶花伤害较小，所以准确识别油茶果，判断果实疏密区域从而确定振动头夹持位置是实现自动摇枝采收方法的重要步骤[6-7]，因而解决自然场景中油茶果果实准确、高效的识别难题对实现油茶果自动化采收具有重大意义。

自然环境中生长的油茶树枝叶茂密，加之油茶果果实较小，因此存在着大量果实重叠、果实被枝叶遮挡的情况，另外受光照条件变化的影响，易出现背光、强光等不利因素，给识别造成困难[8-10]。目前针对果实识别问题主要是基于 RGB 图像的识别方法，陈志健等[11] 为了实现重叠油茶果的定位，将 RGB 图像经过阈值分割、形态学操作和最小二乘法拟合的方法确定图像中油茶果的位置，单张图像平均耗时 0.52 s。陈斌等[12] 将 Faster RCNN 深度学习模型用于油茶果的识别之中，油茶果识别准确率达到 98.92%，平均每幅图像识别时间为 0.2 s。为了提高识别速度，宋怀波等[13] 使用 YOLOv5s 模型进行油茶果果实识别，平均检测精度达到了 98.71%，单幅图像检测时间仅为 12.7 ms，与 YOLOv4-tiny 和 RetinaNet 模型相比，检测时间分别减少了 96.39% 和 96.25%。

当前国内外学者使用 RGB 图像对果实进行识别进行了充分的研究，取得了大量的成果，但大部分集中在模型结构优化与改进，提高检测速度与精度上[14-17]，缺少对多模态数据使用的研究。随着消费级 RGB-D 相机的普及，其正在被越来越多的应用于果实的识别与定位研究中[18-20]，如王文杰等[21] 提出基于 RGB-D 信息融合的番茄识别方法，该方法将 RGB 图像、深度图像和红外图像融合成 5 通道的融合图像，并输入 Mask RCNN 模型进行训练，果实识别准确率为 98.3%，高出只使用 RGB 图像训练的 Mask RCNN 模型 2.9 个百分点。WANG 等[22] 为提高遮挡番茄识别效果，提出一种集合深度信息与彩色图像信息的改进 SSD 模型，该模型在后端融合彩色特征与深度特征进行预测。结果表明，该方法的平均识别精度高于只使用 RGB 图像或深度图像。但是，由于消费级 RGB-D 相机传感器精度与成像原理的限制，导致深度图像的质量不高，存在一些深度值为零的像素点组成的深度孔。而且在室外果园环境中获取的深度图像上难以直接分辨果实与叶片，简单的将其与 RGB 图像进行融合，会忽略不同模态和区域对检测结果的影响，且更容易在

收稿日期：2023-03-09 修订日期：2023-04-10
基金项目：国家林业和草原局应急科技项目（202202-3）
作者简介：周宏平，教授，博士生导师，研究方向为自动化与智能化林业机械。Email：npzhou@njfu.edu.cn。

深度图像噪声区域产生过拟合现象。

本文为了更好地利用多模态数据，提出一种双主干特征提取网络，分别提取彩色特征与深度特征，并在特征层的维度进行多尺度特征融合。为了降低双主干模型大小，本文在 YOLOv5 模型主干的基础上，结合 Inception-Res 模块，提出了一种轻量化的特征提取网络。同时，针对深度图像中存在空洞，图像质量不高的问题，本文使用一种基于卷积注意力机制的特征融合方法，增加可能存在果实区域的特征权重，在特征融合过程中降低深度噪声的影响，提高果实检测精度率。最后通过试验验证所提出模型对自然环境中油茶果果实的识别效果，以期为实现油茶果的自动化采收提供技术支持。

1 试验数据

1.1 数据样本采集

本次试验数据采集地位于南京市江宁区南京金航油茶合作社（31°68′19″，118°89′34″），油茶果颜色多为黄褐色与红色，部分品种为青绿色，形状为圆球形、椭球形或橄榄型，如图 1 所示，果实之间形态差异大，遮挡情况严重，给识别带来了困难。本次试验研究的数据采集于 2022 年 10 月 2 日至 15 日，采集设备是 Intel RealSense d435f 深度相机，用于采集 RGB-D 图像，每组 RGB-D 图像由一张 RGB 图像和对应的深度图像组成。数据采集工作在 Windows10 平台上进行，通过 Intel RealSense 官方提供的 pyrealsense2 函数库在 python3.8 环境中进行编程和程序运行，采集油茶果 RGB-D 图像，并通过函数库中的 align 函数保证 RGB 图像与深度图像之间的保持对应。

图 1 油茶果样本

Fig.1 Samples of camellia oleifera

为确保数据的多样性与可靠性，分别采集了远景、近景、遮挡、重叠、强光、背光和密集等场景中的油茶果 RGB-D 图像，共采集到 8 000 组分辨率为 1 280×720 的 RGB-D 图像。

1.2 数据集构建

从最初的 8 000 组 RGB-D 图像中剔除重复、拖影、无果实的图像后，剩余 1 040 组 RGB-D 图像作为原始数据集。为符合模型输入端 640×640 的尺寸要求，在每组图像上随机生成 10 个 640×640 的方框对 1 280×720 的原始图像进行裁剪，生成 10 400 组 RGB-D 图像。再次筛选掉其中相似、无果实的图像后，得到 1 379 组 RGB-D 图像作为试验数据集，命名为 MCOTDD（multi-modal Camellia oleifera target detection dataset，多模态油茶果目标检测数据集）。另外将 1 379 组 RGB-D 图像中的 RGB 图像取出建立 RGB 单模态数据集，命名为 COTDD（Camellia oleifera target detection dataset，油茶果目标检测数据集），在比较不同输入对模型检测效果的影响时使用。

同时为了降低深度图像中可能存在的远景处过大的深度值对模型训练产生的不利影响[23]，将深度值大于 1.20 m 的像素点的数值置为 0，效果如图 2 所示。本试验使用 YOLO 格式的数据集，采用 LabalImg 图像标注工具在 RGB 图像上进行标注。由于本次研究目的仅是油茶果果实的识别，因此在标记时仅有油茶果一类目标，其余未标注部分由 LabalImg 默认为背景。标记过程中对被严重遮挡的、远处目标过小的果实不予标记，防止模型训练出现错误，最终共标记了 8 419 个果实。将标记好的图像按照 4:1 的比例划分成训练集与验证集，其中训练集图像 1 104 组，验证集图像 275 组。

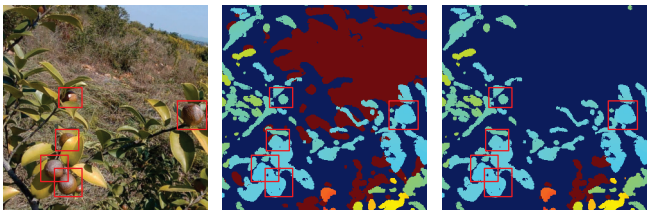


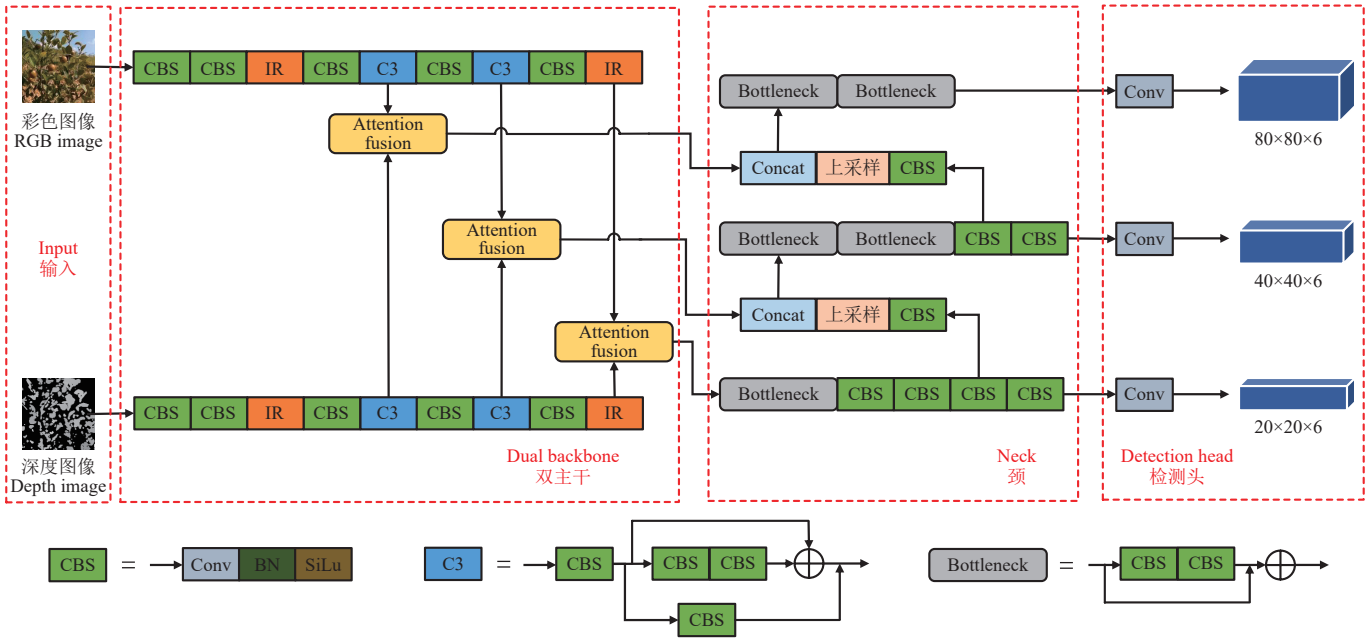
图 2 油茶果 RGB-D 数据的可视化示例

Fig.2 Visual example of camellia oleifera RGB-D data

2 使用多模态图像的双主干油茶果识别模型

2.1 双主干油茶果识别模型

为提高自然环境中油茶果小目标识别精度，使用多模态的 RGB-D 图像作为数据源，提出了一种双主干的油茶果目标识别模型 YOLO-DBM（YOLO- dual backbone model），结构如图 3 所示。该模型的核心思想是使用两个轻量化的特征提取网络作为 RGB-D 图像的特征提取器，分别用来提取 RGB-D 图像中的颜色特征与深度特征，避免模型在特征提取过程中，由于不同模态数据性质不同而发生干扰。其次，为了更好的融合多模态特征，提出了一种基于注意力机制的特征融合模块，来对双主干特征提取网络提取到的不同模态特征进行逐级融合，降深度孔的不利影响，并使不同特征层之间融合更充分。最后，使用 FPN（feature pyramid network，特征金字塔网络）作为颈网络，对经过特征融合后的不同特征层进行多尺度融合，提高对油茶果小目标的识别能力。



注：BN 为批归一化操作；SiLu 为激活函数；⊕为叠加操作； Attention fusion 为注意力融合模块；80×80×6、40×40×6 和 20×20×6 分别代表网络不同输出特征层的大小。
Note：BN is a batch normalization operation; SiLu is the activation function; ⊕ is a superposition operation; Attention fusion is an attention fusion module; 80×80×6, 40×40×6, and 20×20×6 respectively represent the size of different output feature layers in the network.

图 3　YOLO-DBM 网络结构
Fig.3　Structure of YOLO-DBM network

另外，为比较本文提出的双主干模型 YOLO-DBM 的有效性，提出了一种与其对应的单主干网络模型 YOLO-IR（YOLO-InceptionRes），该模型在 YOLO-DBM 的基础上，移除了特征融合模块和一支特征提取网络，仅使用一支主干网络作为特征提取单元，其它结构不变，为后续消融试验提供参照。

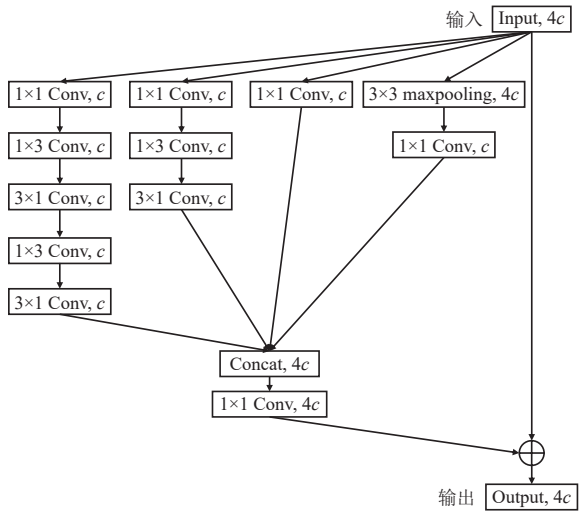
2.2　轻量化的特征提取网络

YOLOv5s 是目前较为常用的一阶段目标检测算法，其在保证较高的检测精度的同时还能保持较快的检测速度，在果实识别领域被大量应用[24-28]。因此，本文在 YOLOv5s 模型的主干网络 CSP-Darknet53 的基础上进行了一些轻量化改进，设计了一种轻量化的特征提取网络，结构如表 1 所示。首先，使用 InceptionRes 特征提取模块替代 CSP-Darknet53 中的第一个和最后一个 C3（concentrated-comprehensive convolution block）特征提取模块，引入多尺度信息。其次，控制网络每层输出的通道数，缩小网络宽度，减少冗余的参数。另外，由于使用的 InceptionRes 模块已经引入了多尺度信息，所以将 CSP-Darknet53 中的 SPPF（spatial pyramid pooling-fast，快速空间金字塔池化）多尺度融合模块移除，降低结构复杂度。

InceptionRes 模块[29] 如图 4 所示，由 4 条不同尺度的分支组合而成。其先利用 1×1 卷积将左边 3 个通道降至 c，降低后续计算量，再分别使用 3 个等效 5×5、3×3 和 1×1 卷积进行特征提取，再添加一条 3×3 最大池化并配合 1×1 卷积降低通道维度，得到 4 个通道数为 c 的不同尺度的特征层。然后，经过 Concat 拼接操作，恢复到原始通道数 $4c$，实现多尺度信息融合，提高网络对不同尺度目标的感知能力。最后，添加残差结构防止深度神经网络训练过程出现梯度爆炸或梯度消失的现象。

表 1　轻量化特征提取网络结构
Table 1　Lightweight feature extraction network structure

层模块 Layer	卷积核大小 Kernel size	步距 Stride	填充 Padding	输出大小 Output size	输出通道数 Output channel	参数量 Parameter
Conv	6x6	2	2	320×320	16	1 760
Conv	3x3	2	1	160×160	32	4 672
InceptionRes				160×160	32	3 424
Conv3	3x3	2	1	80×80	64	18 560
C3				80×80	64	18 816
Conv	3x3	2	1	40×40	128	73 984
C3				40×40	128	74 496
Conv	3x3	2	1	20×20	256	295 424
InceptionRes				20×20	256	206 592

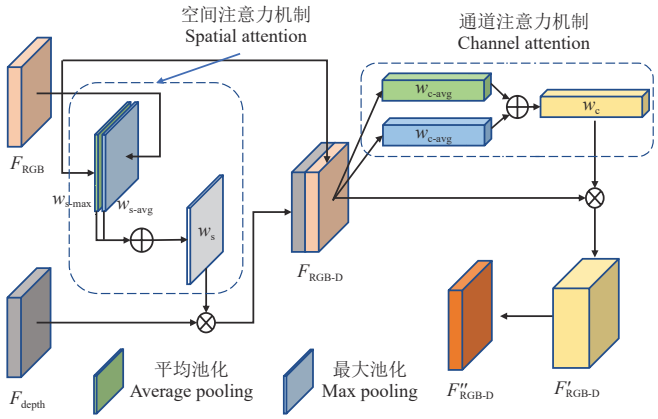


注：c 为通道数量；Conv 为卷积操作；Maxpooling 为最大池化操作；Concat 为拼接操作；⊕为叠加操作。
Note：c is the number of channels; Conv is convolution operation; Maxpooling is the maximum pooling operation; Concat is splicing operation, ⊕ is a superposition operation.

图 4　InceptionRes 模块结构
Fig.4　InceptionRes module structure

2.3 注意力特征融合模块

为了更充分地利用多模态数据，使用特征融合是必要的[30-31]。RGB 图像包含颜色、形状、纹理等二维平面信息，而深度图像更多的表示目标物的空间距离信息，两者包含的信息意义不同，能实现一定程度的信息互补，有助于提升识别效果。然而，由于目前深度图像的质量问题以及不同模态对检测结果作用的占比不同，简单的将彩色特征与深度特征进行相加或叠加操作是不合适的。因此，本文提出了一种注意力融合模块，如图 5 所示。注意力机制可以使模型增加对关键或通道的关注度，过滤一些噪声干扰，提高模型检测精度。



注：F_{RGB} 为彩色特征层；F_{Depth} 为深度特征层；$F_{RGB\text{-}D}$ 为拼接后的特征层；F'_{RGB-D} 为经过通道注意力后的特征层；F''_{RGB-D} 为融合后的特征层；⊗为相乘操作；$w_{s\text{-}max}$ 为最大池化特征；$w_{s\text{-}avg}$ 为平均池化特征，w_s 为空间特征图；$\boldsymbol{w}_{c\text{-}max}$ 为全局最大池化特征；$\boldsymbol{w}_{c\text{-}avg}$ 为全局平均池化特征；$\boldsymbol{w}_c$ 为通道特征权重。
Note: F_{RGB} is a color feature map; $F_{RGB\text{-}D}$ is a depth feature map; $F_{RGB\text{-}D}$ is the feature map after splicing; $F'_{RGB\text{-}D}$ is the feature map after passing the channel attention; $F''_{RGB\text{-}D}$ is the feature map after fusion; ⊗ is a multiplication operation; $w_{s\text{-}max}$ is the maximum pooling feature; $w_{s\text{-}avg}$ is the average pooling feature; w_s is the spatial feature map; $\boldsymbol{w}_{c\text{-}max}$ is the global maximum pooling feature; $\boldsymbol{w}_{c\text{-}avg}$ is a global average pooling feature; $\boldsymbol{w}_c$ is the channel feature weight.

图 5 注意力特征融合模块结构

Fig.5 Attention feature fusion module structure

在图 5 所示的注意力特征融合模块中，F_{RGB} 与 F_{Depth} 分别代表同一尺度的 RGB 与深度特征层，高、宽和通道数分别为 H、W 和 C，F_{RGB} 通过最大池化和平均池化操作后，得到大小为 $H\times W\times 1$ 的 $w_{s\text{-}max}$ 最大特征图与 $w_{s\text{-}avg}$ 平均特征图，将两者相加得到 F_{RGB} 的空间权重 w_s。通过上述操作，增大了 F_{RGB} 中可能存在目标区域的权重，将 w_s 与 F_{Depth} 相乘，强调对深度特征中重要区域的学习。之后，将调整后的深度特征层与原始 F_{RGB} 进行拼接操作，得到大小为 $H\times W\times 2C$ 的 RGB-D 特征层 $F_{RGB\text{-}D}$，通过全局最大池化与平均池化操作，得到长度为 $2C$ 的一维向量 $\boldsymbol{w}_{c\text{-}max}$ 和 $\boldsymbol{w}_{c\text{-}avg}$，相加后得到 $F_{RGB\text{-}D}$ 的通道权重 $\boldsymbol{w}_c$，将其与 $F_{RGB\text{-}D}$ 相乘后，强调了重要通道贡献，削弱无效通道，得到F'_{RGB-D}。最后，利用 1×1 卷积对F'_{RGB-D}进行降维，得到大小为 $H\times W\times C$ 的特征层F''_{RGB-D}，作为模型的预测特征层。

2.4 试验平台配置与训练策略

本次试验使用戴尔 Precision 7 750 工作站进行深度学习部分的训练与验证，硬件配置包括：中央处理器为 Intel（R） Core（TM） i7-10875H CPU @2.30 GHz，运行内存为 64GB，图形处理器为 NVIDIA Quadro RTX A4000 mobile 8GB 版本，1T 固态硬盘。软件运行在 Windows 10（22H2）操作系统，所有程序在 Pytorch1.12 深度学习框架下用 python 语言编写，并使用 NVIDIA CUDA11.6 并行运算驱动加速训练。

经过多次调整参数、测试后，最终确定训练时批处理（batchsize）大小为 16，初始学习率为 0.01，衰减系数为 0.01，动量为 0.9，最大迭代次数为 1 000。为了防止模型在训练初期出现大幅波动，在训练过程使用了热身训练，将前 20 轮的学习率变为从 0.000 5 逐步增加到原来第 20 轮的学习率，使模型从较小的学习率开始学习，学习率变化如图 6 所示。另外，为了提高模型的鲁棒性，在模型训练过程中使用了马赛克数据增强[32]，通过随机裁剪、缩放、翻转、色彩变化等图像增强操作后，再随机拼接成一张图片进行训练，丰富数据的多样性。

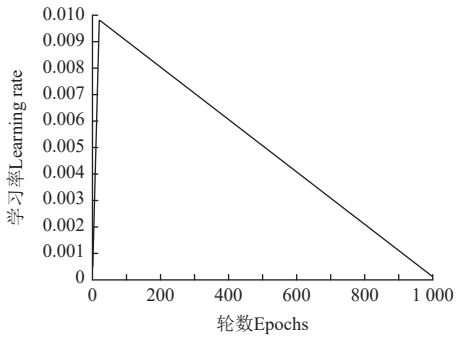


图 6 学习率曲线

Fig.6 Learning rate curve

2.5 评价指标

为了比较模型的性能，设置一些评价指标是必要的，考虑到该模型是针对油茶果识别进行设计的，识别精度与识别速度是衡量识别效果好坏的重要指标，因此使用召回率（Recall，R）、精确率（Precision，P），平均精度（average precision，A_P）和每秒检测图像帧数（frames per second，FPS）作为评价指标。具体计算方法如下：

$$R=\frac{T_P}{(T_P+F_N)}\times 100\% \quad (1)$$

$$P=\frac{T_P}{(T_P+F_P)}\times 100\% \quad (2)$$

$$A_P=\int_0^1 P\cdot(R)\,dR \quad (3)$$

式中 T_P 表示模型将正样本识别为正样本，即正确识别出目标果实的情况，F_P 表示模型将正样本识别为负样本，即未被识别到目标的情况，F_N 表示负样本被识别为正样本，即背景被错误认为是目标的情况。

3 结果与分析

3.1 轻量化特征提取网络的改进效果

为了验证本文设计的轻量化特征提取网络的有效性，使用 COTTD 数据集进行训练与测试。从表 2 中可以看

出，使用轻量化特征提取网络的 YOLO-IR 与 YOLOv5s 相比，模型文件大小减少了 69.27%，模型浮点运算量降低了 70.88%，而模型的平均精度 A_P 仅下降了 0.2 个百分点。改进后的模型在略微损失一些检测进度的情况下，大幅降低了模型的计算量和参数量，说明了轻量化特征提取网络的有效性，为轻量化的双主干网络构建提供了保障。

表 2　YOLO-IR 与 YOLOv5s 模型的检测效果对比

Table 2　Comparison of detection effects between YOLO-IR and YOLOv5s models

模型 Model	平均精度 Average precision/%	模型大小 Model size/MB	计算量 Calculated amount /10^9 次
YOLOv5s	98.3	13.7	15.8
YOLO-IR	98.1	4.21	4.6

3.2　双主干模型消融试验结果

为了证明双主干模型 YOLO-DBM 在多模态数据应用中的优势，本文进行了 4 组对比试验，结果见表 3。其中 YOLO-DBM（Concat）模型是将 YOLO-DBM 模型中的注意力融合模块替换为 Concat 拼接模块，其余结构不变。因此，下文中的 YOLO-DBM 默认代表使用注意力融合的情况。

表 3　不同图像类型和融合策略的检测效果对比

Table 3　Comparison of detection effects of different image types and fusion strategies

模型 Model	图像类型 Image type	融合策略 Fusion strategy	精确率 Precision P/%	召回率 Recall R/%	平均精度 Average precision/%	模型大小 Model size/MB	运行速度 Operating speed/（s·帧$^{-1}$）
YOLO-IR	RGB		93.1	94.5	98.1	4.21	0.015
YOLO-IR	RGB-D	数据层融合	91.9	93.5	96.8	4.21	0.013
YOLO-DBM	RGB-D	特征层融合（拼接融合）	94.6	93.0	98.3	5.72	0.015
YOLO-DBM	RGB-D	特征层融合（注意力机制融合）	94.8	94.6	98.4	6.31	0.016

结果如表 3 所示，在同样使用 YOLO-IR 模型的情况下，使用 RGB-D 图像作为模型输入的检测效果反而低于只使用 RGB 图像作为输入的情况，模型平均精度 A_P 从 98.1% 下降到了 96.8%。与一些类似的研究结果产生了差异，使用多模态数据并没有提高检测精度，反而导致检测精度下降。上述现象的主要原因可能在于本次试验的油茶果园环境复杂，枝叶茂密、遮挡严重，使得所获取的深度图像质量较差，存在大量深度孔，简单的将其在输入端融合会给模型带来噪声，导致模型学习困难。

在同样使用 RGB-D 数据作为输入的情况下，YOLO-DBM 模型的检测效果明显好于数据层融合的 YOLO-IR 模型，模型的精确率 P、召回率 R 和平均精度 A_P 分别增加了 2.9、1.1 和 1.6 个百分点，而模型大小仅为 6.31MB，每幅图片检测速度达到了 0.016 s。而 YOLO-DBM（Concat）模型与 YOLO-IR 相比，精确率有所提高，主要是由于深度图像包含的距离、轮廓等物理信息，使得模型减少了对背景的误判，但是，模型召回率却有所降低，主要原因是深度孔的存在，简单的特征拼接反而会降低深度孔区域的整体权重，导致漏检情况增多。在同样使用特征层融合的情况下，基于注意力机制的特征融合方法的检测效果要优于直接使用拼接融合的方法，精确率与召回率分别提升了 0.2 与 1.6 个百分点，证明了本文提出的注意力融合机制的有效性。

与 YOLO-IR 相比，本文所提出的 YOLO-DBM 的检测效果较好，模型精确率 P、召回率 R 和平均精度 A_P 分别高出 1.7、0.1和 0.3 个百分点。与单主干模型相比，双主干模型使用两个特征提取网络分别提取彩色特征与深度特征，避免了特征提取阶段彩色特征信息与深度特征信息的干扰，并通过注意力融合模块将深度特征与彩色特征相融合，强调了深度特征中有效的信息，而不是简单的特征层叠加或相加。试验结果表明，正确的使用多模态数据，可以提高以单模态数据为基础的目标检测模型的检测效果。

3.3　不同目标检测模型的对比可视化

为了比较本文所提出的基于多模态数据的油茶果识别网络 YOLO-DBM 的检测效果，将其与 YOLOv3、YOLOv5s 以及 YOLO-IR 模型识别效果对比，其中 YOLO-DBM 与 YOLO-DBM（Concat）模型使用的是 RGB-D 图像，其余模型使用的是 RGB 图像，对比结果如图 7 所示。在图 7 背光的情况下，YOLOv3 和 YOLOv5s 将树叶缝隙中的黄色背景误识别为果实，而 YOLO-DBM 可以避免这种误检；在图 7 光照正常的情况下，YOLOv3 与 YOLOv5s 漏检了一些图像边缘的小目标果实，而使用 InceptionRes 结构的 YOLO-IR 与 YOLO-DBM 可以提高小目标的置信度，但同样使用该结构的 YOLO-DBM（Concat）由于深度孔的存在，对边缘小目标也出现了漏检现象；在图 7 的果实密集的情况下，准确识别图中每个果实是困难的，除了 YOLO-DBM 模型检测到了场景中所有果实，其余模型都出现了漏检现象。综上所述，本文所提出的 YOLO-DBM 模型可以较好的利用颜色与深度信息的互补作用，减少对果实和背景的误判，可以准确定位密集生长与被遮挡的油茶果果实。

模型的定量比较分析如表 4 所示，其中 YOLO-DBM 使用的是 MCOTDD 数据集，其余模型使用的是 COTDD 数据集。YOLOv3 与 YOLOv5s 都是较为常用并且先进的目标检测模型，精确率 P 和召回率 R 都达到了 90% 以上，而改进后的单主干模型 YOLO-IR 在体积大幅减小的情况下，取得了 98.1% 的平均精度，比 YOLO-v3 模型高 2.6 个百分点，仅比 YOLOv5s 模型下降了 0.2 个百分点，在保证检测进度的情况下实现了模型的轻量化。但是 3 个模型的精确率均低于召回率，表明模型对相邻果实容易出现误判。双主干模型 YOLO-DBM 在测试集上的精确率 P、召回率 R 和平均精度 A_P 分别达到了 94.8%、94.6% 和 98.4%，对比 YOLOv5s、YOLO-IR，该模型的精确率分别高 1.1 和 1.7 个百分点。YOLO-DBM 模型的文件大小不足 YOLOv5s 的二分之一，浮点运算量下降了 55.7%。

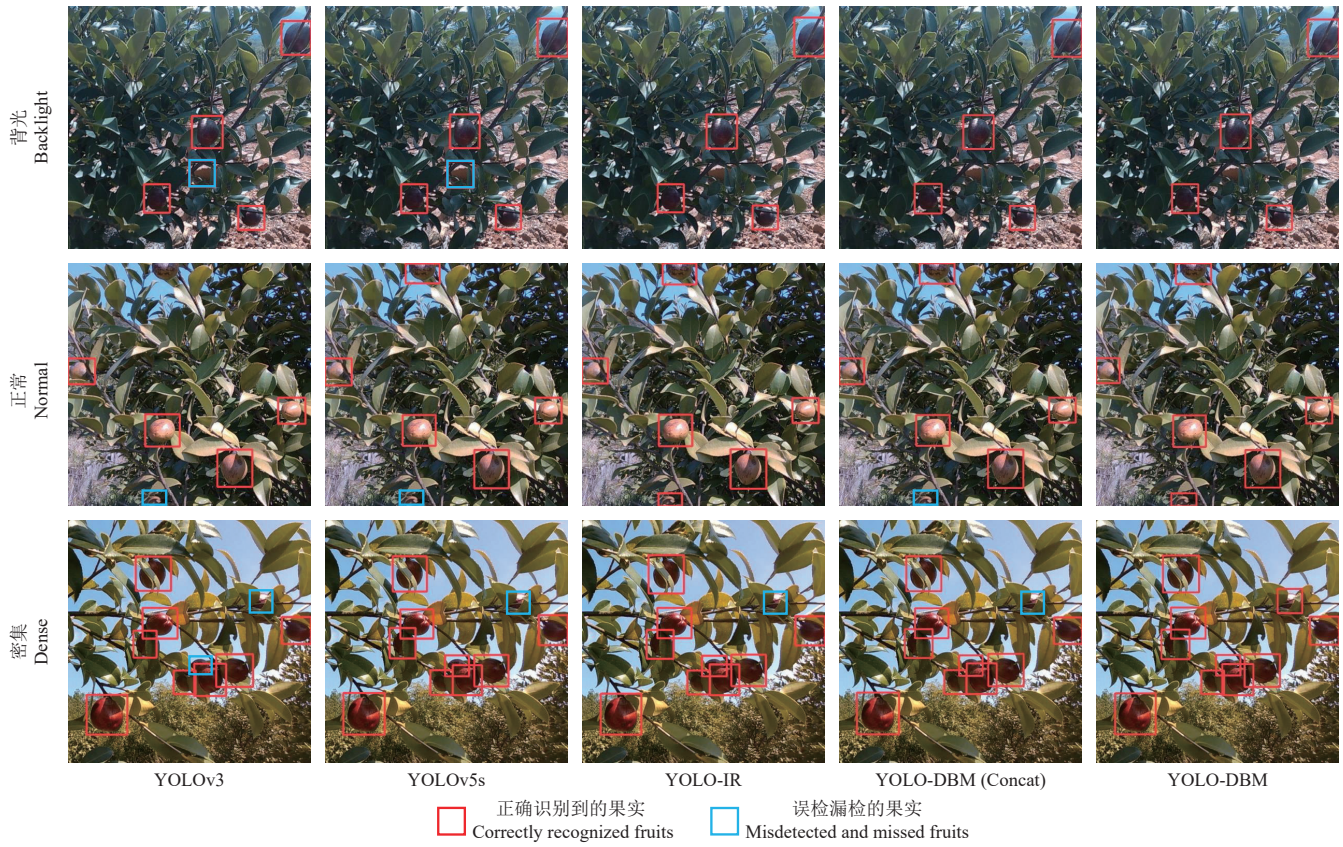


图 7 不同场景下模型检测效果对比

Fig.7 Comparison of model detection effects in different scenarios

表 4 不同检测模型的检测效果对比

Table 4 Comparison of detection effects of different detection models

模型 Model	*P*/%	*R*/%	平均精度 Average precision/%	模型大小 Model size/MB	计算量 Calculated amount/ 10^9 次	运行速度 Operating speed/（s·帧$^{-1}$）
YOLOv3	90.7	91.8	95.5	117.0	154.5	0.035
YOLOv5s	93.7	94.6	98.3	13.7	15.8	0.017
YOLO-IR	93.1	94.5	98.1	4.21	4.6	0.015
YOLO-DBM	94.8	94.6	98.4	6.31	7.0	0.016

4 结 论

1）本文基于 YOLOv5s 主干网络提出了改进的特征提取网络，应用该网络的 YOLO-IR 模型对自然环境下油茶果识别的精确率 *P* 为 93.1%，召回率 *R* 为 94.5%，平均精度为 98.1%，单张图片平均检测耗时仅为 0.015 s，模型仅有 4.21MB。相比于 YOLOv5s 模型，YOLO-IR 虽然在检测性能上略有下降，但模型大小与计算量都有了大幅下降，为搭建轻量化双主干网络提供基础。

2）探讨了多源图像在油茶果识别中的可行性。在多源图像的利用上，本文提出了一种双主干网络 YOLO-DBM，用来分别进行彩色特征与深度特征的提取。相较于只使用彩色特征的 YOLOv5s 模型，YOLO-DBM 模型在检测精确率 *P* 和平均精度上分别提升 1.1 和 0.1 个百分点，模型大小却降低了 53.9%，可有效识别重叠、被遮挡与背光处的目标果实，同时减少误检。

3）在同样使用单主干网络模型进行比较的时候发现，使用 RGB-D 融合图像的检测效果相比只使用 RGB 图像的平均检测精度下降了 1.3 个百分点。与先验知识相违背，更丰富的数据并不能保证检测效果的提升。在后续研究中可以深入探索不同阶段进行特征融合对模型检测效果的影响。

本文提出的 YOLO-DBM 网络模型实现了在实际复杂的果园环境中对油茶果实高精度识别的目标，平均精度达到了 98.4%。且模型大小仅为 6.31MB，可在户外嵌入式设备部署，具备实际应用能力。

［参 考 文 献］

[1] 伍德林，杨俊华，刘芸，等. 我国油茶果采摘装备研究进展与趋势[J]. 中国农机化学报，2022，43(1)：186-194.
WU Delin, YANG Junhua, LIU Yun, et al. Research progress and trend of camellia fruit picking equipment in China[J]. Journal of Chinese Agricultural Mechanization, 2022, 43(1): 186-194. (in Chinese with English abstract)

[2] 陈素传，季琳琳，姚小华，等. 油茶品种果实主要经济性状和营养成分的差异分析[J]. 经济林研究，2022，40(2)：1-9.
CHEN Suchuan, JI Linlin, YAO Xiaohua, et al. Variation analysis on the main economic characters and nutrients of fruit from camellia oleifera varieties[J]. Nonwood Forest Research, 2022, 40(2): 1-9. (in Chinese with English abstract)

[3] Zhu X Y, Shen D Y, Wang R P, et al. Maturity grading and identification of Camellia oleifera fruit based on unsupervised image clustering[J]. Foods, 2022, 11(23): 3800.

[4] 杜小强，宁晨，贺磊盈，等. 履带式高地隙油茶果振动采

收机设计与试验[J]. 农业机械学报，2022，53(7)：113-121.
DU Xiaoqiang, NING Chen, HE Leiying, et al. Design and test of crawler-type high clearance camellia oleifera fruit vibratory harvester[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(7): 113-121. (in Chinese with English abstract)

[5] 伍德林，袁嘉豪，李超，等. 扭梳式油茶果采摘末端执行器设计与试验[J]. 农业机械学报，2021，52(4)：21-33.
WU Delin, YUAN Jiahao, LI Chao, et al. Design and experiment of twist-comb end effector for picking camellia fruit[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(4): 21-33. (in Chinese with English abstract)

[6] 伍德林，李超，曹成茂，等. 摇枝式油茶果采摘装置作业过程分析与试验[J]. 农业工程学报，2020，36(10)：56-62.
WU Delin, LI Chao, CAO Chengmao, et al. Analysis and experiment of the operation process of branch-shaking type camellia oleifera fruit picking device[J]. Transactions of the Chinese Society of Agricultural Engineering (Transaction of the CSAE), 2020, 36(10): 56-62. (in Chinese with English abstract)

[7] Wu D L, Zhao E L, Fang D, et al. Determination of vibration picking parameters of camellia oleifera fruit based on acceleration and strain response of branches[J]. Agriculture, 2022, 12(8): 1222.

[8] Zhou Y H, Tang Y C, Zou X J, et al. Adaptive active positioning of camellia oleifera fruit picking points: Classical image processing and YOLOv7 fusion algorithm[J]. Applied Sciences, 2022, 12(24): 12959.

[9] Zhu X Y, Yu Y, Zheng Y L, et al. Bilinear attention network for image-based fine-grained recognition of oil tea (*camellia oleifera* abel.) cultivars[J]. Agronomy, 2022, 12(8): 1846.

[10] Wu D L, Jiang S, Zhao E L, et al. Detection of camellia oleifera fruit in complex scenes by using YOLOv7 and data augmentation[J]. Applied Sciences, 2022, 12(22): 11318.

[11] 陈志健，伍德林，刘路，等. 复杂背景下油茶果采收机重叠果实定位方法研究[J]. 安徽农业大学学报，2021，48(5)：842-848.
CHEN Zhijian, WU Delin, LIU Lu, et al. Research on overlapping fruit positioning method of camellia fruit harvester in complex background[J]. Journal of Anhui Agricultural University, 2021, 48(5): 842-848. (in Chinese with English abstract)

[12] 陈斌，饶洪辉，王玉龙，等. 基于 Faster-RCNN 的自然环境下油茶果检测研究[J]. 江西农业学报，2021，33(1)：67-70.
CHEN Bin, RAO Honghui, WANG Yulong, et al. Study on detection of camellia fruit in natural environment based on Faster-RCNN[J]. Acta Agriculturae Jiangxi, 2021, 33(1): 67-70. (in Chinese with English abstract)

[13] 宋怀波，王亚男，王云飞，等. 基于 YOLO v5s 的自然场景油茶果识别方法[J]. 农业机械学报，2022，53(7)：234-242.
SONG Huaibo, WANG Ya'Nan, WANG Yunfei, et al. Camellia oleifera fruit detection in natural scene based on YOLO v5s[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(7): 234-242. (in Chinese with English abstract)

[14] Xu Z B, Huang X P, Huang Y, et al. A real-time Zanthoxylum target detection method for an intelligent picking robot under a complex background, based on an improved YOLOv5s architecture[J]. Sensors, 2022, 22(2): 682.

[15] Shi R, Li T X, Yamaguchi Y. An attribution-based pruning method for real-time mango detection with YOLO network[J]. Computers and Electronics in Agriculture, 2020, 169: 105214.

[16] Zheng C, Chen P T, Pang J, et al. A mango picking vision algorithm on instance segmentation and key point detection from RGB images in an open orchard[J]. Biosystems Engineering, 2021, 206: 32-54.

[17] 刘洁，李燕，肖黎明，等. 基于改进 YOLOv4 模型的橙果识别与定位方法[J]. 农业工程学报，2022，38(12)：173-182.
LIU Jie, LI Yan, XIAO Liming, et al. Recognition and location method of orange based on improved YOLOv4 model[J]. Transactions of the Chinese Society of Agricultural Engineering (Transaction of the CSAE), 2022, 38(12): 173-182. (in Chinese with English abstract)

[18] 刘德儿，朱磊，冀炜臻，等. 基于 RGB-D 相机的脐橙实时识别定位与分级方法[J]. 农业工程学报，2022，38(14)：154-165.
LIU De'Er, ZHU Lei, JI Weizhen, et al. Real-time identification, localization, and grading method for navel oranges based on RGB-D camera[J]. Transactions of the Chinese Society of Agricultural Engineering (Transaction of the CSAE), 2022, 38(14): 154-165. (in Chinese with English abstract)

[19] Fu L S, Gao F F, Wu J Z, et al. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review[J]. Computers and Electronics in Agriculture, 2020, 177: 105687.

[20] Tu S Q, Pang J, Liu H F, 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.

[21] 王文杰，贡亮，汪韬，等. 基于多源图像融合的自然环境下番茄果实识别[J]. 农业机械学报，2021，52(9)：156-164.
WANG Wenjie, GONG Liang, WANG Tao, et al. Tomato fruit recognition based on multi-source fusion image segmentation algorithm in open environment[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(9): 156-164. (in Chinese with English abstract)

[22] Wang Y W, Chen Y F, Wang D F. Recognition of multi-modal fusion images with irregular interference[J]. PeerJ Computer Science, 2022, 8: e1018.

[23] Lv J D, Xu H, Xu L M, et al. Recognition of fruits and vegetables with similar-color background in natural environment: A survey[J]. Journal of Field Robotics, 2022, 39(6): 888-904.

[24] 黄彤镔，黄河清，李震，等. 基于 YOLOv5 改进模型的柑橘果实识别方法[J]. 华中农业大学学报，2022，41(4)：170-177.
HUANG Tongbin, HUANG Heqing, LI Zhen, et al. Citrus fruit recognition method based on the improved model of YOLOv5[J]. Journal of Huazhong Agricultural University, 2022, 41(4): 170-177. (in Chinese with English abstract)

[25] 何斌，张亦博，龚健林，等. 基于改进 YOLO v5 的夜间温室番茄果实快速识别[J]. 农业机械学报，2022，53(5)：201-208.
HE Bin, ZHANG Yibo, GONG Jianlin, et al. Fast recognition of tomato fruit in greenhouse at night based on improved YOLO v5[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(5): 201-208. (in Chinese

with English abstract)
[26] Yang R L, Hu Y W, Yao Y, et al. Fruit target detection based on BCo-YOLOv5 Model[J]. Mobile Information Systems, 2022, 2022: 1-8.
[27] 黄硕，周亚男，王起帆，等. 改进 YOLOv5 测量田间小麦单位面积穗数[J]. 农业工程学报，2022，38(16)：235-242.
HUANG Shuo, ZHOU Ya'nan, WANG Qifan, et al. Measuring the number of wheat spikes per unit area in fields using an improved YOLOv5[J]. Transactions of the Chinese Society of Agricultural Engineering (Transaction of the CSAE), 2022, 38(16): 235-242. (in Chinese with English abstract)
[28] 段洁利，王昭锐，邹湘军，等. 采用改进 YOLOv5 的蕉穗识别及其底部果轴定位[J]. 农业工程学报，2022，38(19)：122-130.
DUAN Jieli, WANG Zhaorui, ZOU Xiangjun, et al. Recognition of bananas to locate bottom fruit axis using improved YOLOv5[J]. Transactions of the Chinese Society of Agricultural Engineering (Transaction of the CSAE), 2022, 38(19): 122-130. (in Chinese with English abstract)
[29] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning[C]//AAAI Conference on Artificial Intelligence, Phoenix, USA, 2016.
[30] Wang D C, Chen X N, Yi H, et al. Improvement of non-maximum suppression in RGB-D object detection[J]. IEEE Access, 2019, 7: 144134-144143.
[31] Sun Q X, Chai X J, Zeng Z K, et al. Noise-tolerant RGB-D feature fusion network for outdoor fruit detection[J]. Computers and Electronics in Agriculture, 2022, 198: 107034.
[32] Bochkovskiy A, Wang C Y, Liao H M. YOLOv4: Optimal speed and accuracy of object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA: IEEE, 2020: 1-17.

Recognition of camellia oleifera fruits in natural environment using multi-modal images

ZHOU Hongping, JIN Shouxiang, ZHOU Lei, GUO Ziliang, SUN Mengmeng

(*College of Mechanical and Electronic Engineering*, *Nanjing Forestry University*, *Nanjing* 210037, *China*)

Abstract: An accurate and rapid identification can greatly contribute to the automated harvesting of Camellia oleifera fruits. However, Camellia oleifera grown in the natural environment has the dense branches and leaves, severely obstructed fruits, leading to the overlapping fruits. Only RGB images cannot fully meet the required effectiveness of the fruit recognition in modern agriculture. In this study, a dual backbone network model was proposed to combine the Red Green Blue-Depth (RGB-D) multi-modal images for the recognition and localization of Camellia oleifera fruits. Firstly, the lightweight improved YOLOv5s model was selected to detect the Camellia oleifera fruit targets. The YOLO-IR (YOLO-InceptionRes) was introduced the InceptionRes module into a feature extraction network for the multi-scale information fusion using four convolution operations of different sizes and concatenation. At the same time, the FPN (Feature Pyramid Network) + PAN (Path Aggregation Network) module of YOLOv5s was simplified into an FPN module to reduce the network complexity. Furthermore, the depth and width of the model were compressed to limit the model size for the smaller number of model parameters. The improved YOLO-IR was achieved in an average progress AP decrease of 0.2 percentage points, compared with the YOLOv5s, but the model size decreased by 69%. Provide support for building A lightweight dual backbone model was provided for the building support. Secondly, a dual backbone detection of Camellia oleifera fruit object, YOLO-DBM (YOLO-Dual Backbone Model) was constructed with the RGB-D images, according to the YOLO-IR. Two feature extraction networks were the same as the YOLO-IR to extract the color and depth features. An attention mechanism was constructed with the feature fusion module to fuse the color and depth features, Hierarchical fusion of color features and depth features at different scales. The attention module consisted of the spatial and channel attention mechanism. Specifically, the spatial attention mechanism was used to increase the weight of effective regions in the deep feature layer, but to reduce the interference of deep holes. Then, it was concatenated with the RGB feature layer. As such, the channel attention mechanism was used to emphasize the contribution of effective channels in the fused feature layer. Finally, the fused feature layer was input into the prediction module for the prediction. The experimental results show that the accuracy P, recall R, and average accuracy AP of the YOLO-DBM model using RGB-D images on the test set were 94.8%, 94.6%, and 98.4%, respectively. The average detection time for a single image was 0.016s. Compared with the YOLOv3, YOLOv5s, and YOLO-IR models, the average accuracy of AP was improved by 2.9, 0.1, and 0.3 percentage points, respectively, while the model size was only 6.21MB, which was only 46% of the YOLOv5s size. In addition, the accuracy P, recall R, and average accuracy AP increased by 0.2, 1.6, and 0.1 percentage points, respectively, compared with the YOLO-DBM model with the attention fusion module and the YOLO-DBM model with splicing fusion. The high effectiveness was also verified for the dual backbone network and attention fusion module. The finding can provide a strong reference and a new approach for the fruit recognition tasks in the oil tea fruit automatic harvesters.

Keywords: image recognition; deep learning; models; camellia oleifera; multi-modal; multi-scale fusion