doi:10.6041/j.issn.1000-1298.2024.10.002

基于 SwinS－YOLACT 的番茄采摘机器人实时实例分割算法研究

倪纪鹏[1,2]　朱立成[1,2]　董力中[1,2]　崔学智[1,2]　韩振浩[1,2]　赵　博[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: 在设施番茄种植环境中,果实重叠遮挡等情况会影响识别精度。因此,本文提出了一种基于 YOLACT 的实例分割模型,提高识别精度。首先,对果实重叠遮挡的类别进行细分并增加该类数据集,从而接近真实采摘场景,并在采摘决策中改善重叠遮挡对识别精度的影响;其次,采用 Simple Cope－Paste 数据增强方法提高了模型的泛化能力,降低了环境因素对实例分割效果的干扰;然后,在 YOLACT 基础上,引用多尺度特征提取技术克服了单一尺度特征提取的局限性,并降低了模型复杂度;最后,引入 Swin Transformer 中的 Swin－S 注意力机制,优化了模型对于番茄实例分割的细节特征提取效果。实验结果表明,本文模型能够一定程度上缓解分割结果中出现的漏检、误检问题,其目标检测平均精度为 93.9%,相比于 YOLACT、YOLO v8－x、Mask R－CNN、InstaBoost 分别提升 10.4、4.5、16.3、3.9 个百分点;平均分割精度为 80.6%,相比于上述模型分别提升 4.8、1.5、7.3、4.3 个百分点;推理速度为 25.6 f/s。该模型综合性能有较强的鲁棒性,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参考。

关键词: 番茄果实; 重叠遮挡; 实例分割; YOLACT; Swin Transformer; 数据增强

中图分类号: TP391.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2024)10-0018-13　　**OSID**:

Real-time Instance Segmentation Algorithm for Tomato Picking Robot Based on SwinS－YOLACT

NI Jipeng[1,2]　ZHU Licheng[1,2]　DONG Lizhong[1,2]　CUI Xuezhi[1,2]　HAN Zhenhao[1,2]　ZHAO Bo[1,2]

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

Abstract: In the facility tomato planting environment, the accuracy of automatic fruit picking can be affected by overlapping and occlusion of fruits. An instance segmentation model was proposed based on YOLACT to address this issue. Firstly, the categories of fruit overlap and occlusion were subdivided, and the dataset of this type was increased to simulate real picking scenes and improve recognition accuracy in picking decisions. Secondly, the Simple Copy－Paste data enhancement method was employed to enhance the model's generalization ability and reduce the interference of environmental factors on instance segmentation. Next, based on YOLACT, multi-scale feature extraction technology was used to overcome the limitation of single-scale feature extraction and reducethe complexity of the model. Finally, the Swin－S attention mechanism in Swin Transformer was incorporated to optimize the detailed feature extraction effect for tomato instance segmentation. Experimental results demonstrated that this model can alleviate the problems of missed detection and false detection in segmentation results to a certain extent. It achieved an average target detection accuracy of 93.9%, which was an improvement of 10.4, 4.5, 16.3, and 3.9 percentage points compared with that of YOLACT, YOLO v8－x, Mask R－CNN and InstaBoost, respectively. Additionally, the average segmentation accuracy was 80.6%, which was 4.8, 1.5, 7.3, and 4.3 percentage points higher than that of the aforementioned models, respectively. The inference speed of this model was 25.6 f/s. Overall, this model exhibited stronger robustness and real-time performance in terms of comprehensive performance, effectively addressing both accuracy and speed requirements. It can serve as a valuable reference for tomato picking robots in performing visual tasks.

Key words: tomato fruit; overlap and occlusion; instance segmentation; YOLACT; Swin Transformer; data augmentation

收稿日期: 2023－12－07　修回日期: 2023－12－25

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: 倪纪鹏(1998—),男,硕士生,主要从事机器人视觉研究,E-mail: ni_jipeng@126.com

通信作者: 赵博(1981—),男,研究员,博士生导师,主要从事农业机器人和农业工程技术研究,E-mail: zhaoboshi@126.com

0　引言

采摘是番茄相关生产活动中关键环节之一[1]，但在设施番茄种植园中，仍以人工采摘为主，其劳动强度大，生产成本高的缺点直接影响采摘效率。因此，近年来，基于视觉技术的采摘机器人在番茄采摘方面得到推广应用。番茄采摘机器人需要实时精准定位采摘目标，以确保番茄能被精准夹取。同时，还需要在保证精度的前提下兼顾检测速度，这对采摘效率的提升具有重要的意义。

目前，在计算机视觉领域，深度学习技术已经成为主流，相较于传统机器学习算法[2－4]有一定的优势。基于深度学习[5－8]的实例分割在诸多农业相关领域都得到了广泛应用。实例分割算法能够识别番茄[9－11]的类别并得到具体轮廓，并对同一类别的不同实例对象进行分割。该类算法可归纳为两类：两阶段和单阶段实例分割。

两阶段实例分割算法（先生成候选区域，再对其进行分类和像素级分割）在果蔬识别领域存在着一定的应用[12－16]。单阶段实例分割算法（在一个阶段内完成目标分类和像素级分割）在性能上不弱于两阶段算法，并且相较于两阶段算法，其结构更简洁、计算更高效，训练更迅速[17－19]。

尽管国内外学者已在果实检测方面开展了大量工作，但大多数并未针对机器人采摘任务进行有效改进。在实际设施番茄种植园中，其环境十分复杂，且因果实间存在重叠遮挡（包括果实枝叶、果实与背景等障碍物的遮挡以及果实间相互遮挡）情况，需要精准检测并在采摘决策中降低重叠遮挡对识别精度的影响。应用于采摘任务中的现有实例分割算法仍然存在分割精度低、识别速度慢、鲁棒性差等问题，其性能有待提升。

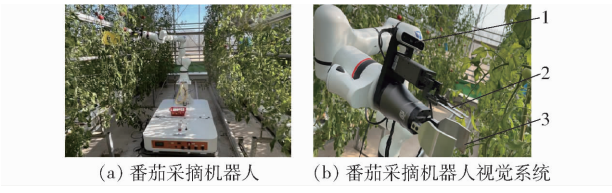
针对上述问题，本文提出一种基于 YOLACT 的改进模型，首先，对果实重叠遮挡的类别进行细分，增加该类数据集，以接近真实采摘场景；其次，采用 Simple Cope－Paste 数据增强方法对数据集图像进行数据增强，以降低环境因素对实例分割效果的干扰；最后，引入 Swin Transformer[20] 中的 Swin－S 注意力机制模块，以处理不同尺度的信息，优化模型对于番茄实例分割的特征细节提取，以提升模型整体检测准确率。

1　材料与方法

1.1　数据集

1.1.1　图像采集与数据集构建

模型训练与测试所用的番茄数据集拍摄于 2023 年 3 月 18 日至 5 月 8 日，拍摄时间为 09:00—11:00 和 14:00—17:00，拍摄地点位于北京市农业机械研究所京鹏植物工厂，共获取含有成熟期番茄在内的有效图像 1 500 幅。图像采集设备为 Intel® RealSense™ D415i 型深度相机，采集图像分辨率为 1 280 像素 ×720 像素。图像采集过程中，相机安装在中国农业机械化科学研究院集团有限公司自主研发的番茄采摘机器人末端位置，通过调整机械臂末端位姿，实现多角度拍摄，较好地模拟番茄采摘机器人视觉系统（图 1）的实际运行状况，通过手动操作机器人进行图像采集。为保证番茄果实样本的多样性，采集了不同时间段、不同光照强度、不同果实数量、不同遮挡程度等[21]多种情况的番茄果实样本图像（图 2）。



(a) 番茄采摘机器人　　(b) 番茄采摘机器人视觉系统

图 1　番茄采摘机器人平台

Fig. 1　Tomato picking robot platform

1. D415i 型深度相机　2. 剪刀　3. 力控夹爪

以成熟度和遮挡情况为分类依据，划分为成熟未遮挡番茄（ripe）、成熟部分遮挡番茄（ripe_partial）、未成熟未遮挡番茄（unripe）以及未成熟部分遮挡番茄（unripe_partial）4 个类别的标签，分别对应图 2a～2d，图 2e～2g 是由上述标签组合而成，其中图 2h～2j 为不同光照条件下采集的番茄图像。考虑到数据集构建中图像随机选取的原则，将所有图像按照比例 8∶1∶1进行随机划分，设置为训练集、验证集和测试集。数据集划分情况及不同光照图像数量如表 1 所示，各类别标签数量如表 2 所示。

1.1.2　数据增强

为降低环境因素对实例分割效果的干扰，采用 Simple Cope－Paste[22] 方法进行数据增强，虽然 MixUp[23]、CutMix[24] 和 Mosaic[25] 可用于组合多个图像以创建新的训练数据，但是这些数据增强方法并不适用于实例级别，对实例分割任务的帮助有限。Simple Copy－Paste 类似于 MixUp 和 CutMix，但仅复制实例像素，不涉及实例检测框中的背景像素。

Simple Cope－Paste 方法的基本原理如下：首先，将训练集数据进行随机图像亮度增强与对比度增强（图 3）；其次，随机选取 2 幅图像进行随机大尺度抖动（Large scale jittering，LSJ），大尺度抖动是相较于标准尺度抖动（Standard scale jittering，SSJ）而言更具有挑战性的一种尺度抖动方法。在 SSJ 中，

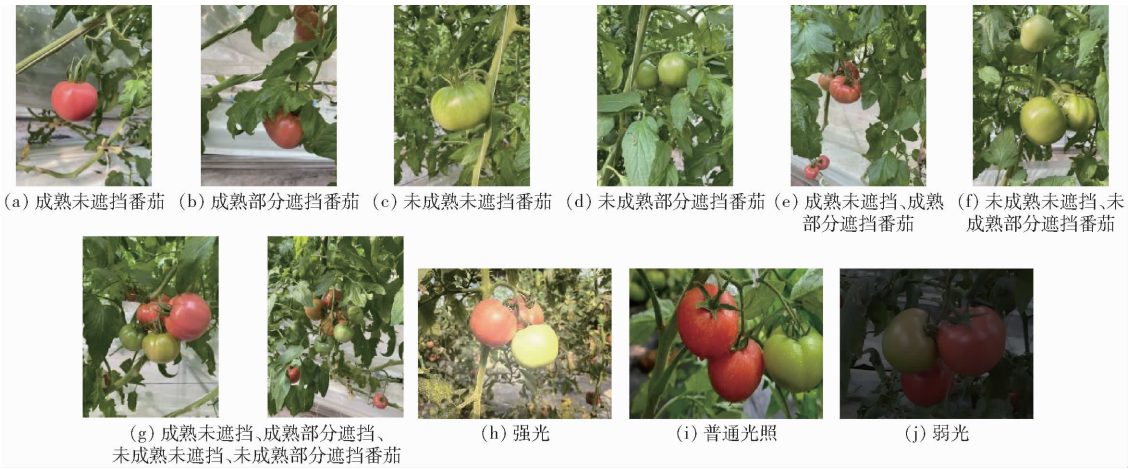


图 2 不同条件下番茄果实样本图像

Fig. 2 Tomato fruit samples under different conditions

表 1 数据集划分情况

Tab. 1 Dataset division 幅

数据集	强光	普通光照	弱光	总计
训练集	375	449	376	1 200
验证集	50	56	44	150
测试集	49	52	49	150
总计	474	557	469	1 500

表 2 各类别标签数量

Tab. 2 Number of labels for each category 个

数据集	成熟未遮挡番茄	成熟部分遮挡番茄	未成熟未遮挡番茄	未成熟部分遮挡番茄	总计
训练集	1 003	1 439	1 020	1 509	4 971
验证集	160	175	156	178	669
测试集	190	260	200	252	902
总计	1 353	1 874	1 376	1 939	6 542

尺度变化范围为 0.8 ~ 1.25,而在 LSJ 中,尺度变化范围为 0.1 ~ 2.0;如此大范围的抖动,会产生对比强烈的抖动效果。此外,LSJ 和 SSJ,都使用了随机水平翻转;最后,从 1 幅图像中随机选取若干实例,并将它们直接粘贴到另 1 幅图像上,同时更新检测框、类别标签和掩码(图 4)。

虽然 Simple Cope - Paste 方法在一定程度上能够增加样本数量,但也存在一些不足。例如,源图像中番茄目标被复制并粘贴到存在未遮挡番茄的目标图像中时,可能覆盖目标图像中的番茄目标区域,导致未遮挡番茄变成部分遮挡番茄。对于本来就有限的番茄图像信息来说,生成的错误类别标签会对模型训练产生负面影响,可能降低模型性能。

因此,本文在复制粘贴番茄实例目标时,通过计算 mask 重叠面积确定是否粘贴,保证番茄实例粘贴在目标图像的背景区域,而不覆盖原有的番茄目标区域(图 4b)。此方法能够确保在增加样本数量的同时,不产生错误标签,从而有效提升模型的泛化性能,使模型能适应实际场景中的复杂环境。在数据增强后,数据集规模增加 50%,由 1 500 幅扩充至 2 250 幅。

1.2 番茄果实实例分割模型

1.2.1 模型简介

相比于传统的 2 阶段实例分割模型,YOLACT

图 3 数据增强

Fig. 3 Data augmentation

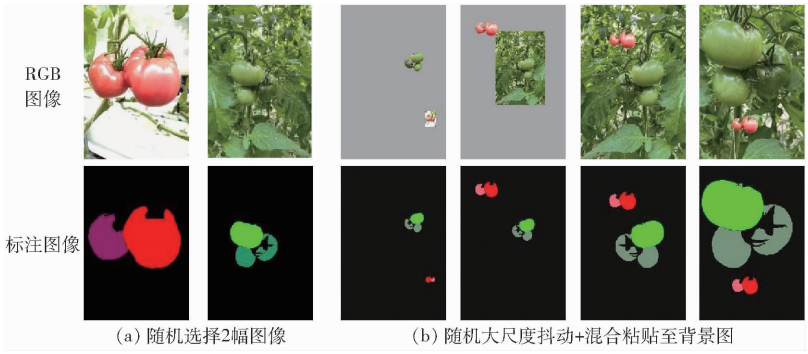


图 4　Simple Cope－Paste 数据增强

Fig. 4　Simple Copy－Paste data augmentation

具有多项优势：首先，基于单阶段目标检测的思想设计模型，使其拥有更快的响应速度，更适用于实时番茄采摘任务；其次，YOLACT 的模型参数量更小，故占用的计算资源更少，因此能够在计算资源有限的设备上高效运行。最后，该算法将实例分割任务划分为生成原型掩码和掩码预测系数 2 个并行的子任务执行，最终通过线性组合生成实例掩码。然而，因其无法有效地捕捉到目标边界和细节信息的缺点，导致其检测精度稍有欠缺。为了在不影响番茄检测速度的前提下提升检测精度，本文对 YOLACT 模型进行了改进(图 5)，替换了原模型骨干网络，采用了特征提取能力更强的 Swin Transformer 网络，为降低网络复杂度，选用了结构复杂度适中的 Swin－S[20] 作为骨干网络。

1.2.2　网络结构

SwinS－YOLACT 整体结构如图 5 所示。图中 H 为输入图像垂直方向上的像素数量，W 为输入图像水平方向上的像素数量，C 为块特征维度，a 表示每层特征图中每个像素点有不同长宽比的锚框数量(此处 a 取 3，比例为 1∶1、1∶2、2∶1)。

针对输入的图像，SwinS－YOLACT 网络进行如下处理：首先，使用 Swin－S 骨干网络提取 C3 到 C5 层级的特征；其次，基于特征金字塔网络(Feature pyramid network，FPN)架构逐层生成特征图 P3、P4、P5，并结合 3×3 卷积和双线性插值对 P5 层进行上采样，从而生成 P6 层和 P7 层，获得不同尺度的特征[26]。

其次，将复杂的实例分割任务分解为 2 个更快的并行分支任务。第 1 个分支任务使用 Protonet 结构对不同位置的实例对象和前景背景进行分割预测，该分支以 P3 层作为输入，通过多个卷积层进行语义分割预测；输出 32 个大小为 138×138 的原型掩码。

第 2 个分支任务是在目标检测过程中添加并行的掩码分支，即用于预测分割物体对应原型掩码的线性组合系数(Mask coefficients)。在预测头部结构(Prediction Head)中，对 P3、P4、P5、P6、P7 层，3 个分支共享 3×3 卷积输出的特征图，并同时处理各自分支。这些分支预测 $(4+N+k)$ 个值，包括锚框的 4 个坐标值、$N(N=4)$ 个类别以及 $k(k=32)$ 个掩码系数，表示用于分割物体对应的原型掩码的线性组合权重。

在并行任务完成后，通过快速非极大值抑制(Fast non-maximum suppression，F－NMS)去除第 2 个分支预测的冗余边界框，并将预测的掩码系数与

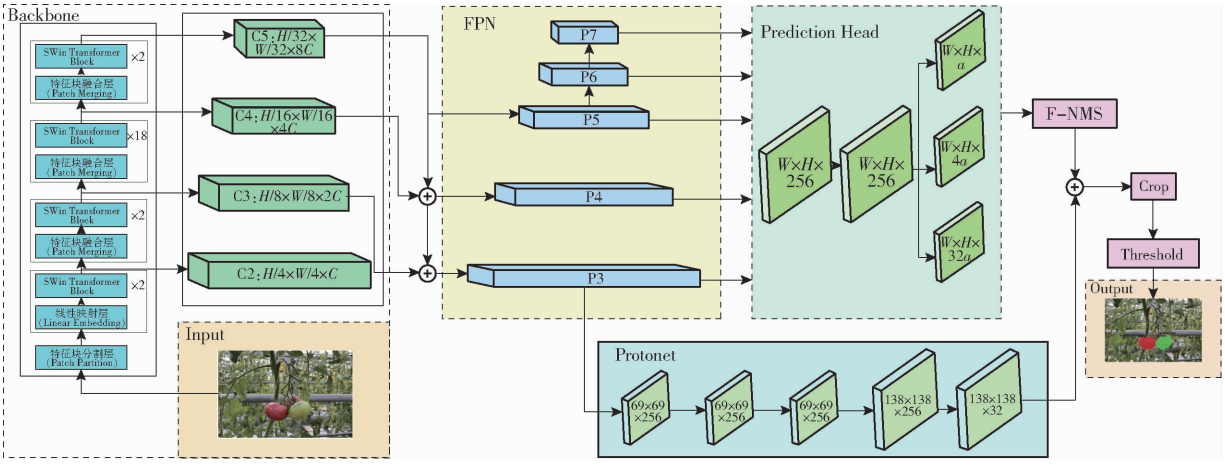


图 5　SwinS－YOLACT 整体结构

Fig. 5　Overall structure of SwinS－YOLACT

第 1 个分支对应的原型掩码进行线性组合,以生成每个锚框对应的分割掩码结果。这种并行处理方式在特征空间中保持了空间一致性,通过结合多个分支任务,能够快速生成分割结果。

1.2.3 Swin-S 特征提取模块

图 6 展示了使用 Swin Transformer 块进行特征下采样的 4 个阶段。第 1 阶段使用 2 个 Swin Transformer 块,第 2 阶段使用 2 个 Swin Transformer 块,第 3 阶段使用 18 个 Swin Transformer 块,第 4 阶段使用 2 个 Swin Transformer 块。具体流程描述如下:首先,输入图像先经过像素块划分层,将输入图像($H \times W \times 3$)划分为不重叠的块(Patch),其中每个块的尺寸为 4×4,每个块的特征维度为 $4 \times 4 \times 3 = 48$,块的数量为 $H/4 \times W/4$,然后进入第 1 个阶段,通过线性嵌入层将划分后的块特征维度转换为 $C(C=96)$,输入到 Swin Transformer 块中,输出为 $H/4 \times W/4 \times C$。第 2 阶段到第 4 阶段的操作类似,通过块合并(Patch merging)将输入相邻块逐对合并,块的数量减半为 $H/8 \times W/8$,特征维度扩展为 $4C$,随后降维至 $2C$,第 2 阶段输出为 $H/8 \times W/8 \times 2C$;类似的,第 3 阶段输出为 $H/16 \times W/16 \times 4C$,第 4 阶段输出为 $H/32 \times W/32 \times 8C$。

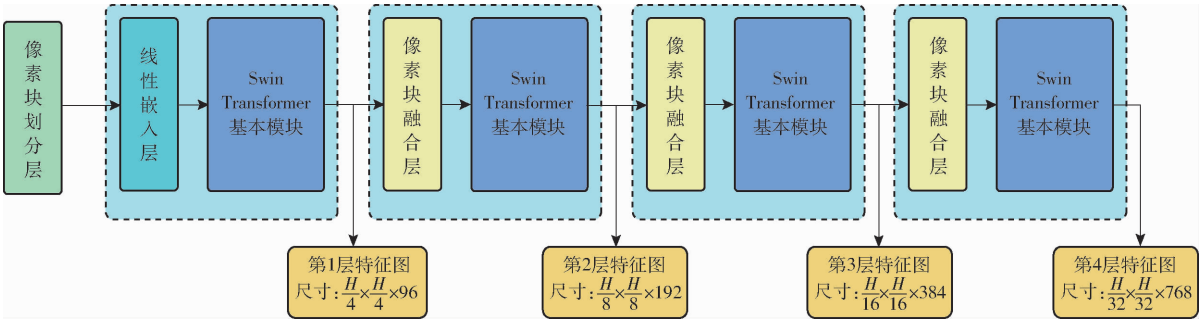


图 6 Swin-S 特征提取模块

Fig. 6 Swin-S feature extraction module

Swin Transformer 的核心是各阶段中的 Swin Transformer 块,它们以成对的方式出现(图 7)。主要结构包括窗口多头自注意力(Window multi-head self-attention,W-MSA)和移位窗口多头自注意力(Shifted-window multi-head self-attention,SW-MSA),均嵌套在多层感知机(Multi layer perceptron,MLP)内。每个 MSA 和 MLP 层之间引入层归一化(LayerNorm,LN)处理,并进行残差连接。这一设计有效地解决了 Transformer 参数量较大的问题。但由于使用 W-MSA 来降低网络计算的复杂性,会导致不同窗口之间缺乏信息交互的能力。为了解决此问题,设计了 SW-MSA,重新计算窗口偏移后的自注意力,以保持不同窗口之间的信息交流,从而减少了网络复杂度,提高模型的检测速度。

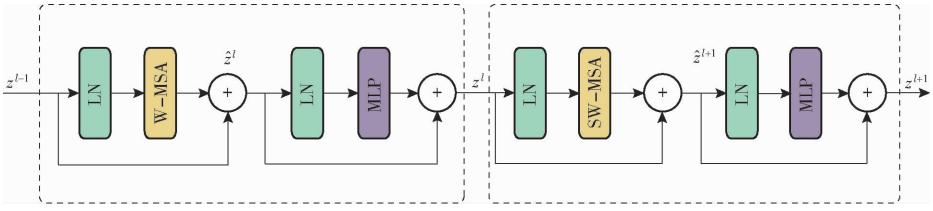


图 7 Swin Transformer 块

Fig. 7 Swin Transformer block

整个过程可以描述如下:

(1)输入特征首先通过层归一化处理,然后通过 W-MSA 进行特征学习,获得相应的特征图。之后,执行一次残差操作(将学习得到的特征图与 z^{l-1} 堆叠),得到[27]

$$\hat{z}^l = \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \tag{1}$$

式中,W-MSA( )为在局部窗口内计算多头自注意力函数,LN( )用于对每层输入进行归一化处理的函数。

(2)输入特征 $\hat{z}^l$ 经过层归一化处理,通过 MLP 进行特征提取,随后进行一次残差操作(将提取得到的特征与 $\hat{z}^l$ 进行堆叠),得到输出特征

$$z^l = \text{MLP}(\text{LN}(\hat{z}^l)) + \hat{z}^l \tag{2}$$

式中,MLP( )由两层全连接层和一个非线性激活函数组成,用于对输入进行非线性变换。

(3)输入特征 z^l 进入新的一层结构,此步骤与步骤(1)类似,但在计算特征部分使用了 SW-MSA,得到输出特征

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^l)) + z^l \tag{3}$$

式中,SW-MSA( )为在 W-MSA 的基础上,引入位移操作。

(4)与步骤(2)类似,输入特征 $\hat{z}^{l+1}$ 经过 LN、

MLP 和残差操作，得到最终的输出特征

$$z^{l+1} = \mathrm{MLP}(\mathrm{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \tag{4}$$

在整个过程中，算法的核心操作是 W－MSA 和 SW－MSA，其中 SW－MSA 是最重要的部分，解决了窗口偏移导致信息交流不足的问题，从而显著降低了网络的复杂度。

1.2.4　W－MSA

为了应对 Transformer 计算量过大的问题，引用了窗口多头自注意力（W－MSA）。

传统的全局自注意力（Multi-head self－attention，MSA）将图像划分为一系列图块（图 8a），然后在这些图块上执行自注意力计算。相比之下，W－MSA 采用了规则的窗口划分方法，将输入图像按照规则划分为尺寸相同但不重叠的窗口，并在不同的窗口上执行自注意力计算。其窗口划分方法（图 8b 中将其划分为 4 个窗口并进行自注意力计算）使得不同窗口之间的计算相互独立，从而提高了计算效率。

(a) MSA　　(b) W-MSA

图 8　W－MSA 示意图

Fig. 8　W－MSA schematics

因此，W－MSA 的窗口划分方法为 Transformer 提供了一种更高效的计算方式，能够满足大规模图像数据的处理需求，减轻计算负担。

1.2.5　SW－MSA

移位窗口多头自注意力通常紧跟在 W－MSA 模块之后，旨在解决 W－MSA 引入的窗口通信缺失问题。在 W－MSA 中，图像被划分为 4 个不同尺寸的窗口并执行自注意力计算。而 SW－MSA（图 9）执行了一次位移操作，通过循环填充，将原始图像向右下方移动，移动距离为窗口尺寸的一半，将原始图像转换为图 9 中第 2 个图中红色矩形框所示的新形式。如图 9 中 Layer1 所示，移动后窗口数量增加，从原来的 4 个变为 9 个，并且同时存在尺寸为 2×2、2×4 和 4×4 的窗口。在图 9 Layer2 中的绿色虚线表示窗口自注意力计算的尺寸，对该窗口进行自注意力计算，实际上是不同窗口自注意力信息的融合，实现了窗口之间的通信。

然而，由于移动窗口操作的存在，实际上图像内容的顺序被打乱，同时窗口的数量也增加，由原来的 4 个窗口增加到当前的 9 个窗口，导致计算量增加 9/4 倍（图 10a）。

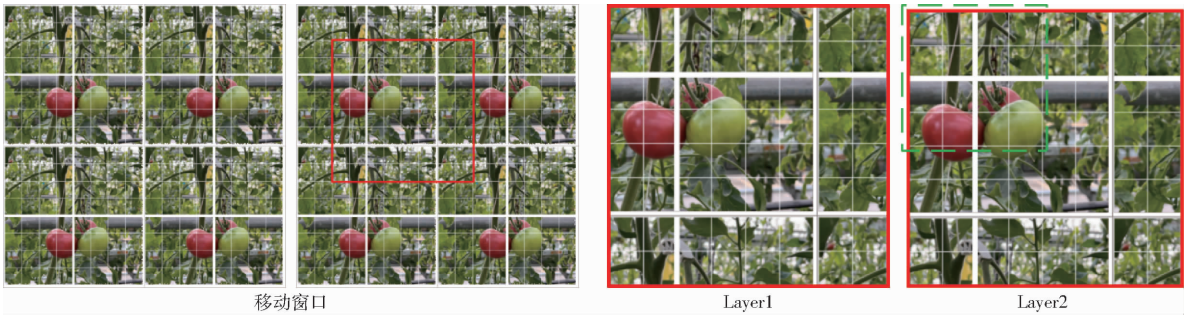
为了解决这个问题并更好地展示 SW－MSA 操作的效果，将对位移操作后的 9 个窗口进行编码，具体如图 10 所示。通过观察可以发现，图像被划分为 4 个窗口：5、(6,4)、(8,2)、(9,7,3,1)。接着，可以根据线性计算获得相应的查询－键－值（Query－Key－Value，$\boldsymbol{Q}-\boldsymbol{K}-\boldsymbol{V}$）[27] 矩阵，用于计算查询－键转置（$\boldsymbol{QK}^{\mathrm{T}}$）矩阵。在这些新窗口的组合中，只有具有相同编码的部分才能进行计算，不同编码位置之间的自注意力计算需要被置零。为此，根据以下公式执行自注意力，并通过掩码（Mask）将不同编码位置之间的自注意力结果加－100。

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{SoftMax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d}} + \boldsymbol{B}\right)\boldsymbol{V} \tag{5}$$

式中　$\boldsymbol{B}$——偏置项　　d——$\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的维度

由于在 SoftMax 计算中，SoftMax（－100）趋近于 0，从而清零了不同编码位置之间的注意力计算结果。计算结果如图 11 所示。

该计算方案保证了在计算自注意力时只考虑具有相同编码的窗口间的关联性，忽略了不同编码位置之间的干扰。通过掩码技术，能够有效消除同编码位置之间的注意力计算结果间的干扰，从而提升计算准确性。图 11 展示了经过掩码处理后的自注意力计算结果，突出了相同编码位置之间的关联性，将不同编码位置间的注意力置零，获得更加准确和可靠的注意力权重。



移动窗口　　Layer1　　Layer2

图 9　SW－MSA 示意图

Fig. 9　SW－MSA schematics

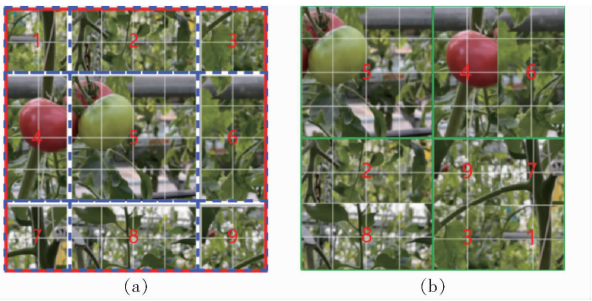


图 10 SW－MSA 编码示意图

Fig. 10 SW－MSA encoding schematics

1.3 实验平台

实验采用配置为 Intel® Core™ i7－12700H 处理器、NVIDIA® GeForce RTX3070Ti Laptop 显卡以及 32 GB 运行内存的便携式计算机进行模型训练和测试。软件平台采用基于 Windows 11 操作系统下的 PyTorch[28] 深度学习框架和 MMDetection[29] 目标检测框架进行实验。

1.4 评价指标

通过测试集对模型的性能进行定量评估，在实例分割问题中，需要综合考虑目标分类和像素级轮廓定位能力。目标检测和实例分割的评估指标通常使用平均精度均值（Mean average precision，mAP）[30] 来衡量模型的性能。mAP 能够全面评估模型在目标检测和实例分割任务中的综合表现。在计算 mAP 的过程中，需要考虑精确率（Precision）和召回率（Recall）这 2 个重要指标。精确率是指在所有被预测为目标的样本中，真正属于目标的样本所占的比例。而召回率则表示被正确定位和识别的目标样本占所有真实目标样本的比例。在多个类别的目标分割中，每个类别可以绘制一条（Precision－Recall，P－R）曲线（以召回率为横坐标，精确率为纵轴），平均精度（Average precision，AP）是指该 P－R 曲线包围的面积，mAP 通过对多个类别的 AP 计算平均值得到。通过以上指标，可以全面评估实例分割模型性能。

2 实验与结果分析

2.1 骨干网络对比实验

首先，使用骨干网络对比实验来验证本文提出的 SwinS－YOLACT 模型的鲁棒性。其次在所有实验中，将使用相应的预训练模型作为初始状态进行初始化。

为了验证改进模型的有效性，将所提出的算法与其他典型骨干网络进行对比测试。在实验中，选择了 3 个骨干网络对比实验模型：ResNet50[31]、

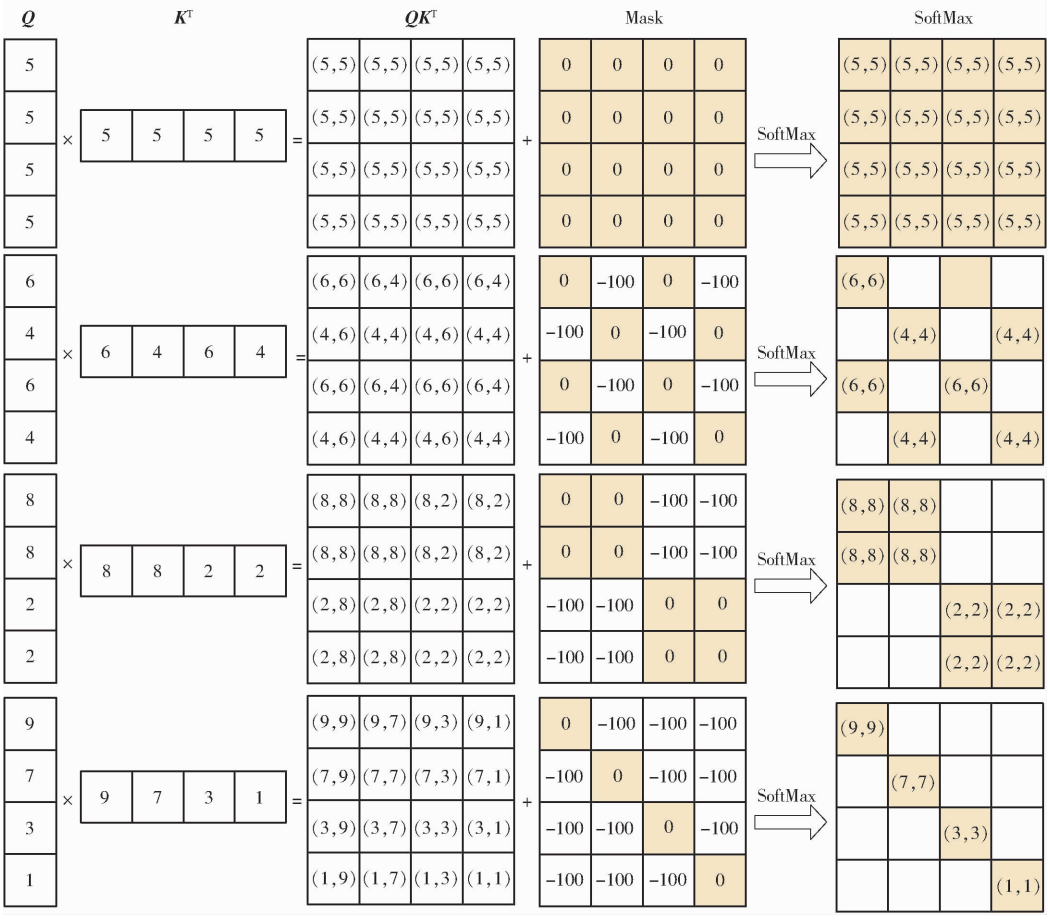


图 11 窗口编码计算结果

Fig. 11 Window encoding calculation results

ResNet101[31]和 Swin－T[20]。这些模型在训练时，除了骨干网络不同外，训练数据集和预处理方法保持一致。所有模型均采用随机梯度下降法作为优化器，设置动量为 0.92，权重衰减系数为 0.000 5。网络的初始学习率为 0.000 7，并且每训练 200 个周期，将学习率降低为原来的 1/10。最大训练周期设定为 1 000。

从图 12 可以看出，本文提出的骨干网络模型 Swin－S 相对更稳定。其在番茄实例分割方面的损失率最低，训练 200 次后基本达到收敛状态。相比其他，其损失值降低 3%，表现出更好的鲁棒性。

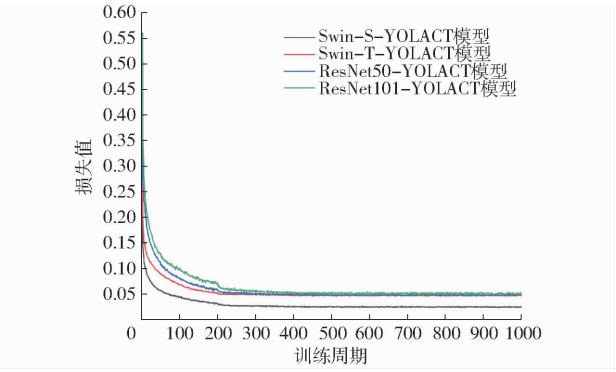


图 12　训练损失曲线

Fig. 12　Training loss curves

以 ResNet101 骨干网络作为本文的基准网络进行实验评估（表 3）。使用 ResNet50 作为骨干网络，与基准网络相比模型的参数有所减少，但 mAP^{box} 和 mAP^{mask} 的精度并未明显提升。使用 Swin－T 骨干网络，相较于基准网络，该模型的 mAP^{box} 和 mAP^{mask} 指标虽没有得到提升，但推理速度高于基准，可能由于 Swin－T 层级较浅，无法有效地捕捉图像的复杂细节特征。Swin－S 引入更深的层级结构与注意力机制，能够更好地捕捉图像的细节特征，与基准网络相比，mAP^{box} 和 mAP^{mask} 的精度提升 10.4 个百分点和 4.8 个百分点，且参数量仅增加 5.2%，虽然增加了模型的复杂度，但 Swin Transformer 基于移位窗口方法有效提升了推理速度，其推理速度与基准相近。

表 3　骨干网络实验结果对比

Tab. 3　Comparison of backbone network experiment results

模型	骨干网络	mAP^{box}/%	mAP^{box}_{50}/%	mAP^{box}_{75}/%	mAP^{mask}/%	mAP^{mask}_{50}/%	mAP^{mask}_{75}/%	推理速度/(f·s^{-1})	浮点运算量	参数量
YOLACT	ResNet101	83.5	98.4	96.6	75.8	98.1	91.6	26.8	8.513×10^{10}	5.374×10^{7}
YOLACT	ResNet50	77.6	97.6	92.9	73.0	97.2	88.5	27.3	6.048×10^{10}	3.475×10^{7}
YOLACT	Swin－T	77.1	97.9	92.5	73.2	97.6	89.3	27.2	6.357×10^{10}	3.524×10^{7}
SwinS－YOLACT	Swin－S	93.9	98.7	96.8	80.6	98.6	94.4	25.6	9.442×10^{10}	5.656×10^{7}

综上，本文的方法在相对于基准网络模型复杂度变化不大的情况下性能有明显提升。同时，这也证明了深层次网络结构和注意力机制可以有效地捕捉图像的复杂细节特征。

2.2　不同网络模型性能指标对比

为验证本文模型的检测效果，在相同测试集下，分别与实例分割领域的主流模型 YOLACT、YOLO v8－x[32]、Mask R－CNN[33]、InstaBoost[34]进行对比，其中 SwinS－YOLACT、YOLACT、YOLO v8－x 属于单阶段实例分割模型，而 Mask R－CNN[33]、InstaBoost[34]采用两阶段实例分割模型。所有模型输入图像尺寸均为 550 像素×550 像素。各模型性能指标如表 4 所示。

从表 4 可以看出，SwinS－YOLACT 模型在一定程度上改善了检测中存在的漏检、误检问题，其目标检测平均精度 mAP^{box} 为 93.9%，相比于 YOLACT、YOLO v8－x、Mask R－CNN、InstaBoost 分别提升 10.4、4.5、16.3、3.9 个百分点；平均分割精度 mAP^{mask} 达到 80.6%，相较于上述模型分别提升 4.8、1.5、7.3、4.3 个百分点。推理速度为 25.6 f/s，虽然略低于 YOLACT 和 YOLO v8－x，但该模型的其他性能指标均优于以上模型，综合模型整体的检测性能指标，SwinS－YOLACT 在精度与速度具有较大的优势。

表 4　模型性能对比

Tab. 4　Model performance comparison

模型	骨干网络	mAP^{box}/%	mAP^{box}_{50}/%	mAP^{box}_{75}/%	mAP^{mask}/%	mAP^{mask}_{50}/%	mAP^{mask}_{75}/%	推理速度/(f·s^{-1})	浮点运算量	参数量
YOLACT	ResNet101	83.5	98.4	96.6	75.8	98.1	91.6	26.8	$8.513\,0\times10^{10}$	5.374×10^{7}
YOLO v8－x	CSPDarkNet[32]	89.4	97.6	96.6	79.1	95.6	92.3	27.9	$3.437\,0\times10^{11}$	7.170×10^{7}
Mask R－CNN	ResNet101	77.6	98.0	93.3	73.3	97.8	90.0	20.8	$1.396\,7\times10^{11}$	6.275×10^{7}
InstaBoost	ResNet101	90.0	98.0	95.7	76.3	97.0	88.2	19.6	$2.447\,8\times10^{11}$	9.580×10^{7}
SwinS－YOLACT	Swin－S	93.9	98.7	96.8	80.6	98.6	94.4	25.6	$9.442\,0\times10^{10}$	5.656×10^{7}

2.2.1　特征提取对比

为更好地理解骨干网络对番茄果实的特征提取过程，选取不同模型的骨干网络 C2～C5 层输出特

征图进行了可视化(图 13)。所有模型浅层特征图具有与 RGB 图相似的轮廓图像但是随着模型网络层数的加深,其更加聚焦于分割目标的细节特征,逐渐忽视背景特征,从而实现对目标的分割。

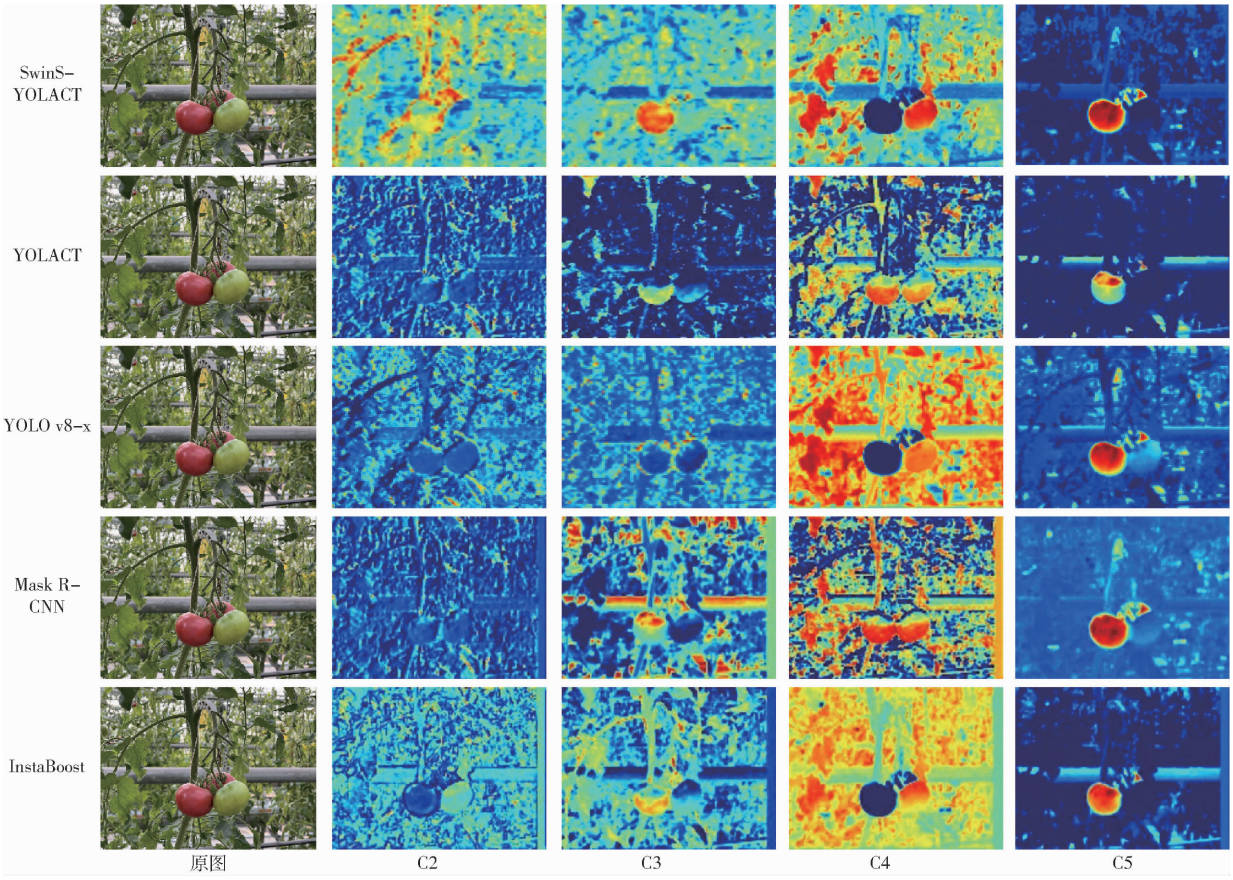


图 13 模型骨干网络特征图对比

Fig. 13 Comparison of feature maps of model backbone networks

SwinS - YOLACT 的骨干网络基于 Swin - S 特征提取结构,且引入分层局部注意力机制和窗口化非局部注意力机制,以处理不同尺度的信息。其将图像分解为多个小窗口进行计算,并在局部和全局级别进行注意力交互。相较于 YOLACT、YOLO v8 - x、Mask R - CNN、InstaBoost,SwinS - YOLACT 对番茄目标具有较大的关注力度,突显了窗口注意力机制的优势,其番茄细节特征提取更加丰富。

2.2.2 实验结果对比

为验证本模型在果园复杂环境下对番茄成熟度检测与分割的准确性,以复杂果园环境为背景进行番茄的成熟度检测实验。图 14 为不同模型在不同条件下的实验结果。从图 14 中可以发现,在复杂条件下,果实间存在重叠遮挡(包括果实枝叶、果实与背景等障碍物的遮挡以及果实间相互遮挡)情况,模型检测的准确率都受到影响,由于光照条件不佳时番茄的表皮会变得暗沉且不清晰,给番茄检测增加了难度;其次,未成熟番茄的表皮与茎干叶表皮的颜色、纹理和光滑度相似,当果实数量很多且互相发生重叠,模型检测的平均精度会受到一定的影响。

在成熟、成熟部分遮挡标签实验中,SwinS - YOLACT 模型的实例分割效果最佳,全部预测正确,而其他 4 种模型均出现误检。YOLO v8 - x 将左上角和右下角成熟部分遮挡番茄误检为成熟未遮挡番茄;YOLACT 和 InstaBoost 将右下角成熟部分遮挡番茄误检为成熟未遮挡番茄,并且把部分番茄叶片误检为未成熟部分遮挡番茄;Mask R - CNN 将左上角成熟部分遮挡番茄误检为成熟未遮挡番茄。

在成熟部分遮挡、未成熟部分遮挡标签实验中,5 种模型均成功预测出成熟部分遮挡的番茄。SwinS - YOLACT 和 InstaBoost 模型的实例分割效果较好,成功预测了右侧未成熟部分遮挡的番茄。

在未成熟、未成熟部分遮挡标签实验中,SwinS - YOLACT 和 YOLO v8 - x 模型的实例分割效果最佳,全部预测正确,而其他 3 种模型均将左上角未成熟番茄误检成未成熟部分遮挡番茄。

在未成熟部分遮挡标签实验中,5 种模型均成功预测,SwinS - YOLACT 模型的实例分割效果较为完整。

在成熟和未成熟混合测试(成熟、成熟部分遮挡,未成熟、未成熟部分遮挡)标签实验中,SwinS - YOLACT 模型的实例分割效果最佳,全部预测正确,而其他 4 种模型均出现误检;YOLO v8 - x 和

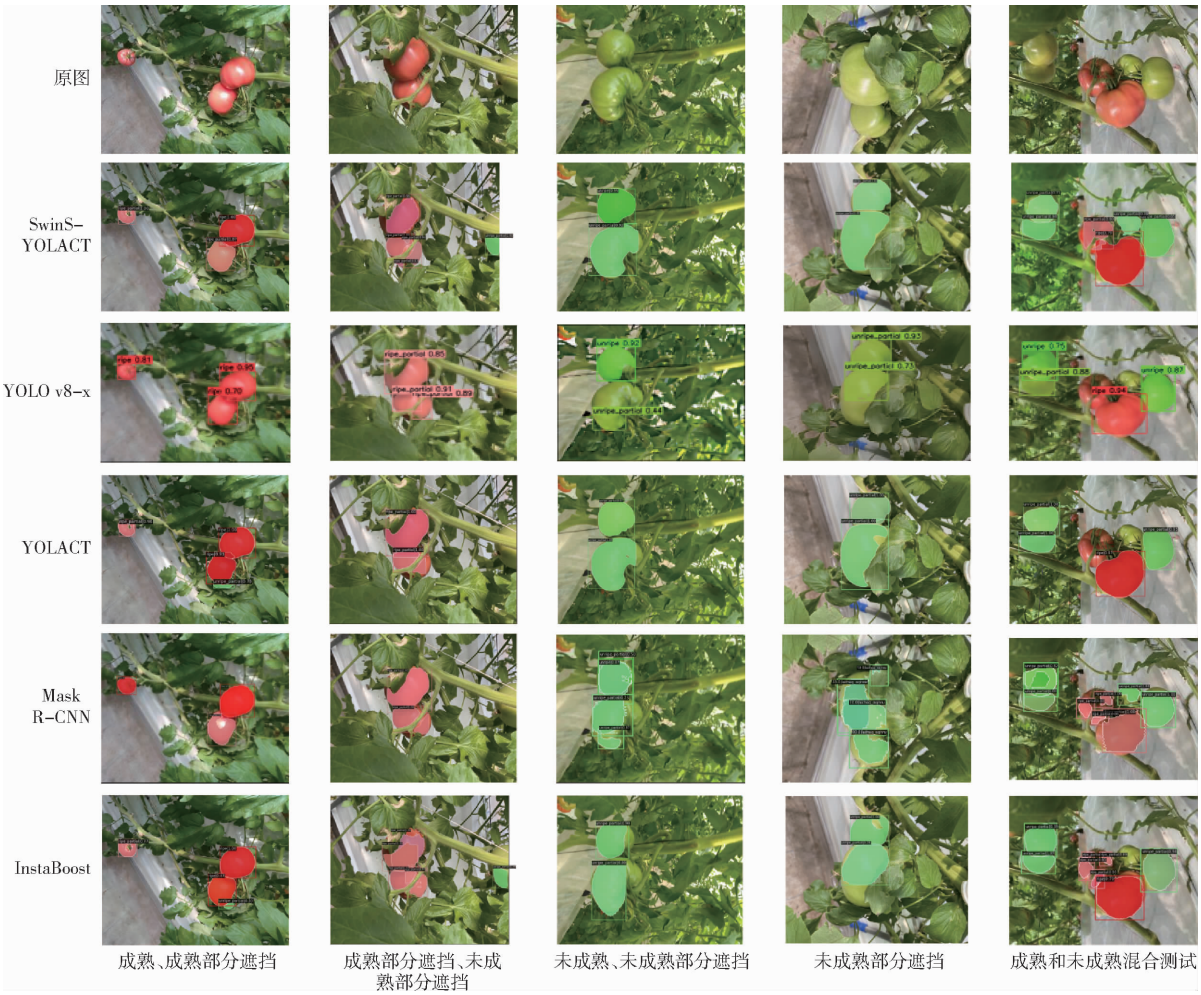


图 14　不同条件下番茄成熟度检测效果对比

Fig. 14　Comparison of tomato maturity detection results under different conditions

YOLACT 未检测出中上方成熟部分遮挡和未成熟部分遮挡番茄，Mask R－CNN 将中下方成熟未遮挡番茄误检为成熟部分遮挡番茄；InstaBoost 未检测出中上方未成熟部分遮挡番茄。

通过以上实验可以发现，SwinS－YOLACT 模型的泛化能力较强，能够一定程度上改善模型预测中存在的漏检、误检的情况，同时也证明了该模型具有较强的鲁棒性，能够实现果园复杂环境下对番茄成熟度的检测。

2.2.3　混淆矩阵结果对比

为了评估不同模型在 4 种番茄类别的实例分割性能，引用混淆矩阵观察模型的误检情况。使用训练后的 5 种模型对同一个测试集进行预测，并生成混淆矩阵，将成熟未遮挡、成熟部分遮挡、未成熟未遮挡、未成熟部分遮挡的番茄和背景 5 种类型标签设置为 ripe、ripe _ partial、unripe、unripe _ partial、background。混淆矩阵的横坐标表示不同类别番茄的预测标签，纵坐标表示真实标签，预测结果如图 15 所示。

通过观察图 15 可以发现：SwinS－YOLACT 对 4 种番茄类别的实例分割性能最佳，预测的精确率均高于 95%，其中，成熟和未成熟的番茄正确率高达 98%，误检率几乎为 0，此外，成熟部分遮挡和未成熟部分遮挡的番茄正确率也达到了 97%，由于未成熟的番茄纹理与番茄茎叶相似，同时设施内环境背景复杂，难免会有红色背景以及远处红色小番茄的干扰，故存在 2% 的成熟部分遮挡番茄类别误检率。

通过以上测试结果可以发现，SwinS－YOLACT 在实例分割任务中表现较好，准确率高且误检率低，适应复杂环境的能力较强。其他模型在不同程度上受到复杂环境干扰，从而存在部分误检情况。

2.3　不同光照情况下结果分析

为了验证不同光照下的模型检测效果，对成熟、未成熟以及遮挡、未遮挡的番茄进行统计区分。选取强光、普通光照、弱光情况下样本数量各 40 幅进行测试。

图 16 为 SwinS－YOLACT 模型在不同的光照条件下的实例分割可视化结果，可以看出，改进模型能够在不同光照强度下较为准确地识别出不同程度遮挡的番茄。

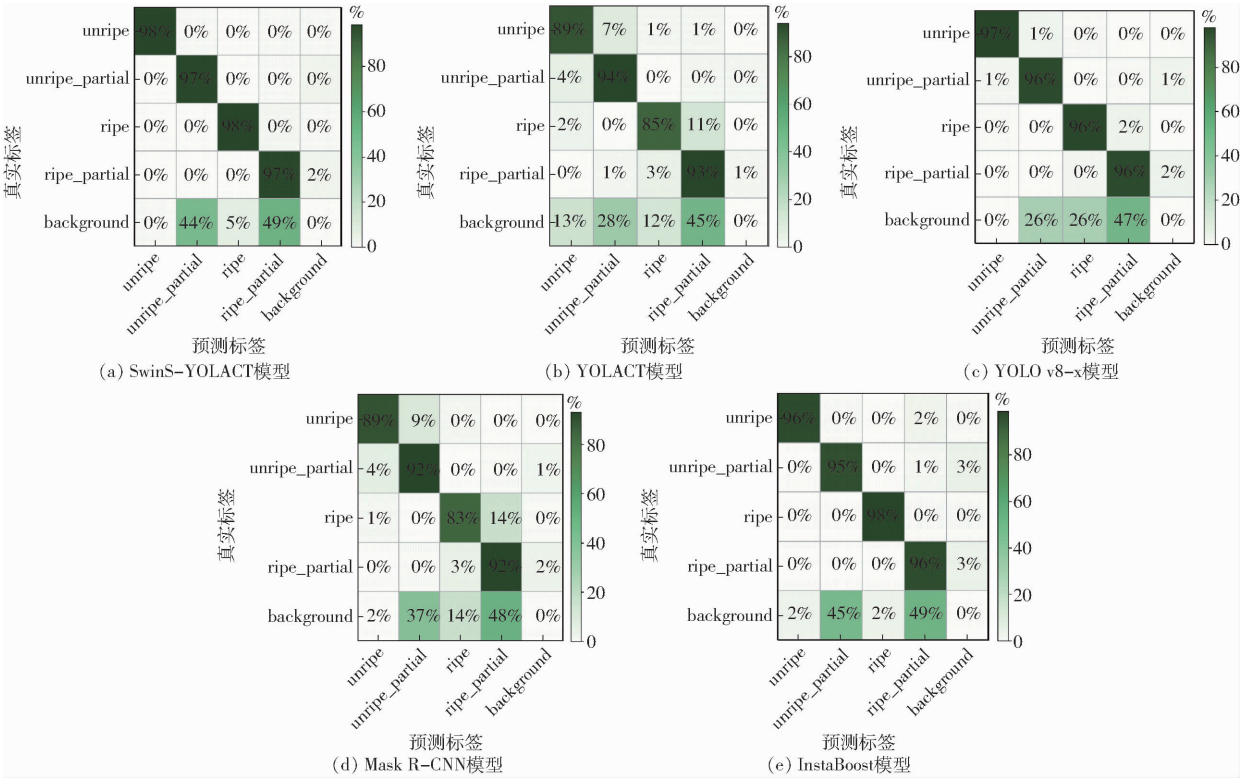


图 15 模型混淆矩阵结果对比

Fig. 15 Comparison of model confusion matrix results

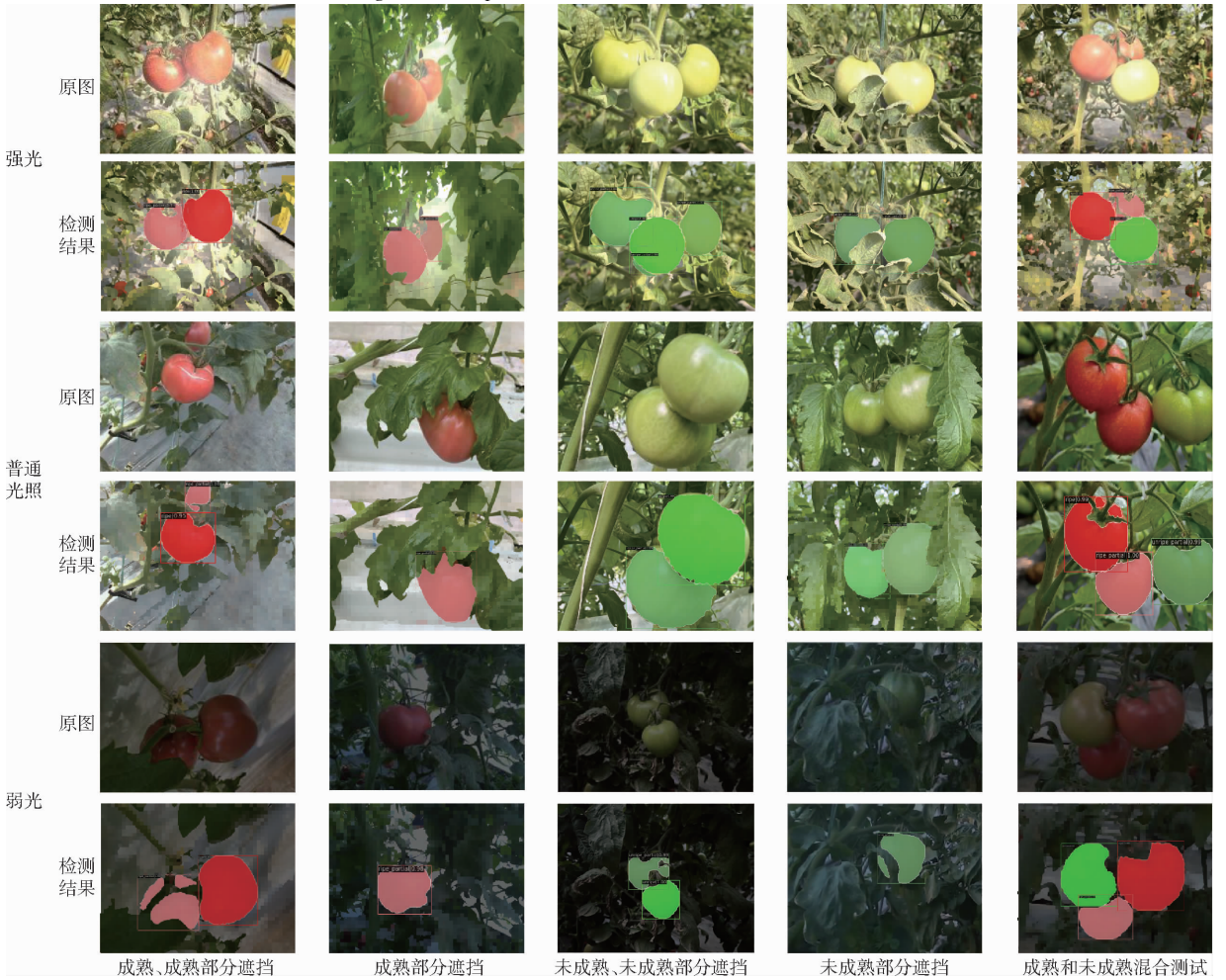


图 16 SwinS - YOLACT 模型不同光照条件下检测可视化

Fig. 16 Visualization of SwinS - YOLACT model detection under different lighting conditions

由表 5 可以看出，在强光和普通光照的情况下，番茄目标检测平均精度和平均分割精度均高于弱光环境下的番茄识别成功率。由于未成熟番茄在弱光环境下颜色容易与叶片、茎秆等混淆，在重叠情况下边界条件区分较为困难，因此弱光环境下番茄平均分割精度略低。

表 5　不同光照条件下检测结果对比

Tab.5　Comparison of detection results under different lighting conditions　%

模型	强光		普通光照		弱光	
	mAP^{box}	mAP^{mask}	mAP^{box}	mAP^{mask}	mAP^{box}	mAP^{mask}
SwinS－YOLACT	94.8	80.6	95.4	83.6	91.5	77.6

3　结论

（1）针对番茄采摘机器人精准采摘任务中存在的分割精度低、鲁棒性差等问题，提出了一种基于 YOLACT 的改进模型。

（2）对于果实重叠遮挡的情况，对该类别进行了细分，并增加了相关数据集，以贴近真实采摘场景，从而提升对重叠遮挡果实的精准识别。在进行采摘任务时，可优先剔除此类目标，为采摘任务的决策提供指导。

（3）采用 Simple Cope－Paste 数据增强方法，避免了番茄采摘机器人在进行实时图像采集时，角度单一以及由于光照等复杂背景对视觉系统带来干扰，增强模型的泛化能力。

（4）通过引入 Swin Transformer 中的 Swin－S 注意力机制模块，用以处理不同尺度的信息。这一机制优化了模型对番茄实例分割的细节特征提取，提升了模型整体检测准确率。

（5）在自建的番茄数据集上测试结果表明，其目标检测平均精度为 93.9%，相比于 YOLACT、YOLO v8－x、Mask R－CNN、InstaBoost 分别提升 10.4、4.5、16.3、3.9 个百分点；平均分割精度达到 80.6%，相比上述模型分别提升 4.8、1.5、7.3、4.3 个百分点；推理速度为 25.6 f/s。在综合性能上，该模型兼有较强的鲁棒性与实时性，因此能够在设施番茄种植环境下，高效完成番茄的实时实例分割任务。

参　考　文　献

[1] 刘洁，高润蕾，崔瑛，等. 番茄产业发展现状与内蒙古巴彦淖尔市番茄产业发展建议[J]. 现代农业，2023，48(3)：47－51.

[2] 孙博，李靖，王静. 机器学习在植物工厂中的研究现状与挑战[J]. 中国农学通报，2023，39(18)：142－150.

[3] 庞文媛，孙意岚，王芹，等. 机器学习在花果茶生产加工中的应用进展[J]. 食品安全质量检测学报，2023，14(11)：181－189.

[4] 马战林，文枫，周颖杰，等. 基于作物生长模型与机器学习算法的区域冬小麦估产[J]. 农业机械学报，2023，54(6)：136－147.
MA Zhanlin，WEN Feng，ZHOU Yingjie，et al. Regional winter-wheat yield estimation based on coupling of machine learning algorithm and crop growth model[J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(6)：136－147.(in Chinese)

[5] 金寿祥，周宏平，姜洪喆，等. 采摘机器人视觉系统研究进展[J]. 江苏农业学报，2023，39(2)：582－595.
JIN Shouxiang，ZHOU Hongping，JIANG Hongzhe，et al. Research progress on visual system of picking robot[J]. Jiangsu Journal of Agricultural Sciences，2023，39(2)：582－595.(in Chinese)

[6] 于丰华，周传琦，杨鑫，等. 日光温室番茄采摘机器人设计与试验[J]. 农业机械学报，2022，53(1)：41－49.
YU Fenghua，ZHOU Chuanqi，YANG Xin，et al. Design and experiment of tomato picking robot in solar greenhouse[J]. Transactions of the Chinese Society for Agricultural Machinery，2022，53(1)：41－49.(in Chinese)

[7] 黄泽霞，邵春莉. 深度学习下的视觉 SLAM 综述[J]. 机器人，2023，45(6)：756－768.

[8] 宋怀波，尚钰莹，何东健. 果实目标深度学习识别技术研究进展[J]. 农业机械学报，2023，54(1)：1－19.
SONG Huaibo，SHANG Yuying，HE Dongjian. Review on deep learning technology for fruit target recognition[J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(1)：1－19.(in Chinese)

[9] 张美娜，王潇，梁万杰，等. 温室远程监控系统人机交互与番茄识别研究[J]. 农业机械学报，2022，53(10)：363－370.
ZHANG Meina，WANG Xiao，LIANG Wanjie，et al. Human-computer interaction and tomato recognition in greenhouse remote monitoring system[J]. Transactions of the Chinese Society for Agricultural Machinery，2022，53(10)：363－370.(in Chinese)

[10] 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)

[11] 张俊宁，毕泽洋，闫英，等. 基于注意力机制与改进 YOLO 的温室番茄快速识别[J]. 农业机械学报，2023，54(5)：236－243.
ZHANG Junning，BI Zeyang，YAN Ying，et al. Fast recognition of greenhouse tomato targets based on attention mechanism and improved YOLO[J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(5)：236－243.(in Chinese)

[12] XU P，FANG N，LIU N，et al. Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation[J]. Computers and Electronics in Agriculture，2022，197：106991.

[13] ZU L，ZHAO Y，LIU J，et al. Detection and segmentation of mature green tomatoes based on Mask R－CNN with automatic

image acquisition approach[J]. Sensors,2021,21(23):7842.
[14] SHINODA R, KATAOKA H, HARA K, et al. Transformer-based ripeness segmentation for tomatoes[J]. Smart Agricultural Technology,2023,4:100196.
[15] 高倩,诸德宏,封浩. 基于深度学习的番茄识别与实例分割[J]. 软件导刊,2023,22(2):75 - 80.
[16] 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.
LONG Jiehua, ZHAO Chunjiang, LIN Sen, et al. Segmentation method of the tomato fruits with different maturities under greenhouse environment based on improved Mask R - CNN[J]. Transactions of the CSAE, 2021, 37(18):100 - 108. (in Chinese)
[17] LI Y, FENG Q, LIU C, et al. MTA - YOLACT: multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting[J]. European Journal of Agronomy,2023,146:126812.
[18] BOLYA D, ZHOU C, XIAO F, et al. YOLACT: real-time instance segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:9157 - 9166.
[19] JIA W, LIU J, LU Y, et al. Polar-Net: green fruit example segmentation in complex orchard environment[J]. Frontiers in Plant Science,2022,13:5176.
[20] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:10012 - 10022.
[21] 张勤,庞月生,李彬. 基于实例分割的番茄串视觉定位与采摘姿态估算方法[J]. 农业机械学报,2023,54(10):205 - 215.
ZHANG Qin, PANG Yuesheng, LI Bin. Visual positioning and picking pose estimation of tomato clusters based on instance segmentation[J]. Transactions of the Chinese Society for Agricultural Machinery,2023,54(10):205 - 215. (in Chinese)
[22] GHIASI G, CUI Y, SRINIVAS A, et al. Simple copy-paste is a strong data augmentation method for instance segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:2918 - 2928.
[23] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[J]. arXiv Preprint, arXiv:1710.09412,2017.
[24] YUN S, HAN D, OH S J, et al. Cutmix: regularization strategy to train strong classifiers with localizable features[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:6023 - 6032.
[25] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[J]. arXiv Preprint, arXiv:2004.10934,2020.
[26] 张赫. 基于深度学习的作物虫害快速识别与分割算法研究[D]. 杭州:杭州电子科技大学,2021.
ZHANG He. Research on fast recognition and segementation algorithm of crop pests based on deep learning[D]. Hangzhou: Hangzhou Dianzi University,2021. (in Chinese)
[27] 鲍嘉伟. 面向空海协同无人搜救的图像识别方法研究[D]. 大连:大连理工大学,2022.
BAO Jiawei. Research on image recognition method for air-sea collaborative unmanned search and rescue[D]. Dalian: Dalian University of Technology,2022. (in Chinese)
[28] PASZKE A, GROSS S, MASSA F, et al. Pytorch: an imperative style, high-performance deep learning library[J]. Advances in Neural Information Processing Systems,2019,32:8024 - 8035.
[29] CHEN K, WANG J, PANG J, et al. MMDetection: open mmlab detection toolbox and benchmark[J]. arXiv Preprint arXiv: 1906.07155,2019.
[30] 王冲. 基于计算机视觉的番茄检测方法研究[D]. 济南:山东大学,2022.
WANG Chong. Research on tomato detection method based on computer vision[D]. Ji'nan: Shandong University,2022. (in Chinese)
[31] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition,2016:770 - 778.
[32] REIS D, KUPEC J, HONG J, et al. Real-time flying object detection with YOLO v8[J]. arXiv Preprint, arXiv:2305.09972, 2023.
[33] HE K, GKIOXARI G, DOLLAR P, et al. Mask R - CNN[C]// Proceedings of the IEEE International Conference on Computer Vision,2017:2961 - 2969.
[34] FANG H S, SUN J, WANG R, et al. Instaboost: boosting instance segmentation via probability map guided copy-pasting[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:682 - 691.