doi:10.6041/j.issn.1000-1298.2026.02.016

基于 Transformer - CNN 特征深度融合的复杂环境香梨目标检测

杨　瑛[1]　谭　忠[2]　郑文轩[1]

(1.江苏第二师范学院物理与电子信息学院,南京 210013;2.厦门大学数学科学学院,厦门 361005)

摘要:为了能在复杂环境下提高香梨目标检测准确性,本研究提出了一种基于 Transformer - CNN 特征深度融合的 TC - ICSA - YOLO v8 香梨目标检测模型。模型有效融合了卷积神经网络 CNN 在提取图像局部(高频)特征信息和 Transformer 在提取图像全局(低频)特征信息的优势,设计了 Inception dilated 卷积模块和自适应细节融合模块 ADI,引入挤压增强轴向注意力机制 SeaAttention 和坐标注意力机制 CA 等提升模型的特征提取能力,采用傅里叶变换 FT 进行数据增强,使用频率斜坡结构以更好地平衡局部(高频)特征信息和全局(低频)特征信息成分,从而优化网络在特征提取过程的性能表现。试验结果表明,TC - ICSA - YOLO v8 模型在验证集上的平均精度均值(mAP)、精确率、召回率和检测速度分别达到 97.01%、97.33%、95.69% 和 81.21 f/s。本文模型对夜间拍摄的图像目标检测精度也优于同等条件的 YOLO v8s 模型;其平均精度均值(mAP)与 Faster R - CNN、YOLO v3、YOLO v7s、YOLO v8s、SwinTransformer、RT - DETR 模型对比分别提高 14.65、3.34、0.52、0.20、6.68、5.45 个百分点;模型内存占用量分别减少 74.60、34.16、9.48、16.81、20.84、13.64 MB,本文模型检测精度更高,参数量更少,更有利于部署在移动端。本文提出的检测模型对香梨具有很好的目标检测效果,为复杂环境下目标检测提供参考,可为香梨自动化采摘提供技术支撑。

关键词:Transformer;香梨;目标检测;傅里叶变换

中图分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1000-1298(2026)02-0161-10　　**OSID**:

Pear Object Detection in Complex Environment Based on Transformer - CNN Feature Deep Fusion

YANG Ying[1]　TAN Zhong[2]　ZHENG Wenxuan[1]

(1. *School of Physics and Electronic Information Engineering*, *Jiangsu Second Normal University*, *Nanjing* 210013, *China*
2. *School of Mathematical Sciences*, *Xiamen University*, *Xiamen* 361005, *China*)

Abstract: In order to improve the accuracy of pear target detection in complex environments, a TC - ICSA - YOLO v8 pear target detection model was proposed based on Transformer - CNN feature deep fusion. The proposed model effectively leveraged the strengths of convolutional neural networks (CNNs) in extracting local (high-frequency) features from images and the capabilities of Transformers in capturing global (low-frequency) features. To further enhance the model's performance, the experimental design incorporated the Inception dilated convolution module and the adaptive detail fusion module (ADI), as well as introduced novel mechanisms such as squeeze-enhanced axial attention (SeaAttention) and coordinate attention (CA), thereby enhancing the model's feature extraction capabilities. Data augmentation was achieved through the use of the Fourier transform (FT) method, with a frequency ramp structure employed to better balance the contributions of local (high-frequency) and global (low-frequency) feature components, optimizing the network's performance in feature extraction. The experimental results demonstrated that the TC - ICSA - YOLO v8 model achieved a mean average precision (mAP) of 97.01%, precision of 97.33%, and recall of 95.69% on the validation set, with a

收稿日期:2024 - 10 - 02　修回日期:2024 - 11 - 14

基金项目:国家自然科学基金项目(51466014)

作者简介:杨瑛(1979—),女,教授,博士,主要从事农业信息技术研究,E-mail:591908603@qq.com

通信作者:郑文轩(1980—),男,教授,博士,主要从事图像分析和计算机视觉技术研究,E-mail:wenxuanzhengdx@163.com

detection speed of 81.21 frames per second. Compared with the YOLO v8s model under the same conditions, the TC - ICSA - YOLO v8 model demonstrated superior target detection precision for night-time images, with an improvement in mAP of 14.65, 3.34, 0.52, 0.20, 6.68, and 5.54 percentage points over Faster R - CNN, YOLO v3, YOLO v7s, YOLO v8s, SwinTransformer, and RT - DETR models, respectively. The model's parameter count was reduced by 74.60, 34.16, 9.48, 16.81, 20.84, and 13.64 MB compared with that of these models. The improved model had high detection accuracy and less number of parameters, which was favorable for deployment on mobile. The detection model proposed had good target detection effect for pear, which can provide reference for object detection in complex environment, and can provide technical support for pear automated picking.

Key words: Transformer; pear; object detection; Fourier transform

0 引言

香梨成熟期较短,采摘工作需在短时间内完成,否则会出现烂果和落果等严重的经济损失问题。目前香梨采收仍以手动采摘为主,需要大量的劳动力,并且采摘效率低、成本高。因此,降低生产成本,提高采摘效率,实现香梨机械化、智能化采摘迫在眉睫。目标检测技术因其具有自动、高效、准确的特点,是实现智能化采摘的关键[1]。

近年来,随着智慧农业的兴起,基于深度学习的目标检测在农产品采摘领域展现出巨大的潜力。卷积神经网络(Convolutional neural network, CNN)和 Transformer[2]模型作为深度学习的两大基石,正逐渐成为智慧农业中不可或缺的技术工具。CNN 以其卓越的图像特征提取和分层表示能力,已经在图像分类[3]、目标检测[4]和语义分割[5]等计算机视觉任务中大放异彩;而 Transformer 这个最初在自然语言处理领域表现优异的模型,凭借其自注意力机制和捕捉长距离依赖关系的强大能力,也被引入到计算机视觉领域,为智慧农业的农作物采摘提供了新的解决方案。TU 等[6]提出一种多尺度的 MS - FRCNN 网络模型,在 Faster R - CNN[7]的基础上添加多尺度特征提取器,再分别训练 RGB 检测器和深度检测器,提升了对百香果小目标的检测能力。GAI 等[8]通过改进 YOLO v4 骨干网络,结合 DenseNet 将模型先验框的层间密度改为适合樱桃果实形状的圆形标记框,增强了模型特征提取和检测能力。高新阳等[9]针对现有柑橘检测算法准确率低、实时性差等问题改进 YOLO v5 模型,其检测精度达到 99.1%。赵鹏飞等[10]针对农田环境下甜椒误检率较高问题,提出了一种改进 YOLO v7 - Tiny 算法的甜椒果实检测方法。QI 等[11]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和番茄串的数量和成熟度水平。CARION 等[12]采用了先 CNN 再 Transformer 的串联拼接方式,提出了 DETR 模型,首次将 Transformer 用于图像目标检测。JOSH 等[13]则是选择先 Transformer 再 CNN,提出了 ViT - FRCNN 模型。该模型在 ViT[14](Vision Transformer, ViT)后串行拼接了 Faster R - CNN 作为模型的目标检测网络,实现利用 Transformer 完成图像目标检测任务的目的。PENG 等[15]选择将 CNN 和 Transformer 并行拼接,提出了 Conformer,实现以并行交互的方式融合 CNN 网络所提取的局部特征信息和 Transformer 网络所提取的全局特征信息,该模型在多个视觉识别任务中取得了优秀的性能。基于 CNN 和 Transformer 串并联拼接的混合方式有很多,这些混合模型的出现为计算机视觉领域带来了新的思路和方法,也为实现更加高效和准确的图像处理任务提供了新的途径。然而,鉴于此类混合模型结构的复杂性,在应用过程中,针对特定任务的选择以及模型架构的设计,必须依据具体应用场景和需求进行精确的定制。

针对现有目标检测模型在复杂环境下水果目标检测精度不高的问题,尤其是将卷积神经网络(CNN)和 Transformer 图像处理优势进行有机融合以实现农产品目标检测所面临的挑战,本研究在 YOLO v8s 模型基础上,设计一种简单高效的香梨目标检测模型 TC - ICSA - YOLO v8。模型采用设计的 TC - ICSNET 网络架构,替换 YOLO v8s 原有的骨干网络,以实现更有效的特征提取。在 TC - ICSNET 网络架构中,设计 InceptionZdilated 卷积模块,同时引入挤压增强轴向注意力机制(Squeeze-enhanced axial attention, SeaAttention)和坐标注意力机制(Coordinate attention, CA)等提升骨干网络的性能。在特征融合层,设计并引入自适应细节融合模块(Adaptive detail infusion, ADI),替代 YOLO v8s 中上采样和下采样层。采用加权交并比(Weighted intersection over union, WIoU)损失函数,替换 YOLO v8s 中使用的 CIoU 损失函数,以期优化模型的性能。采用傅里叶变换(Fourier transform, FT)作为一种数据增强技术[16],对图像数据进行处理。

1　试验数据

1.1　图像采集

所使用数据集于 2024 年 8 月 20—30 日在塔里木大学园艺试验站专属试验区域采集（采用佳能 EOS 77D 型相机对成熟期新梨 7 号香梨进行采集）。考虑到检测目标尺寸、相机分辨率、镜头焦距以及场景复杂性等多方面因素，试验设定 2 种图像拍摄距离：远距离（50 ~ 100 cm）和近距离（10 ~ 30 cm）。拍摄时间为 07：00—08：00、13：00—14：00 和 22：00—23：00 3 个时间段，在图像采集过程中，严格控制各类参数，以确保数据集的一致性与可靠性。剔除重复、拖影以及效果不理想的图像，最终筛选出 4 810 幅图像组成原始数据集，其中近距和远距图像各 2 405 幅，图像采集信息见表 1 和图 1。

表 1　采集图像信息

Tab. 1　Details of acquired images

采集距离	采集条件	图像数量/幅
近距	顺光	600
	逆光	560
	枝叶遮挡	530
	光照不均	415
	夜间	300
远距	顺光	600
	逆光	560
	枝叶遮挡	530
	光照不均	415
	夜间	300

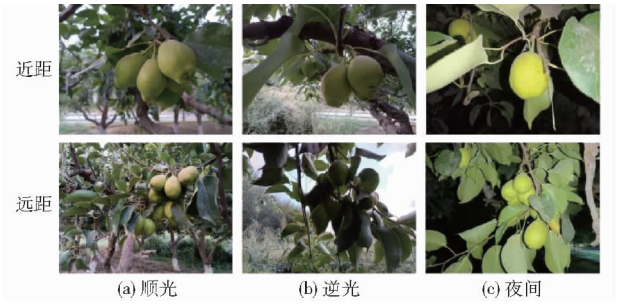

图 1　不同场景香梨图像

Fig. 1　Images of pear in different scenes

1.2　数据集制作

为了提升模型训练效果和提高模型泛化能力，需对数据集进行扩充。传统的数据增广方法主要采用随机翻转、剪裁、缩放、添加噪声、Mixup 算法[17]等，这些算法更注重样本的时域表现，忽略了频域信息。图像频域信息可体现灰度变化速度，保留图像轮廓信息，对图像进行频域数据增强处理可有效降低外部环境对检测精度的影响。为此，本文引入基于傅里叶变换的频域数据增强算法，以增强模型鲁棒性。图像增广后的数据集共 10 000 幅，以比例 6∶2∶2 划分为训练集（6 000 幅）、测试集（2 000 幅）和验证集（2 000 幅）进行模型训练和验证，所有图像在输入模型前尺寸均被调整至 448 像素 ×448 像素。

信号在时域和频域之间的变换依赖离散傅里叶变换（Discrete Fourier transform，DFT），二维离散傅里叶变换被广泛应用于图像变换，能将时域图像转换到频域。傅里叶变换及傅里叶反变换的核函数具有分离性，且指数部分具有周期性。通过傅里叶变换分离图像的相位分量，并构建不同振幅信息生成新图像。新图像与原图像轮廓相同，但幅值不同，以此达到数据增强的目的，从而增加训练样本多样性，提高所训练模型泛化性。需要注意的是，新的振幅信息来自训练样本中的其他图像，并非随机产生。

将 2 幅图像 $f(x_1,y_1)$ 和 $f(x_2,y_2)$ 分别进行快速傅里叶变换，公式为

$$F(\mu_1,v_1)=\mathrm{FFT}(f(x_1,y_1)) \tag{1}$$

$$F(\mu_2,v_2)=\mathrm{FFT}(f(x_2,y_2)) \tag{2}$$

对应的振幅为 $|F(\mu_1,v_1)|$ 和 $|F(\mu_2,v_2)|$，随机产生 1 个 $[0,1]$ 内的随机因子 α，以 α 和 $(1-\alpha)$ 为权重，组合得到 2 个新的振幅为

$$|F'(\mu_1,v_1)|=\alpha|F(\mu_2,v_2)|+(1-\alpha)|F(\mu_1,v_1)| \tag{3}$$

$$|F'(\mu_2,v_2)|=\alpha|F(\mu_1,v_1)|+(1-\alpha)|F(\mu_2,v_2)| \tag{4}$$

对重新组合的相位分量和新的振幅分量进行傅里叶逆变换，还原为时域图像。从数据集中随机选取 2 幅图像，采用傅里叶变换数据增强算法，并取随机因子 α 为 0.65，得到增强样本如图 2 所示。



图 2　傅里叶变换数据增强效果

Fig. 2　Effectiveness diagrams of Fourier transform data augmentation

由图 2 可知,应用傅里叶变换对图像进行数据增强处理后,图像在保持原有轮廓特征的同时,实现了色彩和对比度信息的交换,这些图像由于色彩特征的改变成为新的图像样本,从而达到数据增强的效果。

2 研究方法

2.1 检测模型构建

在 YOLO v8s 的基础上,优化设计 TC-ICSNET 网络架构作为检测模型的骨干特征提取网络。该网络架构的优势在于能有效融合卷积神经网络(CNN)提取图像局部(高频)特征信息的优点和发挥 Transformer 在处理长距离依赖关系方面出色表现,具有捕捉全局(低频)特征信息的优势,可弥补卷积神经网络(CNN)对全局(低频)特征信息的感知能力不足和 Transformer 捕捉局部(高频)特征信息局限的问题。为了进一步提升网络模型的性能,设计 Inception dilated 卷积模块和自适应细节融合模块,引入坐标注意力机制[18]和挤压增强轴向注意力机制[19]等提升网络特征提取能力。试验使用 WIoU 损失函数替换 YOLO v8s 中使用的 CIoU 损失函数,采用傅里叶变换增强数据,从而进一步提升模型泛化能力。为了描述方便,将设计的检测模型命名 TC-ICSA-YOLO v8,结构如图 3 所示。

2.2 TC-ICSNET 骨干特征提取模块

TC-ICSNET 骨干网络特征提取的设计思路来源于 Vision Transformer。在视觉任务中,Transformer 首先将输入的图像分割成一系列的标记(token),每个图像块(patch token)通过一个线性层被投影到隐藏表示向量中,表示为 $\boldsymbol{X}\in\mathbf{R}^{N\times C}$。如图 4 所示,其核心是一个 TC-ICSNET Block,在该模块中设计了一个 ICS-Fusion 特征融合器,分为局部(高频)和全局(低频)特征信息提取。局部(高频)特征信息提取模块采用双分支并行结构,分支 1 由 InD Conv Block 卷积模块串联坐标注意力模块 CA 组成;分支 2 由坐标注意力模块 CA 串联 Max Pool 操作组成。全局(低频)特征信息提取部分是由 Ave Pool 串联压缩增强轴向空间注意力机制 SeaAttention 操作组成。

输入的特征图为 $\boldsymbol{X}\in\mathbf{R}^{N\times C}$,将其沿着通道分为全局(低频)特征信息提取($q$)和局部(高频)特征信息提取($j$)2 个维度,即 $\boldsymbol{X}_q\in\mathbf{R}^{N\times C_q}$和 $\boldsymbol{X}_j\in\mathbf{R}^{N\times C_j}$,其中 $C=C_q+C_j$。将 $\boldsymbol{X}_q$ 和 $\boldsymbol{X}_j$ 分别分配给全局特征提取器和局部特征提取器。

在局部特征提取器中,采用并行结构提取图像的局部特征信息,试验将输入的 $\boldsymbol{X}_j$ 分为 $\boldsymbol{X}_{j1}\in\mathbf{R}^{N\times\frac{C_j}{2}}$和 $\boldsymbol{X}_{j2}\in\mathbf{R}^{N\times\frac{C_j}{2}}$,局部(高频)特征提取部分定义为

$$\boldsymbol{Y}_{j1}=\mathrm{Sigmoid}(\mathrm{InDConv}(\mathrm{CA}(\mathrm{FC}(\boldsymbol{X}_{j1}))))$$

$$\boldsymbol{Z}_{j1}=\boldsymbol{Y}_{j1}\odot\boldsymbol{X}_{j1} \tag{5}$$

$$\boldsymbol{Y}_{j2}=\mathrm{Sigmoid}(\mathrm{FC}(\mathrm{CA}(\mathrm{MaxPool}(\boldsymbol{X}_{j2}))))$$

$$\boldsymbol{Z}_{j2}=\boldsymbol{Y}_{j2}\odot\boldsymbol{X}_{j2} \tag{6}$$

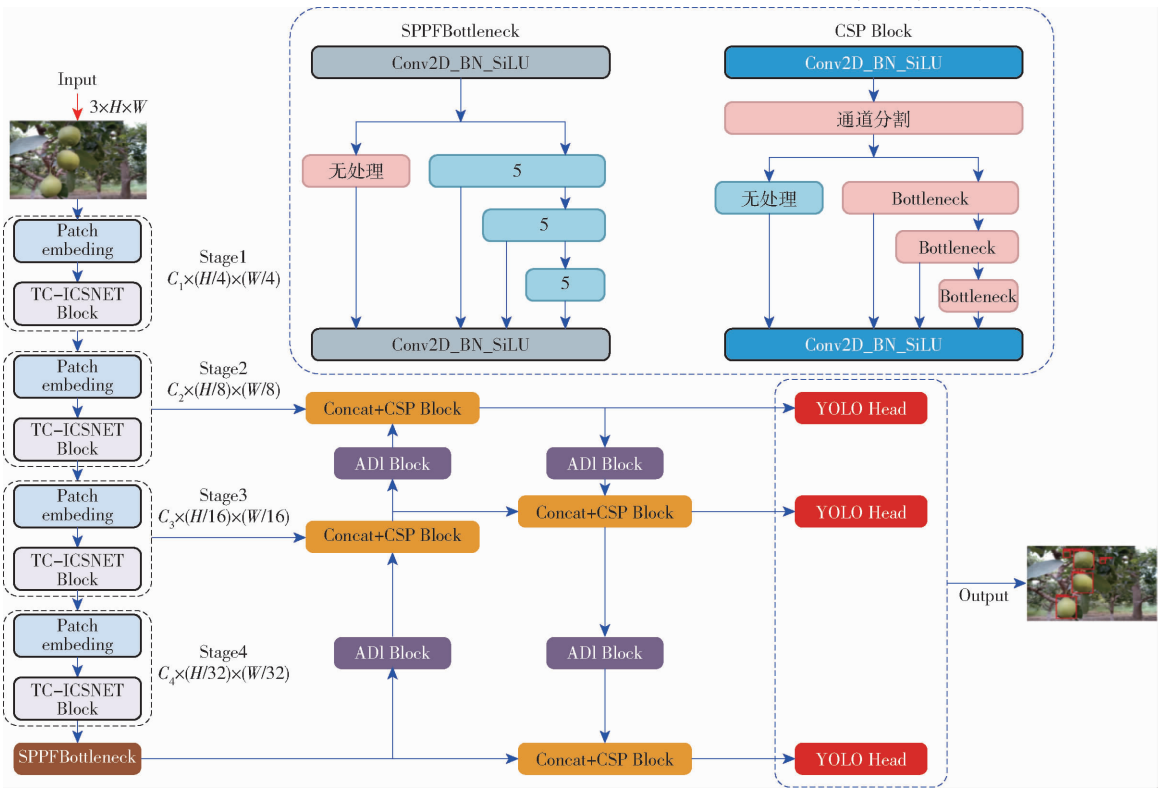

图 3 TC-ICSA-YOLO v8 检测模型结构

Fig. 3 TC-ICSA-YOLO v8 detection model

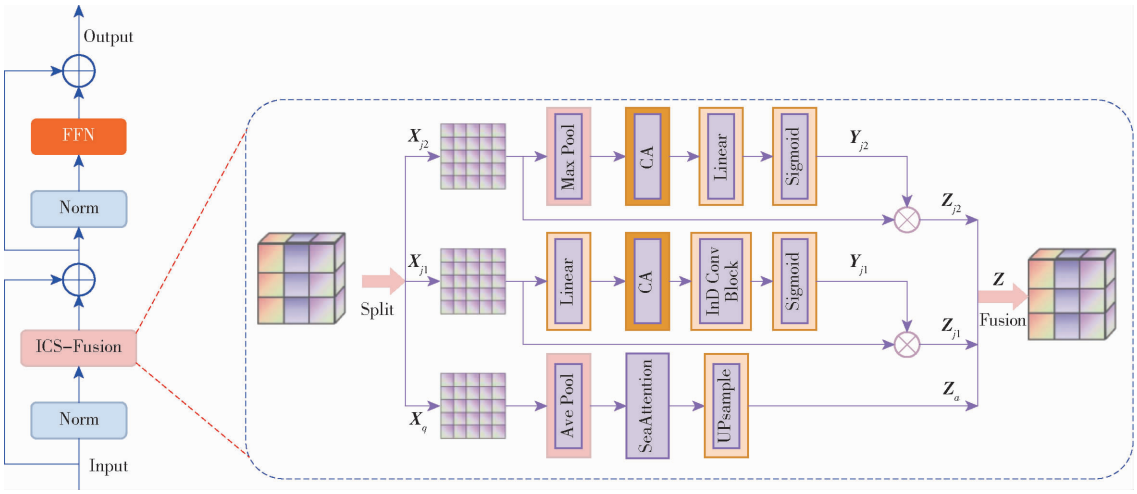

图 4　TC－ICSNET 骨干网络特征提取模块

Fig. 4　TC－ICSNET backbone feature extraction module

传统的自注意力机制的运算成本和内存消耗相对较高，为了降低计算成本，本文在低频特征提取时采用一种在学习全局特征信息强大能力的压缩增强轴向空间注意力机制（SeaAttention），使用前利用平均池化层减小输入特征的空间尺度，并在注意力操作之后，利用上采样来恢复原始的输入维度。这种设计可减少计算量，并使注意力操作集中于嵌入全局信息，低频特征提取部分定义为

$$\boldsymbol{Z}_q = \mathrm{Upsample}(\mathrm{SeaAttention}(\mathrm{AvePool}(\boldsymbol{X}_q))) \quad (7)$$

图像特征提取输出表达式为

$$\boldsymbol{Z} = \mathrm{Concat}(\boldsymbol{Z}_q, \boldsymbol{Z}_{j1}, \boldsymbol{Z}_{j2}) \quad (8)$$

2.3　频率斜坡结构

在一般视觉框架中，底层在捕获高频细节方面扮演着更为重要的角色，而顶层则在建模全局（高频）信息方面更为重要，即 ResNet[20] 的层次化表示。采用一种频率斜坡结构，从底层到顶层逐渐将更多的通道维度从低频混合器分配给高频混合器，从而保留更少的通道维度给高频混合器。如图 3 所示，骨干网络有 4 个阶段，具有不同的通道和空间维度。对于每个阶段，定义一个通道比率，以更好地平衡高频和低频成分，即 $\frac{C_q}{C}$，$\frac{C_j}{C}$，其中 $\frac{C_q}{C}+\frac{C_j}{C}=1$。在提出的频率斜坡结构中，$\frac{C_j}{C}$ 从浅层到深层逐渐减少，而 $\frac{C_q}{C}$ 逐渐增加。因此，随着灵活的频率斜坡结构，TC－ICSNET 骨干网络特征提取能够有效地在整个层面上权衡局部（高频）和全局（低频）特征信息。

2.4　挤压增强的轴向空间注意力机制

为了解决 Transformer 计算成本高，内存消耗大的问题，本文使用挤压增强的轴向空间注意力机制（SeaAttention），如图 5 所示。该机制在特征图的水平和垂直方向上通过压缩操作将全局特征信息保留到单个轴，然后分别应用自注意力机制建模轴上的长程依赖，可降低计算复杂度。

2.5　Inception dilated 卷积模块

为增强模型在复杂场景中的适应性，扩大感受野以及提升计算效率，设计 Inception dilated 卷积模块，如图 6 所示，由 2 个膨胀的 Inception 层构成。第 1 个膨胀 Inception 层后接 tanh 激活函数，作为滤波器使用。第 2 个膨胀 Inception 层后接 Sigmoid 门控单元，该门控单元充当门控滤波器，负

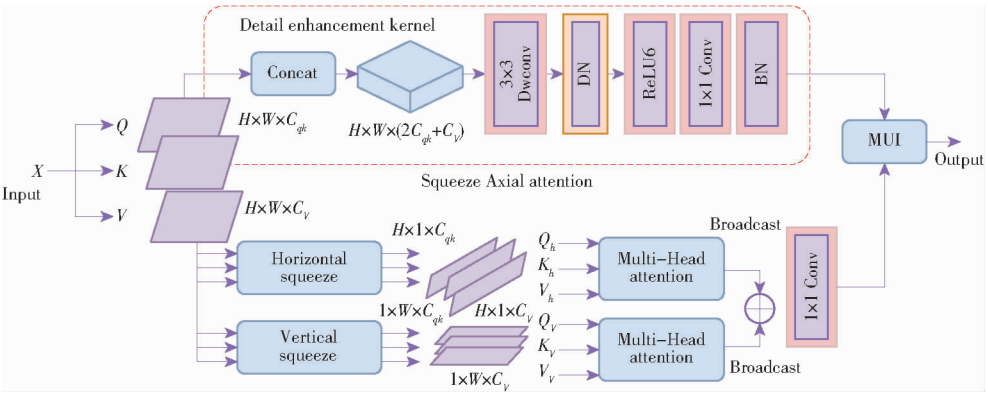

图 5　SeaAttention 注意力机制

Fig. 5　SeaAttention attention mechanism

责调节信息流，仅选择性地传递部分有用信息；在每个膨胀 Inception 层中，采用 3 个并行 Dilated 卷积，其膨胀率分别设置为 1、2 和 4；3 个分支输出被相加后，通过 Sigmoid 门控单元控制信息的输出，这种设计旨在通过多尺度特征融合和门控机制，提高模型对复杂场景的理解能力，同时优化计算效率。

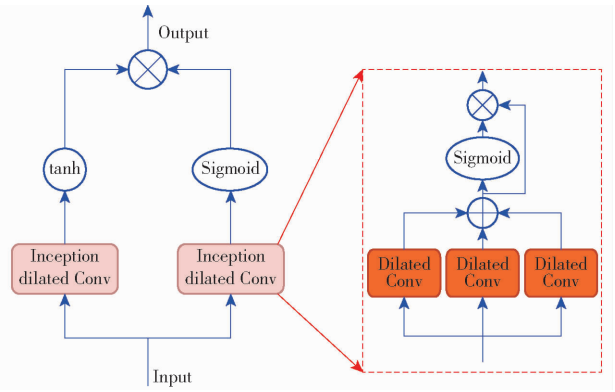

图 6 Inception dilated 卷积模块

Fig. 6 Inception dilated convolution module

2.6 自适应融合模块

在传统的多尺度特征融合模块中，特征图融合通常依赖于为特定尺寸设计的多个上采样和下采样层，这导致计算成本高、结构复杂以及训练困难等问题。针对这些局限性，设计自适应尺寸融合的 ADI 模块，旨在有效地处理和融合不同尺度的特征图。该模块设计灵感来源于 Unetv2[21] 中的分段和细节融合(Segment and detail infusion，SDI)模块，SDI 通过连接不同尺度的特征图减少特征差异，从而在编码器和解码器之间建立有效的信息流通路径。然而，与 SDI 不同，ADI 模块专注于编码过程中不同尺度特征的融合。

ADI 模块将第 1 个输入特征图的尺度定义为参考尺度。如图 7 所示，对于后续输入的特征图，模块根据与参考尺度的关系，通过不同选择匹配参考尺度。

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

(2)当输入特征图尺度与参考尺度匹配时，通过恒等映射直接传递特征图，无需进行尺度调整。

(3)当输入特征图尺度大于参考尺度时，采用自适应池化技术对特征图进行下采样，减小其尺寸以匹配参考尺度。

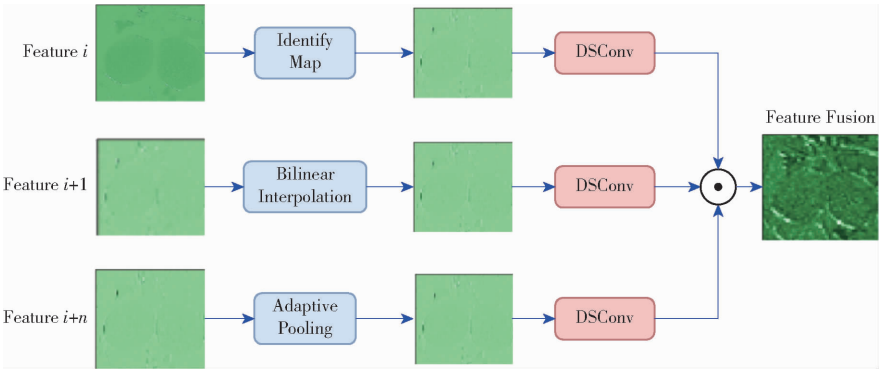

图 7 自适应尺寸融合模块

Fig. 7 Adaptive dimension integration module

调整尺寸后的特征图再经过一个深度可分离卷积[22](Depthwise separable convolution，DSConv)，旨在增强其表征能力。最终，所有调整尺寸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。

为了提高模型在复杂环境下对香梨关键特征的提取能力，减少背景因素的干扰，试验将 CA 注意力机制与 Inception dilated 卷积模块串联使用，这样不仅灵活、轻量，而且能够显式地建模特征的位置信息，这更有助于提高模型泛化能力。

3 试验与结果分析

3.1 环境配置与参数设置

本文使用 Win10 操作系统，处理器型号为 Intel(R) Core(TM) i7 - 12700KF 3.60 GHz，基于 Python 语言，在 PyTorch 框架下实现，具体环境配置如表 2 所示。

表 2 环境配置

Tab. 2 Environment configuration

语言/环境	版本/型号
GPU	NVIDIA GeForce RTX - 4090
CUDA	11.7
Python	3.8.0
PyCharm	2022.3
PyTorch - GPU	2.0

训练过程中采用 Adam 优化器更新参数，动量

设置 0.93，在训练过程中对学习率进行实时衰减处理，训练迭代次数 300，训练参数如表 3 所示。

表 3　参数设置

Tab.3　Parameters settings

参数	数值
迭代次数	300
样本数	16
优化器	Adam
初始学习率	1.0×10^{-4}

3.2　模型评价指标

为了评估 TC－ICSA－YOLO v8 模型在复杂环境下对香梨目标检测的能力，采用平均精度均值（mean Average Precision，mAP）、精确率（Precision）、召回率（Recall）、检测速度（Detection speed）、浮点运算量（FLOPs）和模型内存占用量作为评价指标。

3.3　不同损失函数性能对比

采用 Wise－IoUv3[24] 损失函数替换 YOLO v8s 中使用的 CIoU[25] 损失函数。为了系统评估损失函数 WIoU 在目标检测任务中的有效性，在控制变量的基础上，选取目标检测主流的边界框损失函数 CIoU、DIoU[26]、GIoU[27] 和 SIoU[28]，并将这些损失函数分别集成到 YOLO v8s 模型中，开展对比试验，结果见表 4。由表 4 可知，采用 WIoU 作为模型边界框损失函数时，模型在准确率和检测速度方面均表现更优，说明 WIoU 损失函数的智能权重调整机制能够更有效地平衡图像特征之间的梯度变化，减少由于特征缺陷而导致的误差，提升模型对香梨目标检测任务的适应性。相较于 CIoU 损失函数，采用 WIoU 损失函数的模型在平均精度均值（mAP）、精确率和召回率分别提升 0.71、0.64、0.26 个百分点，检测速度提升 0.47 f/s。与 SIoU 损失函数相比，WIoU 损失函数在平均精度均值（mAP）、精确率和召回率上分别提升 0.03、0.06、0.24 个百分点，检测速度提升 0.23 f/s。此外，从图 8 可以看出，采用 CIoU、DIoU、GIoU、SIoU 和 WIoU 损失函数训练模型，其损失值均能迅速收敛至较低水平，并保持稳定。

表 4　不同损失函数性能比较

Tab.4　Performance comparison of different loss functions

损失函数	平均精度均值/%	精确率/%	召回率/%	检测速度/（f·s^{-1}）
CIoU	96.64	96.85	93.87	94.19
DIoU	95.11	95.13	93.17	93.18
GIoU	95.35	95.61	93.34	93.93
SIoU	97.32	97.43	93.89	94.43
WIoU	97.35	97.49	94.13	94.66

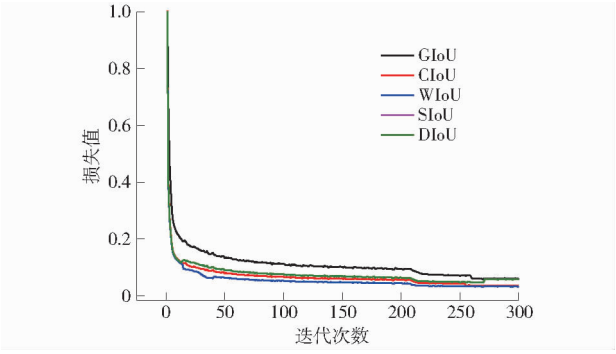

图 8　不同损失函数损失值

Fig.8　Iterations of different loss functions

与 DIoU 和 GIoU 损失函数相比，WIoU 损失函数在平均精度均值（mAP）、精确率、召回率和检测速度上明显优于二者。

3.4　TC－ICSA－YOLO v8 模型消融试验

为了更好理解 TC－ICSA－YOLO v8 网络模型中不同模块的作用，在相同试验条件下，从完整的模型中分别移除 InD ConvBlock、SeaAttenion、CA 和 AID 等模块，采用模型内存占用量、浮点运算量（FLOPs）和平均精度均值（mAP）综合评价各模块检测性能。

试验结果如表 5 所示，TC－ICSA－YOLO v8 网络模型与 YOLO v8s 模型相比，在模型内存占用量基本相同的情况下，平均精度均值（mAP）为 97.23%。

表 5　TC－ICSA－YOLO v8 模型消融试验结果

Tab.5　Ablation experiment results for TC－ICSA－YOLO v8 model

InD Conv Block	SeaAttenion	CA	AID	模型内存占用量/MB	浮点运算量	平均精度均值/%
×	√	√	√	34.83	1.121×10^{10}	92.62
√	×	√	√	34.83	1.121×10^{10}	95.39
√	√	×	√	34.83	1.120×10^{10}	97.17
√	√	√	×	34.84	1.120×10^{10}	97.15
√	√	√	√	34.84	1.115×10^{10}	97.23

注：√表示加入模块，×表示未加入模块。

3.5　香梨目标检测结果

3.5.1　白天强光条件检测结果

图 9 为白天强光条件下不同距离、有无遮挡和不同光照条件 YOLO v8s 和 TC－ICSA－YOLO v8 模型对香梨检测结果。由图 9 可知，在白天强光条件下近距离顺光或逆光香梨有无遮挡的情况下，YOLO v8s 和 TC－ICSA－YOLO v8 模型均可准确识别香梨果实，没有漏检和误检情况。在有枝叶遮挡情况下，YOLO v8s 出现漏检现象，如图 9f～9h 中紫色圆圈标识所示。而 TC－ICSA－YOLO v8 模型均无出现误检和漏检情况，展示了较为优秀的目标检测能力。

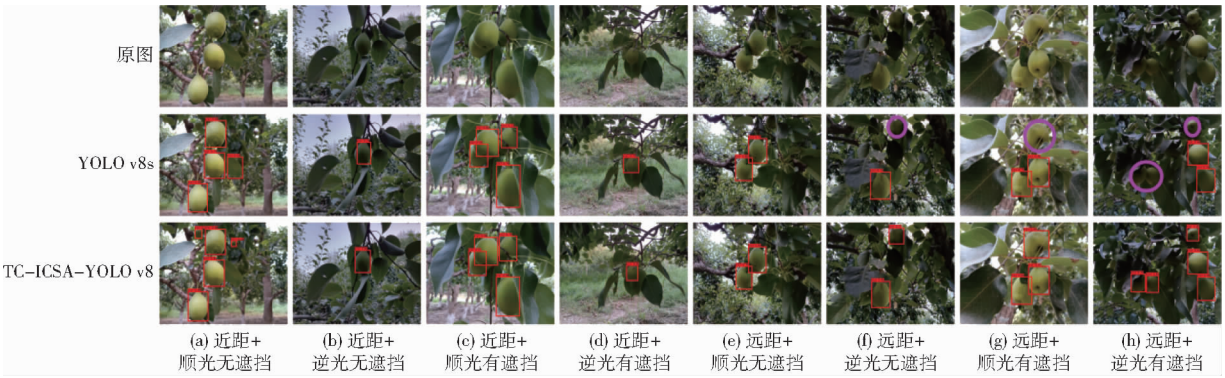

图 9 白天条件香梨果实图像检测结果

Fig. 9 Image detection of pear fruit on day

3.5.2 夜间条件检测结果

图 10 为在夜间条件下不同距离、有无遮挡 YOLO v8s 和 TC - ICSA - YOLO v8 模型对香梨检测结果。由图 10 可知,YOLO v8s 模型无论在近距离还是远距离有无遮挡的情况下,均出现了漏检现象,如图 10a、10c、10d 中紫色圆圈标识所示。TC - ICSA - YOLO v8 模型仅在远距离有遮挡的情况下出现个别香梨漏检和误检情况,如图 10d 中紫色圆圈标识所示。

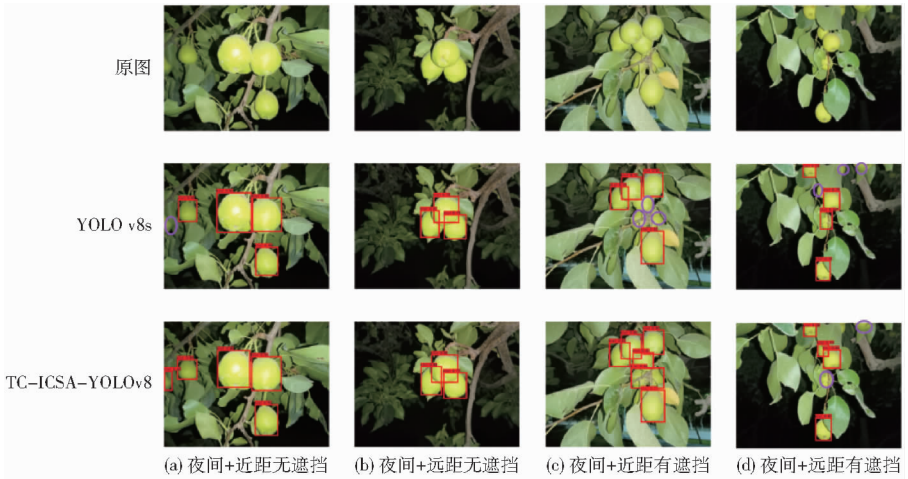

图 10 夜间条件香梨果实图像检测结果

Fig. 10 Image detection of pear fruit on night

3.5.3 不同条件下不同模型目标检测结果

表 6 为不同条件下模型检测结果,由表 6 可知,TC - ICSA - YOLO v8 模型在白天和夜间,近景和远景拍摄图像数据集的测试中,其平均精度均值(mAP)、精确度和召回率均优于 YOLO v8s 模型。

表 6 不同条件下模型检测结果

Tab. 6 Model detection performance under different conditions %

模型	时间	场景	平均精度均值	精确率	召回率
YOLO v8s	白天	远景	96.53	96.79	93.35
		近景	97.33	97.39	93.97
	夜间	远景	96.11	96.57	93.14
		近景	96.97	97.12	93.68
TC - ICSA - YOLO v8	白天	远景	97.19	97.45	93.77
		近景	97.36	97.43	94.21
	夜间	远景	96.13	96.66	93.21
		近景	97.23	97.29	94.15

在白天近景拍摄数据集测试中,TC - ICSA - YOLO v8 模型平均精度均值(mAP)相较于 YOLO v8s 模型提高 0.03 个百分点,而在夜间近景的情况下,平均精度均值(mAP)提高 0.26 个百分点。而在白天远景拍摄数据集的测试中,TC - ICSA - YOLO v8 模型平均精度均值(mAP)相较于 YOLO v8s 模型提高 0.66 个百分点,而在夜间远景的情况下,平均精度均值(mAP)提高 0.02 个百分点。

3.5.4 不同目标检测模型检测结果

为了验证 TC - ICSA - YOLO v8 模型性能,试验选用 Faster R - CNN、YOLO v3[29]、YOLO v7s[30]、YOLO v8s[31]、SwinTransformer[32]、RT - DETR[33] 模型与 TC - ICSA - YOLO v8 模型在验证集上进行性能对比。从表 7 可知,TC - ICSA - YOLO v8 模型在白天强光条件验证集的平均精度均值(mAP)、精确率和召回率分别达到 97.01%、97.33% 和 95.69%,

表 7　不同类型检测模型在验证集性能比较

Tab. 7　Performance comparison of different types of detection models on validation set

模型	平均精度均值/%	精确率/%	召回率/%	模型内存占用量/MB	检测速度/(f·s^{-1})
Faster R－CNN	82.36	85.92	80.55	102.39	28.36
YOLO v3	93.67	94.96	91.51	61.95	49.27
YOLO v7s	96.49	96.82	94.77	37.27	81.67
YOLO v8s	96.81	96.97	95.53	44.60	94.65
SwinTransformer	90.33	92.19	89.77	48.63	20.33
RT－DETR	91.56	91.97	90.33	41.43	41.17
TC－ICSA－YOLO v8	97.01	97.33	95.69	27.79	81.21

检测速度达到 81.21 f/s。其平均精度均值(mAP)与相应模型对比分别提高 14.65、3.34、0.52、0.20、6.68、5.45 个百分点。其模型内存占用量分别减少 74.60、34.16、9.48、16.81、20.84、13.64 MB，TC－ICSA－YOLO v8 模型的检测速度快，内存占用量更少，有利于部署在移动端。

3.5.5　检测热力图分析

利用 Grad－CAM 生成目标检测热力图，如图 11 所示。由图 11 可以看出 TC－ICSA－YOLO v8 模型的目标检测热力图更贴近真实梨果区域。YOLO v8s 目标检测热力值相对 TC－ICSA－YOLO v8 模型的目标检测热力值更低，表现出漏检现象，说明 TC－ICSA－YOLO v8 模型可以从弱语义的目标中提取更强的目标特征，有着较好的鲁棒性和泛化能力。

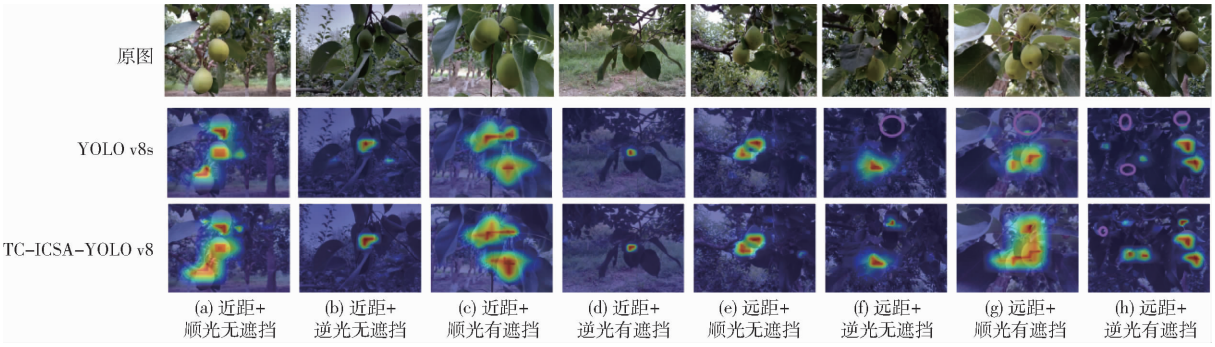

图 11　香梨图像检测热力图

Fig. 11　Heatmap of pear fruit image detection

4　结论

(1)针对复杂自然环境下香梨目标精准检测问题，在 YOLO v8s 的基础上，优化设计了 TC－ICSA－YOLO v8 检测模型。该模型有效地融合了卷积神经网络(CNN)提取图像局部(高频)特征信息和 Transformer 提取图像全局(低频)特征信息的优势。在验证集上模型平均精度均值(mAP)、精确率和召回率分别达到 97.01%、97.33% 和 95.69%，检测速度达到 81.21 f/s。

(2)消融试验结果表明，将 InDConv Block 卷积模块、SeaAttention、CA 和 AID 共同集成到设计的 ICS－Fusion 特征融合器中，其目标检测平均精度均值(mAP)达到 97.23%。

(3)选用 Faster R－CNN、YOLO v3、YOLO v7s、YOLO v8s、SwinTransformer、RT－DETR 模型与 TC－ICSA－YOLO v8 模型在验证集上进行性能对比。试验结果表明，TC－ICSA－YOLO v8 模型平均精度均值(mAP)、精确率、召回率、检测速度和内存占用量等均优于传统的目标检测模型。

参　考　文　献

[1] 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)

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

[3] XIAO T，MINTUN E，DARRELL T，et al. Early convolutions help transformers see better[J]. Advances in Neural Information Processing Systems，2021，34(3)：30392－30400.

[4] FANG Y，LIAO B，WANG X，et al. You only look at one sequence：rethinking transformer in vision through object detection[J]. Advances in Neural Information Processing Systems，2021，34(1)：34－49.

[5] ZHENG S，LU J，ZHAO H，et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2021：6881－6890.

[6] TU S，PANG J，LIU H，et al. Passion fruit detection and counting based on multiple scale Faster R－CNN using RGB-Dimages[J]. Precision Agriculture，2020，21(5)：1072－1091.

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

[8] GAI R, CHEN N, YUAN H. A detection algorithm for cherry fruits based on the improved YOLOv4 model[J]. Neural Computing and Applications, 2023, 35(19): 13895 - 13906.
[9] 高新阳,魏晟,温志庆,等. 改进 YOLOv5 轻量级网络的柑橘检测方法[J]. 计算机工程与应用,2023,59(11):212 - 221.
GAO Xinyang, WEI Sheng, WEN Zhiqing, et al. Citrus detection method based on improved YOLOv5 lightweight network[J]. Computer Engineering and Applications, 2023, 59(11): 212 - 221. (in Chinese)
[10] 赵鹏飞,钱孟波,周凯琪,等. 改进 YOLOv7 - Tiny 农田环境下甜椒果实检测[J]. 计算机工程与应用,2023,59(15):329 - 340.
ZHAO Pengfei, QIAN Mengbo, ZHOU Kaiqi, et al. Improvement of sweet pepper fruit detection in YOLOv7 - Tiny farming environment[J]. Computer Engineering and Applications, 2023, 59(15): 329 - 340. (in Chinese)
[11] 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: 114 - 117.
[12] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]// 16th European Conference of Computer Vision—ECCV, Glasgow, 2020: 213 - 229.
[13] JOSH B, ERIC K, ERIC T, et al. Toward transformer-based object detection[J]. arXiv preprint arXiv: 2012.09958, 2012.
[14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2010.
[15] PENG Z, HUANG W, GU S, et al. Conformer: local features coupling global representations for visual recognition[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, 2021: 357 - 366.
[16] 吕珊. 面对复杂场景的鲁棒目标检测算法研究[D]. 北京:北京交通大学,2022.
LÜ Shan. Research on robust object detection algorithms for complex scenes[D]. Beijing: Beijing Jiaotong University, 2022. (in Chinese)
[17] ZHANG H, CISSE M, DAUPHIN Y, et al. Mixup: beyond empirical risk minimization[C]// ICLR 2018 International Conference on Learning Representations, 2018: 12 - 22.
[18] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and PatternRecognition. Washington, 2021: 13713 - 13722.
[19] WAN Q, HUANG Z, LU J, et al. SeaFormer: squeeze-enhanced axial transformer for mobile semantic segmentation[J]. arXiv preprint arXiv: 2301.13156, 2023.
[20] HE K M, ZHANG X, REN S. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 - 778.
[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: 116 - 129.
[22] FRANCOIS C. Xception: deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1251 - 1258.
[23] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1 - 12.
[24] TONG Z, CHEN Y, XU Z, et al. Wise - IoU: bounding box regression loss with dynamic focusing mechanism[J]. arXiv preprint arXiv: 2301.10051, 2023.
[25] REZATOFIGHI H, TSOI N, AMIR J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2019: 658 - 666.
[26] ZHENG Z, WANG P, LIU W, et al. Distance - IoU loss: faster and better learning for bounding box regression [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 1 - 7.
[27] DANELLJAN M, KHAN F, MICHAEL F, et al. Adaptive color attributes for real-time visual tracking[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2014, 143(1): 1090 - 1097.
[28] GEVORGYAN Z. SIoU Loss: more powerful learning for bounding box regression [J]. Computer Vision and Pattern Recognition, 2022, 1: 1 - 12.
[29] 武星,齐泽宇,王龙军,等. 基于轻量化 YOLOv3 卷积神经网络的苹果检测方法[J]. 农业机械学报,2020,51(8):17 - 25.
WU Xing, QI Zeyu, WANG Longjun, et al. Apple detection method based on light - YOLOv3 convolutional neural network [J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(8): 17 - 25. (in Chinese)
[30] 宋怀波,马宝玲,尚钰莹,等. 基于 YOLOv7 - ECA 模型的苹果幼果检测[J]. 农业机械学报,2023,54(6):233 - 242.
SONG Huaibo, MA Baoling, SHANG Yuying, et al. Detection of young apple fruits based on YOLOv7 - ECA model[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(6): 233 - 242. (in Chinese)
[31] 李慧琴,宋赵铭,刘存祥,等. 基于 YOLOv8n 的番茄果实检测模型改进[J/OL]. 河南农业大学学报,https://doi.org/10.16445/j.cnki.1000-2340.20240511.002.
LI Huiqin, SONG Zhaoming, LIU Cunxiang, et al. Improvement of tomato fruit detection model based on YOLOv8n[J/OL]. Journal of Hennan Agricultural University, 2024, https://doi.org/10.16445/j.cnki.1000-2340.20240511.002. (in Chinese)
[32] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows [C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, 2021: 9992 - 10002.
[33] ZHAO Y, LV W, XU S, et al. DETRs beat YOLOs on real-time object detection[J]. arXiv preprint arXiv: 2304.08069, 2023.