doi:10.6041/j.issn.1000-1298.2026.18.005

融合语义与图像感知的棉花病虫害检测方法

李　阳[1,2]　吴　科[1]　聂　晶[1,2]　方　凯[3]

(1. 石河子大学机械电气工程学院, 石河子 832003; 2. 新疆生产建设兵团现代农业机械重点实验室, 石河子 832003;
3. 浙江农林大学数学与计算机科学学院, 杭州 311300)

摘要: 针对棉花田间病虫害目标尺度变化显著且单一视觉表征泛化能力不足的问题,本文提出一种融合语义先验与多尺度图像感知的多模态检测方法。以冻结 CLIP 模型为特征提取基础,借助文本调整模块与参数高效微调网络,将农艺文本转换为结构化提示,实现农学语义与视觉模态深度对齐;同时设计原型引导注意力机制,以可学习原型为桥梁,动态连接像素分布与语义空间。在检测模型层面,以 VGG19 为主干网络,融合 Swin Transformer 的全局上下文建模能力,并引入合成融合 SFM 模块和内块融合 IFM 模块构建连续尺度特征流,从而缓解传统离散下采样对微小目标特征的截断。试验结果表明,所提模型方法在棉花病虫害数据集上平均精度均值达 92.5%,较其他主流模型 YOLO v5s 和 YOLO v8n 提升 3.6、4.6 个百分点。消融试验结果表明,语义引导与多尺度优化策略在性能提升上具有明显的协同效应,且模型参数量仅为 1.56×10^6,远低于 YOLO v5s 和 YOLO v8n;此外,Grad-CAM 可视化结果表明,模型能够较稳定地聚焦于典型病灶区域。综上,该方法在精度与轻量化之间取得了较好的平衡,可为复杂农情条件下病虫害精准识别及边缘端部署提供参考。

关键词: 棉花; 病虫害; 多模态融合; 语义先验; 注意力机制

中图分类号: S435.62　　**文献标识码**: A　　**文章编号**: 1000-1298(2026)18-0051-11　　**OSID**:

Cotton Pests and Diseases Detection Method Based on Integration of Semantic and Image Perception

Li Yang[1,2]　Wu Ke[1]　Nie Jing[1,2]　Fang Kai[3]

(1. *College of Mechanical and Electrical Engineering, Shihezi University, Shihezi* 832003, *China*
2. *Xinjiang Production and Construction Corps Key Laboratory of Modern Agricultural Machinery, Shihezi* 832003, *China*
3. *College of Mathematics and Computer Science, Zhejiang A&F University, Hangzhou* 311300, *China*)

Abstract: Aiming to address the bottlenecks posed by the significant variation in the target scales of cotton field pests and diseases and the limited generalization capability of single-modal visual representations, a multimodal detection method that integrated semantic prior knowledge with multi-scale image perception was proposed. The method used a frozen CLIP model as the foundation for feature extraction. By leveraging a text adaptation module and a parameter-efficient fine-tuning network, it converted agronomic text into structured prompts, thereby achieving deep alignment between agronomic semantics and visual modalities. Additionally, a prototype-guided attention mechanism was designed to dynamically link pixel distributions with the semantic space based on learnable prototypes. At the detection model, VGG19 served as the backbone network, integrating the global context modeling capabilities of swin transformer. Additionally, SFM and IFM modules were introduced to construct a continuous-scale feature flow, thereby mitigating the truncation of fine-scale object features caused by traditional discrete downsampling. Experiments demonstrated that the proposed model achieved an mAP of 92.5% on the cotton pest and disease dataset, outperforming other mainstream models YOLO v5s and YOLO v8n by 3.6 and 4.6 percentage points, respectively. Ablation results further revealed that the semantic guidance and multi-scale optimization strategies exhibited a significant synergistic effect in performance improvement, while the model's parameter count was only 1.56×10^6, which was much lower than that of YOLO v5s and YOLO v8n. Additionally, Grad-CAM visualization results indicated

收稿日期: 2026-05-20　修回日期: 2026-06-08

基金项目: 石河子大学青年创新拔尖人才计划项目(CXBJ202306)和浙江省尖兵领雁+X 科技项目(2026C04008)

作者简介: 李阳(1991—),男,副教授,博士,主要从事精准农业、图像处理和农业人工智能研究,E-mail: liyang328@shzu.edu.cn

that the model can consistently focus on typical lesion areas. In summary, this method achieved a balance between accuracy and model compactness, providing a reference for the precise identification of pests and diseases under complex agricultural conditions and for edge deployment.

Key words: cotton; pests and diseases; multimodal fusion; semantic prior; attention mechanism

0 引言

棉花作为全球重要的经济作物之一[1-2],其产量与纤维品质直接关系到农业经济效益以及纺织产业链的稳定发展。棉花病虫害的大规模发生不仅会造成减产和纤维品质退化,还可能因长期依赖化学药剂防治而诱发环境污染、农药残留以及害虫抗药性增强等问题[3]。目前,病虫害识别仍较大程度依赖植保人员开展人工田间巡检,但存在劳动成本高、主观经验依赖明显以及实时性不足等局限,难以满足大规模棉田精准监测的实际需求[4]。

近年来,随着深度学习技术的快速发展,在农业视觉任务中特征学习能力不断提升。从卷积神经网络到目前广泛应用的目标检测框架[5],相关方法已逐步应用于植物病虫害识别与智能监测领域[6-7]。蒋心璐等[8]对农业害虫检测相关深度学习方法进行了系统综述,指出当前研究仍面临数据样本不足[9]、小目标难以有效表征[10]、尺度变化剧烈[11]以及目标密集遮挡等关键问题。卷积神经网络[12]在植物病害识别领域取得了较好的应用效果。Chang 等[13]针对小麦锈病小样本细粒度分类问题,结合 Plant Village 与自建数据集,实现了 3 类锈病自动化识别,验证了 CNN 模型在病害纹理特征提取方面的有效性。不过,传统卷积结构在建模长距离依赖关系和全局语义信息时仍存在一定局限,因此研究者开始尝试将 Transformer 结构引入农业病虫害识别任务,以提升模型全局特征建模能力。刘拥民等[14]利用 Swin Transformer 自注意力机制提取番茄病害图像高层语义特征,并结合 Mixup 增强与迁移学习策略,模型识别准确率达 98.40%。此外,Hemalatha 等[15]提出多任务视觉 Transformer 模型,通过融合 co-scale、co-attention 以及 cross-attention 等机制,在 Plant Village 数据集上分类准确率达 99.97%,进一步说明 Transformer 架构在农业视觉任务中的应用潜力。

面对田间复杂背景、光照以及目标尺度差异等问题,研究者将注意力机制[16-17]与多尺度特征融合策略[18]用于病虫害检测任务[19-20]。Ma 等[21]针对低光照环境下棉花害虫目标特征模糊的问题,提出 DCP - YOLO v7x 模型,结合 FFDNet 去噪网络与 EnlightenGAN 低光照增强网络,并引入可变形注意力机制 DAttention 和动态检测头 DyHead,在低照度条件下精确率达 95.9%、mAP@0.5 达 95.4%。此外,在边缘端部署需求增加的背景下,Mo 等[22]提出 RT - DETR - DFSA 模型,通过解耦聚焦自注意力机制将二维自注意力计算分解为一维操作,在保持检测精度的同时降低模型参数量。然而,上述方法在检测性能与轻量化方面取得了一定进展,但大多仍依赖单一图像模态进行特征学习,对农学先验知识与语义信息的利用相对有限。

随着视觉-语言联合建模的发展,多模态融合方法[23-24]为突破单一图像模态表征能力不足的问题提供了新的研究思路[25]。Qiu 等[26]围绕图文多模态融合策略开展研究,利用 ShuffleNetV2 与 TextCNN 提取视觉和文本语义特征,最优模型在混合数据集上分类准确率达 98.33%。Liang 等[27]提出 BLIP - DP 方法,通过动态文本提示增强多模态特征融合能力,在 Plant Village 数据集上实现了较优的图像描述性能。相关研究表明,视觉信息与语义知识的协同建模[28]有助于提升模型对病虫害病理特征的理解能力。不过,现有研究对象仍主要集中于柑橘、番茄等作物,对于棉花这一病虫害类型复杂、田间场景变化明显的大宗经济作物,多模态检测研究仍相对有限。

基于上述问题,本文提出一种融合语义先验与多尺度图像感知的棉花病虫害检测方法。以冻结 CLIP 模型作为视觉-语言基础,通过文本调整模块与参数高效微调网络实现农学语义与视觉特征的结构化对齐,并引入双层原型引导注意力机制以增强跨模态信息交互能力。在检测阶段,模型结合 Swin Transformer 全局建模能力以及合成融合、内块融合模块构建连续尺度特征空间,以缓解传统离散下采样过程微小目标特征易丢失的问题。

1 材料与方法

1.1 试验数据集

围绕棉花病虫害多模态智能管控需求,构建了适用于复杂田间场景的专用数据集,涵盖棉铃虫、棉蚜、红蜘蛛、棉盲蝽、棉蓟马 5 类虫害及黄萎病、炭疽病 2 类病害。

原始样本采集于新疆典型棉区,覆盖苗期至吐絮期 4 个关键生育阶段,并采用多光谱相机与移动

终端田间下获取原始图像 2 840 幅。经剔除模糊及目标缺失等无效样本后，最终保留有效原始图像 2 120 幅，并按比例 7∶1.5∶1.5 划分为训练集、验证集与测试集。针对农业数据存在的长尾分布特性，引入随机旋转、水平翻转、亮度对比度扰动及 MixUp 等 4 类数据增强策略，将样本总量扩充至 7 060 幅，如表 1 所示。

表 1　各类棉花病害虫标签数量

Tab. 1　Number of labels for various cotton pests and diseases　　幅

害虫种类	训练集	验证集	测试集	总计标签数
棉铃虫	850	180	170	1 200
棉蚜虫	720	155	145	1 020
红蜘蛛	680	145	135	960
棉盲蝽	650	140	130	920
棉蓟马	610	130	120	860
黄萎病	780	165	155	1 100
炭疽病	710	150	140	1 000
合计	5 000	1 065	995	7 060

如图 1 所示，增强后数据集仍较好保留了棉花虫害形态特征，为多模态语义融合与少样本学习提供了高质量数据基础。

1.2　语义与图像感知信息融合网络

提出一种语义-图像融合网络，将大语言模型生成的农学语义先验与视觉特征进行深层交互，并借助跨模态协同机制提升模型对病害病理模式及虫害形态特征的细粒度感知能力。如图 2 所示，多模态语义融合网络（MSFN）主要由预训练多模态基础模型（CLIP）、参数高效微调网络（FTN）、文本调整（TA）模块以及原型引导注意力（PGA）机制 4 部分组成。

在保持 CLIP 骨干参数冻结的基础上，网络仅引入轻量级 FTN 完成下游任务适配，有效降低模型在边缘计算设备上部署成本。TA 模块在训练初始化阶段借助大语言模型，将原始类别标签转换为符合棉花植保规范的结构化文本提示；在后续微调阶段，则直接加载预生成文本，以避免推理阶段频繁调用大模型带来的额外延迟。PGA 机制作为多模态知识融合的重要环节，通过可学习原型向量动态关联视觉像素分布与农学语义空间，从而增强模型对病虫害发生规律的可解释性表达能力。

1.2.1　网络总体架构

图像原型精炼过程主要包括权重共享的微调网

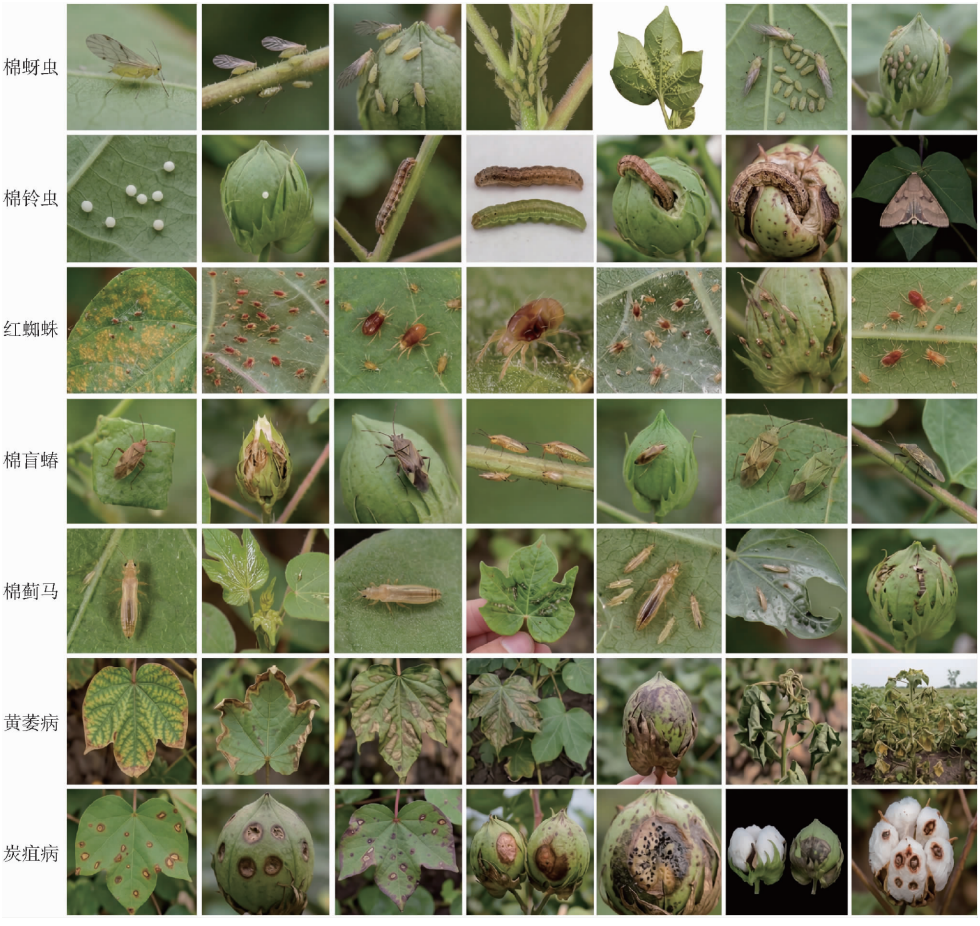

图 1　棉花病虫害图像示例

Fig. 1　Examples of images of cotton pests and diseases

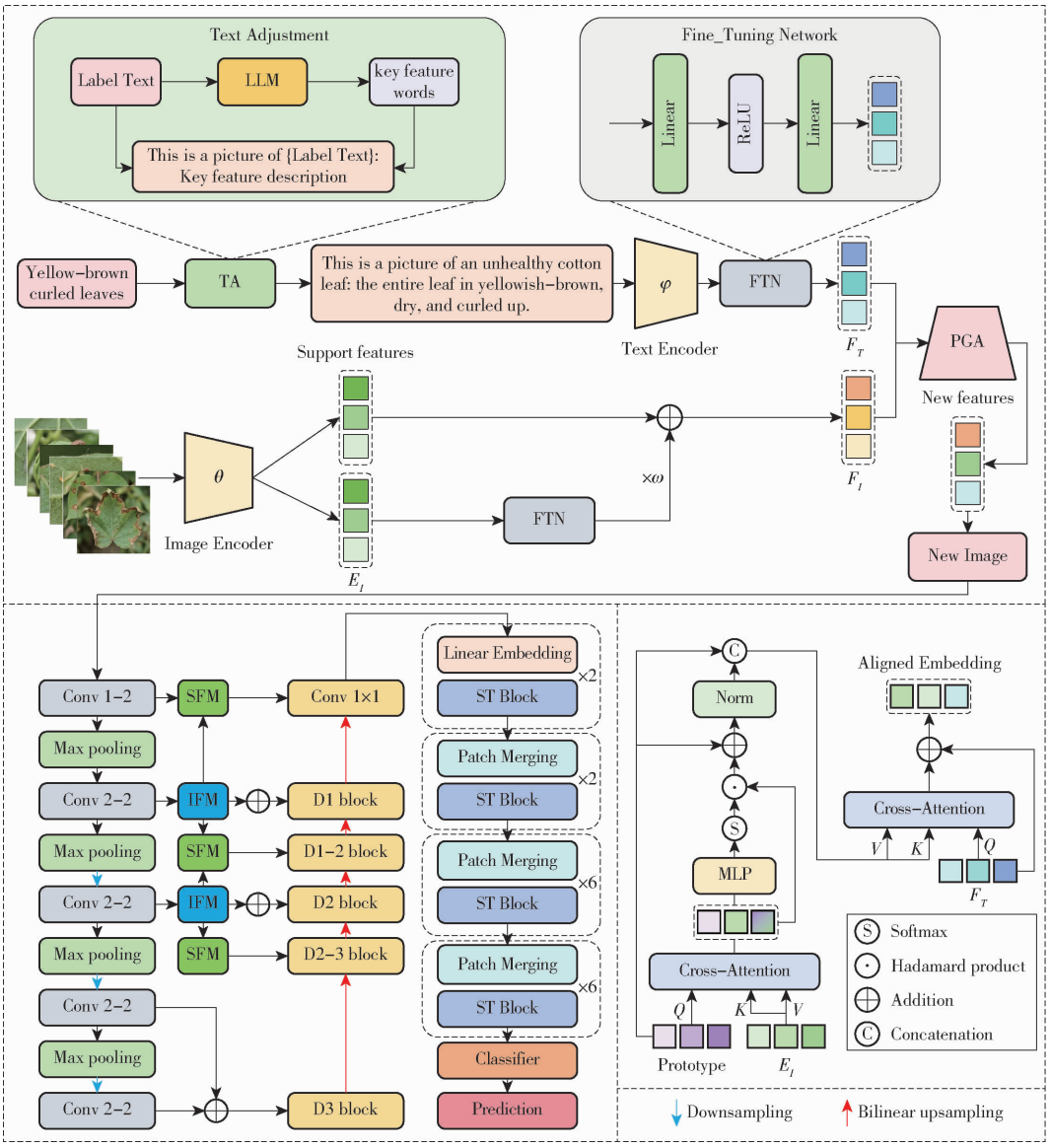

图 2 系统整体架构

Fig. 2 System architecture

络以及融合文本调整的原型对齐策略。微调网络与文本分支采用参数共享方式,实现图像与文本模态之间的语义交互;在此基础上,文本信息被进一步用于优化图像原型,从而增强模型对任务相关语义特征的表征能力。初始文本输入在大语言模型辅助下完成语义精调,以使文本描述更加贴近当前任务的数据分布特征。通过上述双层精炼机制,模态间的语义协同能力得到增强,同时提升了网络对复杂病虫害场景的任务适应性。

在原型精炼阶段,模型冻结 CLIP 主干参数,仅对微调网络进行训练,以保证整体优化过程的稳定性。具体而言,首先利用图像编码器 $E_I(\cdot)$ 提取支持集特征,得到初始视觉特征 F_S。在此基础上,FTN 进一步生成微调特征 F_{ftn},并保留原始视觉特征信息。最终支持集特征由 2 部分特征进行残差融合得到,可表示为

$$F'_S = F_S + \omega F_{ftn}(F_S) \tag{1}$$

式中 F'_S——最终的支持集特征

ω——微调特征权重系数

随后对支持集特征进行均值聚合操作,从而获得图像原型 F_I,可表示为

$$F_I = \frac{1}{K}\sum_{i=1}^{K} F'_{S,i} \tag{2}$$

文本集合首先经由文本调整模块完成语义精炼,随后利用文本编码器 $E_T(\cdot)$ 和 FTN 提取文本特征,从而获得更具判别性的语义表示,可表示为

$$F_T = E_T(T(\Gamma)) \tag{3}$$

由于各类别对应唯一文本描述,因此文本特征可直接作为文本原型,其中 $T(\cdot)$ 表示文本调整模块中语义映射操作。

1.2.2 PGA 机制

如图 2 所示,PGA 模块采用双层交叉注意力结

构，并利用可学习原型向量对图像特征学习过程进行引导。模块内部包含 2 层多层感知机结构，并采用 GeLU 激活函数与 0.1 的 Dropout 正则化策略。第 1 层注意力主要实现原型向量与图像特征之间的交互；在完成视觉特征增强后，第 2 层注意力进一步融合文本语义信息，最终输出 512 维多模态融合特征。其中，初始原型向量维度为 7×512（对应 7 类病虫害），视觉嵌入维度为 $B\times(H/32\times W/32)\times512$（$B$ 为批次大小），经第 1 层交叉注意力后输出维度仍为 7×512；第 2 层交叉注意力中，文本查询维度为 7×512，拼接后原型上下文维度为 $7\times1\,024$，最终输出维度保持 7×512，经全局平均池化后得到 512 维全局语义特征。

模型初始化一组 K 个可学习原型向量，其中 K 表示类别数量，记为 $\boldsymbol{P}\in\mathbf{R}^{K\times D}$。这些原型向量采用标准正态分布随机初始化，能够减少人为先验偏差，使模型能够从病虫害数据中自主学习更优的原型表示。同时，标准正态分布具备较稳定的梯度传播特性，有助于提升反向传播阶段的训练稳定性。原型向量在网络中承担关键病变模式及语义锚点的动态表示作用。为捕获图像特定的上下文信息，将原型向量扩展至与当前批次大小一致，并将其作为查询 Q_P 输入注意力模块。通过交叉注意力机制实现其与视觉嵌入 $E_I\in\mathbf{R}^{B\times N_{img}\times D}$ 之间的特征交互，N_{img} 表示视觉嵌入序列长度形式化地，原型向量 $\boldsymbol{P}$ 被视为 Q_P，视觉嵌入则映射为键 $\boldsymbol{K}_I$ 和值 V_I，原型引导注意力计算式为

$$P'=\mathrm{Softmax}\left(\frac{Q_P\boldsymbol{K}_I^{\mathrm{T}}}{\sqrt{D}}\right)V_I \tag{4}$$

式中　D——嵌入维度

P'——图像感知原型

经过上述过程，初始原型 P 被进一步精炼为图像感知原型 P'，从而能够更加关注 E_I 中关键视觉区域。图像感知原型 P' 进一步经过 MLP 与 Softmax 激活处理，以生成更加突出判别性原型信息的注意力权重。原型增强表示为

$$P_{enh}=\mathrm{LayerNorm}(\mathrm{MLP}(P')\odot P'+P) \tag{5}$$

式中　$\odot$——Hadamard 乘积

P_{enh}——图像增强原型

为进一步融合大语言模型中语义知识，网络继续为第 2 层交叉注意力构建对应键值对。此时，注意力机制中键值对由检索参考向量构成，并通过拼接初始原型 P 与图像感知原型嵌入 P_{enh} 生成多模态上下文表示。由文本编码器输出的 F_T 则被用作查询 Q_T。随后计算文本查询与原型增强视觉上下文之间的交叉注意力，表达式为

$$T_{com}=\mathrm{CrossAttn}(Q_T,[P,P_{enh}],[P,P_{enh}]) \tag{6}$$

式中　T_{com}——经多模态知识补全后文本特征

该机制能够有效实现文本语义与原型化视觉表示之间的对齐，从而提升模型在复杂田间背景下抗干扰能力与特征鲁棒性。

1.3　多尺度检测模型

多模态语义融合网络在病虫害类别判别与特征表达方面取得了较好效果，但对于病害斑块与微小虫害目标的精确定位，仍需依赖高性能目标检测框架的进一步支持。因此，在改进 VGG19 检测框架的基础上，引入 Swin Transformer 模块与多尺度特征优化策略，构建了一种适用于复杂棉田环境的高精度病虫害检测模型。如图 2 下半部分所示，模型以固定输入尺寸为基础，通过特征提取层、多尺度融合层以及检测头层的协同构建，实现了从全局农学语义信息到局部病灶区域的逐层特征聚焦。

1.3.1　模型网络结构

棉花病虫害检测模型核心任务为对图像中微小虫害目标及弥散型病斑进行准确定位与识别。采用以 VGG19 为主干网络的检测架构，并结合 Swin Transformer 模块完成多尺度特征融合与全局上下文建模。VGG19 依托连续堆叠的 3×3 卷积核结构，能够较好提取图像中低层纹理特征与中层形态信息，从而为病虫害模式识别提供稳定的特征基础。模型整体结构主要由输入层、特征提取层、多尺度融合层以及检测分割层组成。输入层首先对预处理后的田间图像进行尺寸统一与数据标准化处理。特征提取层延续 VGG19 的多阶段卷积块结构，在提取多层次局部特征基础上，进一步借助 Swin Transformer 模块获取全局上下文信息，实现不同尺度特征之间的有效整合。

MSFN 输出的 512 维多模态融合特征将通过 1×1 卷积层压缩至 256 维后，与检测主干 VGG19 第 3、4、5 卷积块输出的多尺度特征进行逐通道拼接融合，拼接后特征分别输入后续 SFM 与 IFM 模块进行连续尺度优化，最终将融合了全局语义信息多尺度特征传递至检测头，完成目标定位与分类。合成融合模块（SFM）通过构建不同尺度的特征映射并进行融合处理，以增强模型对小目标细节信息的保留能力。内块融合模块（IFM）则在卷积块内部完成特征交互，从而促进局部细粒度信息的充分传递。检测与分割层采用锚框机制完成候选区域定位，并结合非极大值抑制（NMS）算法过滤冗余检测结果，最终通过解码器网络输出病虫害目标边界框及类别置信度。

1.3.2 SFM 模块

在传统卷积神经网络中,池化操作通常会将图像尺寸缩减为原来的 1/2,进而使特征向量仅能在 1/2、1/4、1/8 等离散尺度上完成缩放。然而,这种离散化尺度变化会限制连续尺度特征的表达能力,尤其在红蜘蛛、棉蓟马等微小目标检测过程,易出现细粒度特征缺失问题。为缓解上述问题,引入 SFM 模块,使特征图能够在 1/2、1/3、1/4、1/6 等连续尺度之间实现更加平滑的缩放。通过构建更加连续的尺度空间,不仅能够适应田间目标尺度持续变化的实际情况,同时也减弱了尺度截断带来的特征信息衰减。

如图 3 与图 4 所示,SFM 通过插值生成额外特征层,并引入 1/3 尺度变化作为缓冲层,以减轻池化操作导致的尺度突变问题。具体而言,第 i 层中维度为 $C\times4W\times4H$ 的特征图经缩放操作后,被调整为 $C\times3W\times3H$。在此过程中,输入特征首先被划分为规则的 4×4 图像补丁,随后利用核大小为 2×2、步幅为 1 的卷积操作生成 3×3 的新特征块。同时,在第($i+1$)层中,插值放大操作会先将维度为 $C\times2W\times2H$ 的特征图上采样至 4×4 尺度,再通过相同卷积映射得到 3×3 特征图。随后,将 2 部分特征进行拼接,并通过 1×1 卷积完成融合,最终生成维度为 $C\times3W\times3H$、对应 $1/3\times2^{i-1}$ 尺度的新特征层,C 统一取 64、128、256(对应检测主干第 3、4、5 卷积块输出通道数),输入特征图尺寸分别为 128×128、64×64、32×32,经 SFM 模块处理后输出尺寸分别为 96×96、48×48、24×24,实现了 1/2、1/3、1/4 尺度的连续过渡。

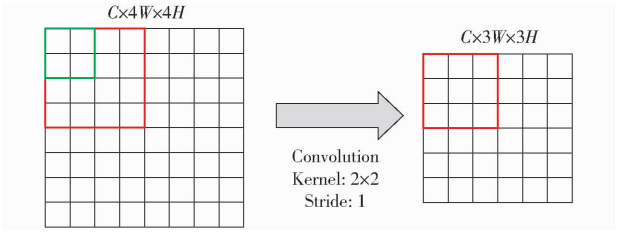

图 3 缩放过程

Fig. 3 Zooming process

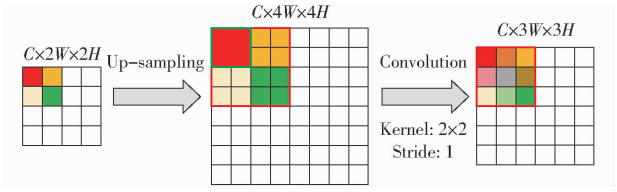

图 4 插值上采样过程

Fig. 4 Pocess of upsampling by interpolation

1.3.3 IFM 模块

如图 5 所示,在大多数经典 CNN 架构中,虽网络通过多层卷积提取不同层级的特征信息,但各卷积块通常仅将最后 1 层特征图传递至下一阶段网络。除最终输出层外,卷积块内部中间特征层同样包含大量细粒度判别信息。

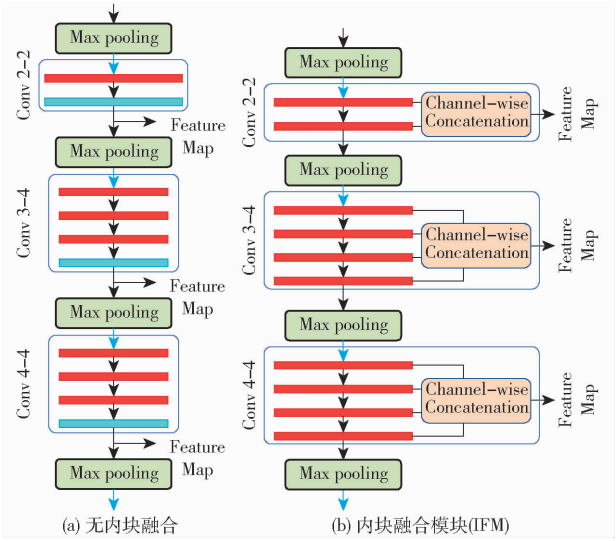

图 5 有无内块融合对比

Fig. 5 Comparison with and without IFM

因此,引入 IFM 模块对中间特征进行进一步激活与融合,以获得更加准确的病虫害特征表征。如图 5 所示,在特征图向后续网络传递前,IFM 首先利用通道卷积对同一卷积块中的多层特征进行跨通道融合。

1.3.4 Swin-Transformer 模块

如图 6 所示,Swin-Transformer 模块的核心思想为利用局部窗口自注意力机制替代传统全局注意力计算,从而在保持感受野连续性的同时,实现计算复杂度线性降低。输入图像经过 Patch Partition 处理后,每个 4×4×3 图像块会被线性映射至 C 维特征空间,进而生成尺寸为 $H/4\times W/4\times C$ 的初始特征图。随后,初始特征图依次经过由 W-MSA 与 SW-MSA 组成的 Swin-Transformer Block。其中,W-MSA 首先将特征图划分为互不重叠的局部窗口,并仅在窗口内部执行注意力交互,有效降低整体计算规模;而 SW-MSA 则借助循环位移策略打破窗口边界限制,实现不同窗口之间的特征信息交互流动。

2 结果与分析

2.1 试验

2.1.1 试验平台

所有试验均在统一的软硬件环境下完成,包括模型对比试验与消融试验。试验平台基于搭载 NVIDIA RTX 4090 GPU 的高性能计算服务器,并采用 PyTorch 深度学习框架完成模型训练、验证与推理过程。在训练阶段,模型优化器采用 AdamW,初始学习率设置为 1×10^{-4},并结合余弦退火学习率

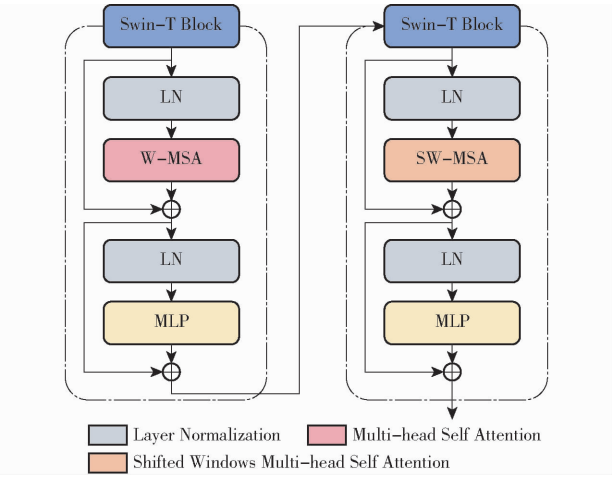

图 6　两个相连的 Swin－T 块的网络结构

Fig. 6　Network structure consisting of two interconnected Swin－T blocks

调度策略进行动态更新；批次大小设为 32，总训练轮次为 150。

2.1.2　评价指标

采用精确率（Precision）、召回率（Recall）、F1 分数以及平均精度均值（mAP@0.5）作为主要评价指标，以综合衡量模型在目标识别完整性与误检控制方面的检测性能。此外，为进一步评估模型在边缘设备中的部署潜力，试验同步统计模型参数量、浮点运算量以及单幅图像平均检测时间等指标。

2.2　结果分析

2.2.1　7 类棉花病虫害检测性能

针对 7 类典型棉花病虫害，包括棉铃虫、棉蚜虫、红蜘蛛、棉盲蝽、棉蓟马、黄萎病以及炭疽病，开展了细粒度单类别检测性能分析，结果如表 2 所示。从表 2 可以看出，所提出的语义–图像融合模型在各类别检测任务中均表现出优于对比模型整体性能。在棉铃虫检测任务中，本文模型平均精度达 94.7%，相比 YOLO v8n 提升 3.5 个百分点；在棉蚜虫检测中，模型平均精度达 92.8%，相比 YOLO v8n 的 89.5% 同样具有一定优势。对于红蜘蛛这类微小目标，由于其虫体尺寸较小且易受叶片纹理干扰，传统检测模型往往容易出现漏检现象。相比之下，本文模型平均精度为 89.7%，较 YOLO v8n 提升 4.4 个百分点，表明所构建的连续尺度特征空间能够有效增强微小目标的特征感知能力。此外，本文模型针对红蜘蛛这类典型小目标漏检率为 6.3%，误检率为 4.0%，较 YOLO v8n 分别降低 3.1、2.2 个百分点，验证了连续尺度特征流对小目标检测性能的显著提升作用。

表 2　不同模型对各类别识别平均精度

Tab. 2　Average accuracy of different models for each category　%

模型	棉铃虫	棉蚜虫	红蜘蛛	棉盲蝽	棉蓟马	黄萎病	炭疽病
Faster R－CNN	86.4	83.1	78.5	81.2	76.8	89.3	85.6
RetinaNet	88.2	85.4	80.1	83.5	79.3	90.5	87.2
YOLO v3	84.6	81.2	76.8	79.4	74.1	87.3	83.5
YOLO v5s	90.1	88.3	84.6	87.2	82.4	91.7	89.1
YOLO v8n	91.2	89.5	85.3	88.4	83.6	92.4	90.2
本文模型	94.7	92.8	89.7	91.4	87.5	95.1	93.6

图 7 给出了本文模型与 Faster R－CNN、RetinaNet、YOLO v3、YOLO v5s 以及 YOLO v8n 等主流目标检测模型的混淆矩阵检测结果。可以发现，本文模型在多数棉花病虫害类别上的预测结果更加集中，对相似病斑与虫害类别的误分类现象得到明显缓解。

数据集采集过程中已覆盖正午强光、傍晚弱光、叶片部分遮挡等典型复杂田间场景，占总样本量的 32%。在该类复杂场景子集中，本文模型 mAP@0.5 仍达 88.2%，较 YOLO v8n 高 5.1 个百分点，表明模型在复杂光照与遮挡条件下具有较好的鲁棒性，能够适应实际棉田多变环境。

2.2.2　MSFN 网络消融试验

为验证多模态语义融合网络（MSFN）中各核心模块的有效性，开展了模块消融试验，结果如表 3 所示，仅使用 CLIP 骨干网络时，模型 mAP@0.5 仅为 86.3%，说明预训练多模态模型难以直接适应棉花病虫害检测任务。引入 TA 模块后，mAP@0.5 提至 88.9%，精确率与召回率分别提高 2.1、1.5 个百分点，表明结构化农学文本提示能够增强模型对农业专业语义的理解能力。加入 FTN 模块后，模型性能继续提升，mAP@0.5 达 90.7%，F1 分数提升至 88.5%，说明参数高效微调策略能够在冻结 CLIP 骨干参数的条件下有效适配下游任务。当进一步集成 PGA 机制后，模型性能达到最优，mAP@0.5 提至 92.5%，F1 分数达 91.1%，相比基线模型提升 6.2 个百分点。结果表明，PGA 机制能够通过可学习原型向量实现视觉特征与农学语义之间的动态关联，从而增强模型对病虫害特征的综合表征能力。

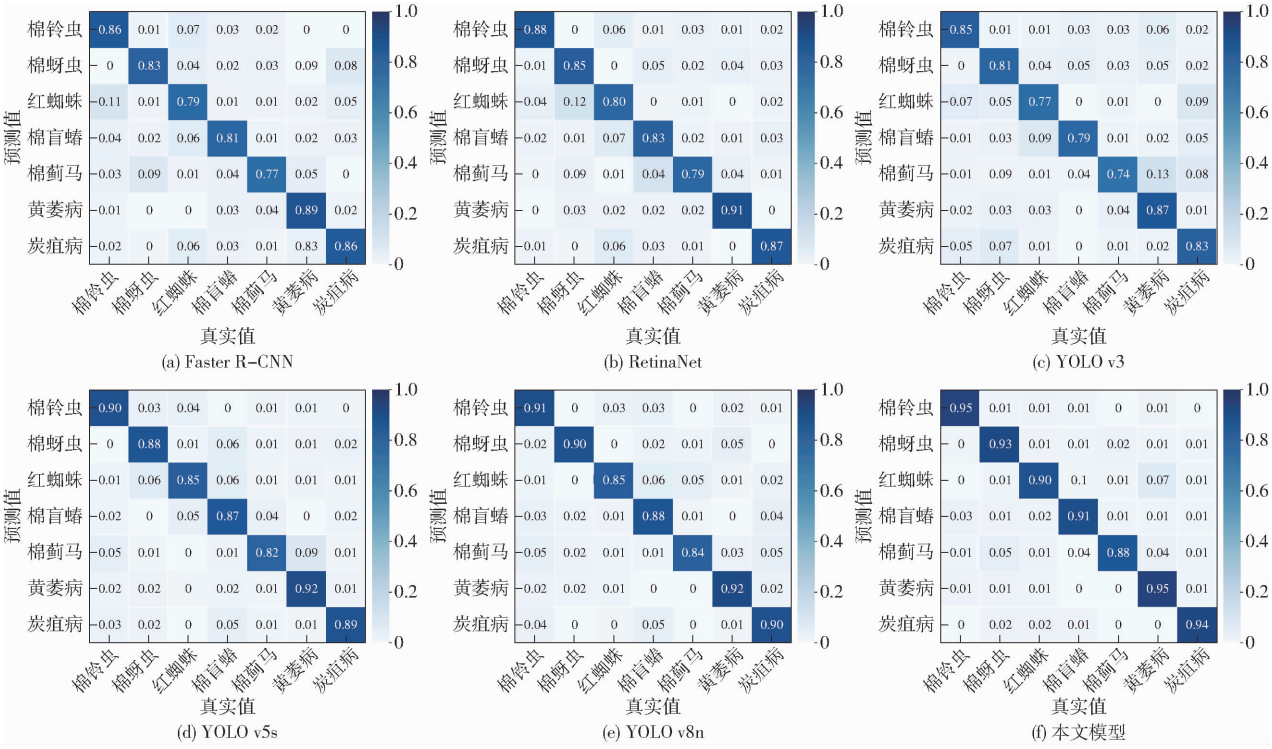

图 7 混淆矩阵结果

Fig. 7 Results of confusion matrices

表 3 MSFN 网络核心模块消融试验结果

Tab. 3 Results of MSFN network core module ablation experiment

CLIP	TA	FTN	PGA	精确率/%	召回率/%	F1 分数/%	mAP@0.5/%	参数量	浮点运算量	平均检测时间/ms
√	×	×	×	85.2	84.6	84.9	86.3	1.12×10^6	3.85×10^9	12.4
√	√	×	×	87.3	86.1	86.7	88.9	1.15×10^6	3.92×10^9	15.1
√	√	√	×	89.6	87.4	88.5	90.7	1.34×10^6	4.41×10^9	16.7
√	√	√	√	92.5	90.3	91.1	92.5	1.56×10^6	5.23×10^9	21.2

注:√ 表示使用该模块,× 表示未使用该模块。

2.2.3 多尺度目标检测模型消融试验

针对棉花田间病虫害目标尺度变化显著且形态多样的特点,开展了多尺度核心模块消融试验,由表 4 可知,基础模型(Baseline)性能有限,mAP@0.5 仅为 81.8%,反映出传统 CNN 架构在农业图像处理中的不足。集成 MSFN 网络后,模型性能明显提升,mAP@0.5 增至 84.2%,说明多模态语义融合对增强病虫害识别能力具有基础性贡献。

表 4 多尺度模型核心模块消融试验结果

Tab. 4 Results of ablation experiments on core modules of multiscale model

序号	模型	精确率/%	召回率/%	F1 分数/%	mAP@0.5/%	参数量	浮点运算量	平均检测时间/ms
T1	Baseline	79.2	76.5	80.3	81.8	9.80×10^5	2.09×10^9	13.1
T2	Baseline + MSFN	83.4	82.1	82.7	84.2	1.15×10^6	3.12×10^9	14.6
T3	Baseline + MSFN + PGA + SFM	88.1	86.3	87.2	89.4	1.34×10^6	4.41×10^9	16.7
T4	Baseline + MSFN + PGA + SFM + IFM	90.1	88.7	89.4	91.2	1.49×10^6	4.95×10^9	19.8
T5	本文模型	92.5	90.3	91.1	92.5	1.56×10^6	5.23×10^9	21.2

当进一步引入 PGA 机制与 SFM 模块时,mAP@0.5 显著升至 89.4%,精确率与召回率分别达 88.1% 和 86.3%,验证了 SFM 模块在处理尺度连续变化目标时的有效性。随着 IFM 模块引入,mAP@0.5 进一步提至 91.2%,F1 分数达 89.4%,较仅使用 SFM 模块提升 1.8 个百分点。最终,当各模块完整集成时,模型性能达到峰值,mAP@0.5 为 92.5%,F1 分数为 91.1%,较基础模型提升 10.7 个百分点。

2.2.4　与主流目标检测模型性能对比

为系统评估所提出模型在棉花病虫害检测任务中的性能表现，将其与当前主流目标检测方法进行了对比，包括 Faster R－CNN、RetinaNet、YOLO v3、YOLO v5s 和 YOLO v8n，结果见表 5。由表 5 可知，本文模型在精确率、召回率、F1 分数及 mAP@0.5 方面均优于所有对比模型。较 YOLO v8n，本文模型 mAP@0.5 高 4.6 个百分点，F1 分数提升 1.8 个百分点。尽管本模型计算复杂度略高于 YOLO v8n，参数量仍远低于 YOLO v5s 和 YOLO v3，显示模型在保持轻量化的同时能够实现高性能。本文模型平均检测时间为 21.2 ms，略高于 YOLO v8n 的 18.5 ms，但明显低于 Faster R－CNN 的 28.4 ms。这种轻微的实时性牺牲换来了显著的精度提升，对于以精度优先的农业病虫害检测任务是可接受的。

在边缘设备适配性方面，本文模型参数量仅为 1.56×10^{6}，远低于主流检测模型，以常见的农业边缘设备 NVIDIA Jetson Nano 为例，其浮点运算量为 4.72×10^{11}，本文模型在该设备上推理时间约为 110 ms，可满足田间病虫害实时巡检需求，具备良好的边缘部署潜力。

表 5　不同模型在棉花病虫害数据集上性能

Tab.5　Model performance on a cotton pest and disease dataset

模型	精确率/%	召回率/%	F1 分数/%	mAP@0.5/%	参数量	浮点运算量	平均检测时间/ms
Faster R－CNN	82.1	83.5	83.8	83.1	4.13×10^{6}	1.85×10^{9}	28.4
RetinaNet	88.3	85.2	85.7	84.9	3.68×10^{6}	1.42×10^{9}	12.1
YOLO v3	86.4	80.1	81.2	83.0	6.19×10^{6}	2.18×10^{9}	14.3
YOLO v5s	89.1	87.4	88.2	88.9	7.03×10^{6}	2.65×10^{9}	18.2
YOLO v8n	90.2	88.5	89.3	87.9	3.25×10^{6}	3.94×10^{9}	18.5
本文模型	92.5	90.3	91.1	92.5	1.56×10^{6}	5.23×10^{9}	21.2

2.2.5　热力图可视化

为增强模型可解释性并深入理解对图像关键区域的注意机制，采用基于梯度加权的类别激活映射（Grad－CAM）技术对模型关注区域进行可视化与分析。Grad－CAM 通过计算分类输出对特定卷积层特征图的梯度影响生成可视化，反映图像各区域对最终预测的贡献。本文模型热力图效果如图 8 所示，本文模型能够充分学习棉花病害虫的特征，并有效识别目标区域。



图 8　本文模型热力图效果

Fig.8　Heatmap results of proposed model

3　结论

（1）提出了一种融合语义信息与图像感知的多模态检测方法。通过构建涵盖 7 类棉花病虫害高质量数据集，并设计语义-图像融合网络与多尺度检测架构，有效应对田间复杂环境下病虫害精准识别难题。

（2）试验结果表明，所提出的 MSFN 网络通过文本调整模块将原始类别标签转换为符合棉花植保规范的结构化文本提示，结合参数高效微调网络实现对农学语义的深度理解；同时，原型引导注意力机制通过 2 层交叉注意力结构，动态桥接视觉像素分布与农学语义空间，模型 mAP@0.5 达到 92.5%，较最优对比模型 YOLO v8n 提升 4.6 个百分点。在多尺度特征处理上，SFM 模块通过连续尺度缩放有效缓解微小目标特征丢失，IFM 模块则激活卷积块中间层的细粒度判别信息。两者协同作用使红蜘蛛等微小目标检测精度提至 89.7%。

参　考　文　献

[1]　张友昌，张教海，牛林涛，等. 棉花病虫害智能识别技术研究进展[J]. 湖北农业科学，2025，64(10)：213－218，232.
Zhang Youchang, Zhang Jiaohai, Niu Lintao, et al. Research progress on intelligent identification technology for cotton diseases and pests[J]. Hubei Agricultural Sciences, 2025, 64(10): 213－218, 232. (in Chinese)

[2]　刁兆科. 基于深度学习算法的棉花病虫害图像识别技术[J]. 农业工程技术，2025，45(23)：69－70.

Diao Zhaoke. Image recognition technology for cotton pests and diseases based on deep learning algorithms[J]. Agricultural Engineering Technology, 2025, 45(23): 69 - 70. (in Chinese)

[3] 李亚金,昝庆安,曹轶乐,等. 农业蓟马种类鉴定技术研究进展[J/OL]. 环境昆虫学报,2026 - 06 - 13. https://link.cnki.net/urlid/44.1640.Q.20260507.1730.072.

[4] 蔡国庆,吴建军,祝玉华,等. 基于知识蒸馏的棉花病害分类轻量化算法改进[J]. 江苏农业学报,2025,53(5):195 - 202.
Cai Guoqing, Wu Jianjun, Zhu Yuhua, et al. Improvements to a lightweight algorithm for cotton disease classification based on knowledge distillation[J]. Journal of Jiangsu Agriculture, 2025, 53(5): 195 - 202. (in Chinese)

[5] 熊顺,周瑞,石炼,等. 基于目标检测的农业害虫计数方法综述[J]. 现代信息科技,2026,10(4):60 - 66,72.
Xiong Shun, Zhou Rui, Shi Lian, et al. A review of agricultural pest counting methods based on object detection[J]. Modern Information Technology, 2026, 10(4): 60 - 66,72. (in Chinese)

[6] 蔡金程. 物联网与深度学习在智慧农业中的应用[J]. 智能物联技术,2026,58(4):6 - 9.
Cai Jincheng. Applications of the internet of things and deep learning in smart agriculture[J]. Technology of IoT & AI, 2026, 58(4): 6 - 9. (in Chinese)

[7] Sykes J, Denby K, Franks D. Computer vision for plant pathology: a review with examples from cocoa agriculture[J]. Applications in Plant Sciences, 2024, 12(2): e11559.

[8] 蒋心璐,陈天恩,王聪,等. 农业害虫检测的深度学习算法综述[J]. 计算机工程与应用,2023,59(6):30 - 44.
Jiang Xinlu, Chen Tianen, Wang Cong, et al. Survey of deep learning algorithms for agricultural pest detection[J]. Computer Engineering and Applications, 2023, 59(6): 30 - 44. (in Chinese)

[9] 苏闪闪,吴建军,王松. DropBoost - MAML:基于元学习的小样本农业病害识别研究[J]. 江苏农业科学,2026,54(4):245 - 251.
Su Shanshan, Wu Jianjun, Wang Song. DropBoost - MAML: a meta-learning approach for small-sample agricultural disease recognition[J]. Jiangsu Agricultural Sciences, 2026, 54(4): 245 - 251. (in Chinese)

[10] 哈里旦木·阿布都克里木,张耿荣,梁宇涵,等. 面向小目标优化的棉叶病害智能诊断模型[J/OL]. 计算机应用,2026 - 06 - 13. https://link.cnki.net/urlid/51.1307.TP.20260120.1436.020.

[11] 曹英丽,陈晓安,蔺雨桐,等. 基于多尺度信息融合的杂草语义分割方法[J]. 沈阳农业大学学报,2024,55(6):743 - 751.
Cao Yingli, Chen Xiaoan, Lin Yutong, et al. Semantic segmentation of weeds based on multiscale information fusion[J]. Journal of Shenyang Agricultural University, 2024, 55(6): 743 - 751. (in Chinese)

[12] 廖晓娟,刘雷,唐锡雷. 基于改进 CenterNet 的农作物害虫智能检测算法[J]. 江苏农业科学,2026,54(4):277 - 288.
Liao Xiaojuan, Liu Lei, Tang Xilei. An intelligent crop pest detection algorithm based on an improved CenterNet[J]. Jiangsu Agricultural Sciences, 2026, 54(4): 277 - 288. (in Chinese)

[13] Chang S, Cheng J, Fan Z, et al. Recognition of wheat rusts in a field environment based on improved DenseNet[J]. Biosystems Engineering, 2024, 238:10 - 21.

[14] 刘拥民,刘翰林,石婷婷,等. 一种优化的 Swin Transformer 番茄叶片病害识别方法[J]. 中国农业大学学报,2023,28(4):80 - 90.
Liu Yongmin, Liu Hanlin, Shi Tingting, et al. Tomato leaf disease recognition based on an optimized Swin Transformer[J]. Journal of China Agricultural University, 2023, 28(4): 80 - 90. (in Chinese)

[15] Hemalatha S, Jayachandran J. A multitask learning-based vision transformer for plant disease localization and classification[J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 188.

[16] 时庆涛,王佳贺,司君蕊. 融合注意力机制与对抗生成网络的水稻虫害小样本识别模型研究[J]. 黑龙江粮食,2025(12):40 - 42.

[17] 鲍文霞,赵诗意,黄林生,等. 基于轻量级密集多尺度注意力网络的小麦叶部锈病识别方法[J]. 农业机械学报,2024,55(11):21 - 31.
Bao Wenxia, Zhao Shiyi, Huang Linsheng, et al. Lightweight dense multi-scale attention network for identification of rust on wheat leaves[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(11): 21 - 31. (in Chinese)

[18] 于丽敏,吴奇峰,曲荣芳,等. 融合多尺度特征与注意力机制的改进 YOLO v12 农作物虫害检测方法[J]. 湖北农业科学,2026,65(3):190 - 196.
Yu Limin, Wu Qifeng, Qu Rongfang, et al. Improved YOLO v12 with multi-scale feature fusion and attention mechanism forcrop pest detection[J]. Hubei Agricultural Sciences, 2026, 65(3): 190 - 196. (in Chinese)

[19] Zhang T, Zhu J, Zhang F, et al. Residual swin transformer for classifying the types of cotton pests in complex background[J]. Frontiers in Plant Science, 2024, 15: 1445418.

[20] 赵晓,杨梦婷,张懿丹. 基于改进 YOLO v10n 的苹果叶片病害检测方法[J]. 农业机械学报,2025,56(9):657 - 666,728.
Zhao Xiao, Yang Mengting, Zhang Yidan. Apple leaf disease detection method based on improved YOLO v10n[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(9): 657 - 666, 728. (in Chinese)

[21] Ma Y, Wei Y, Ma M, et al. DCP - YOLO v7x: improved pest detection method for low-quality cotton image[J]. Frontiers in Plant Science, 2024, 15: 1501043.

[22] Mo H, Wei L. Lightweight detection of cotton leaf diseases using StyleGAN2 - ADA and decoupled focused self-attention[J].

Journal of King Saud University Computer and Information Sciences, 2025, 37(4): 41.
[23] 李浩,张海明,聂珊,等. 融合知识图谱与动态门控机制的作物病虫害多模态识别方法[J/OL]. 微电子学与计算机, 2026-06-13. https://link.cnki.net/urlid/61.1123.tn.20260303.1503.006.
[24] 郝霞,周子钰,宋扬,等. 基于轻量化多模态 Blend-CNN 模型的小麦病虫害识别方法[J]. 农业机械学报,2025,56(12): 490-498,559.
Hao Xia, Zhou Ziyu, Song Yang, et al. Wheat disease and pest recognition method based on lightweight multimodal Blend-CNN model[J]. Transactions of the Chinese Society for Agricultural Machinery,2025,56(12):490-498,559. (in Chinese)
[25] 朱莉凯,柯灵. 基于数字图像的农业害虫精准检测算法及应用分析[J]. 南方农机,2026,57(4):63-66.
Zhu Likai, Ke Ling. Analysis of precision detection algorithms and applications for agricultural pests based on digital images [J]. Southern Agricultural Machinery, 2026, 57(4): 63-66. (in Chinese)
[26] Qiu X, Chen H, Huang P, et al. Detection of citrus diseases in complex backgrounds based on image-text multimodal fusion and knowledge assistance[J]. Frontiers in Plant Science, 2023, 14: 1280365.
[27] Liang F, Huang Z, Wang W, et al. Dynamic text prompt joint multimodal features for accurate plant disease image captioning [J]. The Visual Computer, 2025, 41(8): 5405-5419.
[28] 尚华胜,甘淑,袁希平,等. 级联语义分割和边缘检测的 GF-2 影像耕地提取[J]. 遥感信息,2024,39(4):134-143.
Shang Huasheng, Gan Shu, Yuan Xiping, et al. Extraction of cultivated land from GF-2 images based on level wise semantic segmentation and edge detection[J]. Remote Sensing Information, 2024, 39(4): 134-143. (in Chinese)

(上接第 15 页)

[101] Maneesha A, Suresh C, Kiranmayee B V. Prediction of rice plant diseases based on soil and weather conditions[C]// Proceedings of International Conference on Advances in Computer Engineering and Communication Systems, 2021.
[102] Xu C, Zhang Y, Zhao L, et al. Knowledge-guided adaptive spatial-temporal graph contrastive learning framework: regional crop diseases prediction based on electronic medical records[J]. Neural Networks, 2025, 189: 107597.
[103] Su J, Yi D, Su B, et al. Aerial visual perception in smart farming: field study of wheat yellow rust monitoring[J]. IEEE Transactions on Industrial Informatics, 2021, 17(3): 2242-2249.
[104] Mangewa L J, Ndakidemi P A, Munishi L K. Integrating UAV technology in an ecological monitoring system for community wildlife management areas in Tanzania[J]. Sustainability, 2019, 11(21): 6116.
[105] Marek K, Jakub S, Kniećˊ W, et al. Survey of drones for agriculture automation from planting to harvest[C]//2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), 2018: 353-358.
[106] Ravlin F W. Development of monitoring and decision-support systems for integrated pest management of forest defoliators in North America[J]. Forest Ecology and Management, 1991, 39: 3-13.
[107] Jakubowska M, Bocianowski J, Nowosad K, et al. Decision support system to improve the effectiveness of chemical control against cutworms in sugar beet[J]. Sugar Tech., 2020, 22: 911-922.
[108] Jin Y, Li J, Gu T, et al. Efficient multimodal large language models: a survey[J]. Visual Intelligence, 2025, 3: 27.
[109] Wang Y, Pan Y, Su Z, et al. Large model based agents: state-of-the-art, cooperation paradigms, security and privacy, and future trends[J]. IEEE Communications Surveys & Tutorials, 2026, 28: 1906-1949.
[110] Sun J, Zheng C, Xie E, et al. A survey of reasoning with foundation models: concepts, methodologies, and outlook[J]. ACM Computing Surveys, 2025, 57(11): 1-43.
[111] Qu G, Chen Q, Wei W, et al. Mobile edge intelligence for large language models: a contemporary survey[J]. IEEE Communications Surveys & Tutorials, 2025, 27(6): 3820-3860.
[112] Xu M, Cai D, Yin W, et al. Resource-efficient algorithms and systems of foundation models: a survey[J]. ACM Computing Surveys, 2025, 57(5): 1-39.
[113] Din M, Akoram W, Saoud L, et al. Multimodal fusion with vision-language-action models for robotic manipulation: a systematic review[J]. Information Fusion, 2026,129:104062.
[114] Kazmierczak R, Berthier E, Frehse G, et al. Explainability and vision foundation models: a survey[J]. Information Fusion, 2025, 122: 103184.