杨国亮,王吉祥,聂子玲. 基于改进型 YOLOv5s 的番茄实时识别方法[J]. 江苏农业科学,2023,51(15):187 – 193.
doi:10.15889/j.issn.1002 – 1302.2023.15.026

基于改进型 YOLOv5s 的番茄实时识别方法

杨国亮,王吉祥,聂子玲
(江西理工大学电气工程与自动化学院,江西赣州 341000)

摘要:针对现有番茄检测精度低、没有品质检测和部署难度高等问题,提出基于 YOLOv5s 改进的番茄及品质实时检测方法,并与原始 YOLOv5 模型及其他经典模型进行对比研究。结果表明,针对番茄大小不同的问题,采用 K – Means ++ 算法重新计算先验锚框提高模型定位精度;在 YOLOv5s 主干网络末端添加 GAM 注意力模块,提升模型检测精度并改善鲁棒性;应用加权双向特征金字塔网络(BiFPN)修改原有结构,完成更深层次的加权特征融合;颈部添加转换器(transformer),增强网络对多尺度目标的检测能力。改进后的 YOLOv5s 番茄识别算法检测速度达到 72 帧/s。在测试集中对番茄检测均值平均精度(mAP)达到 93.9%,分别比 SSD、Faster – RCNN、YOLOv4 – Tiny、原始 YOLOv5s 模型提高 17.2、13.1、5.5、3.3 百分点。本研究提出的番茄实时检测方法,在保持检测速度的同时,可降低背景因素干扰,实现复杂场景下对番茄的精准识别,具有非常好的应用前景,为实现番茄自动采摘提供相应技术支持。

关键词:番茄检测;YOLOv5s;K – means ++ ;GAM 注意力模块;加权双向特征金字塔
中图分类号:TP391.41　　**文献标志码**:A　　**文章编号**:1002 – 1302(2023)15 – 0187 – 07

番茄作为世界上非常重要的蔬菜作物,每年全球的总产量可以达到 1.7 亿 t,其在蔬菜作物中常常位居前列。我国新鲜番茄的出产量常年居于全球首位,经过加工后的番茄产量则名列全球第二或第三[1]。随着我国社会老龄化程度的不断加深,用工难问题也日渐凸现了出来。在番茄生产及销售链中,采摘工作是一个非常重要的环节,目前采摘工作仍然是以人工采摘为主,无论是工作环境还是劳动强度都不尽人意,用时和用工成本也居高不下,番茄自动采摘应运而生[2-3]。国内外对果蔬自动采摘的研究大同小异,先通过深度学习进行图像识和定位,再通过执行机构进行采摘[4]。提高对番茄及其品质的检测,对采摘效率和存储运输都有非常重要的意义。

由于计算机科学的进步,基于卷积式神经网络的深度学习得以蓬勃发展。和传统机器学习相比,

收稿日期:2022 – 11 – 28
基金项目:江西省教育厅科技计划(编号:GJJ190450、GJJ180484)。
作者简介:杨国亮(1973—),男,江西宜春人,博士,教授,主要从事人工智能和模式识别研究。E – mail:ygliang30@126.com。
通信作者:王吉祥,硕士研究生,主要从事模式识别研究。E – mail:1661270181@qq.com。

[18]易　翔,张立福,吕　新,等. 基于无人机高光谱融合连续投影算法估算棉花地上部生物量[J]. 棉花学报,2021,33(3):224 – 234.
[19]陶惠林,冯海宽,徐良骥,等. 基于无人机高光谱遥感数据的冬小麦生物量估算[J]. 江苏农业学报,2020,36(5):1154 – 1162.
[20]周　萌,韩晓旭,郑恒彪,等. 基于参数化和非参数化法的棉花生物量高光谱遥感估算[J]. 中国农业科学,2021,54(20):4299 – 4311.
[21]石雅娇,陈鹏飞. 基于无人机高光谱影像的玉米地上生物量反演[J]. 中国农学通报,2019,35(17):117 – 123.
[22]邓　江,谷海斌,王　泽,等. 基于无人机遥感的棉花主要生育时期地上生物量估算及验证[J]. 干旱地区农业研究,2019,37(5):55 – 61,69.
[23]刘　杨,冯海宽,黄　珏,等. 基于无人机高光谱特征参数和株高估算马铃薯地上生物量[J]. 光谱学与光谱分析,2021,41(3):903 – 911.
[24]Dong J W,Xiao X M,Wagle P,et al. Comparison of four EVI – based models for estimating gross primary production of maize and soybean croplands and tallgrass prairie under severe drought[J]. Remote Sensing of Environment,2015,162:154 – 168.
[25]Majasalmi T,Rautiainen M,Stenberg P. M odeled and measured fPAR in a boreal forest:validation and application of a new model[J]. Agricultural and Forest Meteorology,2014,189/190:118 – 124.
[26]李龙伟. 基于时间序列遥感数据的毛竹林物候监测、分类和地上生物量估测研究[D]. 杭州:浙江农林大学,2020.
[27]朱吉祥. 基于光谱信息的夏玉米水氮状况诊断及产量评估[D]. 泰安:山东农业大学,2021.
[28]赵　涵. 杨树水力学特性与生长速率及生物量的关系[D]. 杨凌:西北农林科技大学,2021.

不论是在工作效率,还是在准确度方面,深度学习方法都有着巨大的优越性,使得基于深度学习方法的目标检测效率得以显著提高[5-9],同时也在农业相关方面得到广泛的应用。目标检测算法大致分为 2 种:一种是先生成候选框,再对候选框中的目标进行分类的 two - stage 目标检测方法,包括 R - CNN[10]、Fast - RCNN[11]、Faster - RCNN[12] 等。此类算法鲁棒性高,识别错误率较低,但其需要运行较长的时间,难以满足实际生产的实时性要求。例如,张文静等提出的改进 Faster R - CNN 算法对番茄的识别方法,检测每张样本需要 245 ms 的时间[13];龙洁花等提出改进 Mask R - CNN 的方法,以 CSP - Rest50 为骨干,识别准确率达到 90%[14]。另一种是不出现候选框的 one - stage 目标检测方法,包括 SSD[15] 和 YOLO[16] 等。此种方法不仅可以达到第 1 种方法的准确度,并且识别速度快,完全可以满足实时性的要求。例如,文斌等针对三七叶片病害改进 YOLOv3,提升了病害检测精度和鲁棒性[17];张兆国等提出对 YOLOv4 模型改良对复杂环境条件下的马铃薯进行测试,其检测准确率达到 91.4%[18];黄彤镔等针对柑橘识别改进 YOLOv5,添加注意力机制改善了遮挡问题[19]。

上述检测手段不能实现对果蔬真正的实时检测,检测效率低下,无法适应实际农业生产活动的需要,同时针对当前对成熟、未成熟和腐坏的番茄检测研究较少,本研究将以 YOLOv5s 算法为前提加以完善,通过融合注意力等新内容,提出一种改进型 YOLOv5s 的番茄识别方法,通过识别番茄品类自动采摘,降低采摘成本,研究结果将为实现番茄自动采摘提供技术支持。

1　材料与方法

1.1　数据集

本试验所用番茄图像数据集主要来源于公开数据集和实地拍摄。为了接近番茄生长的真实环境,图像数据包括番茄数量、密集度和遮挡度不同的各种情况,同时为了剔除腐败的番茄,减少养分的浪费,数据集还包括大量的腐败番茄的图像。尽可能保证数据的准确性,还需要人为进行标注,在标注的同时尽量将框内的背景减小到最小。为更好地模拟真实情况,对图像数据进行线性数据增强,通过旋转、缩放和添加噪声,增加样本的多样性。通过数据增强后得到 4 428 张图像,将数据集以 4∶1 的比例分割为训练集和验证集。番茄检测任务分为以下 3 类:1 类为成熟的番茄(Ripe_tomatoes),指可以进行采摘的番茄;2 类为未成熟的番茄(Unripe_Tomatoes),指不能进行采摘的番茄;3 类为腐败的番茄(Diseased),指需要采摘并丢弃的番茄。数据集类别标签数量见图 1。

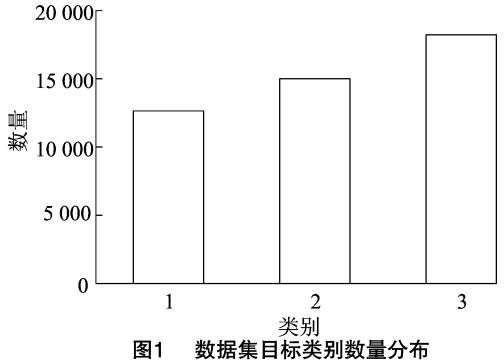


图1　数据集目标类别数量分布

1.2　YOLOv5s 网络模型

YOLOv5s 网络模型一般由输入端、躯干网络(backbone)、颈部(neck)和头部(head)4 个部分构成(图 2)。输入端通常由 3 个部分组成,分别为数据增强、图像锚框运算与缩放。主干网络主要由卷积(CONV)、卷积层与瓶颈层模块 C3 和空间金字塔池化(SPPF)构成,负责图像特征的获取。颈部通过金字塔构造实现特征融合。头部采用 CIOU_Loss 损失函数和非极大值抑制(non maxi - mum suppression,简称 NMS)进行预测。

1.3　模型改进

1.3.1　K - Means ++ 进行锚框优化　YOLOv5s 网络的初始先验锚框是通过 COCO 数据集得到的(表 1)。COCO 数据集共有 80 个类别,本研究中使用的数据集与之存在比较大的差异,最终会影响网络的整体性能。本研究采用了 K - Means ++ 算法对锚框进行聚类分析,相比于 K - Means 算法,它进一步优化了初始点的选取,首先通过随机选取一个样本作为聚类中心,随后再计算每个样本到达聚类中心的最短距离,然后再计算出每个样本被选为后一个聚类中心的概率,概率公式为

$$P = \frac{D(x_i)^2}{\sum_{i=1}^{n} D(x_i)^2}。\tag{1}$$

其中:$D(x_i)$ 表示第 i 个样本与当前已有聚类中心之间的最短距离;n 为样本总数;P 表示每个样本点被选为下一个聚类中心的概率。

通过 K - Means ++ 聚类算法,产生不同大小和

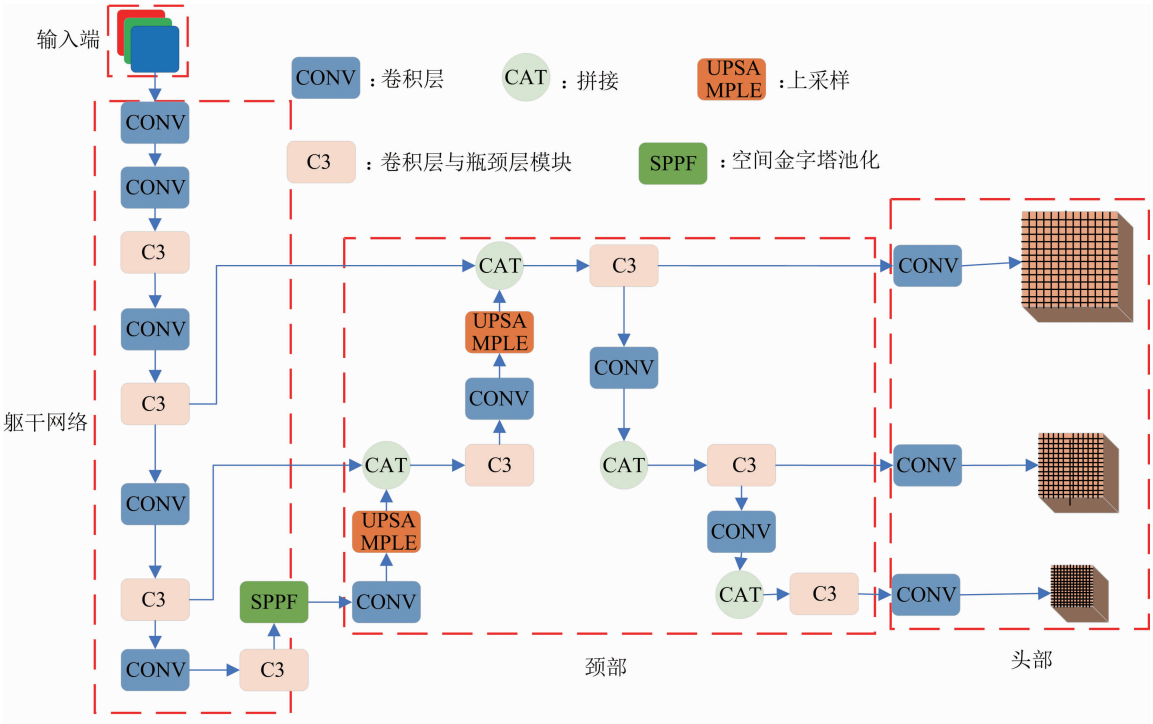


图2　YOLOv5s 网络结构图

表 1　原始锚框

特征图尺度	锚框 1	锚框 2	锚框 3
小尺度	(10,13)	(16,30)	(33,23)
中尺度	(30,61)	(62,45)	(59,119)
大尺度	(116,90)	(156,198)	(373,326)

数量的先验锚框，使之尽可能与实际目标框相匹配，从而提高系统检测的准确度，最终确定的锚框尺寸见表 2。

表 2　改进后锚框

特征图尺度	锚框 1	锚框 2	锚框 3
小尺度	(33,41)	(60,117)	(72,60)
中尺度	(103,152)	(132,88)	(138,246)
大尺度	(185,152)	(246,232)	(282,338)

1.3.2　引入 Vision Transformer 转换器　(transformer)已成为自然语言处理方面的主流模型，在图像处理方面更是大放异彩。在目前以卷积神经网络为核心的电脑视觉技术任务的重大背景下，Vision Transformer(ViT)的应用对卷积神经网络的地位产生了冲击。Dosovitskiy 等将一个图像分割成数个固定大小的图像块，并将其编码成序列向量作为 transformer 输入，成功解决图像处理领域在 transformer 中的输入问题。同时经过试验证明，当预训练数据更丰富时，transformer 在图像处理领域的性能会超越卷积神经网络[20]。本试验所用番茄图像包括尺度不同的目标，故在检测网络中融入 transformer 模块解决尺度问题，ViT 图像处理流程如图 3 所示。

ViT 和普通 Transformer 在输入上有所区别，后者将标记嵌入的一维序列作为输入，而前者在处理二维图形时，要把图形 $x \in \mathbb{R}^{H\times W\times C}$ 重塑为一组二维的扁平序列 $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$，$\mathbb{R}$ 表示维度，H 和 W 是原始图形的高和宽，C 是图形通道数量，P 是每个图形块的高宽，$N=\mathrm{HW}/P^2$ 既是图形块的总量，又是 ViT 输入序列的有效长度。从 ViT 的每个层中产生一个恒定维度为 D 的特征向量，通过利用可训练的线性投影可以把找平的像素块映射到 D 维度上，如公式(2)所示。随后在图像序列($z_D^0=x_{\mathrm{class}}$)前加入一个具有学习能力的嵌入，其在 Transformer 编码器输出时的状态 z_L^0 用 y 作图像表示，如公式(5)所示。

$$z^0=[x_{class};x_p^1E;x_p^2E;\cdots;x_p^NE]+E_{pos},\quad E\in R^{(P^2\cdot C)\times D},E_{pos}\in \boldsymbol{R}^{(N+1)\times D};\tag{2}$$

$$z_l'=MSA[LN(z_{l-1})]+z_{l-1},l=1\cdots L;\tag{3}$$

$$z_l=MLP[LN(z_l')]+z_l',l=1\cdots L;\tag{4}$$

$$y=LN(z_L^0)。\tag{5}$$

其中：E 表示线性变换；E_{pos} 表示在 pos 处的线性变换；z_l 表示第几个图像序列；z_l' 表示操作完成后的第几个序列；MSA 表示多头部自注意力；LN 表示归一

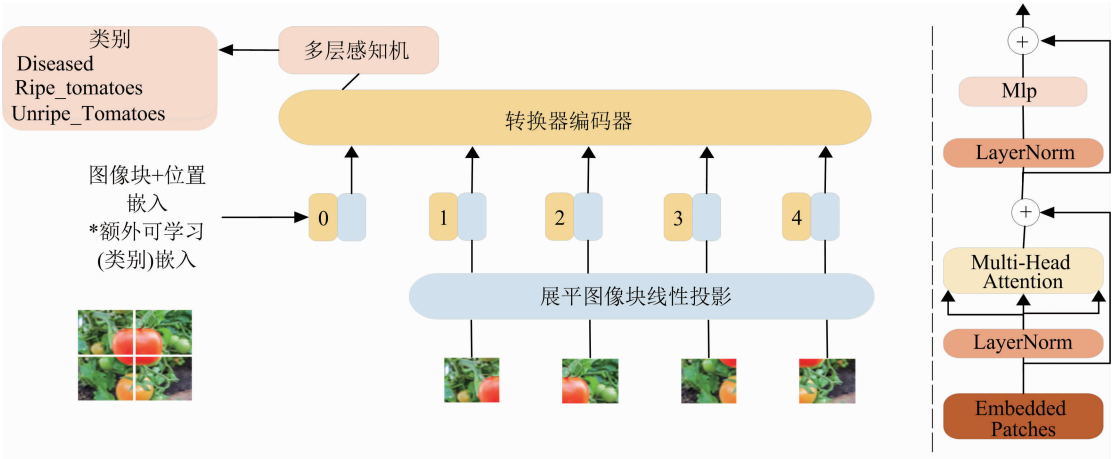


图3　ViT 图像处理流程

化;MLP 表示多层感知机。

但 Transformer 也有不足之处,提取到的特征鲁棒性较弱,经过研究证明,卷积神经网络能够通过 Transformer 提高性能。本研究通过将 C3 模块中的 BottleNeck 替换为 TransformerBlock 实现二者的有机结合构成 C3TB,C3 和 C3TB 结构如图 4 所示。

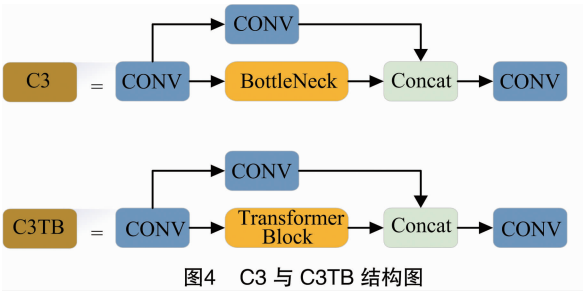


图4　C3 与 C3TB 结构图

1.3.3　添加 GAM 注意力模块　注意力机制的添加能使网络关注到图像中的关键点,有助于提高番茄检测任务的性能。不论是挤压激励网络(squeeze and excitation network,简称 SENet),还是之后的卷积注意力模块(convolutional block attention module,简称 CBAM),都没有注意到空间 - 通道之间的相互作用,而削弱了跨纬度的交互。鉴于上述问题,本研究在 Backbone 末端使用全局注意力机制(global attention mechanism,简称 GAM)[21],使网络关注更重要的区域,减少背景因素的影响,保留更多的特征信息,提升网络检测准确度,GAM 模块整体结构如图 5 所示。

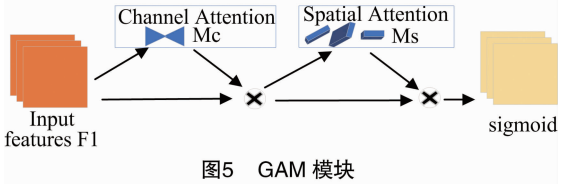


图5　GAM 模块

输入特征先经过通道注意力进行校正,再通过空间注意力继续校正。图 6 是通道注意力结构图。首先将纬度大小为 $C \times W \times H$ 的输入特征经过三维排列保存 3 个纬度上的信息,其中 C 是特征通道数量,W 和 H 分别是输入特征的宽和高。随后将输出信息通过 2 层的多层感知器,第 1 层将 C 压缩为 C/R,R 为压缩比,再经由第 2 层恢复到 C,最后再经由反三维排列操作,通过 Sigmoid 激活函数得到一个新的特征图。

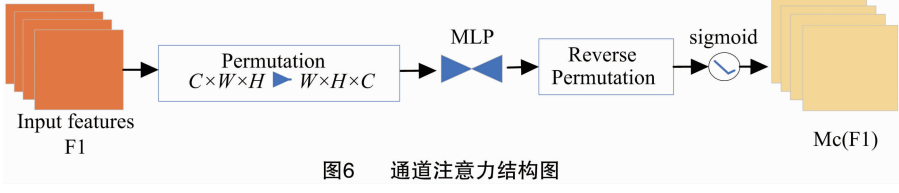


图6　通道注意力结构图

图 7 是空间注意力结构图,输入特征纬度大小为 C × W × H,通过 2 个卷积核为 7 × 7 的卷积层,实现空间信息的融合,同时进行通道的编码和解码操作,然后通过 Sigmoid 激活函数得到新的特征图。

1.3.4　特征金字塔网络改进　在卷积神经网络中,图像特征容易受浅层网络的影响,而语义特征容易受深层网络的影响,从而在目标检测中因卷积神经网络的这个特性而影响精度。根据这些现象,特征

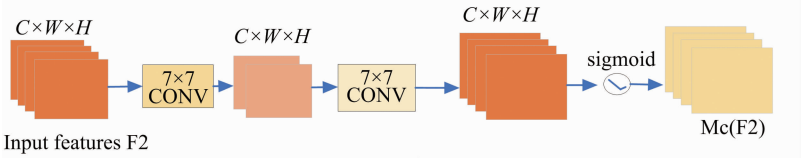


图7　空间注意力结构图

金字塔网络(feature pyramid networks,简称 FPN)随之被提出,结构如图 8 - a 所示,通过使不同维度的特征图中含有尽可能多的语义信息,再通过上采样把顶层信息与下层信息加以结合,从而实现相应的目的,并且每层都是独立进行预测。但是 FPN 这种设计有种缺陷,只增加了特征图的语义信息,定位信息并不能进行传输。为了解决相关问题,又建立了一个由底往顶的金塔,即与 FPN 操作相反的路径聚合网络(path aggregation network,PANet),结构如图 8 - b 所示。通过 2 种结构的结合,检测精度有了明显的提升。

加权双向特征金字塔网络(bidirectional feature pyramid network,简称 BiFPN)最先在 EfficientDet 中被提到,通过在输入与输出节点中间增加一个直接相连路径,可以使得在不提高计算量的前提下,能够融入更多需要的特性。与 PANet 中仅有一条自顶向下和一条自底向上路线有所不同的是,把所有双向路线视作一条特征网络层,并多次重复同一层来进行更深层次的特性融合,BiFPN 如图 8 - c 所示。在此操作中加快了计算的速度,如公式(6)所示:

$$O = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_j} \times I_i 。 \quad (6)$$

式中:权重 $\omega_i \geqslant 0$,$\omega_j \geqslant 0$;I_i 为输入其中的特征;ε 表示学习率;O 表示结果。鉴于标量权重没有边界,为保证训练稳定,应用 softmax 实现归一化运算。把 Backbone 中 P_3、P_4、P_7 这 3 个不同尺度的特征都输入到 BiFPN 中,然后即可建立 20×20、40×40、80×80 这 3 个纬度的预测分支。以 P_6 节点为例说明融合过程,如下所示:

$$P_6^{td} = \mathrm{Conv}\left[\frac{\omega_1 P_6^{in} + \omega_2 \mathrm{Resize}(P_7^{in})}{\omega_1 + \omega_2 + \varepsilon}\right]; \quad (7)$$

$$P_6^{\mathrm{out}} = \mathrm{Conv}\left[\frac{\omega'_1 P_6^{in} + \omega'_2 P_6^{td} + \omega'_3 \mathrm{Resize}(P_5^{out})}{\omega'_1 + \omega'_2 + \omega'_3 + \varepsilon}\right]。 \quad (8)$$

式中:P_6^{td} 表示第 6 节点自顶向底的中间特征;P_6^{in} 表示第 2 节点输入的特征;P_7^{in} 表示第 7 节点输入的特征;P_5^{out} 表示第 5 节点自底向顶的输出特征;P_6^{out} 表示第 6 节点自底向顶的输出特征;Resize 表示上取样或下取样;Conv 表示卷积处理。根据上述优势,把 YOLOv5s 模型里的金字塔模块修改为 BiFPN,以增强特征融合,并提高测速率。

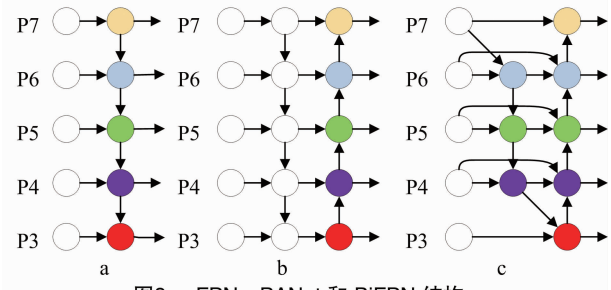


图8　FPN、PANet 和 BiFPN 结构

1.4　试验环境

本试验使用的运行系统为 Windows 10,并使用了 Pytorch 作为深度学习结构,详细试验环境设置见表 3。训练时优化器使用随机梯度下降法(stochastic gradient descent,简称 SGD),初始的学习率参数调整为 0.01,余弦退火超参数设置成 0.1,动量因子调整为 0.937,权重衰减系数最终确定为 0.000 5。网络图像输入大小为 640×640,Batchsize 设置为 16,总训练 300 个 epoch。此次试验于 2022 年 11 月 1 日在江西理工大学电气学院 315 实验室完成。

表 3　试验环境配置

项目	配置
CPU	Intel® Core™ i9 - 9900CPU @ 3.10 GHz 16 G
GPU	GeForce RTX 2070SUPER 8 G
系统环境	Windows 10
框架	Pytorch 1.11.0
语音	Python 3.8
加速环境	CUDA 11.3

1.5　评价指标

基于量化判断方法并分析试验结论,本研究选择在目标测试中使用的精度(precision,简称 P)、召回率(recall,简称 R)和均值平均精度(mean average precision,简称 mAP)作为相关衡量指标。P 是用来表示真正的正样本在检测结果为正样本中所占的

比例，R 是表示被检测到的正样本在真正的正样本中的占比，mAP 表示各个类别平均精度的均值，相关公式如下所示：

$$P = \frac{TP}{TP + FP}; \tag{9}$$

$$R = \frac{TP}{TP + FN}; \tag{10}$$

$$A_P = \int_0^1 P(r)\,\mathrm{d}r; \tag{11}$$

$$mAP = \frac{1}{C}\int_0^1 P(r)\,\mathrm{d}r。 \tag{12}$$

式中：TP 为正确分配的正样本，即番茄成熟并且检测结果正确；FP 为分配错误的正样本，即番茄成熟但被检测为不成熟或者腐败的；FN 为分类错误的负样本；A_P 表示平均精准度；C 为类别数。

2　结果与分析

2.1　训练结果

将原始模型与改进后的模型在相同环境下训练 300 轮，mAP 曲线对比如图 9 所示，橘色曲线为改进前，蓝色曲线为 YOLOv5s 改进后。其中横坐标为 300 轮训练次数，纵坐标为 mAP。由图 9 可知，在训练 30 轮前模型收敛速度极快，经过 100 轮训练 2 个模型都趋于稳定，同时改进后的模型在 mAP 上相较于原模型得到明显提升，表明模型改进可行。

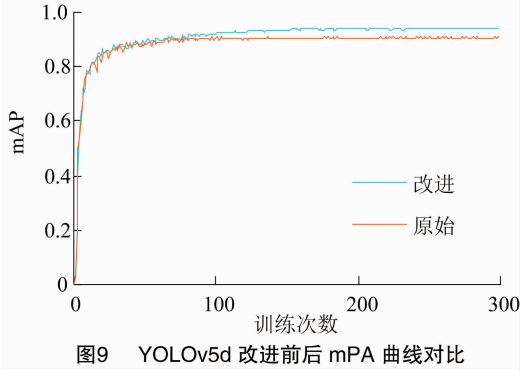


图9　YOLOv5d 改进前后 mPA 曲线对比

2.2　对比试验

为进一步评价本研究中改进方法对番茄的检验能力，本试验将经过优化的 YOLOv5s 与 SSD、Faster－RCNN、YOLOv4－tiny 以及 YOLOv5s 目标检测方法进行比较，并采用相同的数据划分和试验设置。由表 4 可知，改进的 YOLOv5s 算法在均值平均精度和召回率上比其他算法有更好的表现，相较于 YOLOv5s，分别提升了 3.3、5.6 百分点，分别达到了 93.9% 和 92.7%。同时，由于本算法召回率的提升，导致会对每幅图像检测更多的目标，从而帧率下降了 18，但仍快于 SSD、Faster－RCNN 和 YOLOv4－tiny，满足实时性的要求。

表 4　试验对比结果

模型	mAP（%）	R（%）	帧率（帧/s）
SSD	76.7	73.9	18
Faster－RCNN	80.8	88.5	9
YOLOv4－tiny	88.4	85.3	52
YOLOv5s	90.6	87.1	90
改进的 YOLOv5s	93.9	92.7	72

2.3　消融试验

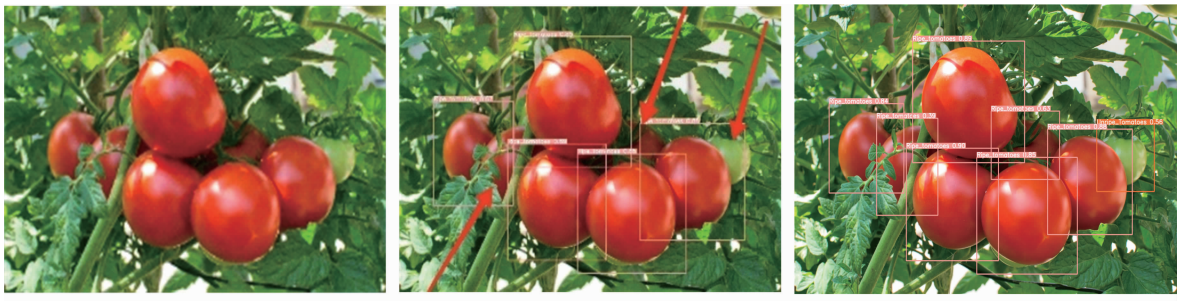
对经过优化的 YOLOv5s 模型，通过消融对比试验结果来证明每个改进模块对模型的优化效果，试验结果见表 5。其中改进模型 1 是通过使用 K－means ++ 修改了先验锚框，从而使该锚框的匹配性提高，均值平均精度也增加了 1.3 百分点；改进模型 2 是改变金字塔网结构为加权双向金字塔网络，均值平均精度增加 1.7 百分点；改进模型 3 是改变主干网络增加 GAM 注意力，均值平均精度增加 2.5 百分点；改进模型 4 是改变颈部网络 C3 结构为 C3TB，均值平均精度增加 2.1 百分点。把 4 个优化方案同时融入到一个模型，均值平均精度相较于原 YOLOv5s 模型整体增加 3.3 百分点。

表 5　消融试验结果

模型	K－means ++	BiFPN	GAM	C3TB	mAP(0.5)（%）
YOLOv5s	×	×	×	×	90.6
改进模型 1	√	×	×	×	91.9
改进模型 2	×	√	×	×	92.3
改进模型 3	×	×	√	×	93.1
改进模型 4	×	×	×	√	92.7
改进的 YOLOv5s	√	√	√	√	93.9

2.4　试验结果分析

为更好地检验经优化后的 YOLOv5s 方法的测试效果，选择了测试集中的一些图片进行了检测，番茄测试效果如图 10 所示，图 10－a 是原始图像；图 10－b 是原始 YOLOv5s 算法的检测结果，其中红色箭头表示漏检的番茄；图 10－c 是优化后 YOLOv5s 方法的测试结果。通过图 10－b 和图 10－c 对比可知，原始 YOLOv5s 算法对图 10－b 中红色箭头所指番茄漏检，改进后的 YOLOv5s 算法能准确地检测出这些目标，并且置信度得到提高，能



a. 原图　　b. YOLOv5s　　c. 改进的 YOLOv5s

图10　YOLOv5s 改进前后检测效果对比

够捕捉到关键信息进而对遮挡目标也有较好的检测效果。

3　讨论与结论

本研究根据目前对番茄的传统检测方法以及对密集目标漏检的测量精度较差的情况，给出一个更完善的 YOLOv5s 检测模型。通过使用 K－means ++ 算法对自制番茄数据集提高先验锚框匹配度、对 YOLOv5s 主干网络增加注意力模块、设计 C3TB 模块替换 C3 模块、优化特征金字塔网络等提高模型的检测能力。通过对比试验证明，完善后的 YOLOv5s 模型相比于原始的模型，mAP 提升了 3.3% 且置信度更高，对遮挡目标的辨识度提高减少了漏检，虽然检测速率有所下降，但本模型精度能够满足实际采摘的需求，为番茄自动采摘提供技术支持。

参考文献：

[1]李君明，项朝阳，王孝宣，等. “十三五”我国番茄产业现状及展望[J]. 中国蔬菜，2021(2)：13－20.

[2]王海楠，弋景刚，张秀花. 番茄采摘机器人识别与定位技术研究进展[J]. 中国农机化学报，2020，41(5)：188－196.

[3]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.

[4]阮承治，赵德安，陈　旭，等. 双指型农业机器人抓取球形果蔬的控制器设计[J]. 中国农机化学报，2019，40(11)：169－175.

[5]陈科圻，朱志亮，邓小明，等. 多尺度目标检测的深度学习研究综述[J]. 软件学报，2021，32(4)：1201－1227.

[6]赵立新，邢润哲，白银光，等. 深度学习在目标检测的研究综述[J]. 科学技术与工程，2021，21(30)：12787－12795.

[7]包晓敏，王思琪. 基于深度学习的目标检测算法综述[J]. 传感器与微系统，2022，41(4)：5－9.

[8]邵延华，张　铎，楚红雨，等. 基于深度学习的 YOLO 目标检测综述[J]. 电子与信息学报，2022，44(10)：3697－3708.

[9]李　萍，邵　彧，齐国红，等. 基于跨深度学习模型的作物病害检测方法[J]. 江苏农业科学，2022，50(8)：193－199.

[10]Girshick R，Donahue J，Darrell T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus，2014：580－587.

[11]Girshick R. Fast R－CNN[C]//2015 IEEE International Conference on Computer Vision. Santiago，2016：1440－1448.

[12]Ren S Q，He K M，Girshick R，et al. Faster R－CNN：towards real－time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(6)：1137－1149.

[13]张文静，赵性祥，丁睿柔，等. 基于 Faster R－CNN 算法的番茄识别检测方法[J]. 山东农业大学学报(自然科学版)，2021，52(4)：624－630.

[14]龙洁花，赵春江，林　森，等. 改进 Mask R－CNN 的温室环境下不同成熟度番茄果实分割方法[J]. 农业工程学报，2021，37(18)：100－108.

[15]Liu W，Anguelov D，Erhan D，et al. SSD：single shot multiBox detector[C]//European Conference on Computer Vision. Cham：Springer，2016：21－37.

[16]Redmon J，Divvala S，Girshick R，et al. You only look once：unified，real－time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，NV，USA. IEEE，2016：779－788.

[17]文　斌，曹仁轩，杨启良，等. 改进 YOLOv3 算法检测三七叶片病害[J]. 农业工程学报，2022，38(3)：164－172.

[18]张兆国，张振东，李加念，等. 采用改进 YoloV4 模型检测复杂环境下马铃薯[J]. 农业工程学报，2021，37(22)：170－178.

[19]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.

[20]Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16x16 words：transformers for image recognition at scale[EB/OL]. 2020：arXiv：2010.11929. https://arxiv.org/abs/2010.11929.

[21]Liu Y C，Shao Z R，Hoffmann N. Global attention mechanism：retain information to enhance channel－spatial interactions[EB/OL]. 2021：arXiv：2112.05561. https://arxiv.org/abs/2112.05561.