doi:10.6041/j.issn.1000-1298.2024.10.006

面向樱桃番茄采摘识别的轻量化 Transformer 架构优化研究

赵　博[1,2]　柳苏纯[1,2]　张巍朋[1,2]　朱立成[1,2]　韩振浩[1,2]　冯旭光[1,2]　王瑞雪[1,2]

(1.中国农业机械化科学研究院集团有限公司,北京 100083;2.农业装备技术全国重点实验室,北京 100083)

摘要:为进一步提升穗收型樱桃番茄识别准确率和识别速度,实现设施环境番茄自动采摘,提出了一种基于改进 Transformer 的轻量化樱桃番茄穗态识别模型。首先,构建了包含不同光照环境和采摘姿态的樱桃番茄数据集,并对樱桃番茄果穗姿态进行了划分。然后,提出了一种基于改进 RE-DETR 的轻量化穗收樱桃番茄识别模型,通过引入一个轻量级的骨干网络 EfficientViT 替换 RT-DETR 原有的骨干网络,显著减少了模型参数和计算量;同时设计了一个自适应细节融合模块,旨在高效处理并融合不同尺度特征图,并进一步降低计算复杂度。最后,引入加权函数滑动机制和指数移动平均思想来优化损失函数,来处理样本分类中的不确定性。实验结果表明,该轻量化模型在保持高识别准确率(90%)的同时,实现了快速检测(41.2 f/s)和低计算量(8.7×10^9 FLOPs)。与原始网络模型、Faster R-CNN 和 Swin Transformer 相比,平均识别准确率提高 1.24% ~15.38%,每秒处理帧数(FPS)提高 25.61% ~255.17%,同时浮点运算量实现了 69.37% ~92.37% 的大幅降低。该模型在综合性能上有着较强的鲁棒性,兼顾了精度与速度,可为番茄采摘机器人完成视觉任务提供技术支撑。

关键词:樱桃番茄;目标检测;深度学习;RT-DETR;轻量化;Transformer

中图分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1000-1298(2024)10-0062-10　　**OSID**:

Performance Optimization of Lightweight Transformer Architecture for Cherry Tomato Picking

ZHAO Bo[1,2]　LIU Suchun[1,2]　ZHANG Weipeng[1,2]　ZHU Licheng[1,2]　HAN Zhenhao[1,2]
FENG Xuguang[1,2]　WANG Ruixue[1,2]

(1. *Chinese Academy of Agricultural Mechanization Sciences Group Co., Ltd., Beijing* 100083, *China*
2. *China National Key Laboratory of Agricultural Equipment Technology, Beijing* 100083, *China*)

Abstract: To further improve the recognition accuracy and speed of truss-harvested cherry tomatoes, targeting the scenario of automated tomato harvesting in facility environments, a lightweight cherry tomato truss recognition model was proposed based on an improved transformer. Firstly, a cherry tomato dataset encompassing various lighting conditions and harvesting postures was constructed, and the postures of cherry tomato trusses were categorized. Then a lightweight truss-harvested cherry tomato recognition model based on an improved RE-DETR was proposed. This model introduced a lightweight backbone network, EfficientViT, to replace the original backbone of RT-DETR, which significantly reduced model parameters and computational complexity. Additionally, an adaptive detail fusion module was designed to efficiently process and merge feature maps of different scales while further lowered computational complexity. Finally, a weighted function sliding mechanism and exponential moving average concept were introduced to optimize the loss function, which addressed uncertainties in sample classification. Experimental results demonstrated that this lightweight model achieved high recognition accuracy (90.00%) while enabled fast detection (41.2 f/s) and low computational cost (8.7×10^9 FLOPs). Compared with that of the original network model, Faster R-CNN, and Swin Transformer, the average recognition accuracy was improved by 1.24% ~15.38%, the frames processed per second

收稿日期:2024-06-03　修回日期:2024-07-08

基金项目:山东省重点研发计划项目(2023CXGC010715)和中国机械工业集团有限公司科技专项(ZDZX2023-2)

作者简介:赵博(1981—),男,研究员,博士生导师,主要从事智能农业机器人研究,E-mail:zhaoboshi@126.com

通信作者:朱立成(1984—),男,正高级工程师,主要从事农业机械化技术研究,E-mail:zhulicheng@caams.org.cn

(FPS) was increased by 25.61% ~255.17%, while simultaneously achieved a substantial reduction of 69.37% ~ 92.37% in floating-point operations. The model exhibited strong robustness in overall performance, balancing accuracy and speed, and can serve as a reference for tomato harvesting robots in completing visual tasks.

Key words: cherry tomatoes; object detection; deep learning; RT - DETR; lightweight; Transformer

0　引言

樱桃番茄的生长周期较短,在温室中进行多周期栽培可一年内收获多批次,具有较高的经济价值[1]。随着农业劳动力市场紧缩和农村人口减少的影响,樱桃番茄产业的发展受到人工采摘效率低下和成本高昂的制约。因此研发樱桃番茄采摘机器人有助于推进农业向可持续、高效和智能化方向发展[2]。

农业采摘机器人作为用先进技术、装备和科技应用于传统农业装备的产物,与新质生产力发展密不可分[3]。在机器人系统中,视觉系统发挥着至关重要的作用。通过处理深度相机捕捉的数据,机器人能够精确地识别出樱桃番茄的成熟状态以及空间位置等关键信息,从而确保机器人可以准确地采摘目标果实。传统的图像处理方法主要依赖于手工设计的特征提取技术,如颜色特征、边缘检测、纹理特征和局部特征等。冯青春等[4]通过分析成熟番茄果串图像色彩特征,采用 $R-G$ 色差模型强调目标与背景之间的差异,并通过色差灰度逐列统计来锁定果串图像区域。尽管传统方法在一定程度上能够有效地解决一些图像处理问题,但其局限性在于难以捕捉复杂的图像模式,对图像的变形和光照变化不够鲁棒。

为了克服这些局限性,研究者们开始探索基于深度学习的图像处理方法[5-6]。卷积神经网络(Convolution neural network, CNN)通过多个卷积层和池化层提取图像的局部特征[7-9],并利用全连接层将这些局部特征组合成高维的特征表示,从而直接从特征图上检测物体并计算损失函数[10]。荣佳诚[11]采用 YOLO 算法对图像中的番茄果实和番茄串进行识别与定位,并通过聚类判定和番茄串内最近距离方法,优化了机器人的采摘顺序。李兴旭等[12]提出了一种联级的视觉检测系统,该系统引入 YOLO v5 目标检测模型以检测番茄串和番茄果实,并利用 MobileNetv3 网络模型进行果实与果梗相对位置关系判断。该系统平均推理用时仅为 22 ms,且对樱桃番茄果实和果串的平均检测精度达到 89.9%。TAEHYEONG 等[13]提出了基于 OpenPose 的番茄植株关键点检测方法,通过引入额外的深度图,实现了对关键点检测结果的多番茄姿态估计。QI 等[14]提出了一个基于 YOLO v8 的检测跟踪框架,该框架能够同时估计温室中樱桃番茄串的数量和成熟度水平。

DETR[15]首次将 Transformer 机制[16](一种在自然语言处理领域广泛应用的架构)引入目标检测任务中,相较于基于卷积神经网络的方法,其网络结构更为简洁。在 DETR 中,编码器利用注意力机制提取深层次的图像特征,并将这些特征传递给解码器。在解码器阶段,一组可学习的目标查询(Object query)通过自注意力机制和交叉注意力机制与图像特征进行交互,其输出通过二分图匹配(Bipartite matching)算法与真实标签(Ground truth, GT)实现匹配,从而完成正负样本的识别。目前,基于 Transformer 的樱桃番茄识别与定位研究较少,基于 Transformer 的模型参数量较大,不易部署。因此,本文提出一种基于 RT - DETR 轻量化的樱桃番茄模型,主要用于穗收型樱桃番茄采摘姿态的识别。对樱桃番茄整穗的姿态进行简单的划分,并建立包含不同光照环境和不同采摘姿态的樱桃番茄数据集;针对 Transformer 模型参数大等问题,对 RT - DETR 模型进行轻量化处理,引入 EfficientViT 骨干网络替换原来的骨干,以优化模型的计算和参数效率;为增强模型在处理细节信息方面的能力,拟改进原有的编码器结构,在多尺度融合模块 CCFM 中引入一个全新的自适应细节融合(ADI)模块,以增强模型对于不同姿态和尺寸果穗的适应性;拟对损失函数进行改进,通过指数移动平均机制动态调整交并比(Intersection over union,IoU)阈值,以显著提高模型对不同姿态果穗的检测准确性和鲁棒性。

1　材料与方法

1.1　樱桃番茄果穗姿态划分

樱桃番茄的收获模式主要分为散粒收获和整穗收获两种,其中整穗收获能够一次性采摘一整穗番茄,既降低了劳动成本,又提高了收获效率,更适用于机械化、自动化的收获系统。樱桃番茄在采摘时,果农一般贴着主茎对花序主轴进行剪切,如图 1 所示。

樱桃番茄果穗的悬挂在主茎上表现出随机性,

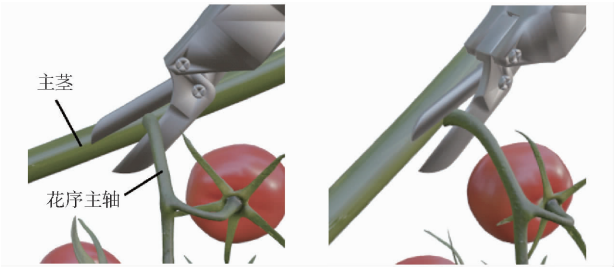

图 1 樱桃番茄人工剪切方式示意图

Fig. 1 Schematic of artificial cutting method for cherry tomatoes

导致其在自动化采摘过程中呈现多样化的姿态。为了更好地描述这些姿态,本研究将花序轴的第 1 个分岔点到主茎这段花序主轴定义为剪切区(图 2),并将其方向作为关键的参考。根据剪切区的方向,将樱桃番茄的串收采摘姿态分为 4 种:①L(Left):花序轴、果实和萼片之间存在前后遮挡,剪切区向左倾斜。②R(Right):花序轴、果实和萼片之间存在前后遮挡,剪切区向右倾斜。③B(Backward):花序轴和萼片无遮挡,可直接被观察到,果实被萼片遮挡,剪切区向后倾斜。④F(Forward):能够观察到花序轴、果实,但萼片被果实遮挡,主茎对剪切区有一定的遮挡,剪切区向前倾斜。

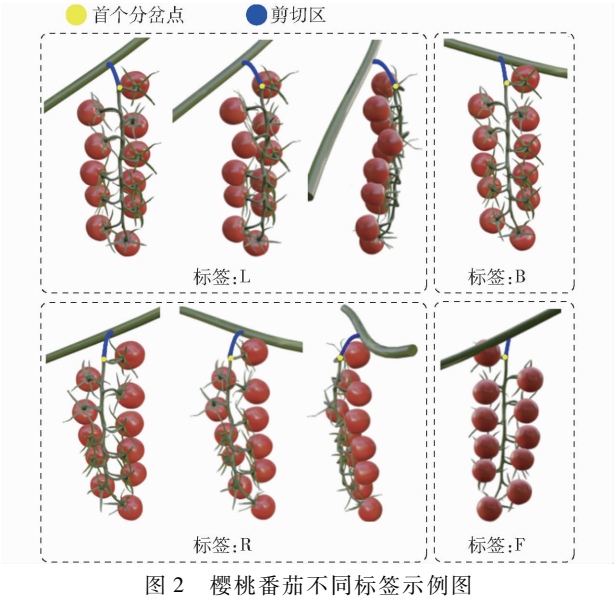

图 2 樱桃番茄不同标签示例图

Fig. 2 Example of different labels for cherry tomatoes

1.2 试验数据采集

试验数据采集于北京市平谷区某智慧蔬菜工场,所采集的樱桃番茄品种为德澳特,该品种果穗长而一致,可进行穗收或粒收。图像采集设备包括 1 台 DJ Osmo Pocket 型相机和 1 台 Intel RealSense D405 型相机,其参数如表 1 所示,其中由 DJ Osmo Pocket 型相机拍摄的图像分辨率为 4 000 像素 × 3 000 像素,由 Intel RealSense D405 型相机拍摄的图像分辨率为 1 280 像素 ×720 像素。在图像采集过程中,研究者在不同时间(上午、中午、下午)和光照条件下,由人工手持设备对番茄果穗进行多角度拍摄,以降低光照条件对识别准确率的影响。共采集 973 幅图像,涵盖了顺光、逆光、单一目标和多目标等不同类型的图像。为后续模型训练和评估提供了充足的样本,有助于全面评估模型在不同场景下的性能表现。

表 1 图像采集设备参数

Tab. 1 Image acquisition device parameters

规格	DJ Osmo Pocket 型相机	Intel RealSense D405 型相机
尺寸/(mm × mm × mm)	121.9 × 36.9 × 28.6	42 × 42 × 23
质量/g	116	60
最大分辨率/(像素 × 像素)	4 000 × 3 000	1 280 × 720
RGB 传感器	彩色 CMOS 传感器	彩色 CMOS 传感器
深度传感器	无	主动红外结构光
FOV 场视角/(°)	80(水平)	69 × 42(水平 × 垂直)

1.3 数据集构建

为确保养分集中供应至果穗,在结果期会对番茄植株采取枝叶修剪措施。修剪后,果穗裸露在外,有利于机器人视觉系统采集数据。对采集到的图像采用色彩变换、噪声等手段进行数据增强,得到共计 2 750 幅樱桃番茄图像,所有图像的分辨率均设置为 1 280 像素 × 720 像素。其中按照比例 7∶2∶1 将 1 925 幅图像划分为训练集,550 幅图像划分为验证集,275 幅图像划分为测试集。

采用标注软件 Labelme 对数据集中的樱桃番茄果穗进行采摘姿态标注,由于采摘时需要机器人识别到每穗的花序主轴。因此对于花序主轴被遮挡的果穗不予标注。表 2 展示了数据集中 4 种采摘姿态的标签数量分布。图 3 展示了数据集中不同类别边界框大小分布。

表 2 番茄果穗姿态的划分与标注数量

Tab. 2 Division and annotation number of tomato ear pose in dataset

标签	训练集	验证集	测试集	总计
L	967	244	114	1 325
R	849	118	84	1 051
F	829	99	60	988
B	811	121	78	1 010
总计	3 456	582	336	4 374

1.4 樱桃番茄目标检测模型

针对穗收型樱桃番茄在采摘过程中的目标姿态识别任务需求,在现有研究的基础上,提出了一种新型的轻量级端到端实时目标检测模型。基于 RT-DETR(Real time detection transformer)框架[17],通

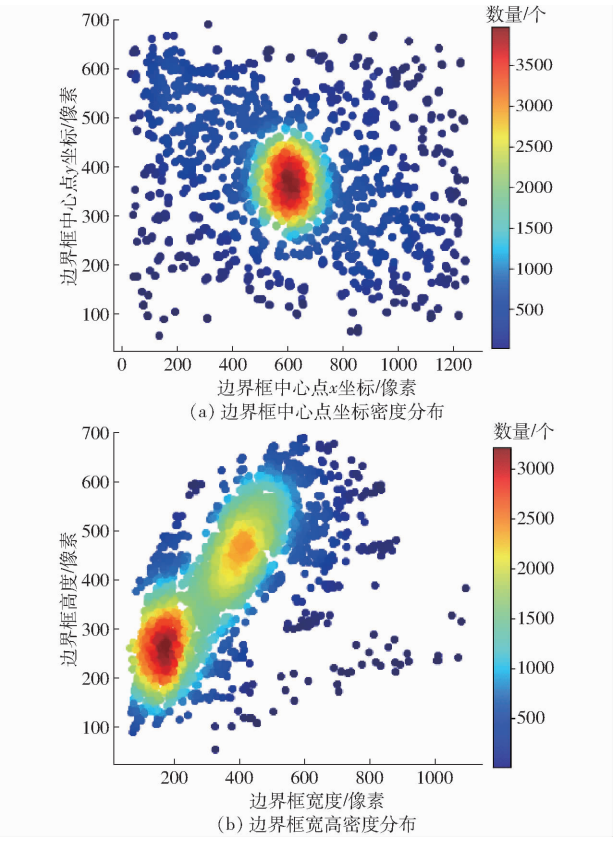

图 3　边界框大小分布图

Fig. 3　Distributions of bounding box size

过将其主干网络更换为 EfficientVit 网络，实现了对计算和参数效率的显著优化。EfficientVit 采用一系列优化策略，如深度可分离卷积、逆残差结构等，大大减少了网络参数量和计算复杂度。这一替换不仅降低了模型复杂度，而且提高了模型实时性，使其更适用于实际采摘场景。此外，为了进一步提升模型对目标细节信息处理的能力，在原有的编码器结构中，对跨尺度融合模块（Cross-scale feature fusion module，CCFM）进行了改进，将原有用于统一尺寸的上采样和下采样层替换为自适应细节融合模块（Adaptive detail infusion，ADI），以增强模型对于细节信息处理的能力。这一改进有助于模型更好地捕捉和利用目标的细节特征，提高检测精度。提出了一种自适应 Varifocal 损失函数，命名为 EMASlideVarifocalLoss。传统的 Varifocal 损失函数在目标检测中广泛应用，但固定了 IoU 阈值，这在处理不同复杂度的样本时可能不够灵活。EMASlideVarifocalLoss 通过指数移动平均（Exponential moving average，EMA）机制动态调整 IoU 阈值，使得模型在不同阶段能够自适应地优化其性能，从而显著增强了目标检测任务中的模型鲁棒性。图 4 展示了改进后的模型结构。

1.4.1　EfficientViT 骨干网络

为了实现网络模型的轻量化并高效提取图像特征，采用 EfficientViT[18] 作为特征提取的骨干网络。EfficientViT 结构由 1 个初始模块（Stem）和 4 个阶段（Stages）组成，每个阶段在逐步降低特征图尺寸的同时增加通道数，如图 5a 所示。与传统的 CNN 骨干网络不同，EfficientViT 在其第 3 和第 4 阶段引入了 Efficient Block 模块，如图 5b 所示。Efficient Block 模块包含 1 个轻量级的多尺度线性注意模块（Multi-scale linear attention，MSA）以及 1 个具有深度卷积（Depthwise convolution，DWConv）的前馈神经网络（Feedforward neural network，FNN）[19]。MSA 模块借鉴了 Transformer 架构中的自注意力机制，能够有效建模特征图中不同位置之间的长距离依赖关系。通过计算特征图中每个位置与其他位置之间的注意力权重，MSA 可以动态地聚合全局上下文信息，使得模型能够更好地理解图像的整体结构和语义内容。FNN 与 DWConv 的结合则负责提取局部细节信息。FFN 通过逐点非线性变换，增强了特征的表达能力；而深度可分卷积通过将标准卷积拆分为深度卷积和逐点卷积，减少了参数量和计算量，使

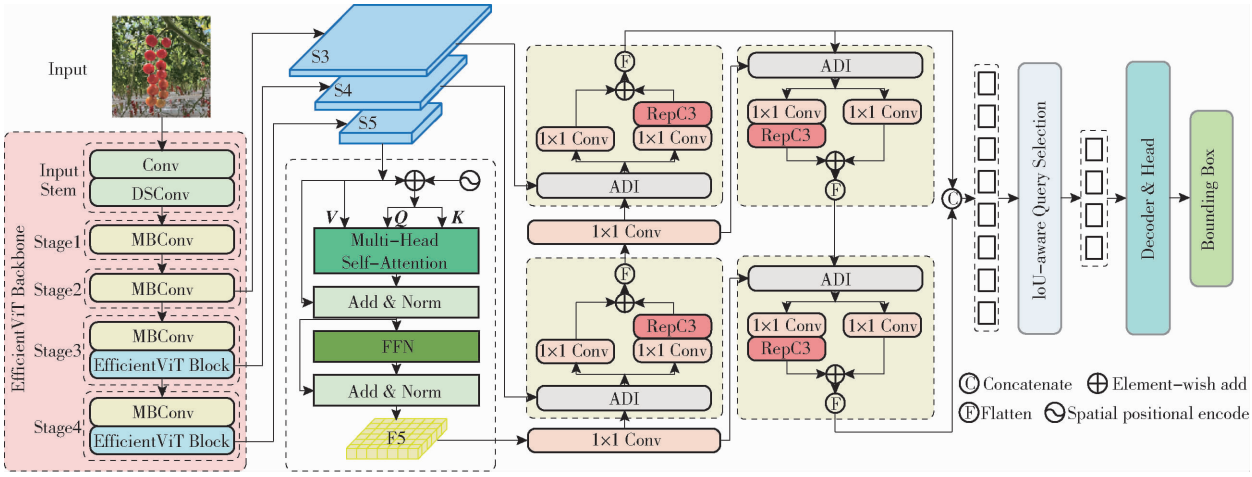

图 4　改进后的网络模型结构图

Fig. 4　Improved network model structure diagram

得模型更加轻量高效。FFN + DWConv 模块专注于捕捉特征图中的局部模式和细节,与 MSA 模块形成互补,共同增强了特征表示的质量。如图 5c 所示,MSA 模块采用轻量级的 ReLU 线性注意力替代传统的 Softmax 注意力,有效扩展了模型的全局感受野。此外,MSA 通过小核深度卷积对邻近的 ***Q/K/V*** tokens 进行卷积聚合,生成了具有多尺度特征的 tokens。这一过程不仅增强了 ReLU 线性注意力的多尺度学习能力,而且通过引入群卷积(Group convolution,GConv)[20] 减少了聚合操作的频率,进一步提高了模型效率和性能。

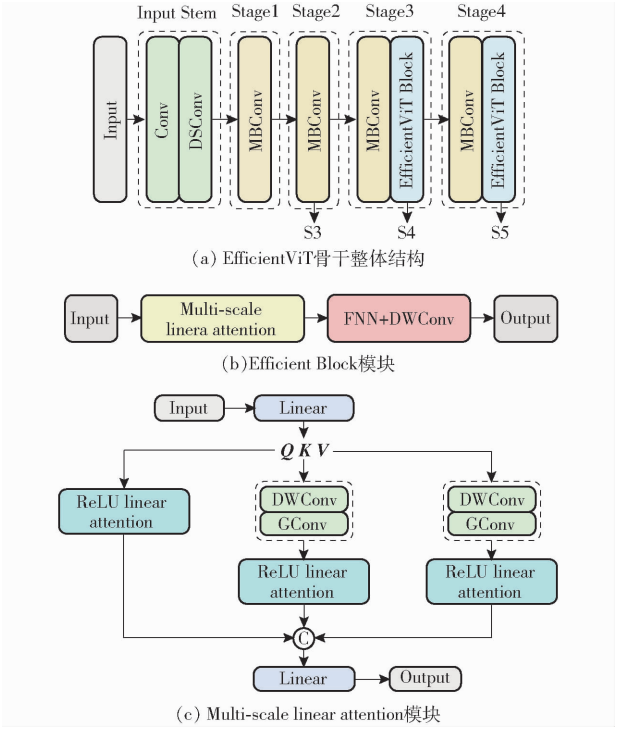

图 5 EfficientViT 骨干结构图

Fig. 5 EfficientViT backbone structure diagram

1.4.2 自适应细节融合模块

在传统的特征融合模块中,不同尺度特征图的融合依赖于专为其尺寸设计的多个上采样和下采样层。存在计算成本高、结构复杂、信息损失、灵活性不足和训练困难等局限性。本文设计了 ADI 模块,以处理并融合不同尺度的特征图。ADI 模块的设计灵感来源于 Unetv2[21] 中的 SDI(Segment and detail infusion)模块,SDI 通过连接不同尺度的特征图来减少特征差异,从而在编码器和解码器之间建立了一个有效的信息流通路径。与 SDI 不同,ADI 模块专注于编码过程中不同尺度特征的融合。具体而言,ADI 模块将第 1 个输入特征图尺度定义为目标尺度。如图 6 所示,对于后续输入的特征图,模块会根据其与目标尺度的关系,采取以下 3 种操作之一:

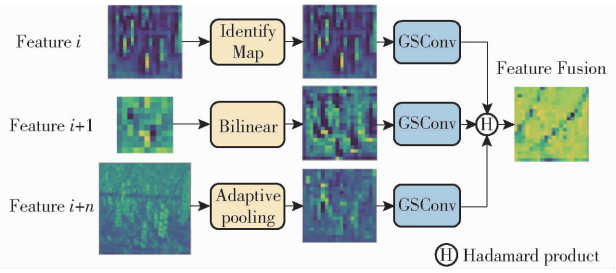

图 6 Adaptive detail infusion 模块结构图

Fig. 6 Adaptive detail infusion module structure diagram

(1)当输入特征图尺度小于目标尺度时,采用双线性插值法对特征图进行上采样,扩大其尺寸以匹配目标尺度。

(2)当输入特征图尺度大于目标尺度时,采用自适应池化技术对特征图进行下采样,减小其尺寸以匹配目标尺度。

(3)当输入特征图尺度已经与目标尺度匹配时,通过恒等映射直接传递特征图,无需进行尺度调整。

首先,通过自适应地调整特征图尺度,ADI 模块能够更好地保留不同尺度下的重要信息,提高了特征融合质量。其次,ADI 模块设计具有很强的灵活性,可以方便地插入到现有的网络结构中,而无需对网络进行大幅修改。最后,由于 ADI 模块减少了专门的上采样和下采样层,它使得模型参数量和计算量显著减少,提高了模型效率。

调整尺寸后的特征图随后会经过一次轻量级的分组可分离卷积(Group-wise separable convolution,GSConv)[22],旨在增强其表征能力,同时避免了给网络带来过重的计算负担。GSConv 卷积是一种高效的卷积操作,它通过将卷积核分组,并在不同组之间共享参数,实现了参数量的减少和计算效率的提升。最终,所有调整尺度后的特征图通过逐元素的 Hadamard 乘法[23] 进行融合,以实现不同特征图间高效且灵活的特征整合。Hadamard 乘法是一种高效的元素级乘法运算,通过将不同特征图对应位置的元素相乘,实现了特征之间的交互和整合。与简单的拼接或相加相比,Hadamard 乘法能够更加灵活地建模特征图之间的相关性,增强了融合后特征的表达能力。同时,由于 Hadamard 乘法的计算简单,它也保证了融合过程的高效性。对于 2 个相同维度的矩阵 ***A*** 和 ***B***,它们的 Hadamard 乘积 ***C*** 可以用符号表示为

$$\boldsymbol{C} = \boldsymbol{A} \circ \boldsymbol{B} \tag{1}$$

式中 $\circ$——Hadamard 乘积符号

矩阵 ***C*** 的每个元素 c_{ij} 定义为

$$c_{ij} = a_{ij} b_{ij} \tag{2}$$

式中　a_{ij}——矩阵 $\boldsymbol{A}$ 在第 i 行、第 j 列的元素

b_{ij}——矩阵 $\boldsymbol{B}$ 在第 i 行、第 j 列的元素

c_{ij}——矩阵 $\boldsymbol{C}$ 在第 i 行、第 j 列的元素

在原本的 CCFM 中，为了对齐输入特征图的尺寸，需要在特征融合前进行上采样/下采样操作，如图 7a 所示。这种方式虽然能够实现特征图尺寸的统一，但也带来了额外的计算开销和信息损失。为了减少浮点运算量并提高特征融合的效率，通过引入 ADI 模块去除了原有的上采样/下采样层。同时，将原本用于融合 2 幅特征图的 Concatenate 操作替换为 ADI 模块中的 Hadamard 乘法。新的 CCFM 模块结构如图 7b 所示。与 Concatenate 操作不同，Hadamard 乘法在更好地保留原始特征图中的局部信息和纹理特征的同时，不会增加输出的通道数。这是因为 Hadamard 乘法是一种元素级的乘法操作，对应位置的元素相乘，输出的特征图与输入的特征图具有相同的通道数。这意味着后续层的输入尺寸保持不变，从而减少了必需的浮点运算。通过这种细致且巧妙的处理机制，ADI 模块既提高了模型对细节的捕捉能力，又保障了特征融合的灵活性和效率。

1.4.3　EMASlideVarifocalLoss

在目标检测领域，正负样本的准确划分对于提升模型性能至关重要。传统方法会设定一个固定的 IoU 阈值 μ，以区分正负样本：若样本的 IoU 小于 μ，则判定为负样本；反之，则判定为正样本。然而，这一方法面临的主要挑战是边界附近的样本可能因分类损失较大而受到影响，导致模型对这些关键样本的分类边界产生模糊。为了解决这个问题，2022 年的 CVPR（Conference on computer vision and pattern recognition）会议上提出了一种滑动损失机制[24]，通过引入一种加权函数滑动策略，特别聚焦于正负样本分界线附近的样本。这种方法能够精细调整边界附近的样本损失，从而提高模型对这些关键样本的分类性能。图 8 展示了这一机制的具体实现和效果。

基于滑动损失机制，本研究提出了一种全新的自适应损失函数——EMASlideVarifocalLoss。该损失函数融合了指数移动平均机制来动态调整 IoU 阈值 μ，而且结合了 Varifocal 损失的优点[25]，进一步提高了模型在樱桃番茄姿态检测任务中的鲁棒性和准确性。EMASlideVarifocalLoss 的计算过程为：

首先给定衰减系数 a_{decay} 和时间常数 τ，衰减函数 $f(x)$ 定义为

$$f(x) = a_{\text{decay}}(1 - \mathrm{e}^{-x/\tau}) \tag{3}$$

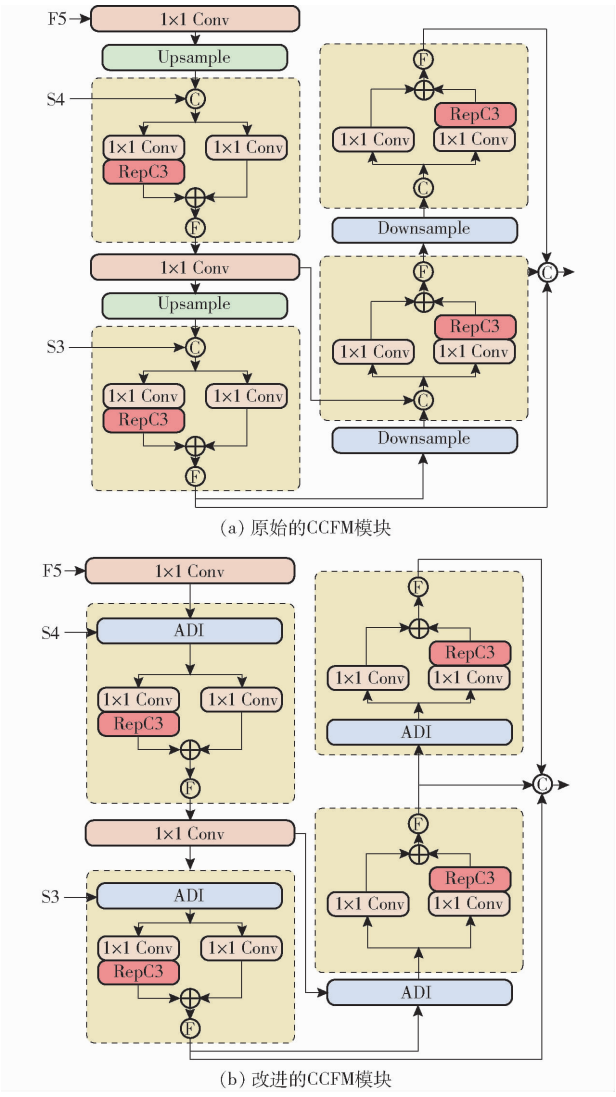

图 7　原始的 CCFM 模块与改进的 CCFM 模块对比

Fig. 7　Comparison between original CCFM module and improved CCFM module

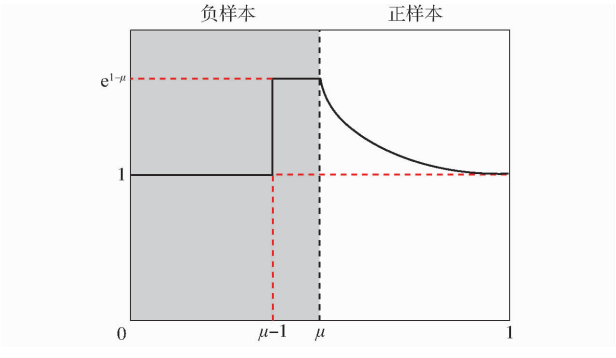

图 8　滑动损失函数示意图

Fig. 8　Schematic of sliding loss function

在每个更新步骤中，采用指数移动平均机制对当前 IoU 阈值进行更新，更新公式为

$$u_i = f(i)\,mIoU_i + (1 - f(i))\,mIoU_{i-1} \tag{4}$$

式中　i——执行的更新迭代次数

根据真实 IoU 值（g_{t_IoU}）和调整后的 IoU 阈值 μ，调节权重 w 计算式为

$$w(g_{t\_IoU})=\begin{cases}1.0 & (g_{t\_IoU}\leqslant\mu-0.1)\\ e^{1.0-\mu} & (\mu-0.1<g_{t\_IoU}<\mu)\\ e^{-(g_{t\_IoU}-1.0)} & (g_{t\_IoU}\geqslant\mu)\end{cases} \tag{5}$$

最后通过将调节权重 w 应用于 Varifocal 损失，并计算所有类别损失的平均和求和，得到 EMASlideVariFocalLoss

$$L=\frac{1}{N}\sum_{i=1}^{N}L_{\text{VarifocalLoss}}w_i \tag{6}$$

其中

$$L_{\text{VarifocalLoss}}=\sum_i -\mu_i[\hat{p}_i+(1-\hat{p}_i)^{\gamma}\ln(\hat{p}_i)]\cdot \ln(p_i)-(1-\alpha_i)p_i^{\gamma}\ln(1-\hat{p}_i) \tag{7}$$

式中 p_i——第 i 个样本为正类的概率

$\hat{p}_i$——第 i 个样本真实概率，通常取 0 或 1

μ_i——第 i 次更新的阈值

α_i——用于平衡正负样本权重的调整因子

γ——用于平衡不同难度样本的调整因子

总的来说，所提出的 EMASlideVarifocalLoss 是对传统损失函数的一项重要拓展。该损失函数不仅提供了一种有效机制来应对样本分类中的不确定性，而且为目标检测任务中提升模型的泛化能力和准确率开辟了新的途径。通过对 IoU 阈值进行自适应调整，并特别关注边界样本，这一方法在处理各种目标检测场景中的复杂变化方面展现出巨大的潜力。

2 实验与分析

2.1 实验设置

在配备 NVIDIA V100 Tensor Core GPU（具有 32 GB 运行内存）的高性能计算机上实施实验，CPU 配置为 Intel(R) Xeon(R) Platinum 8352V，频率为 2.10 GHz。选择 Ubuntu 20.04 作为操作系统，采用 Python 3.10 进行编程，利用 Pytorch 作为主要的深度学习框架进行模型的搭建和训练。

在模型训练的过程中，所有输入图像分辨率均被调整为 640 像素 ×640 像素，以统一数据输入的规格。为了充分利用硬件资源，在不超过 GPU 内存限制的前提下，设置图像的批量大小为 24。在训练策略上，采用 AdamW 优化器来训练网络，初始学习率被设定为 0.01，迭代次数为 200。

2.2 评价指标

为全面评估模型在樱桃番茄采摘姿态识别任务上的性能及其效率，采用查准率（Precision，P）、查全率（Recall，R）和平均精度均值（mean Average Precision，mAP）来衡量模型的性能。查准率反映模型预测结果的准确性，即在所有预测为正样本的结果中，真正为正样本的比例。

查全率反映模型预测结果的完整性，即在所有真实的正样本中，被模型正确预测为正样本的比例。

mAP 则是在不同交并比阈值下，查准率-查全率曲线下的面积，是一个综合考虑查准率和查全率的性能指标，其计算公式为

$$mAP=\frac{1}{N}\sum_{i=1}^{N}AP_i \tag{10}$$

其中

$$AP_i=\sum_{r\in 0,0.1,\cdots,1}P_i(r)\Delta r \tag{11}$$

式中 N——类别数量

AP_i——类别数量第 i 个类别的平均精度

$P_i(r)$——查全率为 r 时第 i 个类别的查准率

Δr——查全率变化步长

在模型复杂度方面，采用模型参数量、浮点运算次数（FLOPs）、内存占用量和帧速率（Frames per second，FPS）作为衡量指标。

2.3 对比实验

2.3.1 消融实验

消融实验是评估模型组件有效性的重要手段。为了验证本文所提出模块和损失函数的效果，本研究遵循控制变量法的原则，设计和实施了消融实验。实验结果见表 3，从表 3 可以直观地观察到关键性能指标的比较数据。实验中，将模型骨干网络替换为 EfficientViT 骨干网络，并引入本文提出的 ADI 模块和 EMASlideVarifocalLoss。结果显示，这些改动均使得模型在识别性能上相对于原始模型有所提升。

表 3 基于自建数据集的消融实验对比结果

Tab.3 Comparison results of ablation experiments based on self-built datasets

组合	主干网络		ADI 模块	FocalLoss	EMASVLoss	AP/%				mAP/%	浮点运算量	内存占用量/MB
	RT-DETR	EfficientViT				L	R	F	B			
1	√			√		92.7	83.6	89.7	89.5	88.9	5.58×10^{10}	65.8
2		√		√		92.6	82.5	88.5	90.3	88.5	2.84×10^{10}	25.7
3		√			√	93.2	84.2	91.0	90.5	89.8	2.72×10^{10}	25.6
4		√	√	√		94.1	83.2	91.0	90.6	89.7	8.70×10^{9}	22.6
5		√	√		√	93.0	84.8	93.0	89.1	90.0	8.70×10^{9}	21.6

注：“√”表示模型中使用该模块。

从表 3 中的组合 2 和组合 4 可以看出，相较于仅替换了骨干网络的模型，在引入 ADI 模块后，平均识别精度提升 1.35%，模型浮点运算量下降 68.01%，内存占用量下降 11.72%，这一改善归因于用 Hadamard 乘积替换了原始 CCFM 中的串联操作，从而减少浮点运算量。ADI 模块在特征融合过程中促进了特征间的直接交互，使得“L”、“R”、“F”、“B”4 种姿态的识别精度均有所提高。将改进后的 CCFM 特征融合效果进行可视化，如图 9 所示。通过观察热力图，可以发现模型的注意力主要集中在与主脉相连接的主梗、整串番茄的脊梗、分岔、萼片以及与萼片相连的果实等部位。说明 ADI 模块可以增强模型对于图像中各种尺度物体的识别能力，引入 ADI 模块后，模型关注的区域正确，没有存在偏差或误解。

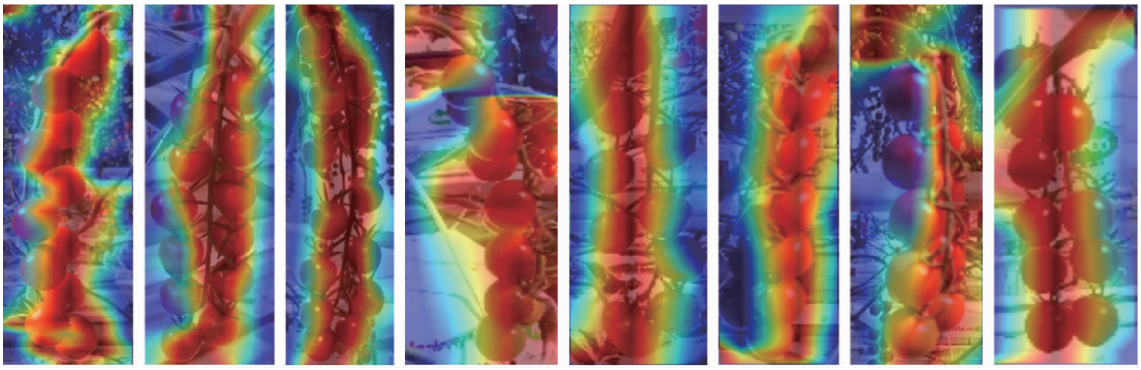

图 9　ADI 细节融合模块效果可视化

Fig. 9　Visualization of ADI detail fusion module effect

图 10 为组合 4 和组合 5 的损失函数曲线，从图中可以看出，在引入 EMASVLoss 后，模型在较少的迭代次数内就能达到较低的损失值，且平均识别准确率有小幅度的提升。尽管模型对于“B”一类的识别精度虽有小幅度的下降，但另外 3 类的识别精度均有不同程度提升。为了进一步验证损失函数对类别不平衡的处理能力，图 11 展示了消融实验中 5 种组合对 4 类采摘姿态的识别精度标准差条图，采用 EMASVLoss 的组相比于采用 FocalLoss 的组，变异性更小，且平均识别精度有所提升，证实了所优化的损失函数在解决目标检测中的类别不平衡问题上具有实际可行性。

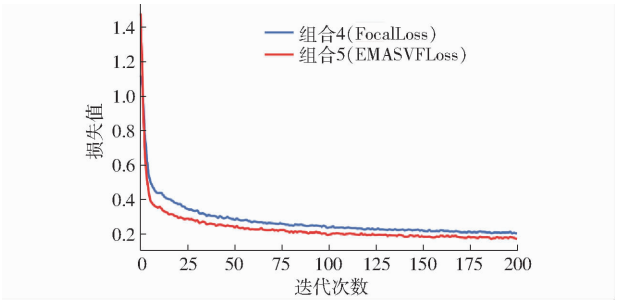

图 10　组合 4 和组合 5 的训练损失曲线对比

Fig. 10　Comparison of training loss between group 4 and group 5

综上所述，通过消融实验，证实了本研究所提出的方法能够实现高精度的樱桃番茄穗识别，为实现樱桃番茄的自动化采摘提供了有力的技术支撑。

2.3.2　不同模型实验结果对比

在上述的实验条件下，将本研究所提出的网络

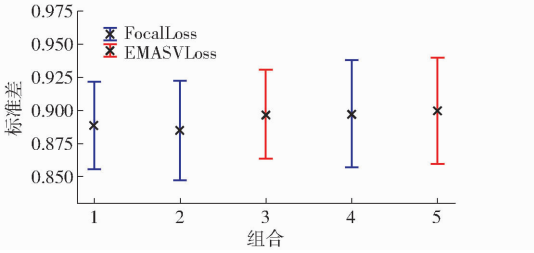

图 11　消融实验识别准确率标准差条图

Fig. 11　Comparison of standard deviation of identification accuracy in ablation experiments

与原始的 RTDETR 网络、Faster R－CNN 以及 Swin Transformer 模型进行比较，对比结果详见表 4。由表 4 表明，本文提出的轻量化模型与 RTDETR 网络、Faster R－CNN 以及 Swin Transformer 模型相比，识别平均精确率分别提高 1.24%、15.38% 和 3.45%。模型浮点运算量均有大幅度下降，分别下降 84.4%、92.37% 和 91.16%。帧速率最高可达 41.2 f/s，相比于其他模型分别提升 122.7%、25.61% 和 255.17%。这些结果充分展示了模型在保持高识别精度的同时，实现了模型轻量化和推理速度的显著提升，为实际应用中的高效目标检测提供了有力支持。

为了清晰地展示不同目标检测模型之间的性能差异，对各模型的检测结果进行了可视化，如图 12 所示。从检测结果图可以看出，对于示例图像 A，RTDETR 模型和 Swin Transformer 模型均表现出对番茄串姿态的敏感识别力，但在精确定位目标边界方面还有一定的提升空间。Faster R－CNN 模型表现则明显逊色，可以明显观察到其对应图像的目标

表 4 不同模型的对比实验结果

Tab. 4 Comparative experimental results of different models

模型	P/%	R/%	mAP50/%	浮点运算量	内存占用量/MB	帧速率/(f·s^{-1})
RTDETR	89.5	89.9	88.9	5.58×10^{10}	65.8	18.5
Faster R-CNN	79.5	77.9	78.0	1.14×10^{11}	98.5	32.8
Swin Transformer	90.2	83.7	87.0	9.84×10^{11}	70.2	11.6
本文模型	90.5	86.9	90.0	8.70×10^{9}	21.6	41.2

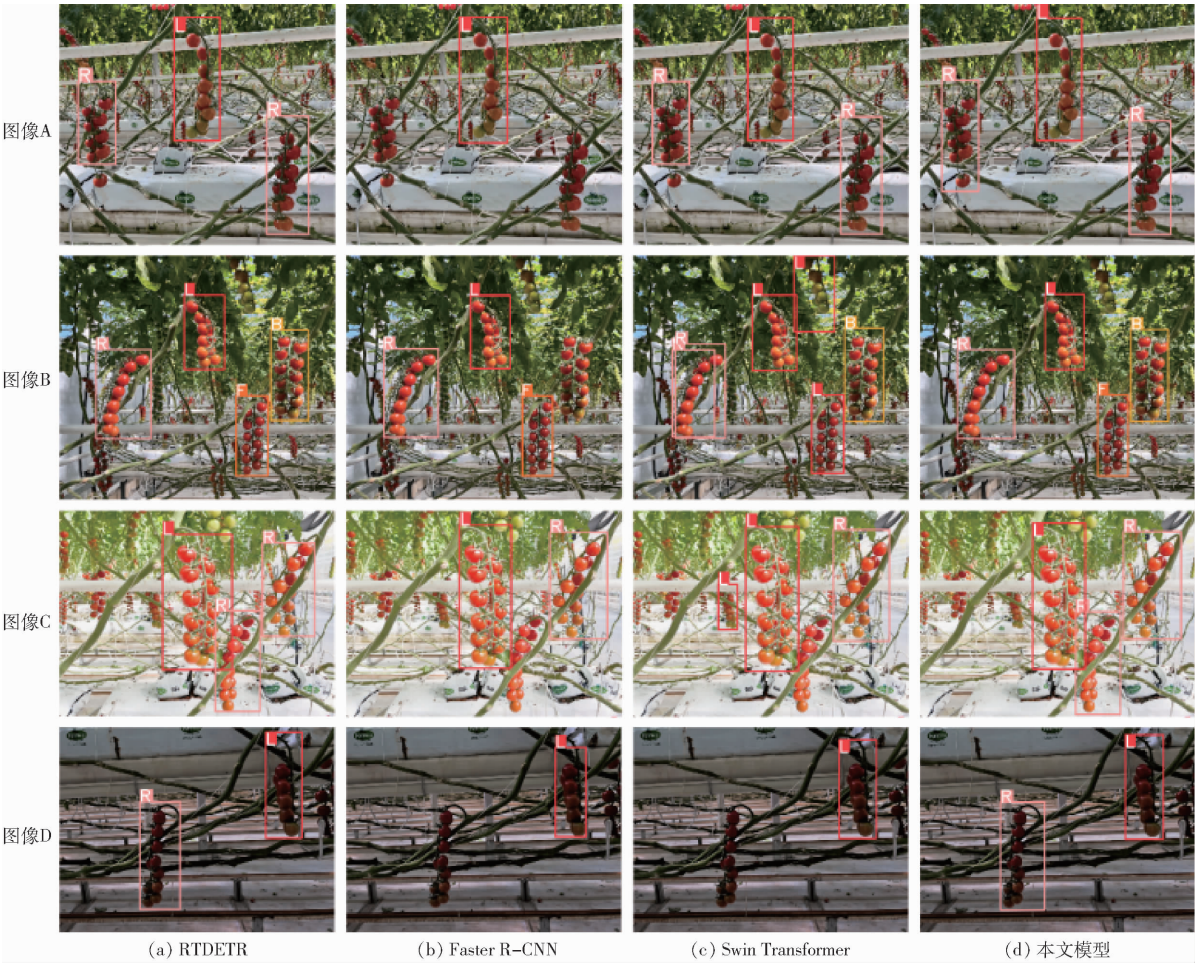

图 12 不同模型检测结果可视化

Fig. 12 Visualizations of detection results of different models

漏检问题。进一步观察示例图像 B,针对“B”姿态的番茄串,Faster R-CNN 模型的不足之处再次显现——漏检情况依然存在。与此同时,Swin Transformer 模型却存在判别失误的案例,将“F”识别成了“L”。此外,示例图像 C 和示例图像 D 可以看出所提出的模型对于不同光照条件具有较强的鲁棒性。在这一比较中,RTDETR 模型与本文所提出的模型检测结果均良好。通过上述可视化结果,可以得出结论:相较于 Faster R-CNN 和 Swin Transformer 模型,本文所提出的模型在精度、可靠性以及识别多样姿态的复杂对象方面已取得显著优势。

3 结论

(1)本研究基于穗收型樱桃番茄采摘机器人,提出了一种基于 RT-DETR 的轻量化穗收樱桃番茄识别模型。

(2)对樱桃番茄的采摘姿态进行了划分,并建立了包涵不同光照环境和不同采摘姿态的樱桃番茄数据集。

(3)通过将 RT-DETR 原始模型中的骨干网络替换为轻量级的 EfficientViT 骨干网络,在保持高精度的同时大幅减少了计算和参数量,使得模型更加适合部署在资源受限的设备上,如嵌入式系统或移动端。

(4)引入了创新的自适应细节融合模块促进特征间的直接交互,能够有效地结合不同尺度下的语义信息,还能够突出目标区域的细节特征,增强了模型了模型的鲁棒性。

(5)将原始的 FocalLoss 损失函数替换为 EMASlideVarifocalLoss 损失函数,通过加权函数滑动机制和指数移动平均的思想,动态地调整样本权重,使得模型能够更加关注那些难以区分的样本。

(6)在自建的樱桃番茄数据集上的实验结果表明,与其他先进模型相比,所提出的模型在平均识别准确率显著提升 1.24% ~15.38%,同时图像处理速度(FPS)提升幅度介于 25.61% ~255.17%。这些性能提升是在模型浮点运算量减少 69.37% ~92.37% 的条件下实现的。

参 考 文 献

[1] 刘潭, 朱洪锐, 袁青云, 等. 基于多模型融合策略的温室番茄光合速率预测方法[J]. 农业机械学报, 2024, 55(4): 337 - 345.
LIU Tan, ZHU Hongrui, YUAN Qingyun, et al. Prediction of photosynthetic rate of greenhouse tomatoes based on multi-model fusion strategy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(4): 337 - 345. (in Chinese)

[2] 窦汉杰, 陈震宇, 翟长远, 等. 果园智能化作业装备自主导航技术研究进展[J]. 农业机械学报, 2024, 55(4): 1 - 22.
DOU Hanjie, CHEN Zhenyu, ZHAI Changyuan, et al. Research progress on autonomous navigation technology for orchard intelligent equipment[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(4): 1 - 22. (in Chinese)

[3] 毛世平, 张琛. 以发展农业新质生产力推进农业强国建设[J]. 农业经济问题, 2024(4): 36 - 46.
MAO Shiping, ZHANG Chen. Promote the construction of an agricultural power through the development of new quality agricultural productivity[J]. Issues in Agricultural Economy, 2024(4): 36 - 46. (in Chinese)

[4] 冯青春, 赵春江, 王晓楠, 等. 基于视觉伺服的樱桃番茄果串对靶测量方法[J]. 农业工程学报, 2015, 31(16): 206 - 212.
FENG Qingchun, ZHAO Chunjiang, WANG Xiaonan, et al. Fruit bunch measurement method for cherry tomato based on visual servo[J]. Transactions of the CSAE, 2015, 31(16): 206 - 212. (in Chinese)

[5] GIRSHICK R. Fast R - CNN [C] //Proceedings of the IEEE International Conference on Computer Cision, 2015: 1440 - 1448.

[6] REN S, HE K, GIRSHICK R, et al. Faster R - CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137 - 1149.

[7] HU H, GU J, ZHANG Z, et al. Relation networks for object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3588 - 3597.

[8] TIAN Z, SHEN C, CHEN H, et al. Fcos: fully convolutional one-stage object detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9627 - 9636.

[9] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779 - 788.

[10] HE Y, ZHANG X, SAVVIDES M, et al. Softer-nms: rethinking bounding box regression for accurate object detection[J]. arXiv Preprint, arXiv:1809.08545, 2018.

[11] 荣佳诚. 串番茄目标精准识别定位与采摘策略研究[D]. 苏州: 苏州大学, 2022.
RONG Jiacheng. Research on accurate identification and localization of tomato target and harvesting strategy[D]. Suzhou: Suzhou University, 2022. (in Chinese)

[12] 李兴旭, 陈雯柏, 王一群, 等. 基于级联视觉检测的樱桃番茄自动采收系统设计与试验[J]. 农业工程学报, 2023, 39(1): 136 - 145.
LI Xingxu, CHEN Wenbai, WANG Yiqun, et al. Design and experiment of an automatic cherry tomato harvesting system based on cascade vision detection[J]. Transactions of the CSAE, 2023, 39(1): 136 - 145. (in Chinese)

[13] TAEHYEONG K, DAE-HYUN L, KYOUNG-CHUL K, et al. 2D pose estimation of multiple tomato fruit-bearing systems for robotic harvesting[J]. Computers and Electronics in Agriculture, 2023, 211: 108004.

[14] QI Z, ZHANG W, YUAN T, et al. An improved framework based on tracking-by-detection for simultaneous estimation of yield and maturity level in cherry tomatoes[J]. Measurement, 2024, 226: 114117.

[15] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C] //Computer Vision-ECCV 2020: 16th European Conference, 2020: 213 - 229.

[16] VASWANI A, SHAZEEER N, PARMAR N, et al. Attention is all you need[C] //Advances in Neural Information Processing Systems, 2017: 6000 - 6010.

[17] ZHAO Y, LV W, XU S, et al. DETRs beat YOLOs on real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2024: 16965 - 16974.

[18] LIU X, PENG H, ZHENG N, et al. EfficientViT: memory efficient vision transformer with cascaded group attention[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2023: 14420 - 14430.

[19] FRANCOIS C. Xception: deep learning with depthwise separable convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017:1800 - 1807.

(下转第 105 页)

[15] 杨福增,雷小燕,刘志杰, 等. 基于 CenterNet 的密集场景下多苹果目标快速识别方法[J]. 农业机械学报,2022,53(2):265 - 273.
YANG Fuzeng, LEI Xiaoyan, LIU Zhijie, et al. Fast recognition method for multiple apple targets in dense scenes based on CenterNet[J]. Transactions of the Chinese Society for Agricultural Machinery,2022,53(2):265 - 273. (in Chinese)

[16] LU R, DICKNSON N, LAMMERS K, et al. Design and evaluation of end effectors for a vacuum-based robotic apple harvester [J]. Journal of the ASABE, 2022, 65(5): 963 - 974.

[17] WANG M, YAN B, ZHANG S, et al. Development of a novel biomimetic mechanical hand based on physical characteristics of apples[J]. Agriculture, 2022, 12(11): 1871.

[18] 郑华栋, 王宏, 张洲, 等. 苹果采摘机器人手部结构设计与分析[J]. 液压与气动, 2023(7): 100 - 105.
ZHENG Huadong, WANG Hong, ZHANG Zhou, et al. Design and analysis of hand structure for apple picking robot[J]. Chinese Hydraulics & Pneumatics, 2023(7): 100 - 105. (in Chinese)

[19] 徐丽明, 刘旭东, 张凯良, 等. 脐橙采摘机器人末端执行器设计与试验[J]. 农业工程学报, 2018, 34(12): 53 - 61.
XU Liming, LIU Xudong, ZHANG Kailiang, et al. Design and test of end-effector for navel orange picking robot[J]. Transactions of the CSAE, 2018, 34(12): 53 - 61. (in Chinese)

[20] 李想,俞经虎.一种欠驱动苹果采摘机器人末端执行器的设计[J].机械传动,2015,39(9):74 - 77.
LI Xiang, YU Jinghu. Design of an underdriven apple picking robot end-effector[J]. Journal of Mechanical Transmission, 2015, 39(9): 74 - 77. (in Chinese)

[21] 马履中,杨文亮,王成军,等.苹果采摘机器人末端执行器的结构设计与试验[J].农机化研究,2009,31(12):65 - 67.
MA Lüzhong, YANG Wenliang, WANG Chengjun, et al. Structure design and experiment of end-effector of apple picking robot[J]. Journal of Agricultural Mechanization Research,2009,31(12):65 - 67. (in Chinese)

[22] 刘晓敏,田德宝,宋懋征, 等.气动球果采摘柔性手爪设计与实验[J].农业机械学报,2021,52(2):30 - 43.
LIU Xiaomin, TIAN Debao, SONG Maozheng, et al. Design and experiment on pneumatic flexible gripper for picking globose fruit[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(2):30 - 43. (in Chinese)

[23] XIE Z, DOMEL A G, AN N, et al. Octopus arm-inspired tapered soft actuators with suckers for improved grasping[J]. Soft Robotics, 2020, 7(5): 639 - 648.

(上接第 71 页)

[20] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6):84 - 90.

[21] PENG Y, SONKA M, CHEN D Z. U - Net v2: rethinking the skip connections of U - Net for medical image segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2023.

[22] LI H, LI J, WEI H, et al. Slim-neck by GSConv: a better design paradigm of detector architectures for autonomous vehicles [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2022.

[23] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[24] YU Z, HUANG H, CHEN W, et al. YOLO - FaceV2: a scale and occlusion aware face detector[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2022.

[25] ZHANG H, WANG Y, DAYOUB F, et al. VarifocalNet: an IoU-aware dense object detector[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8510 - 8519.