doi:10.6041/j.issn.1000-1298.2026.18.010

面向无人机巡检的玉米病虫害图文协同检测模型

承达瑜[1,2]　苏　皓[3]　梁　优[1]　赵　伟[4]　王建东[5]　宋　辞[2]

(1. 河北工程大学矿业与测绘工程学院, 邯郸 056038; 2. 地理信息科学与技术全国重点实验室, 北京 100101;
3. 河北工程大学信息与电气工程学院, 邯郸 056038; 4. 华南师范大学北斗研究院, 佛山 528225;
5. 中国农业科学院农业环境与可持续发展研究所, 北京 100081)

摘要: 结合无人机巡检与深度学习检测玉米病虫害是智慧农业发展趋势。然而深度学习模型检测性能不仅受图像样本数量影响,还受田间环境、成像质量及模型泛化能力等因素制约,导致单一图像模态难以全面表征病虫害特征。针对上述问题,本文基于 YOLO-World 模型,融合文本语义信息,提出一种面向无人机巡检的玉米病虫害图文协同检测模型。该模型结合重参数化视觉几何网络、注意力机制及多尺度融合模块,重构了图像特征提取网络,提高了多尺度图像特征提取能力;通过在文本注入图像中加入特征线性调制层和多类注意力,在图像嵌入文本中加入前馈模块、门控机制和残差结构,改进可重参数化的视觉-语言路径聚合网络,提升了图像与文本特征的语义对齐能力与表达精度。从无人机巡检视角采集 500 幅玉米病虫害图像,包含早期玉米黄叶病和虫害导致的烂叶病 2 类,并通过数据增强扩展至 2 000 幅,结合对应文本描述构建 COCO 格式的多模态数据集。试验结果表明,在 500 幅数据集条件下(分别取 25-shot、50-shot 训练样本,固定 400 幅验证样本),本文模型 mAP@0.5 较单模态 YOLO v8 系列提升 9.40~10.40 个百分点,较 YOLO-World 提升 2.80~4.40 个百分点;在 2 000 幅数据集上(按比例 8∶2 划分训练集与验证集),mAP@0.5 较单模态 YOLO v8 系列提高 5.40~10.70 个百分点,较 YOLO-World 提升 3.10 个百分点。该模型可为小样本条件下无人机巡检识别玉米病虫害提供新的研究思路。

关键词: 玉米病虫害; 无人机巡检; YOLO-World; 多模态; 小样本检测

中图分类号: S435.13; TP391.41　　**文献标识码**: A　　**文章编号**: 1000-1298(2026)18-0106-11　　**OSID**:

Image-Text Collaborative Detection Model for Corn Pests and Diseases in Drone-based Inspections

Cheng Dayu[1,2]　Su Hao[3]　Liang You[1]　Zhao Wei[4]　Wang Jiandong[5]　Song Ci[2]

(1. *School of Mining and Geomatics, Hebei University of Engineering, Handan* 056038, *China*
2. *State Key Laboratory of Geographic Information Science and Technology, Beijing* 100101, *China*
3. *School of Information and Electrical Engineering, Hebei University of Engineering, Handan* 056038, *China*
4. *Beidou Research Institute, South China Normal University, Foshan* 528225, *China*
5. *Institute of Environment and Sustainable Development in Agriculture, Chinese Academy of Agricultural Sciences, Beijing* 100081, *China*)

Abstract: Combining drone inspection with deep learning for maize pest and disease detection represents a trend in the development of smart agriculture. However, the detection performance of deep learning models is influenced not only by the number of image samples but also constrained by factors such as field environment, imaging quality, and model generalization capability, making it difficult for a single image modality to fully characterize pest and disease features. To address the above issues, a vision-language collaborative detection model for maize pests and diseases under drone inspection was proposed based on the YOLO-World framework, integrating textual semantic information. The model reconstructed the image feature extraction network by incorporating a re-parameterized visual geometry network, attention mechanisms, and multi-scale fusion modules, thereby improving the multi-scale image feature extraction capability. By adding a feature linear modulation layer and multi-class attention to the text-injected

收稿日期: 2026-02-16　修回日期: 2026-03-10

基金项目: 河北省重大科技成果转化专项(22287401Z)、国家自然科学基金项目(42171212)和地理信息科学与技术全国重点实验室开放基金项目

作者简介: 承达瑜(1980—),男,副教授,博士,主要从事地理大数据挖掘、作物长势遥感监测及计算机视觉研究,E-mail: yuyumails@126.com

image, and introducing a feed forward module, gating mechanism, and residual structure to the image-embedded text, the re-parameterizable vision-language path aggregation network was enhanced, which improved the semantic alignment capability and expression accuracy between image and text features. In the experiments, totally 500 corn pest and disease images were captured from a drone inspection perspective, including two categories: early corn yellow leaf disease and leaf rot caused by pests. The dataset was augmented to 2 000 images, and a multi-modal dataset in COCO format was constructed by incorporating corresponding textual descriptions. Experimental results showed that on the 500 image dataset (with 25-shot and 50-shot training samples respectively and a fixed validation set of 400 images), the proposed model improved mAP@0.5 by 9.40 ~ 10.40 percentage points over the single modal YOLO v8 series and by 2.80 ~ 4.40 percentage points over YOLO - World. On the 2 000 image dataset (with an 8∶2 training-validation split), mAP@0.5 was improved by 5.40 ~ 10.70 percentage points over the single modal YOLO v8 series and by 3.10 percentage points over YOLO - World. This model can provide a research approach for identifying corn pests and diseases under few-shot conditions in drone based inspection.

Key words: corn diseases and pests; unmanned aerial vehicle inspection; YOLO - World; multi-modal; few-shot object detection

0　引言

玉米作为全球主要粮食作物之一,产量与品质常受病虫害的影响[1]。传统病虫害检测方法多依赖人工目视和专家经验,效率低、主观性强,难以满足大规模农田快速、精准监测的需求[2]。随着计算机视觉技术的发展,基于视觉信息处理的病虫害检测方法已成为农业领域的研究热点[3-4],特别是深度学习与无人机巡检技术结合,可高效获取作物全生育期图像数据,实现对多种病虫害的精准识别,有助于减轻病虫害对农作物的危害,提高农业生产效益[5-7]。当前,面向无人机巡检的病虫害图像识别主要分为 2 类:①飞后识别,即基于无人机采集的 RGB 或红外图像,结合农田复杂背景,通过改进现有目标检测模型实现[8-10]。②实时识别,通过将无人机与边缘计算设备结合,并借助轻量化深度学习模型,实现巡检过程实时识别,如文献[11 - 12]通过模型轻量化改进,实现了该应用。尽管现有方法取得明显成效,但深度学习图像识别模型精度普遍依赖于大量样本。而在现实的大田环境,大规模、高质量的标注样本采集困难,而且仅单一图像信息,也难以全面描述农作物病虫害情况[13-14]。

当前,针对小样本条件下农作物病虫害图像识别,已有诸多学者展开研究并取得一系列进展。Li 等[15]利用农业场景中大量未标注图像,提出半监督小样本学习方法,通过划分源域与目标域试验,结合置信区间自适应选择未标注样本并赋予伪标签,有效提升了跨域、小样本条件下病害识别精度;Egusquiza 等[16]引入 Siamese 网络结构,结合 Triplet Loss 进行深度度量学习,对 5 种作物的 17 种真菌病害实现聚类与分类,增强了类内紧凑性与类间分离度;Liang[17]通过图像分割、度量学习与小样本卷积神经网络结合,提升了棉花叶斑病识别准确性;Wu 等[18]提出的 MSUN 模型,通过多表示建模、子域语义对齐与不确定性控制机制,在无监督跨域植物病害识别中表现优异。尽管上述方法通过数据增强以扩充训练样本,或借助在大规模数据集上预训练模型进行特征迁移,在一定程度上缓解了样本稀少问题,但大多仍局限于单一模态的图像特征提取,缺乏对农作物整体信息理解,易受单模态噪声干扰[12,19-21]。近年来,随着传感器技术快速发展,农作物监测已具备更丰富的数据来源,除图像数据外,还包括光谱、雷达、热红外以及文本描述等多模态信息。已有研究开始探索多模态融合在病虫害识别、分类与预测中的应用。Yang 等[22]融合柑橘图像与高光谱信息,基于 ResNet50 提取图像特征,并引入注意力机制对高光谱数据进行降维,通过双线性融合实现了对柑橘黄龙病的有效预测;Cao 等[23]结合图像与文本模态,采用对比学习方式挖掘模态间相关性,进而识别黄瓜病害。上述研究结果表明,充分利用不同模态间的信息互补性,发展多模态的农作物病虫害识别方法,正成为小样本条件下农作物病虫害精准识别的重要趋势。

基于此,本文在玉米病虫害图像样本稀少情况下,结合病虫害文本描述信息,构建图像-文本的多模态训练数据集,并通过改进 YOLO - World 提出一种玉米病虫害图文协同检测模型。引入重参数化视觉几何网络、注意力机制与多尺度融合模块,重构图像特征提取网络,增强多尺度目标表征能力;改进视觉-语言特征对齐路径,在图像与文本双向交互中引入特征调制、注意力及门控残差结构,提升跨模态语义对齐精度与检测性能,以期在少样

本情况下为无人机智能监测农作物病虫害提供新的技术思路。

1 材料与方法

1.1 多模态数据集获取与构建

研究区域位于河北省邯郸市邱县古城营镇中部中型农场(36°52′55″N,115°06′05″E,图 1),种植作物以夏玉米与冬小麦为主,总面积约 51.85 hm^2。研究区地势平坦、排灌条件良好,便于无人机低空飞行与影像采集。

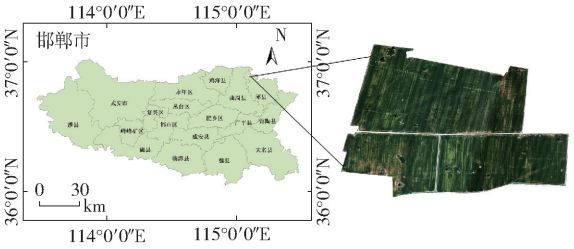

图 1 数据采集区概况

Fig. 1 Overview of data acquisition zone

现有玉米病虫害公开数据集(如 PlantVillage、Corn Leaf Disease Dataset 等)多在实验室或单一背景板环境中拍摄,背景单一、光照均匀,与无人机巡田时复杂的自然背景差异显著,限制了模型在真实农田环境下泛化能力。为此,本文聚焦于无人机巡检视角下玉米病虫害检测,考虑到无人机主要获取叶片上表面的可见光影像,所以模型重点针对叶部病虫害展开研究,以无人机航拍影像为主构建玉米病虫害图像数据集,采集早期玉米黄叶病和虫害导致的烂叶病 2 类样本。二者虽由不同胁迫因素引起,但均在叶片表面表现出明显的颜色与结构退化特征,适合视觉检测与语义表征。

数据采集使用大疆 M300 RTK 型无人机作为机载平台,搭载高清可变焦 RGB 相机(3 840 像素 × 2 160 像素),飞行高度控制在 3 ~ 4 m,以获取清晰的叶部细节并避免因过低飞行造成的视角畸变。选择晴朗、微风条件下进行航拍,并在上午、正午与下午 3 个时段取样,以模拟不同巡检光照环境,确保数据真实性与多样性。

考虑到模型训练对样本多样性与场景复杂度的需求,在无人机影像基础上进一步引入 2 类补充数据:①手机拍摄数据:利用智能手机从近地垂直或倾斜视角拍摄玉米叶部病虫害样本,以模拟无人机低空飞行或变焦观测时视角特征,用于补充近尺度条件下叶片病斑与虫害细节信息。②公开数据筛选:从 Kaggle 玉米病虫害数据集中剔除实验室环境及纯色背景图像,仅保留在自然光条件下、背景复杂且贴近实际农田场景的样本,并筛选与本文研究对象一致的 2 类玉米病虫害图像;同时对图像分辨率进行约束,保留分辨率为 1 224 像素 × 1 632 像素和 1 728 像素 × 2 304 像素样本,以保证成像质量与尺度一致性。3 类数据分别占总体图像比例约为:无人机影像 60%(300 幅)、手机拍摄数据 20%(100 幅)、Kaggle 数据 20%(100 幅),共计 500 幅原始图像,在训练过程与无人机影像等权使用,未单独设置采样权重,亦未针对不同成像设备来源引入显式域差异建模或对齐策略。多源且贴近实际农业场景的成像数据混合使用,有助于验证模型在真实复杂条件下鲁棒性。为克服样本量不足的问题,进一步采用随机翻转、旋转、高斯模糊等小样本常用的数据增强方法,将图像数量扩充至 2 000 幅,从而在保证样本真实性的同时提升数据多样性与模型泛化能力。不同来源及增强后玉米病虫害图像样例如图 2 所示,涵盖多种光照条件、叶片角度及遮挡情况,充分反映真实田间环境下多样场景特征。其中图 2i 为局部放大图。考虑到全文缩小展示时模糊效果不明显,选取局部进行呈现,其模糊区域采用高斯模糊技术处理——该技术通过卷积运算对图像进行平滑处理,有效弱化图像细节信息,呈现区域像素遮盖效果。



图 2 不同来源及增强后玉米病虫害图像样例

Fig. 2 Examples of corn pest and disease images from different sources and after enhancement

为进一步提升模型在小样本场景下学习能力,引入文本语义信息辅助图像特征学习,进而增强模型在低样本条件下识别精度与泛化性能。数据集标注采用 COCO 格式,该格式是微软团队提出的标准化数据标注体系,适用于目标检测、实例分割、关键点检测等任务,核心字段包含 categories、annotations、images 等。在其基础上扩展了病虫害类别及文本描述字段,实现玉米病虫害图像与语义信息的关联标注,同时达成图像与文本标注信息的统一管理与高效调用,相关数据以 JSON 文件存储,具体标注示例如表 1 所示。

表 1　玉米病虫害数据标注示例

Tab. 1　Examples of data annotation for corn pests and diseases

数据类型	内容实例	数据格式
文本	[[“玉米叶片表现出多种叶绿素缺失症状，颜色由深绿色逐渐褪为黄色甚至褐色。部分叶片在叶脉之间或叶缘出现发黄现象，并伴有叶缘卷曲和叶尖枯焦。在严重情况下，叶片会变得干燥易碎。症状的严重程度常在整个植株冠层中分布不均，可在中下部叶片上呈现孤立或成簇发生的情况。”],[“玉米叶片上可见不规则的孔洞和参差不齐的裂口，叶缘处的缺刻痕迹显示出被害虫取食的迹象。咀嚼痕的形状与深度各不相同，邻近组织可能因感染或干旱而萎蔫或变色。这类损伤是典型的咀嚼型害虫危害特征，常同时出现在植株的上下部叶片上。”]]	JSON
多模态	{“categories”:[{“id”:0,“name”:“Corn leaf Yellowing Disease”, “supercategory”:“disease”},{“id”:1,“name”:“CornEarworm Damage”,“supercartegory”:“disease”}],“annotations”:[{“id”:0,“image_id”:0,“category_id”:1,“area”:128904.66439833601,“bbox”:[2096.99904,914.00076, 635.0016, 202.99896],“iscrowd”:0}],“images”:[{“file_name”:“1.jpg”,“id”:0,“width”:3840,“height”:2160}]}	COCO（JSON）

1.2　研究方法

1.2.1　改进 YOLO－World 模型

YOLO－World 是一种融合文本和图像特征的开放场景目标检测模型，主要包含 YOLO 检测器、文本编码器和可重参数化的视觉语言路径聚合网络（Re-parameterizable vision-language path aggregation network，RepVL－PAN）3 部分[24]。其中，YOLO 检测器基于 YOLO v8 架构，采用 Darknet 骨干提取多尺度图像特征，并通过 PAN 构建特征金字塔以融合多层信息，预测头用于边界框回归与目标分类[25]；文本编码器基于 CLIP 预训练，将输入文本编码为文本嵌入向量，实现将自然语言描述向图像特征融合的语义对齐，为跨模态融合提供语义基础[26]；RepVL－PAN 则负责多尺度图像与文本特征的对齐与融合，提升视觉语义表示能力。

针对玉米病虫害图像中细粒度特征复杂、背景干扰强及目标尺度差异显著的问题，在 YOLO－World 框架基础上构建了结构增强的多模态检测模型。首先，在图像特征提取阶段，引入基于结构重参数化思想的 RepVGG 主干网络[27]，其原始设计旨在通过“训练期多分支、推理期单分支”的等价变换，在不增加推理计算量前提下提升特征表达能力，本文将其用于增强无人机图像中多尺度病害特征表达能力；同时结合多尺度特征融合与注意力机制，以强化对复杂田间背景下关键视觉特征提取。其次，在跨模态特征交互阶段，针对图像与文本语义对齐不足的问题，对 RepVL－PAN 结构进行改进：在文本注入图像路径中引入特征线性调制层（Feature-wise linear modulation，FiLM[28]），其原始设计用于利用外部条件信息对特征分布进行自适应调制，本文将文本语义作为条件信号以引导图像特征重加权；同时结合卷积块注意力模块（Convolutional block attention module，CBAM[29]），通过通道与空间注意力增强病害区域的判别性表达。在图像嵌入文本路径中，引入前馈模块（Feed-forward module，FFM[30]）以提升跨模态非线性映射能力，并结合门控机制（Graph-based adaptive tool evolution，GATE[31]）与残差结构，实现对多模态信息流的自适应筛选与稳定融合。上述模块间的协同，有效增强了视觉–语言特征交互深度与语义一致性，提升病虫害检测精度，整体结构如图 3 所示。

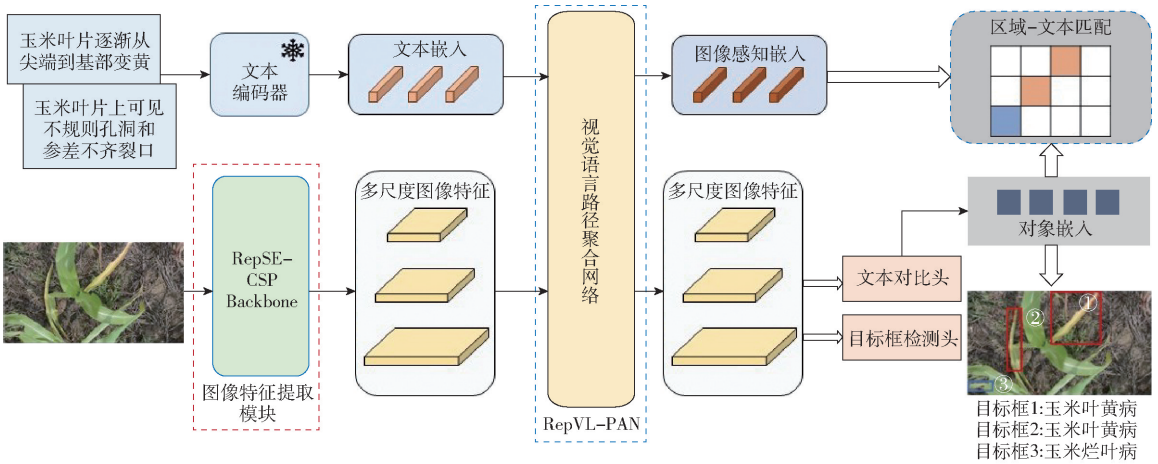

图 3　改进 YOLO－World 模型架构

Fig. 3　Improved YOLO－World model architecture

1.2.2 改进图像特征提取模块

图像特征提取模块基于 YOLO v8 骨干网络进行改进。YOLO v8 骨干网络采用 CSPDarknet 结构，通过快速下采样完成边缘、纹理和颜色等底层特征的初步提取，随后在浅层、中层与深层逐步构建更高语义层次的特征表示，并最终利用 SPPF 模块进行多尺度信息融合与特征增强。针对小样本条件下病虫害目标纹理细微、尺度差异大且背景复杂等难点，在原有 CSPDarknet 基础上设计了结构增强型骨干网络 RepSE－CSP Backbone（图 4）。

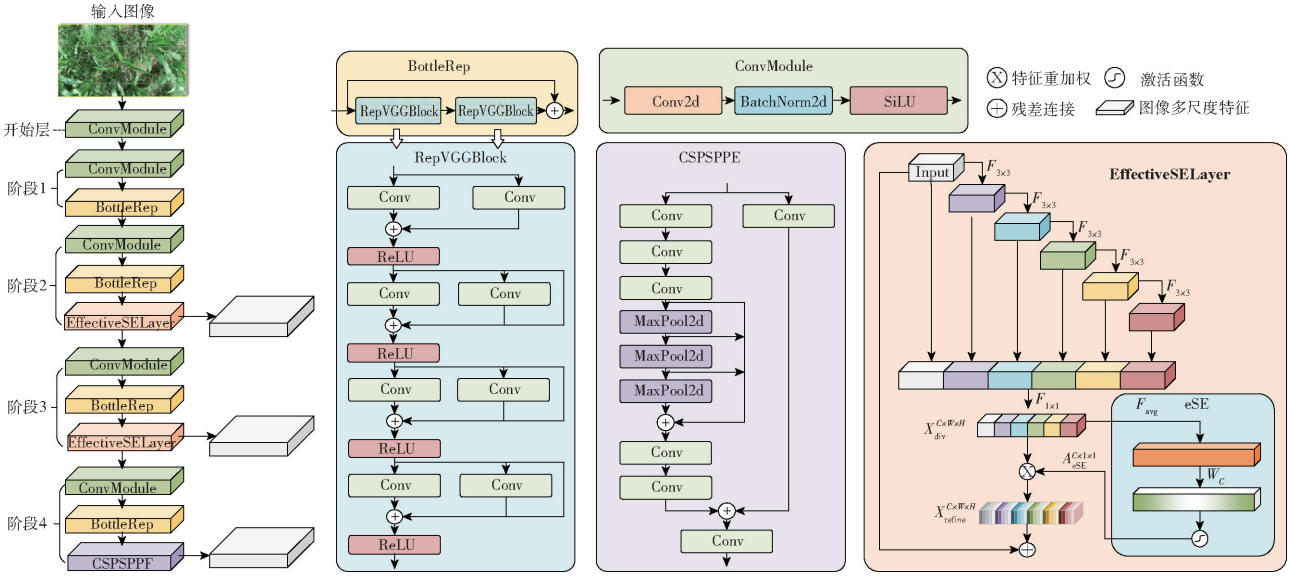

图 4 RepSE－CSP Backbone 网络结构

Fig. 4 RepSE－CSP Backbone network architecture

该骨干网络结构融合了 RepVGG 重参数化策略、轻量级注意力机制与高效的多尺度特征融合模块，在保持结构轻量性的同时提升感受野与细粒度特征建模能力。整体改进主要体现在 3 方面：

（1）骨干 C2f 模块替换为 BottleRep：将 YOLO v8 骨干网络结构中 CSPLayerWith－TwoConv 模块替换为更具表达能力的 BottleRep 模块。BottleRep 模块融合 ResNet 的残差思想与 RepVGG 的重参数化策略，通过构建更深层次的瓶颈结构，增强信息流动与特征重用能力，以提高骨干网络特征提取能力，适用于玉米病虫害的细粒度目标检测任务[27,32]。在训练阶段，BottleRep 模块采用典型的多分支残差结构，包含 3 个并行路径：3×3 卷积分支、1×1 卷积分支与 Identity 恒等映射分支，每条路径后均接有独立的 Batch Normalization 层。设输入特征为 x，则前向传播计算式为

$$y = x + \sum_{i=1}^{n} \mathrm{BN}_i(f_i(x)) \tag{1}$$

式中 n——分支数

$f_i(x)$——3 类卷积变换（3×3 卷积、1×1 卷积和恒等映射）

BN_i——第 i 条分支对应的 BatchNorm 操作

在推理阶段，通过参数重构将并行分支结构融合为 1 个等效的单分支 3×3 卷积结构，实现精度与速度兼顾，计算式为

$$f_{rep}(x) = \sum_{i=1}^{n} \mathrm{Conv}_i(x) \tag{2}$$

式中 $f_{rep}(x)$——重参数化的 3 类卷积变换融合表示

Conv_i——第 i 个分支融合后等效卷积操作

（2）引入轻量注意力机制：在进行层级特征提取时引入 EffectiveSELayer 通道注意力模块，自适应地调整不同通道的重要性，使模型聚焦于病虫害区域的显著特征，有效提升在复杂背景下目标定位和识别能力[33]。令输入特征图为 $F \in X^{C\times H\times W}$，$C$、$H$、$W$ 分别表示通道数、特征图高度、特征图宽度。通道注意力权重 a 为

$$a = \sigma(W_2\delta(W_1\mathrm{AVG}(F))) \tag{3}$$

式中 $\mathrm{AVG}(F)$——特征图 F 全局平均池化

W_1、W_2——全连接层参数

δ——ReLU 激活函数

σ——Sigmoid 激活函数，用于归一化输出

输出加权特征图 F_{att} 为

$$F_{att} = F \otimes a \tag{4}$$

式中 $\otimes$——通道级别的逐元素乘法

（3）引入 CSPSPPF 替换 SPPF：在骨干网络的最终特征融合阶段，采用 CSPSPPF 结构替换 YOLO v8 中原 SPPF 模块，以实现更充分的跨阶段特征交互。CSPSPPF 引入 CSP 的双分支路径设计，其中恒等映射分支用于保留输入特征的原始细节，SPPF 分支则负责进行多尺度感受野扩展与语义特征聚合。双分

支输出在拼接后形成兼具细节信息与高层语义的融合特征，有助于提升模型对不同尺度病虫害目标的表达能力。该结构在设计上旨在保持骨干网络轻量性的同时强化多尺度特征建模。

1.2.3　改进视觉语言路径聚合网络

为提升视觉与文本信息交互能力，增强模型对复杂农业场景适应性，提出了多类型注意力融合的可重参数化视觉－语言特征融合模块（Multi-type attention re-parameterizable vision-language PAN，MTAtt－RepVL－PAN）。该结构在 YOLO－World 架构基础上，聚焦于注意力机制的优化设计，核心目标是提升多模态特征的语义对齐能力与表达精度。整体结构如图 5 所示，遵循自上而下与自下而上的双向路径构建特征金字塔结构，基于多尺度图像特征｛C3，C4，C5｝提取融合特征｛P3，P4，P5｝，并分别在特征增强与图文融合阶段引入 TCAM－CSPBlock 与 GI－Pooling Attention 2 个关键模块。

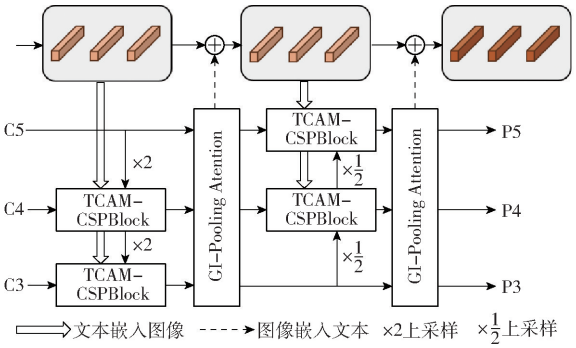

图 5　MTAtt－RepVL－PAN 网络结构

Fig. 5　MTAtt－RepVL－PAN network architecture

（1）TCAM－CSPBlock：采用模块化设计使不同类型注意力机制协同工作，在使用 MaxSigmoidAttnBlock 将文本嵌入作为引导以增强与文本描述相关的视觉特征的基础上，通过 CBAM 注意力机制自适应增强重要特征并抑制无关特征，同时引入 FiLM 层，借助文本嵌入生成条件参数对视觉特征进行线性调制，以实现更精细的语义对齐。如图 6 所示，这种串行注意力增强策略允许每个类注意力机制在不同阶段分别发挥作用，逐步提升特征表达能力。

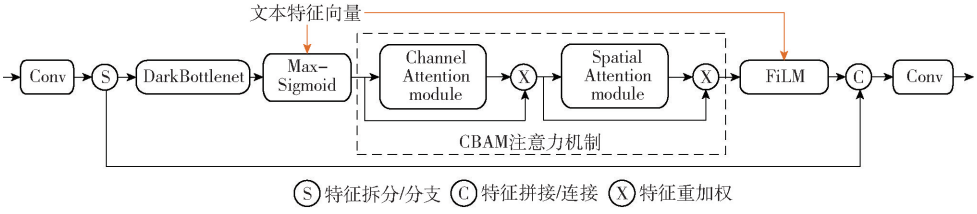

图 6　TCAM－CSPBlock 网络结构

Fig. 6　TCAM－CSPBlock network architecture

（2）GI－Pooling Attention：结构如图 7 所示，在 YOLO－World 原 I－Pooling Attention 基础上，对图像与文本特征进行多头交叉注意力机制（Multi-head cross attention，MHCA）计算前插入 FFM 模块，通过非线性变换强化跨尺度特征关联性，提升语义表达能力，再引入动态门控融合机制，利用 Sigmoid 激活函数生成自适应权重，依据文本与图像匹配程度实时调整融合比例增强模型对模糊描述和复杂场景的处理能力，最后设计双层残差调节结构，将门控层局部融合比例调节与缩放因子的全局增强强度控制相协同，在优化特征融合效果的同时减少噪声干扰并进一步提升模型鲁棒性。

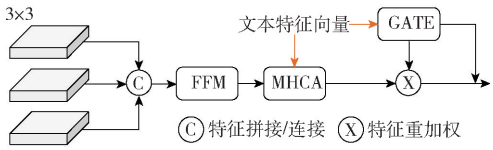

图 7　GI－Pooling Attention 网络结构

Fig. 7　GI－Pooling Attention network architecture

2　结果分析

2.1　试验环境及评价指标

试验基于 MMYOLO[34] 和 MMDetection[35] 工具箱开发。所有试验在 2 台服务器上进行，分别配备 NVIDIA Quadro RTX 8000 和 NVIDIA RTX A5000 GPU，CPU 均为 Intel（R）Xeon（R）Gold 6248R@3.00 GHz。文本编码器采用开源的 CLIP（ViT－B/32），训练批处理大小设置为 8，其余超参数包括模型均采用 YOLO－World－v2－M 官方默认配置。

采用 mAP@0.5 和 mAP@0.5：0.95 作为核心评估指标，全面量化目标检测模型在玉米病虫害识别任务中分类精度与定位性能。mAP@0.5 衡量模型在 IoU 阈值大于等于 0.5 下对病虫害目标的检测性能，贴近农业生产实际需求；mAP@0.5：0.95 则评估模型在多个严格定位 IoU 阈值（0.5～0.95）下平均表现。

2.2　小样本条件下模态性能试验

为系统评估所提多模态模型在小样本条件下检测性能，构建一个统一的小样本试验体系，并从样本规模与增强策略 2 个维度开展多组对比试验。包括 4 种小样本设置：25-shot、50-shot、500 幅原始样本集以及 2 000 幅增强样本集。其中，shot 为图像幅数，

依据小样本学习中常用的“N-way K-shot Q-query”规则,N 个类别随机选取 K 幅图像样本用于训练,剩余 Q 幅图像样本不参与训练,仅用于验证测试[36]。500 幅与 2 000 幅样本集按比例 8∶2 划分为训练集与验证集。各类别训练样本通过随机抽样方式选取,并在不同试验轮次中固定随机种子,以确保数据划分与结果的可复现性。每组模型训练重复运行 5 次,最终结果取 5 次试验平均值,以减小随机性对性能评估的影响,实现统一小样本条件下全面模型检测。

由表 2 可知,多模态模型在 4 种小样本设置下均明显优于单模态系列模型。以 YOLO v8 系列的最优检测结果为基线,在 25-shot 与 50-shot 任务中,多模态融合相较最优单模态 mAP@0.5 提升 5.10 ~ 9.50 个百分点,mAP@0.5∶0.95 提升 3.70 ~ 4.70 个百分点;进一步改进的多模态结构提升幅度扩大至 9.40 ~ 10.40 个百分点(mAP@0.5)和 6.40 ~ 6.50 个百分点(mAP@0.5∶0.95)。在 500 幅原始样本集上,改进模型较最佳单模态模型 mAP@0.5 提升 3.90 个百分点,mAP@0.5∶0.95 提升 2 个百分点;在 2 000 幅增强样本集上,改进模型较最佳单模态模型提升 5.40 个百分点(mAP@0.5)和 1.90 个百分点(mAP@0.5∶0.95)。结果表明,多模态融合策略在小样本条件下能够稳定提升检测精度,且结合结构优化与数据增强后,模型泛化能力得到进一步增强,充分验证了所提方法在小样本农业病虫害检测中的有效性与鲁棒性。

表 2 小样本数据集模态性能试验

Tab. 2 Modal performance testing for few-shot datasets %

模型	25-shot		50-shot		500 幅		2 000 幅	
	mAP@0.5	mAP@0.5∶0.95	mAP@0.5	mAP@0.5∶0.95	mAP@0.5	mAP@0.5∶0.95	mAP@0.5	mAP@0.5∶0.95
YOLO v8n	15.40	5.53	21.80	9.18	37.30	17.40	46.50	22.70
YOLO v8s	12.00	5.20	21.20	8.54	39.50	18.00	48.80	24.70
YOLO v8m	11.60	5.18	21.30	8.54	39.50	19.20	51.30	26.80
YOLO v8l	3.11	1.40	18.10	7.79	39.70	19.40	51.80	27.50
YOLO v10n	15.50	5.66	22.60	10.60	38.80	18.20	48.10	23.41
YOLO 11n	16.30	6.21	21.30	9.56	39.10	18.51	49.90	24.40
RT-DETR	15.50	5.40	22.90	11.60	38.30	17.30	48.80	23.40
YOLO-World-v2-M	21.40	10.30	28.40	12.90	41.30	20.10	54.10	28.10
本文模型	25.80	12.01	31.20	15.60	43.60	21.40	57.20	29.40

为直观展示多模态方法优势,图 8 对比了在 2 000 幅样本数据集上单模态最优检测模型 YOLO v8l 与本文改进多模态模型在农作物病虫害检测任务中的表现。相比 YOLO v8l 在无人机视角下对部分玉米病虫害目标存在漏检的问题,改进 YOLO-World-v2-M 模型在文本特征支持以及特征提纯与图文融合注意力机制作用下,表现出更优的检测能力,漏检现象相对减少。



图 8 识别可视化结果

Fig. 8 Visual result of object recognition between YOLO v8l and improved YOLO-World-v2-M

2.3 图文多模态检测模型性能试验

在 2.2 节小样本模态性能试验基础上,为进一步验证所提多模态检测框架在跨模型场景中的竞争力,在小样本条件下补充开展了多模态模型对比试验,系统评估不同主流视觉-语言多模态检测模型在农业病虫害任务中的适配性与性能表现。试验选取具有代表性的跨模态检测体系,包括对齐驱动的 MDETR、基于大规模图文预训练的 GLIPv1-T 与 GLIPv2-T、具备更强跨模态表达能力的 Grounding DINO-T,以及本文最终采用改进的 YOLO-World。在试验设置上,数据集使用 2.2 节中 500 幅原始样本集以及 2 000 幅增强样本集,使用相同的数据集划分与训练设置。各模型均输入同一套农业病害类别描述,以消除语义差异带来的偏置。在训练策略方面,YOLO-World 同时采用“预编码”与“后编码”2 种模式:其中,预编码模式在推理前将文本提示编

码为特征向量并进行缓存，从而在图像处理阶段移除文本编码器；后编码模式则在每幅图像推理时实时运行文本编码器。训练策略均遵循各模型官方建议配置。最终结果如表 3 所示。

表 3　小样本数据集多模态模型试验结果

Tab. 3　Test result of multimodal models for few-shot datasets

数据集	模型	mAP@0.5/%	推理速度/(f·s^{-1})	权重文件大小/GB
500 幅样本	MDETR	32.50		3.02
	GLIPv1 - T	36.00	0.11	3.90
	GLIPv2 - T	38.20	0.11	3.71
	Grounding DINO - T	39.50	1.24	0.65
	YOLO - World - v2 - M	41.30	20.45(64.22)	0.69
	本文模型	43.60	20.33(63.85)	0.70
2 000 幅样本	MDETR	44.80		3.05
	GLIPv1 - T	49.00	0.10	3.90
	GLIPv2 - T	51.00	0.10	3.75
	Grounding DINO - T	53.50	1.24	0.65
	YOLO - World - v2 - M	54.10	20.49(64.34)	0.69
	本文模型	57.20	20.45(64.22)	0.70

注：YOLO - World 的推理速度包含 2 种模式：括号外数值为后编码模式，需为每幅图像实时运行文本编码器；括号内数值为预编码模式，该模式在推理前将文本提示预编码为特征向量并缓存，从而在图像处理过程移除文本编码器，以获得更高的吞吐量。

由表 3 可知，各类跨模态检测模型在无人机农业病虫害识别任务中表现存在差异。整体来看，MDETR 与 GLIP 系列模型在本研究构建的小样本农业数据集上检测精度相对较低：在 500 幅样本条件下，mAP@0.5 为 32.50% ~38.20%；当样本规模扩展至 2 000 幅时，检测精度提升至 44.80% ~51%。Grounding DINO - T 在 2 000 幅样本条件下 mAP@0.5 为 53.50%，表现出一定的检测能力，但推理速度仅为 1.24 f/s，在实时无人机巡检场景中仍存在一定应用限制，且其性能提升对样本规模依赖较为明显。相比之下，YOLO - World - v2 - M 在 2 种样本规模下均保持了较为均衡的检测精度与推理效率，在后编码模式下分别为 20.45、20.49 f/s，预编码模式下分别为 64.22、64.34 f/s，体现了在工程部署与实时性方面的优势。

在此基础上，本文模型在保持与 YOLO - World - v2 - M 相同主干结构与参数规模前提下，在 500、2 000 幅样本下 mAP@0.5 分别为 43.60% 与 57.20%，相较于 YOLO - World - v2 - M 提升 2.30、3.10 个百分点，均为同组模型中最高值。结果表明，本文所引入的多模态特征融合与语义对齐策略在无人机农业病虫害场景下能够更有效地挖掘有限样本中的判别信息，在保证推理效率的同时进一步提升检测精度，验证了方法设计的有效性与针对性。

2.4　消融试验

为系统评估改进图像特征提取模块（RepSE - CSP Backbone）与视觉语言路径聚合模块（MTAtt - RepVL - PAN）在小样本目标检测中性能贡献，对 25-shot、50-shot、500 幅原始样本集以及 2 000 幅增强样本集 4 组小样本数据设置消融试验。每组试验均采用相同的数据划分与训练配置，并分别测试单独启用与联合启用 2 个模块的性能变化，以验证各模块的独立效果及其协同增益。

试验结果如表 4 所示，整体趋势表明：在 4 种小样本设置下，同时启用 RepSE - CSP Backbone 与 MTAtt - RepVL - PAN 模块模型均取得最优性能。

表 4　RepSE - CSP Backbone 与 MTAtt - RepVL - PAN 模块消融试验

Tab. 4　Ablation testing of RepSE - CSP Backbone and MTAtt - RepVL - PAN module　%

数据集	RepSE - CSP Backbone	MTAtt - RepVL - PAN	mAP@0.5	mAP@0.5:0.95
25-shot	×	×	21.40	10.30
	×	√	23.62	11.21
	√	×	24.15	11.70
	√	√	25.80	12.01
50-shot	×	×	28.10	12.90
	×	√	29.74	14.11
	√	×	29.20	14.10
	√	√	31.20	15.60
500 幅样本 (8:2)	×	×	41.30	20.10
	×	√	42.30	21.10
	√	×	42.80	21.83
	√	√	43.60	21.40
2 000 幅样本 (8:2)	×	×	54.10	28.10
	×	√	55.31	29.10
	√	×	56.20	28.90
	√	√	57.20	29.40

注：√表示启用模块，×表示未启用模块。

其中,mAP@0.5 提升幅度最高可达 4.40 个百分点,mAP@0.5:0.95 提升最高为 2.70 个百分点,优于未启用改进模块的基线模型。进一步分析发现,单独引入 MTAtt - RepVL - PAN 模块 mAP@0.5 提升 1.0 ~ 2.2 个百分点和 mAP@0.5:0.95 提升 0.9 ~ 1.2 个百分点;而单独引入 RepSE - CSP Backbone 模块则 mAP@0.5 提升 1.10 ~ 2.70 个百分点与 mAP@0.5:0.95 提升 0.8 ~ 1.7 个百分点。

结果表明,2 种模块在小样本条件下均能有效提升检测精度,并且联合启用时具有较好的互补与增强作用。如表 4 所示,随着样本规模增加,各改进模块在不同数据设置下性能提升趋势更加明显。其中,联合启用 2 模块的模型在 mAP@0.5 上始终表现最优,在 4 种小样本场景下均较基线模型有明显优势,验证了所提结构优化方案的有效性与鲁棒性。

3 讨论

试验结果表明,传统基于单模态图像的作物病虫害检测方法在真实农田环境中仍面临显著挑战。一方面,实验室或受控环境下采集的数据难以覆盖田间复杂光照、背景干扰及遮挡等情况,模型在实际应用中的泛化能力受限;另一方面,在小样本条件下,单一视觉模态难以充分表征病虫害的细粒度差异特征,检测性能明显下降。尽管已有研究通过数据增强[37]、度量学习[38]及改进卷积神经网络结构[39-40]等方式对上述问题进行探索,但整体仍主要依赖视觉信息,未能充分利用语义层面的互补特征。本文从多模态角度出发,引入病虫害文本语义信息并设计改进的图文融合注意力机制,有效增强了模型在复杂农田背景下鲁棒性与判别能力。同时,基于 YOLO - World 框架的重参数化设计,使模型在推理阶段无需加载文本编码器即可完成检测,有利于无人机平台上的高效部署与实时应用。

从应用场景看,本研究聚焦于无人机巡检视角下玉米叶部病虫害的早期检测。考虑无人机主要获取叶片上表面的可见光影像,模型围绕黄叶病与烂叶病 2 类典型症状开展建模与验证。试验结果表明,多模态方法在无人机航拍条件下具备良好适应性,验证了图文协同建模在农业无人机场景中可行性与推广潜力。然而,需要指出的是,本文模型目前仅针对有限类别、单一作物及 RGB 可见光数据进行验证,对多病害类型、多作物及跨光谱条件下泛化能力尚缺乏系统评估。

从理论层面看,本文验证了图像与文本模态融合在农业病虫害检测中的有效性,为小样本条件下目标识别提供了一种新的建模思路;同时,将跨模态特征交互机制引入无人机农业监测任务中,突破了传统单模态检测框架在语义表达能力上的限制,拓展了多模态学习在农业信息学领域的应用边界。

在实践层面上,依托无人机平台的高效巡航与实时成像能力,本文方法能够实现大面积农田的快速覆盖,为病虫害早期预警提供技术支持。尽管当前模型在病害类别覆盖范围及细粒度识别能力方面仍存在不足,尚难完全替代人工诊断,但作为智能预警与辅助决策模块具有较高应用价值。未来研究可进一步引入多光谱或多时相遥感数据,扩展病虫害类别与作物类型,并结合更轻量化、自适应的多模态网络结构,以提升模型在复杂田间环境下稳定性与实用性。

4 结论

(1)提出了一种基于改进 YOLO - World 的无人机巡检玉米病虫害检测模型:首先采用 YOLO v8 骨干网络提取玉米病虫害的多尺度图像特征,在此基础上引入 RepVGG 结构、注意力机制与多尺度融合模块,进一步提升玉米病虫害图像特征的提取能力与表达能力;随后通过 RepVL - PAN 实现跨模态融合,并在文本注入图像分支中引入 FiLM 层与多类注意力,在图像嵌入文本分支中引入门控机制、前馈模块及残差结构,有效增强图像与文本特征的语义对齐能力及表达精度,提升了农田生产场景下小样本玉米病虫害检测精度。

(2)构建了涵盖玉米叶黄病与虫害导致的烂叶病 2 类病害的多模态数据集,用于验证所提方法在数据稀少条件下的性能。试验结果表明,在 25-shot、50-shot 小样本设置下,本文模型 mAP@0.5 较单模态 YOLO v8 系列模型最优结果提升 9.40 ~ 10.40 个百分点;在采用数据增强策略的 2 000 幅样本数据集上,mAP@0.5 较单模态 YOLO v8 系列模型结果提升 5.40 ~ 10.70 个百分点,最终检测精度达到 57.20%。在基本满足小样本场景下快速部署的农业病虫害检测需求的同时,模型还具有良好的边缘设备部署能力,为农业病虫害智能监测提供了实用方案,并为后续无人机自动化检测病虫害研究提供了可行参考。

参 考 文 献

[1] 朱惠斌, 王明鹏, 白丽珍, 等. 基于视觉识别的玉米病虫害检测与精准变量喷药系统研究[J]. 农业机械学报, 2024,

55(增刊 2)：210 – 221.
Zhu Huibin, Wang Mingpeng, Bai Lizhen, et al. Maize pest and disease detection and precise variable spraying system based on visual recognition[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(S2): 210 – 221. (in Chinese)

[2] 王兴旺, 查海涅, 卢浩男, 等. 基于 YOLO v8 – STSF 的多类别害虫识别算法与监测系统研究[J]. 农业机械学报, 2025, 56(6): 228 – 236.
Wang Xingwang, Zha Hainie, Lu Haonan, et al. Multi-category pest identification algorithm and monitoring system based on YOLO v8 – STSF[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(6): 228 – 236. (in Chinese)

[3] 翟肇裕, 曹益飞, 徐焕良, 等. 农作物病虫害识别关键技术研究综述[J]. 农业机械学报, 2021, 52(7): 1 – 18.
Zhai Zhaoyu, Cao Yifei, Xu Huanliang, et al. Review on key technologies for crop pest and disease identification[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(7): 1 – 18. (in Chinese)

[4] 慕君林, 马博, 王云飞, 等. 基于深度学习的农作物病虫害检测算法综述[J]. 农业机械学报, 2023, 54(增刊 2): 301 – 313.
Mu Junlin, Ma Bo, Wang Yunfei, et al. Review of crop disease and pest detection algorithms based on deep learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(S2): 301 – 313. (in Chinese)

[5] Chin R, Catal C, Kassahun A. Plant disease detection using drones in precision agriculture[J]. Precision Agriculture, 2023, 24(5): 1663 – 1682.

[6] Bhandari M, Chang A, Jung J, et al. Unmanned aerial system-based high-throughput phenotyping for plant breeding[J]. The Plant Phenome Journal, 2023, 6(1): e20058.

[7] Anam I, Arafat N, Hafiz M S, et al. A systematic review of UAV and AI integration for targeted disease detection, weed management, and pest control in precision agriculture[J]. Smart Agricultural Technology, 2024, 9: 100647.

[8] 吕培鑫. 基于无人机航拍图像的大豆病虫害目标检测算法研究[D]. 济南: 山东大学, 2022.
Lü Peixin. Research on detection algorithm of soybean pest based on UAV images[D]. Ji'nan: Shandong University, 2022. (in Chinese)

[9] Zhang J, Zhang D, Liu J, et al. DSCONV – GAN: a UAV – BASED model for *Verticillium wilt* disease detection in Chinese cabbage in complex growing environments[J]. Plant Methods, 2024, 20(1): 186.

[10] Antolinez G A, Caceres C J W. Identification of pathogens in corn using near-infrared UAV imagery and deep learning[J]. Precision Agriculture, 2023, 24(2): 783 – 806.

[11] 赵伟. 面向无人机巡检图像的玉米病虫害识别方法研究[D]. 邯郸: 河北工程大学, 2024.
Zhao Wei. Research on corn pest and disease detection method for the images of UAV inspection[D]. Handan: Hebei University of Engineering, 2024. (in Chinese)

[12] Truong-Dang V L, Thai H T, Le K H. TinyResViT: a lightweight hybrid deep learning model for on-device corn leaf disease detection[J]. Internet of Things, 2025, 30: 101495.

[13] Li E, Wang L, Xie Q, et al. A novel deep learning method for maize disease identification based on small sample-size and complex background datasets[J]. Ecological Informatics, 2023, 75: 102011.

[14] 肖伟, 冯全, 张建华, 等. 基于小样本学习的植物病害识别研究[J]. 中国农机化学报, 2021, 42(11): 138 – 143.
Xiao Wei, Feng Quan, Zhang Jianhua, et al. Research on plant disease identification based on few-shot learning[J]. Journal of Chinese Agricultural Mechanization, 2021, 42(11): 138 – 143. (in Chinese)

[15] Li Y, Chao X. Semi-supervised few-shot learning approach for plant diseases recognition[J]. Plant Methods, 2021, 17(1): 68.

[16] Egusquiza I, Picon A, Irusta U, et al. Analysis of few-shot techniques for fungal plant disease classification and evaluation of clustering capabilities over real datasets[J]. Frontiers in Plant Science, 2022, 13: 813237.

[17] Liang X. Few-shot cotton leaf spots disease classification based on metric learning[J]. Plant Methods, 2021, 17(1): 114.

[18] Wu X, Fan X, Luo P, et al. From laboratory to field: unsupervised domain adaptation for plant disease recognition in the wild[J]. Plant Phenomics, 2023, 5: 0038.

[19] Wang H, He M, Zhu M, et al. WCG – VMamba: a multi-modal classification model for corn disease[J]. Computers and Electronics in Agriculture, 2025, 230: 109835.

[20] Li J, Zhao F, Zhao H, et al. A multi-modal open object detection model for tomato leaf diseases with strong generalization performance using PDC – VLD[J]. Plant Phenomics, 2024, 6: 0220.

[21] 李道亮, 赵晔, 杜壮壮. 农业领域多模态融合技术方法与应用研究进展[J]. 农业机械学报, 2025, 56(1): 1 – 15.
Li Daoliang, Zhao Ye, Du Zhuangzhuang. Advances in multi-modal fusion techniques and applications in agricultural field[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(1): 1 – 15. (in Chinese)

[22] Yang D, Wang F, Hu Y, et al. Citrus Huanglongbing detection based on multi-modal feature fusion learning[J]. Frontiers in Plant Science, 2021, 12: 809506.

[23] Cao Y, Chen L, Yuan Y, et al. Cucumber disease recognition with small samples using image-text-label-based multi-modal language model[J]. Computers and Electronics in Agriculture, 2023, 211: 107993.

[24] Cheng T, Song L, Ge Y, et al. Yolo-world: real-time open-vocabulary object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16901 – 16911.

[25] Jocher G, Chaurasia A, Qiu J. Ultralytics YOLO v8[EB/OL]. [2026-01-16]https://github.com/ultralytics/ultralytics.
[26] Radford A, Kim J W, Halia Cy C, et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning, 2021: 8748-8763.
[27] Ding X, Zhang X, Ma N, et al. RepVGG: making VGG-style convnets great again[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13733-13742.
[28] Perez E, Strub F, De Vries H, et al. Film: visual reasoning with a general conditioning layer[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 3942-3951.
[29] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.
[30] Sonkar S, Baraniuk R G. Investigating the role of feed-forward networks in transformers using parallel attention and feed-forward net design[PP/OL]. V2. arXiv(2023-05-25)[2026-01-16]. https://arXiv.org/abs/2305.13297.
[31] Luo J, Huang Y, Meng J, et al. GATE: graph-based adaptive tool evolution across diverse tasks[PP/OL]. V1. arXiv(2025-02-20)[2026-01-16]. https://arXiv.org/abs/2502.14848.
[32] 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.
[33] Lee Y, Park J. Centermask: real-time anchor-free instance segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 13906-13915.
[34] MMYOLO Contributors. MMYOLO: OpenMMLab YOLO series toolbox and benchmark[EB/OL]. [2026-01-16]https://github.com/open-mmlab/mmyolo.
[35] Chen K, Wang J, Pang J, et al. MMDetection: open MMlab detection toolbox and benchmark[PP/OL]. V1. arXiv(2019-06-17)[2026-01-16]. https://arxiv.org/abs/1906.07155.
[36] Zhan Z, Zhou J, Xu B. Fabric defect classification using prototypical network of few-shot learning algorithm[J]. Computer in Industry, 2022,138:103628.
[37] Riou K, Zhu J, Ling S, et al. Few-shot object detection in real life: case study on auto-harvest[C]// 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2020: 1-6.
[38] 杨信廷, 周子洁, 李文勇, 等. 基于小样本学习的鳞翅目害虫图像识别方法[J]. 农业机械学报, 2025, 56(2): 402-410.
Yang Xinting, Zhou Zijie, Li Wenyong, et al. Image recognition method for lepidoptera pests based on few-shot learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(2): 402-410. (in Chinese)
[39] Pan Q, Fu K, Wang G. Study on few-shot object detection approach based on improved RPN and feature aggregation[J]. Applied Sciences, 2025, 15(7): 3734.
[40] 李显娜, 吴强, 张一丹, 等. 自监督学习下小样本番茄叶片病害检测[J]. 中国农机化学报, 2024, 45(7): 172-179.
Li Xianna, Wu Qiang, Zhang Yidan, et al. Detection of tomato leaf disease in small sample under self-supervised learning [J]. Journal of Chinese Agricultural Mechanization, 2024, 45(7): 172-179. (in Chinese)