农作物病虫害自监督适应性多模态特征融合识别方法

叶鹏林[1,2], 闵 超[1,2,3*], 苟良杰[2,3], 王鹏程[1,2], 黄小鹏[1,2],
李 鑫[1,2], 蒙玉平[4]

(1. 西南石油大学 理学院,四川成都 610500,中国; 2. 西南石油大学 人工智能研究院,四川成都 610500,中国; 3. 西南石油大学 油气藏地质及开发工程全国重点实验室,四川成都 610500,中国; 4. 中国石化中原油田分公司,河南濮阳 457001,中国)

摘 要: **[目的/意义]** 传统农作物病虫害识别普遍依赖单模态图像,信息利用不充分,导致识别精度受限。针对该问题,本研究提出融合图像与文本的多模态识别方法,旨在显著提升分类准确率与模型鲁棒性,为农业精准防控提供数据驱动的新路径。**[方法]** 构建基于自监督适应性特征融合的识别模型。首先,利用大语言模型结合权威农业指南进行提示工程,将标签转化为细粒度病理语义描述;其次,基于图像文本对比学习(Contrastive Language-Image Pre-training, CLIP)提取图像与文本双流特征,设计跨模态平衡对齐模块解决样本不对称问题;再次,通过适应性融合机制动态分配模态权重,实现深层语义交互;最后,引入自监督特征重构任务以增强特征表征的鲁棒性。**[结果和讨论]** 在 PlantVillage 标准数据集上的实验表明:模型分类准确率达 99.67%,较 ResNet50(96.51%)、Swin-Transformer(97.48%)和基础 CLIP(98.23%)准确率高;同时,精确率、召回率与 F_1 分数均超过 99.00%,验证了方法的有效性与稳定性。**[结论]** 该方法通过融合文本语义与视觉特征,有效突破了单模态识别局限,显著提升了模型在细粒度分类任务中的准确率与泛化能力。

关键词: 病虫害识别;多模态融合;适应性特征融合;自监督学习

中图分类号: S435 **文献标识码:** A **文章编号:** SA202509032

引用格式:叶鹏林, 闵超, 苟良杰, 王鹏程, 黄小鹏, 李鑫, 蒙玉平. 农作物病虫害自监督适应性多模态特征融合识别方法[J]. 智慧农业(中英文), 2026, 8(1): 72-85. DOI: 10.12133/j.smartag.SA202509032

YE Penglin, MIN Chao, GOU Liangjie, WANG Pengcheng, HUANG Xiaopeng, LI Xin, MENG Yuping. Self-Supervised Adaptive Multimodal Feature Fusion Recognition of Crop Diseases and Pests[J]. Smart Agriculture, 2026, 8(1): 72-85. DOI: 10.12133/j.smartag.SA202509032 (in Chinese with English abstract)

0 引 言

据联合国粮食及农业组织(Food and Agriculture Organization of the United Nations, FAO)报告,全球每年因农作物病虫害导致的损失高达40%,经济损失超过2 200亿美元[1]。农作物病虫害智能识别技术可以快速识别病害类型,指导防治措施。传统病虫害检测方案多依赖于人工设计图像特征提取算法,制定病虫害的图像特征[2-4]。由于自然环境的复杂多变性,这些预设算法在实际应用中表现出较低的鲁棒性[4,5]。

近年来,随着人工神经网络的发展,卷积神经网络(Convolutional Neural Network, CNN)和Transformer等深度学习模型逐渐应用于农作物病虫害识别领域。这些模型能够在训练过程中自动完成特征提取,从而识别出传统人工设计方法难以发现的潜在特征[4,5]。彭红星等[6]在ShuffleNetV2基础上引入注意力机制SimAM,在不额外增加网络参数的同时增强重要特征的有效提取。冯峰等[7]基于ResNet50提出LSE-ResNet50,在玉米病虫害识别的任务中收敛速度与准确率均高于ResNet50。孙杨俊等[8]提出了一种基于双线性卷积宽度网络的

收稿日期:2025-09-20

基金项目:国家自然科学基金项目(52574048);四川省科技计划项目(2025NSFTD0016)

作者简介:叶鹏林,硕士研究生,研究方向为多模态模式识别。E-mail:yplin0907@outlook.com

*通信作者:闵 超,博士,教授,研究方向为最优化方法与不确定性的数学理论、可解释机器学习等领域。E-mail: minchaosc@126.com

copyright©2026 by the authors (登录www.smartag.net.cn免费获取电子版全文)

水稻病虫害识别模型，通过结合双线性CNN和宽度学习系统增强双线性特征，提高了模型的识别准确率。王杨等[9]改进了Vision-Transformer（ViT），通过增强分块序列化和引入掩码多头注意力，提高了模型在噪声环境下的识别能力。刘拥民等[10]将Swin-Transformer与Mixup混合增强算法进行结合，在番茄叶片病虫害的识别任务上准确率相比Vit、ResNet50等基础模型有所提升。尽管上述研究在农作物病虫害识别中取得了较好的效果，但依赖单一模态的数据输入，导致模型缺乏对不同模态之间关联的深度理解。而在实际应用中，图像和文本信息常常是密切相关的[11]。

为提高图像识别的准确性和鲁棒性，已有研究开始将传统的单模态图像分类任务拓展到多模态学习框架中，通过整合文本信息来实现跨模态数据的优势互补。RADFORD等[11]提出了图像文本对比学习（Contrastive Language-Image Pre-training, CLIP）模型，通过将图像类别转化为文本描述并进行对比学习，显著提升了模型的性能及零样本推理能力。陈燕等[12]将CLIP模型用于多模态情感分析，取得了比基础模型更好的效果。FU等[13]提出的CMA-CLIP（Cross-Modality Attention CLIP）模型，通过结合序列注意力和模态注意力，增强了图像与文本之间的交互和融合，显著提升了分类效果。许睿等[14]利用图文对比学习并结合变分自编码器，提高了预训练特征在广义零样本图像分类任务上的应用能力。这些方法在农作物病虫害识别领域也显示出巨大的潜力，通过引入文本信息丰富图像特征，能进一步提升模型的识别精度和鲁棒性。然而，CLIP的对比学习机制仅实现了模态特征在特征空间的对齐，未实现深入的模态融合。为克服这一限制，LI等[15,16]提出了Align before Fuse（ALBEF）和Bootstrapping Language-Image Pre-training（BLIP）系列模型。谢润峰等[17]基于BLIP-2的Q-Former模块，提出基于视觉语言模型的跨模态多级融合情感分析方法，将图像和文本结合进行情感分析，在情感分析任务中取得了较优性能。这些模型在CLIP特征对齐的基础上，利用注意力机制实现了图像与文本的特征交互与融合，从而提升了模型在图像文本任务中的性能。

基于通用多模态大模型，研究者围绕特定农业病虫害领域的融合机制开展研究，旨在解决CLIP等模型仅做特征对齐的局限性。例如，FENG等[18]提出了一种双向跨模态Transformer（Bidirectional Cross-Modal Transformer，BiCMT）模型，通过结合YOLOv5s与文本信息，有效解决了图像与文本序列长度不一致的问题，显著提升了蔬菜病害在复杂背景下的识别精度。针对小样本场景，CAO等[19]构建了基于图像-文本-标签的多模态语言模型，利用多模态对比学习在黄瓜病害识别任务中实现了优异的泛化性能。此外，LIU等[20]利用大模型生成包含全局、局部病灶及颜色纹理的详细文本描述，并通过交叉注意力机制进行特征融合，以轻量化的参数量在多个数据集上超越了纯图像模型。为解决在嘈杂数据下鲁棒性不足的问题，DENG等[21]提出了Fourier-guided Attention Mechanism（FCMNet）模型，利用其核心的跨视觉-语言对齐模块（Cross Vision-Language Alignment, CVLA），通过图像块分割与特征映射技术，实现了图像与文本描述之间更深度的“语义交互”。尽管上述方法验证了深度融合在特定场景下的优势，但如何设计一个更优的自适应融合机制，以平衡模态间的贡献并进一步提升模型的泛化能力，仍是当前研究的关键难点。

针对现有农作物病虫害识别模型过度依赖单模态图像数据且常规多模态方法仅停留在浅层特征对齐的问题，本研究提出了一种自监督适应性多模态融合框架（Self-supervised Adaptive Multimodal Feature Fusion Recognition-CLIP, SAFusion-CLIP）。首先，使用适应性特征融合机制，通过门控单元实现图像与文本特征的深度交互与互补。其次，引入自监督特征重构任务，迫使模型学习更鲁棒的融合特征表示，突破传统单模态及简单多模态识别的局限性。最后，引入设计跨模态平衡对齐模块解决下游任务中图像样本与文本标签数量不对称的问题。

1　基于自监督的适应性特征融合模型

1.1　模型结构

本研究提出的SAFusion-CLIP，其整体结构如图1所示。模型主要由基于CLIP的跨模态对齐、模态融合和自监督学习三个部分组成。

1.2　基于CLIP的跨模态对齐

首先，为将传统图像识别任务转换为多模态的识别任务，通过提示工程将原始标签转换为文本描述，为每个类别增加病理特征描述，增强每个类别标签的文本语义信息。如图2所示，将“苹果-黑腐病”转换为“这是一张农作物苹果的图片，图中的

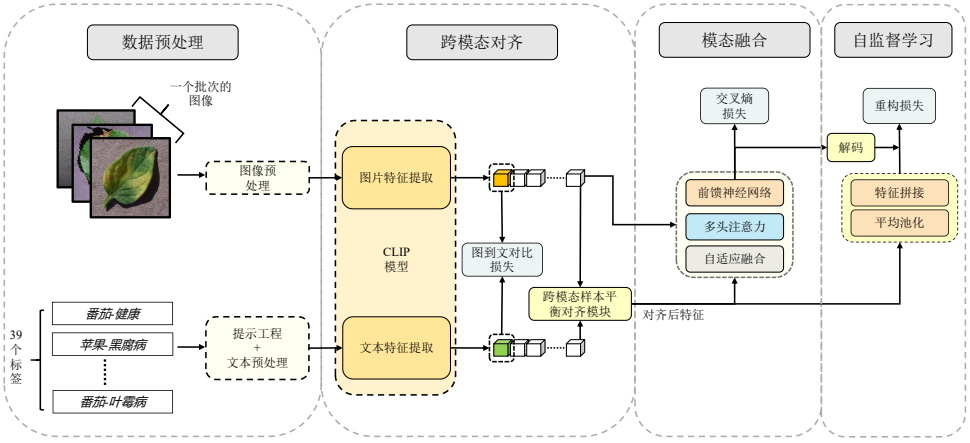


图 1 SAFusion-CLIP 整体结构图

Fig. 1 Overall architecture diagram of SAFusion-CLIP

作物患有黑腐病，其病理特征为：叶片出现边缘发黄、枯焦的现象，病斑表现为不规则的褐色或黑色斑点。叶片可能出现萎缩、卷曲和变形。”

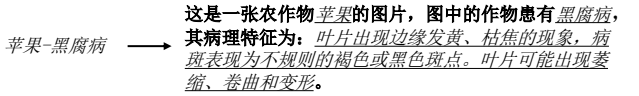


图 2 使用类别标签构建文本描述的提示工程示意图

Fig. 2 Illustration of prompt engineering for constructing text descriptions using category labels

由于CLIP利用了大量的图像和文本数据进行对比学习的预训练，从而能够在下游任务中更有效地学习图像与文本之间的关联关系。所以本研究将CLIP作为图像和文本的特征编码器，如图3所示。

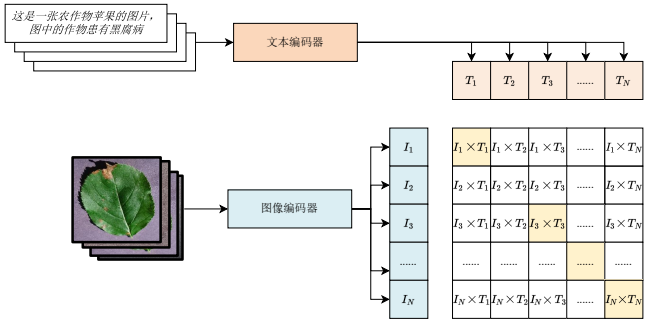


图 3 CLIP模型结构示意图

Fig. 3 Schematic diagram of the CLIP model architecture

基于CLIP的跨模态对齐模块通过图文对比学习可以实现图像和文本的嵌入特征对齐，但是将农作物病虫害识别由图像分类任务转换为多模态任务后，图像和文本化之间存在样本数量不一致的问题，这将导致图像和文本特征在交互时无法形成有效的一一对应关系，使得图像和文本模态无法进行有效的交互与融合。因此，本研究设计了跨模态平衡对齐模块，如图4所示，依靠CLIP的优秀图像识别能力，基于点积注意力[22]的自适应机制，实现图像和文本样本的对齐。假设图像编码结果为$\boldsymbol{I}$（$\boldsymbol{I}\in\mathbb{R}^{b\times p\times h}$），文本编码结果为$\boldsymbol{T}$（$\boldsymbol{T}\in\mathbb{R}^{39\times len\times h}$），其中$b$表示批量大小，$p$表示图像经过视觉编码器处理后的通道数量，$h$表示隐藏层大小，数字39和$len$分别表示有39个文本描述，文本通过填充和截断统一为长度为len的词元序列。

首先分别提取出图像和文本的全局特征（［CLS］标记），如公式（1）所示。

$$\begin{cases}\boldsymbol{I}_{\text{cls}}=\boldsymbol{I}[:,0,:], & \boldsymbol{I}_{\text{cls}}\in\mathbb{R}^{b\times h}\\ \boldsymbol{T}_{\text{cls}}=\boldsymbol{T}[:,0,:], & \boldsymbol{T}_{\text{cls}}\in\mathbb{R}^{39\times h}\end{cases}\tag{1}$$

式中：$\boldsymbol{I}_{\text{cls}}$是图像全局特征；$\boldsymbol{T}_{\text{cls}}$是文本全局特征。

将全局特征分别投影到特征空间，如公式（2）所示。

$$\begin{cases}\boldsymbol{I}_{\text{proj}}=Vision_proj\left(\boldsymbol{I}_{\text{cls}}\right), & \boldsymbol{I}_{\text{proj}}\in\mathbb{R}^{b\times h}\\ \boldsymbol{T}_{\text{proj}}=Text_proj\left(\boldsymbol{T}_{\text{cls}}\right), & \boldsymbol{T}_{\text{proj}}\in\mathbb{R}^{39\times h}\end{cases}\tag{2}$$

式中：$\boldsymbol{I}_{\text{proj}}$表示图像全局特征经过投影后的张量；$\boldsymbol{T}_{\text{proj}}$表示文本全局特征经过投影后的张量；$Vision_proj$表示图像投影层；$Text_proj$表示文本投影层。

进一步，使用点积矩阵衡量图像与文本特征之间的相关性，如公式（3）所示。

$$dot\left(\boldsymbol{I},\boldsymbol{T}\right)=\boldsymbol{I}_{\text{proj}}\times\boldsymbol{T}_{\text{proj}}^{\text{T}};\ dot\left(\boldsymbol{I},\boldsymbol{T}\right)\in\mathbb{R}^{b\times 39}\tag{3}$$

对点积矩阵沿文本维度进行$softmax$归一化，图像对应的文本注意力权重如公式（4）所示。

$$\boldsymbol{w}_{i2t}=softmax\left(dot\left(\boldsymbol{I},\boldsymbol{T}\right),\ \dim=-1\right);\ \boldsymbol{w}_{i2t}\in\mathbb{R}^{b\times 39}\tag{4}$$

式中：$\boldsymbol{w}_{i2t}$为每个图片在所有文本标签上的注

意力权重。

根据注意力权重对文本特征进行加权求和，生成对齐后的文本特征，如公式（5）所示。

$$\begin{cases} \boldsymbol{W}` = Unsqueeze\left(\boldsymbol{w}_{i2t}, 1\right);\ \boldsymbol{W}` \in \mathbb{R}^{(b \times 1 \times 39)} \\ \boldsymbol{T}` = Unsqueeze\left(\boldsymbol{T}, 0\right);\ \boldsymbol{T}` \in \mathbb{R}^{(1 \times len \times 39 \times h)} \\ \tilde{\boldsymbol{T}} = matmul\left(\boldsymbol{W}`, \boldsymbol{T}`\right);\ \tilde{\boldsymbol{T}} \in \mathbb{R}^{(b \times len \times h)} \end{cases} \quad (5)$$

式中：$Unsqueeze(\cdot, \mathrm{k})$表示升维操作，即在张量的第$k$维位置插入一个尺寸为1的维度，目的是将原始张量$\boldsymbol{w}_{i2t}$和$\boldsymbol{T}$变换为满足广播机制要求的中间变量$\boldsymbol{W}`$与$\boldsymbol{T}`$；$matmul(\cdot, \cdot)$表示批量矩阵乘法，该操作首先利用广播机制将输入张量在维度1和维度0上分别进行复制扩展，随后沿大小为39的公共特征维度进行矩阵相乘与缩并，最终得到形状为$b \times len \times h$的对齐后的文本特征$\tilde{\boldsymbol{T}}$。

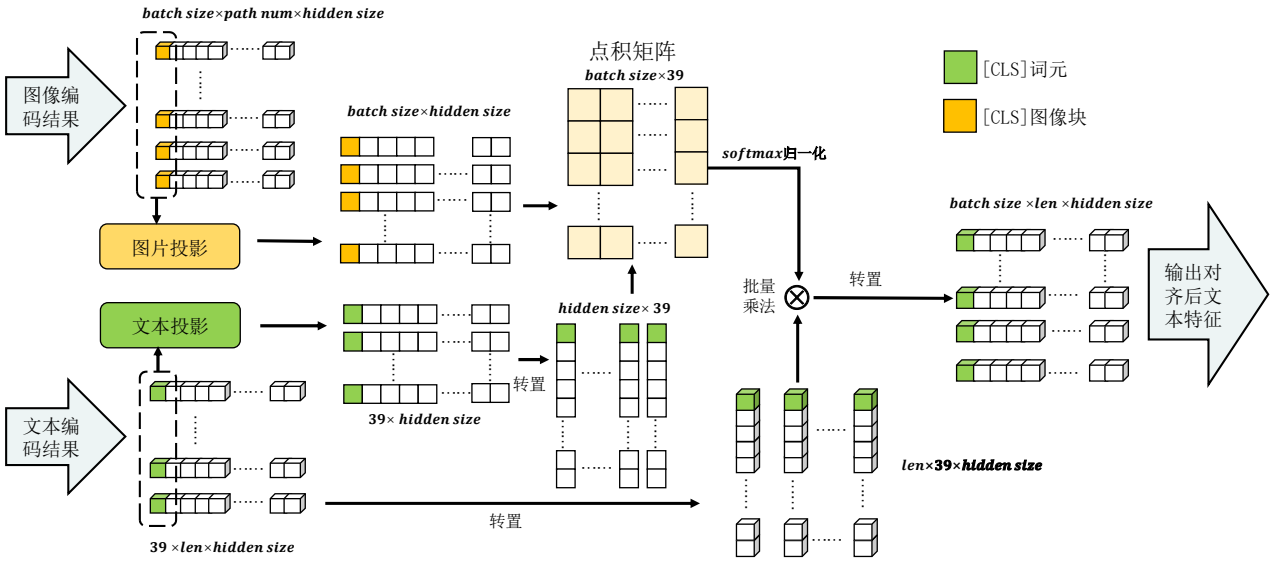


图4　跨模态平衡对齐模块示意图

Fig. 4　Schematic diagram of the Cross-Modal Balanced Alignment module

1.3 适应性多模态特征融合

多模态特征融合则通过适应性特征融合模块实现图像和文本特征的深度融合。模块结构如图5所示。

图5利用线性变换和自适应权重分配，对模态间互补信息进行充分融合，从而生成包含两模态关键信息的高质量融合特征。适应性多模态特征融合基于缩放点积注意力[22]，如公式（6）和公式（7）所示。

$$\boldsymbol{\alpha}\left(\boldsymbol{I}^{\mathrm{q}}, \boldsymbol{T}^{\mathrm{k}}\right)=\alpha_{ijl\ b\times p\times len};\ i\in[1,b], j\in[1,p], l\in[1,len] \quad (6)$$

$$\boldsymbol{\alpha}_i = \frac{\boldsymbol{I}_i^{\mathrm{q}} \cdot (\boldsymbol{T}_i^{\mathrm{k}})^{\mathrm{T}}}{\sqrt{h}};\boldsymbol{\alpha}_i \in \mathbb{R}^{p \times len} \quad (7)$$

式中：$\boldsymbol{\alpha}_i$表示第i个经过跨模态平衡对齐后的图片文本对特征的注意力分数矩阵，其中α_{ijl}表示在第i个图片文本对中，第j个图片通道在第l个文本token上的注意力分数；$\boldsymbol{I}^{\mathrm{q}} \in \mathbb{R}^{b \times p \times h}$为$\boldsymbol{I}$经过投影得到的图像查询矩阵；$\boldsymbol{T}^{\mathrm{k}}$，$\boldsymbol{T}^{\mathrm{v}} \in \mathbb{R}^{b \times len \times h}$为$\tilde{\boldsymbol{T}}$经过投影的键值对矩阵，$\boldsymbol{\alpha}\left(\boldsymbol{I}_{\mathrm{q}}, \boldsymbol{T}_{\mathrm{k}}\right) \in \mathbb{R}^{b \times p \times len}$为注意力分数矩阵。

注意力的计算结果如公式（8）所示。

$$\boldsymbol{T}_{\mathrm{o}} = softmax\left(\boldsymbol{\alpha}\left(\boldsymbol{I}_{\mathrm{q}}, \boldsymbol{T}_{\mathrm{k}}\right)\right) \times \boldsymbol{T}_{\mathrm{v}};\ \boldsymbol{T}_{\mathrm{o}} \in \mathbb{R}^{b \times p \times h} \quad (8)$$

式中：$\boldsymbol{T}_{\mathrm{o}}$为注意力的输出矩阵。

将经过注意力机制交互的文本嵌入特征和图片嵌入特征进行拼接得到拼接矩阵$\boldsymbol{C}$，如公式（9）所示。

$$\boldsymbol{C} = concatenate\left(\left[\boldsymbol{I}, \boldsymbol{T}_o\right], -1\right);\ \boldsymbol{C} \in \mathbb{R}^{b \times p \times 2h} \quad (9)$$

式中：$concatenate([\cdot], -1)$表示张量拼接操作，参数-1指沿最后一个维度进行拼接。

将F_1进行线性降维并计算门控值，如公式（10）所示。

$$\boldsymbol{gate} = \frac{1}{1 + e^{-(\boldsymbol{C} \times \boldsymbol{w} + \boldsymbol{b})}};\ \boldsymbol{gate} \in \mathbb{R}^{b \times p \times h} \quad (10)$$

式中：$w \in \mathbb{R}^{b \times p \times 2h \times h}$；$b \in \mathbb{R}^{b \times p \times h}$；$w$，$b$分别表示解码器的权重，以及偏置项。$\boldsymbol{C} \times w + b$表示对矩阵进行降维。

依据门控值$\boldsymbol{gate}$对图片特征和文本特征加权求和，如公式（11）所示。

$$\boldsymbol{F}'_{\text{fuse}} = \boldsymbol{gate} \times \boldsymbol{I} + (1 - \boldsymbol{gate}) \times \boldsymbol{T}_o;\ \boldsymbol{F}'_{\text{fuse}} \in \mathbb{R}^{b \times p \times h} \tag{11}$$

将 $\boldsymbol{F}'_{\text{fuse}}$ 经过多头自注意力和前馈神经网络的计算就得到了最后的融合特征 $\boldsymbol{F}_{\text{fuse}}$。

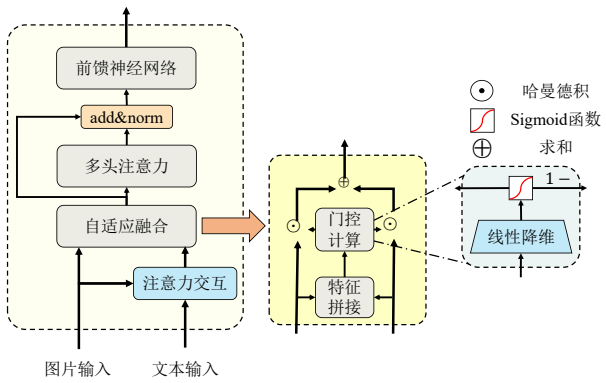


图 5 适应性特征融合模块示意图

Fig. 5 Schematic diagram of the adaptive feature fusion module

1.4 自监督的融合特征学习模块

在自监督学习部分，融合后的特征通过一个结构简单的解码器重构出原始的图像和文本嵌入特征，该部分引入特征重构任务构成自监督学习[23, 24]的架构，从而进一步提高模型对多模态特征的融合能力。SAFusion-CLIP 通过跨模态对齐、模态融合与自监督学习三者的结合，显著增强了多模态任务中的信息交互和特征表征能力。实验结果表明，本研究提出的模型在农作物病虫害识别任务中取得了优异的性能，充分验证了所设计方法的有效性。

SAFusion-CLIP 将一个单层前馈网络作为解码器。由于解码器的结构与特征融合部分相比较为简单，能够迫使模型在特征融合时学习更优质的融合特征，从而能在重构过程中还原出接近原始分布的图像和文本的嵌入特征，确保融合特征的质量和模型的泛化能力。

对图像和文本的嵌入特征进行平均池化，得到池化后图像特征 $\boldsymbol{I}_{\text{pool}}$ 和文本特征 $\boldsymbol{T}_{\text{pool}}$，如公式（12）所示。

$$\boldsymbol{I}_{\text{pool}} = \sum_{i=1}^{p} \boldsymbol{I}_i,\ \boldsymbol{T}_{\text{pool}} = \sum_{j=1}^{len} \boldsymbol{T}_j \tag{12}$$

式中：$\boldsymbol{I}_{\text{pool}}$，$\boldsymbol{T}_{\text{pool}} \in \mathbb{R}^{b \times h}$ 为池化后的图像特征与文本特征。

池化后的特征张量进行拼接得到目标特征 $\boldsymbol{F}_{\text{label}}$，如公式（13）所示。

$$\boldsymbol{F}_{\text{label}} = concatenate\left(\left[\boldsymbol{I}_{\text{pool}}, \boldsymbol{T}_{\text{pool}}\right], \dim = -1\right);\ \boldsymbol{F}_{\text{label}} \in \mathbb{R}^{b \times 2h} \tag{13}$$

使用解码器（*Decoder*）依据融合特征 $\boldsymbol{F}_{\text{fuse}} \in \mathbb{R}^{b \times p \times h}$ 预测目标特征 $\boldsymbol{F}_{\text{label}}$，如公式（14）和公式（15）所示：

$$\boldsymbol{F}_{\text{Pred}} = Decoder\left(\boldsymbol{F}_{\text{fuse}}[:, 0, :]\right);\ \boldsymbol{F}_{\text{pred}} \in \mathbb{R}^{b \times (2h)} \tag{14}$$

$$Decoder(\boldsymbol{X}) = \boldsymbol{X} \times \boldsymbol{w} + \boldsymbol{b};\ w \in \mathbb{R}^{b \times 2h \times h},\ b \in \mathbb{R}^{b \times h} \tag{15}$$

式中：$\boldsymbol{F}_{\text{Pred}}$ 表示 $\boldsymbol{F}_{\text{fuse}}$ 经过解码器计算后预测的原始嵌入特征张量（标签为 $\boldsymbol{F}_{\text{label}}$）；$\boldsymbol{X}$ 表示被解码的张量。

1.5 目标函数

模型的损失函数由图文对比损失（l_{itc}）、重构损失（l_{re}），以及分类损失（l_{cls}）三部分组成，如公式（16）所示。

$$L = \alpha l_{\text{itc}} + \beta l_{\text{re}} + \gamma l_{\text{cls}} \tag{16}$$

式中：α、β 和 γ 分别为控制各项损失对总体目标贡献的超参数。

参考 CLIP 模型优化方法，图文对比损失的目标是优化图像和文本嵌入特征在特征空间中的对齐程度。具体表达如公式（17）。

$$l_{\text{itc}} = -\sum_{i=1}^{n_i} \log \frac{\exp\left(\text{sim}_{\cos}\left(\boldsymbol{I}_i, \boldsymbol{T}_+\right) \times \exp(\tau)\right)}{\sum_{j=1}^{n_t} \exp\left(\text{sim}_{\cos}\left(\boldsymbol{I}_i, \boldsymbol{T}_j\right) \times \exp(\tau)\right)} \tag{17}$$

式中：n_i，n_t 分别表示图片数量和描述文本数量；$\boldsymbol{I}_i$，$\boldsymbol{T}_j$ 分别表示第 i 张图片和第 j 条描述文本的嵌入特征；$\boldsymbol{T}_+$ 表示图片 $\boldsymbol{I}_i$ 所对应的正样本（即标签对应的描述文本）；$\text{sim}_{\cos}$ 是余弦相似度计算函数；τ 为可学习的温度参数，用于调节对比学习的敏感性。

为实现多模态融合特征的高质量，本研究采用均方误差（Mean Squared Error, MSE）重构损失[23, 24]对融合特征进行优化。重构损失基于 MSE 计算，定义如公式（18）所示。

$$l_{\text{re}} = \frac{1}{N}\sum_{i=1}^{N}\left|\left|\left(\boldsymbol{I}_{\text{pool}}\right)_i - \widehat{\boldsymbol{I}}_i\right|\right|^2 + \frac{1}{M}\sum_{j=1}^{M}\left|\left|\left(\boldsymbol{T}_{\text{pool}}\right)_j - \widehat{\boldsymbol{T}}_j\right|\right|^2 \tag{18}$$

式中：$\widehat{\boldsymbol{I}}_i$ 和 $\widehat{\boldsymbol{T}}_i$ 分别为通过解码器重构的图像和文本特征，由 $\boldsymbol{F}_{\text{Pred}}$ 拆分得到，$\|\cdot\|^2$ 表示二范数。

分类损失用于完成最终的病虫害识别任务。考虑到数据集标签不平衡的问题，本研究采用加权交叉熵作为分类损失的计算方式，如公式（19）所示。

$$l_{\text{cls}} = -\sum_{i=1}^{N}\sum_{c=1}^{M} w_c y_{ic} \log\left(\widehat{p_{ic}}\right) \tag{19}$$

式中：N表示样本数量；M表示类别数量；$\hat{p}_{ic}$为模型预测的类别y_{ic}的概率；

w_c表示类别c的权重用来缓解数据集中标签不均衡的问题，表达如公式（20）所示。

$$w_c = N_T/\left(M + n_c\right) \tag{20}$$

式中：N_T为数据集样本总数；M表示标签数量；n_c表示批次类别为c的样本数量。

2　实验设计

2.1　实验数据集和评估指标

本研究实验数据来自PlantVillage[25]数据集，该数据集包含14种植物的图像，其中包括苹果、南瓜、马铃薯、柑橘等，可分为12种健康类植物和26类患病植物，如图6所示。包括没有农作物的背景图像在内共39类标签，各类别数据分布如图7所示。本研究采用准确率、精准率、召回率，以及F_1分数作为评价指标。

2.2　文本模态构建与质量评估

在利用提示工程将标签转换为详细文本描述后，为了确保生成的病理特征描述符合植物病理学事实，本研究引入了客观指标对文本模态的质量进行了量化评估。

为确保语义的一致性与准确性，本研究选取PlantVillage数据集源头机构——宾夕法尼亚州立大学（The Pennsylvania State University, Penn State University）下属农业推广服务发布的官方病害诊断指南[26]，以及美国植物病理学会（American Phytopathological Society, APS）的标准定义作为参考文本集[27]。这些文献详细定义了各类病害在田间的典型视觉表型（如病斑形态、颜色渐变及纹理



a. 苹果的健康与病害样本示例



b. 番茄的健康与病害样本示例

图6　PlantVillage数据集中苹果和番茄的健康与病害样本示例

Fig. 6　Examples of healthy and diseased samples of apple and tomato from the PlantVillage dataset

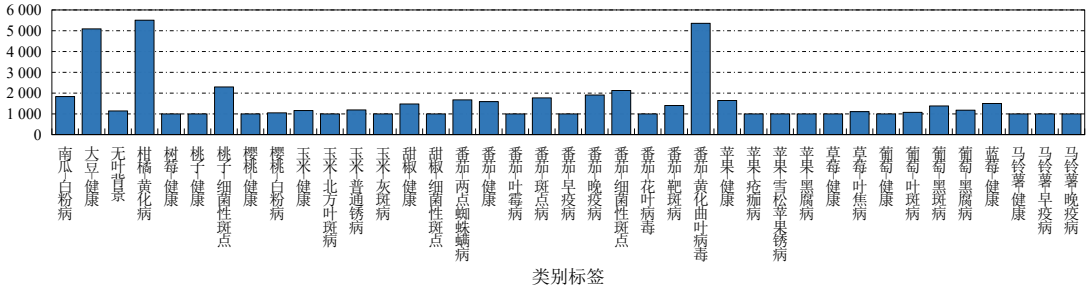


图7　PlantVillage数据集中各类病害的数据分布

Fig. 7　Data distribution of various disease classes in the PlantVillage dataset

特征)，为评估生成文本的准确性提供了权威基准。

本研究采用BERTScore[28]和ROUGE-L[29]两个自然语言处理领域的关键指标计算模型生成的视觉描述文本与官方参考文本之间的语义一致性。其中，BERTScore利用预训练的上下文嵌入来衡量深层语义的相似度，而ROUGE-L则侧重于评估关键术语和句法结构的最长公共子序列匹配程度。

为了验证文本构建策略的有效性，本研究从PlantVillage测试集的39个类别中，每个类别随机生成10个样本，共计390对“生成文本-参考文本”作为评估对象。针对每一对文本，分别计算其BERTScore和ROUGE-L分数，最后将所有样本的得分取平均值作为该模态的最终质量评估结果，如表1所示。计算过程并未参与模型训练，确保了评估的客观性与独立性。

表1 基于BERTScore和ROUGE-L的文本描述质量评估

Table 1 Evaluation results for text description quality based on BERTScore and ROUGE-L

评价指标	平均得分
BERTScore	0.912
ROUGE-L	0.685

实验数据表明，经过提示工程优化的文本描述在保持语义准确性的同时，兼顾视觉特征的精简表达，能够为后续的多模态特征融合提供高质量的语义引导。

2.3 实验环境和参数设置

本研究的实验环境如下：CPU型号为16 vCPU Intel（R） Xeon（R） Platinum 8352V CPU @ 2.10 GHz，显卡为Nvidia A100（80 G）*1，框架及软件版本为PyTorch 2.0.0，Python 3.8，Cuda 11.8。优化算法采用AdamW（Adam with Decoupled Weight Decay）[30]算法，将学习率设置为1×10^{-4}，学习率的衰退率设置为1×10^{-5}。学习率调整策略选择的是余弦退火法[31]，将学习率调整周期设置为1，迭代次数为5，训练批量大小为128，图像预处理分辨率设为224 × 224，文本最大长度为128，按照标签分布对训练集和测试集进行随机采样，训练集和测试集比例为8∶2，α、β和γ分别为0.8，0.2和1。

3 结果与分析

3.1 对比实验结果

分别选取基于卷积神经网络的ResNet58[32]、基于注意力机制的Swin-Transformer[33]，以及CLIP作为对比实验的基础模型。实验结果如图8所示。ResNet50在训练集上准确率达到了96.51%，Swin-Transformer达到了97.48%，CLIP模型的准确率最高，达到了98.23%。从图8可以看出，在CLIP、ResNet50、Swin-Transformer三个基础模型中，CLIP模型在4个评价指标上（准确率、精确率、召回率和F_1分数）均优于其他两个基础模型。CLIP模型由于在训练时考虑了文本模态的信息，因此其性能在三种基础模型中表现最好，这表明在传统的图像分类任务中，加入其他模态信息（如文本）能够增强模型对图像的识别能力。

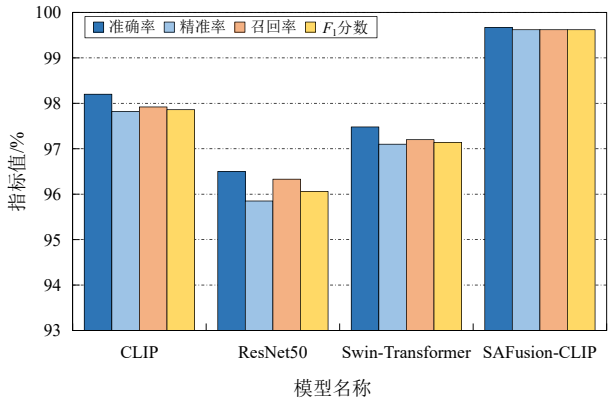


图8 不同模型在测试集上的准确率、精确率、召回率和F_1分数的性能对比

Fig. 8 Performance comparison of different models on the test set in terms of accuracy, precision, recall, and F_1-Score

SAFusion-CLIP通过结合自监督学习与适应性特征融合策略，有效整合了图像和文本模态信息，在实验中优于ResNet50、Swin-Transformer和CLIP模型，这表明自监督学习与多模态融合策略的结合能够显著提升模型对农作物病虫害的识别能力。

为了全面评估模型在实际场景中的计算效率，本研究在统一的硬件环境（NVIDIA 4060 8 G）下对各模型进行了推理速度测试，具体结果详见表2。实验数据显示，结构相对简单的单模态网络ResNet50表现出最快的推理速度（1.88 ms/img），而SAFusion-CLIP的单张推理耗时略高，为8.07 ms/img。同时，对比计算吞吐量，SAFusion-CLIP也是所有对比模型中最低的，为123.95 img/s。这种计算开销的增加主要源于模型对多模态信息的深度处理：一方面，双流编码架构要求模型同时并行执行图像与文本的特征提取；另一方面，为了实现更精细的语义交互，适应性特征融合模块及跨模态对齐计算引入了额外的网络深度与参数量。

表2 不同模型单张图像计算耗时及吞吐量对比

Table 2 Comparison of single-image inference time and throughput across different models

模型	单张图像耗时/(ms/img)	吞吐量/(img/s)
ResNet50	1.88	530.62
Swin-Transformer	4.31	231.92
CLIP	6.29	159.10
SAFusion-CLIP	8.07	123.95

3.2 类激活热力图对比

为深入探究模型关注的区域，采用梯度加权类激活映射（Gradient-weighted Class Activation Mapping, Grad-CAM）[34, 35]方法对模型进行可视化分析。为了确保不同架构模型之间对比的公平性，本研究均选取了各自视觉编码器中负责输出高层语义特征的关键层作为梯度计算的目标层。对于ResNet50，提取了最后一个卷积块（Layer4）的特征图；对于Swin-Transformer，选取了最后一个归一化层（LayerNorm）之前的输出；对于CLIP和本研究提出的SAFusion-CLIP，均选取了ViT图像编码器中最后一个Transformer模块的输出特征。

为直观比较各模型的关注区域差异，本研究在原图与热力图中以白色虚线标注病灶的真实范围（Ground Truth）。如图9所示，ResNet50与Swin-Transformer的高响应区域（红/黄）常分布在虚线范围之外的背景或叶片边缘，说明单模态模型易受非特异性视觉线索干扰、关注较分散；引入文本监督的CLIP模型能够覆盖到病灶区域，但其激活范围相对外扩、对病灶边界外区域存在溢出；而SAFusion-CLIP的热力图并非与虚线病灶范围“高度重合”，模型将更多的总体注意力（红/黄区域的总覆盖）分配在虚线圈定的白色区域内。相比其他模型，病灶内的注意力占比更高，但并不完全集中于病灶核心、局部仍存在一定扩散。这表明SAFusion-CLIP在关注分配上更偏向真实病灶区域，提升了病灶内的关注覆盖度与密度，同时在注意力收敛性上仍有优化空间。尤其在图9d番茄晚疫病识别场景下（小样本、背景干扰强），对比模型出现显著的背景误关注，而SAFusion-CLIP通过多模态融合抑制了部分背景噪声，使总体注意力更多地落在病灶范围内，从而在复杂场景下体现出更强的病灶区域偏置与鲁棒性。

尽管SAFusion-CLIP在整体性能上表现优异，但在处理视觉特征高度重叠或存在主导干扰特征的样本时仍面临挑战。为了深入探究误判原因，本研究选取了三个典型失败案例进行Grad-CAM可视化，并引入了模型预测错误类别的典型样本作为参照进行横向对比，如图10所示。

1）细粒度特征的视觉高度相似性。如图10a所示，模型将“玉米北方叶斑病”误判为“玉米灰斑病”。对比左侧原图与右侧的“玉米灰斑病”参考样本可知，两者均表现为叶片上的褐色坏死条斑，纹理特征极度相似。热力图显示模型关注了病灶区域，但未能有效捕捉到北方叶斑病“长梭形”与灰斑病“矩形”边缘的细微形态差异，导致分类混淆。

2）非特异性物理损伤的干扰。在图10b中，真实标签为“番茄早疫病”，但被误判为“番茄靶斑病”。对比右侧参考样本发现，“番茄靶斑病”同样伴随叶片边缘枯死和卷曲特征。Grad-CAM结果表明，模型过度聚焦于叶片边缘的物理损伤（非特异性特征），而忽略了早疫病核心的“同心轮纹”纹理，从而被两类病害共有的边缘坏死特征误导。

3）全局主导特征掩盖局部病理。图10c展示了一个被误判为“番茄黄化曲叶病毒”的“番茄细菌性斑点”样本。输入图像（左1图）呈现出显著的整体黄化特征，这与右侧“番茄黄化曲叶病毒”参考样本的视觉表现高度一致。虽然真实样本中存在微小的细菌斑点，但热力图显示模型的注意力完全被散落在背景和叶片轮廓上的“黄化”这一全局主导特征所吸引，导致局部微小病灶信息被掩盖。

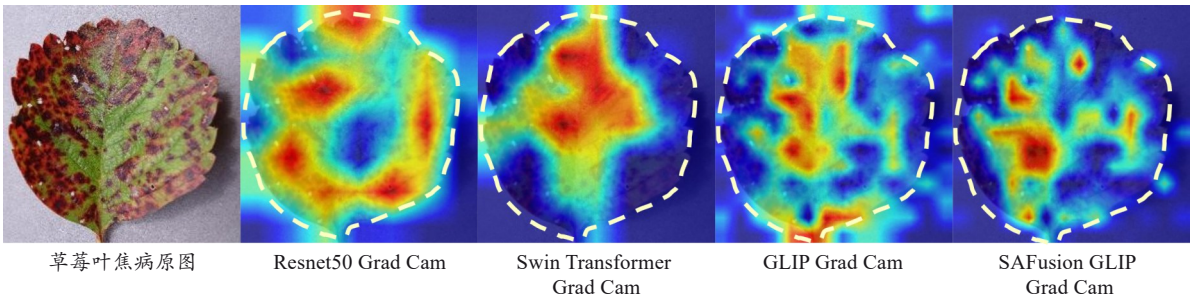
3.3 消融实验

本研究通过对模型依次去除或加入不同模块来探讨其对SAFusion-CLIP的作用。实验中构建的各个子模型及其模块组成如表3所示。

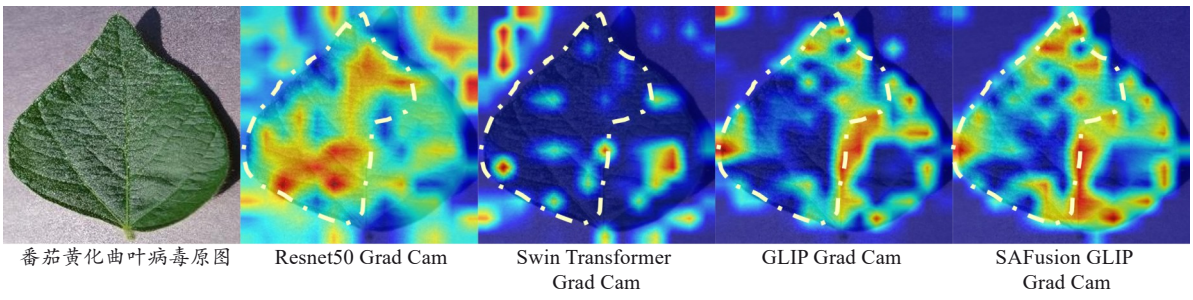
从表4和表5可以看出，不同模块对模型性能的影响显著。Model 3在测试集上的表现最佳，准确率达到了99.67%。

Model 3包含自监督学习和适用性特征融合，其四项评价指标（准确率、精确率、召回率和F_1分数）均超过99.00%。相比之下，Model 1和Model 2未包含自监督模块，表现较差，特别是Model 2去除了适应性融合机制，导致其在所有指标上的性能都低于其他模型。同时，自监督学习和适应性特征融合的结合显著提升了模型的分类性能。

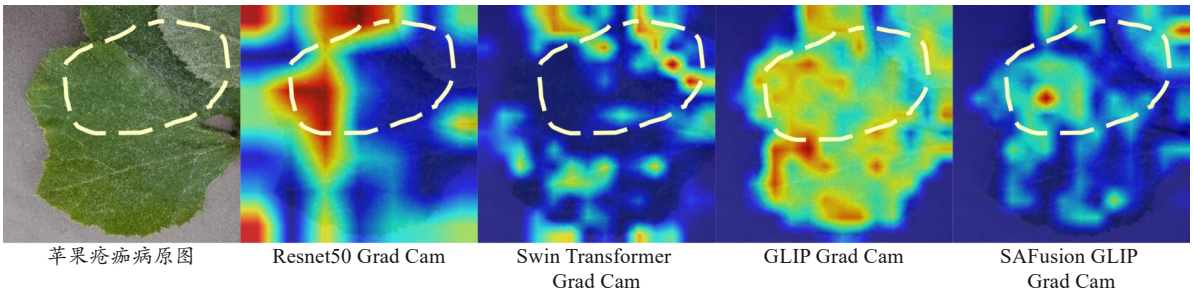
进一步计算得出各个模块对分类准确率的具体



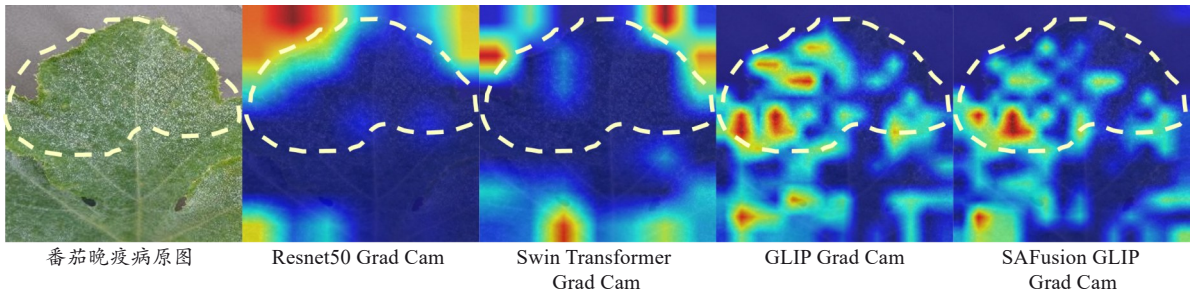
a. 草莓叶焦病 Grad-Cam 对比



b. 番茄黄化曲叶病毒 Grad-Cam 对比



c. 苹果疮痂病 Grad-Cam 对比



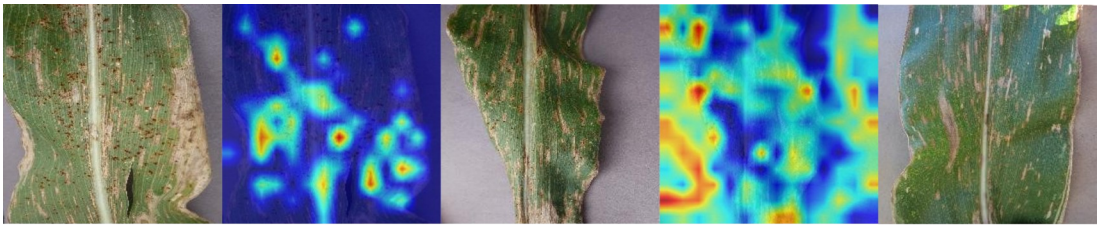
d. 番茄晚疫病 Grad-Cam 对比

图 9 不同类别梯度加权类激活映射图(Grad-CAM)对比

Fig. 9 Comparison of class activation heatmaps (Grad-CAM)

影响，如图 11 所示，引入文本模态信息后，模型准确率较传统单一模态模型（基准）提升了 0.75 个百分点，表明多模态信息的融合有助于增强图像内容的理解能力。在此基础上，自监督学习架构与文本模态结合后，准确率较基准提升了 1.25 个百分点。当同时引入自监督学习和适应性特征融合模块后，模型性能达到最优，准确率较基准提升了 2.46 个百分点，显示出二者的协同作用显著增强了多模态信息整合能力。此外，单独引入适应性融合模块使模型准确率提升了 1.21 个百分点，验证了其在多模态任务中的关键作用。

在引入文本模态的同时加入适应性融合和自监督学习后，SAFusion-CLIP 在测试集上的准确率相较于与其视觉编码器结构相似的 Swin-Transformer



a. 玉米北方叶斑病预测为玉米灰斑病样本



b. 番茄早疫病预测为番茄靶斑病　　c. 番茄细菌性斑点预测为番茄黄化曲叶病毒样本

图 10　基于 Grad-CAM 的典型失败案例可视化分析

Fig. 10　Visualization and analysis of typical failure cases based on Grad-CAM

表 3　SAFusion-CLIP 消融实验模型配置表

Table 3　SAFusion-CLIP ablation experiment model configuration table

模型名称	自监督	适应性融合机制	备注
Model 1	×	√	去除自监督模块
Model 2	×	×	去除自监督模块，去掉适应性融合
Model 3（SAFusion-CLIP）	√	√	去掉适应性融合

注：√表示包含；×表示不包含。

表 4　消融实验的模型配置与性能对比

Table 4　Model configurations and performance comparison of ablation settings

模型	消融实验设置	多头注意力头数	模态交互及融合层数	隐藏层大小	测试集准确率/%
Model 1	无自监督	12	4	768	98.42
Model 2	无自监督，无自适应融合	12	4	768	97.21
Model 3	有自监督	12	4	768	99.67

提升了 2.51%，进一步证明了多模态视角下的自监督适用性融合在提升农作物病虫害识别任务性能中的显著作用。

3.4　跨模态平衡对齐模块效果分析

通过随机选取测试集中 16 个样本对跨模态平衡对齐模块点积矩阵进行可视化分析，如图 12 所示，点积矩阵中的深色区域表示更高的注意力权重，对于一个样本，模型能够将更高的注意力集中于与每个样本图像对应的真实文本标签上，验证了模型在样本对齐过程中的区分能力。

表 5　消融实验下模型的准确率、精确率、召回率以及 F_1 分数结果对比表

Table 5　Comparison table of model's accuracy, precision, recall, and F_1-Score under ablation experiment

模型	评价指标/%			
	准确率	精准率	召回率	F_1 分数
Model 1	98.42	98.26	98.09	98.14
Model 2	97.21	97.05	97.05	96.95
Model 3	99.67	99.62	99.62	99.62

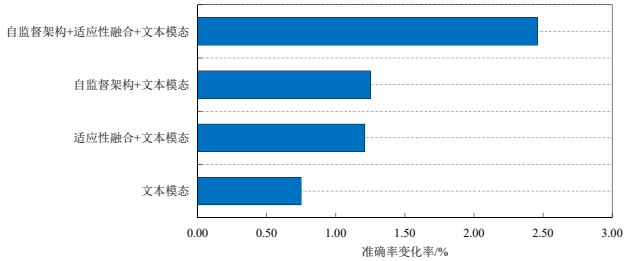


图 11　各模块组合对模型准确率变化的影响分析图

Fig. 11　Analysis chart of the impact of each module combination on model accuracy changes

通过消融实验部分的结果可知，跨模态平衡对齐模块依托于基础 CLIP 模型较为优秀的图像识别

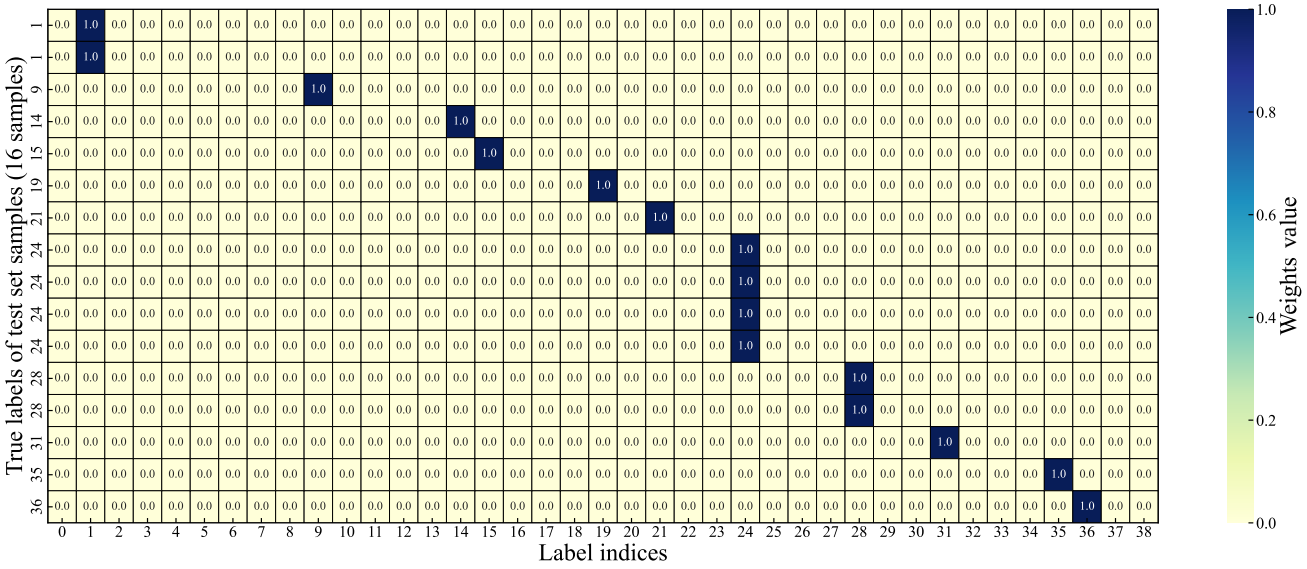


图 12 点积矩阵在跨模态平衡对齐模块中的可视化展示

Fig. 12 Visualization of dot product matrix in cross-modal balanced alignment module

能力，能够较为准确地识别出图像和文本标签的对应关系，从而能够为图像对应的真实文本标签分配更高的权重，使得最后对文本样本进行线性组合时每个图像对应的文本特征能够更多地包含其真实标签文本的信息。因此，跨模态平衡对齐模块有效增强了图像和文本模态之间的样本关联性，尤其在各模态样本数量不对称时，依托CLIP模型优秀的图像识别能力，能够实现对标签文本分配更高的权重，实现不对称图像和文本样本之间的平衡对齐。

3.5 多模态特征融合效果分析

使用核密度估计（Kernel Density Estimation, KDE）[36]曲线对模型的重构特征与未经模态交互和融合的真实特征进行对比，以此评估特征融合的效果。从测试集中随机选取了6个样本进行可视化，结果如图13所示。

从图13可见，6个样本的重构特征与真实特征在形态上高度一致，融合特征解码后得到的重构特征能够较好地拟合真实特征的分布，表明SAFusion-CLIP模型所得到的融合特征包含了丰富的关键信息，进一步验证了自监督学习对于SAFusion-CLIP的有效性。

4 结 论

通过在农作物病虫害识别任务中加入文本信息，相比于单模态ResNet50和Swin-Transformer模型，本研究提出的SAFusion-CLIP模型有着更优秀的识别性能，表明在农作物病虫害识别任务中引入文本模态将其转换为多模态识别任务，能够有效弥补单一模态特征的不足。和CLIP相比，SAFusion-CLIP通过引入自监督机制提升模型对融合特征的学习能力，对文本和图像特征进行自适应融合进一步提升了病虫害识别性能。SAFusion-CLIP在PlantVillage数据集上病虫害分类准确率达到99.67%，优于ResNet50、Swin-Transformer和CLIP模型，显著提升了农作物病虫害识别效果。

但值得注意的是，双流架构的计算开销仍高于单模态网络。未来将致力于利用模型剪枝或知识蒸馏技术压缩参数规模，在保持多模态高精度的前提下，实现模型在农业无人机、手持终端等边缘设备上的轻量化部署。再者，为解决“细粒度纹理混淆”与“全局特征干扰”导致的误判，未来工作将引入细粒度局部注意力机制，旨在增强模型在强背景噪声和高相似度类别下的特征解耦能力，使其能更精准地捕捉微小的病理纹理，而非过度依赖全局视觉特征。

此外，鉴于当前公开数据集缺乏多模态任务所需的标准化病理文本描述，限制了模型泛化能力的充分验证，未来将会尝试构建涵盖更多作物种类与复杂场景的大规模多模态数据集，为多模态技术在农业领域的深入应用提供数据支撑。未来将探索视觉识别与大语言模型的深度结合，构建集病害识别、成因溯源与治理决策于一体的多模态智能农业系统，推动技术从单一的“精准识别”向“智能决策”服务闭环迈进。

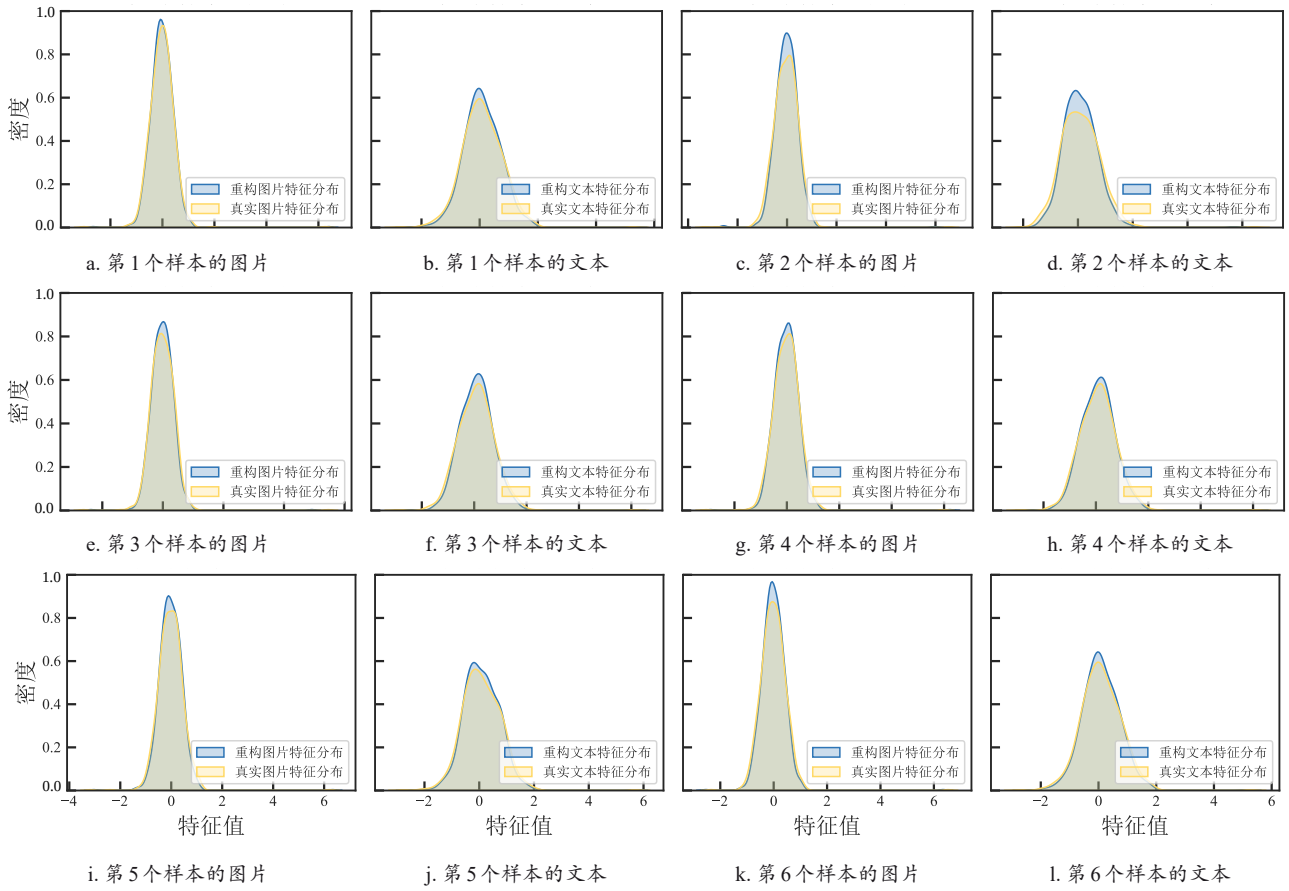


图 13　自监督任务的重构特征核密度估计可视化分析

Fig. 13　Feature reconstruction visualization in self-supervised architecture

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] FAO. Global crop losses due to pests and diseases: Annual impact on food security and economy[R]. Food and Agriculture Organization of the United Nation. [2025-09-02]. https://www.fao.org/plant-production-protection/about/zh.

[2] 王聃, 柴秀娟. 机器学习在植物病害识别研究中的应用[J]. 中国农机化学报, 2019, 40(9): 171-180.
WANG D, CHAI X J. Application of machine learning in plant diseases recognition[J]. Journal of Chinese Agricultural Mechanization, 2019, 40(9): 171-180.

[3] 邵明月, 张建华, 冯全, 等. 深度学习在植物叶部病害检测与识别的研究进展[J]. 智慧农业(中英文), 2022(1): 29-46.
SHAO M Y, ZHANG J H, FENG Q, et al. Research progress of deep learning in detection and recognition of plant leaf diseases [J]. Smart agriculture, 2022(1): 29-46.

[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(S2): 301-313.
MU J L, MA B, WANG Y F, et al. A review of deep-learning-based algorithms for crop pest and disease detection[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54 (S2): 301-313.

[5] 杨锋, 姚晓通. 基于改进YOLOv8的小麦叶片病虫害检测轻量化模型[J]. 智慧农业(中英文), 2024(1): 147-157.
YANG F, YAO X T. Lightweighted wheat leaf diseases and pests detection model based on improved YOLOv8[J]. Smart Agriculture, 2024(1): 147-157.

[6] 彭红星, 何慧君, 高宗梅, 等. 基于改进ShuffleNetV2模型的荔枝病虫害识别方法[J]. 农业机械学报, 2022, 53(12): 290-300.
PENG H X, HE H J, GAO Z M, et al. Litchi diseases and insect pests identification method based on improved ShuffleNetV2[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(12): 290-300.

[7] 冯峰, 周鑫, 陈诗瑶, 等. 一种基于改进神经网络算法ResNet50的玉米病虫害识别模型[J]. 江苏农业科学, 2024, 52(16): 239-244.
FENG F, ZHOU X, CHEN S Y, et al. A maize pest and disease identification model based on improved neural network algorithm ResNet50[J]. Jiangsu Agricultural Sciences, 2024, 52(16): 239-244.

[8] 孙杨俊, 陈滔, 刘志梁, 等. 基于双线性卷积宽度网络的水稻病虫害识别[J]. 计算机应用, 2024, 44(S1): 314-318.
SUN Y J, CHEN T, LIU Z L, et al. Rice pests and diseases recognition based on bilinear convolutional broad network[J]. Journal of Computer Applications, 2024, 44(S1): 314-318.

[9] 王杨, 李迎春, 许佳炜, 等. 基于改进Vision Transformer网络的农作物病害识别方法[J]. 小型微型计算机系统, 2024, 45(4): 887-893.
WANG Y, LI Y C, XU J W, et al. Crop disease recognition method based on improved vision transformer network[J]. Journal of Chinese Computer Systems, 2024, 45(4): 887-893.

[10] 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 Y M, LIU H L, SHI T T, et al. Tomato leaf disease recognition based on an optimized Swin Transformer[J]. Journal of China Agricultural University, 2023, 28(4): 80-90.

[11] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. arXiv: 2103.00020, 2021.
[12] 陈燕, 赖宇斌, 肖澳, 等. 基于CLIP和交叉注意力的多模态情感分析模型[J]. 郑州大学学报(工学版), 2024, 45(2): 42-50.
CHEN Y, LAI Y B, XIAO A, et al. Multimodal sentiment analysis model based on clip and cross-attention [J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(2): 42-50.
[13] FU J M, XU S Y, LIU H D, et al. CMA-CLIP: Cross-modality attention clip for text-image classification[C]// 2022 IEEE International Conference on Image Processing (ICIP). Piscataway, New Jersey, USA: IEEE, 2022: 2846-2850. https://doi. org/10.1109/ICIP46576.2022.9897323.
[14] 许睿, 邵帅, 曹维佳, 等. 基于重构对比的广义零样本图像分类[J]. 模式识别与人工智能, 2022, 35(12): 1078-1088.
XU R, SHAO S, CAO W J, et al. Generalized zero-shot image classification based on reconstruction contrast [J]. Pattern Recognition and Artificial Intelligence, 2022, 35(12): 1078-1088.
[15] LI J N, SELVARAJU R R, GOTMARE A D, et al. Align before fuse: vision and language representation learning with momentum distillation[EB/OL]. arXiv: 2107.07651, 2021.
[16] LI J N, LI D X, XIONG C M, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[EB/OL]. arXiv: 2201.12086, 2022.
[17] 谢润锋, 张博超, 杜永萍, 等. 基于视觉语言模型的跨模态多级融合情感分析方法[J]. 模式识别与人工智能, 2024, 37(5): 459-468.
XIE R F, ZHANG B C, DU Y P, et al. Cross-modal multi-level fusion sentiment analysis method based on visual language model [J]. Pattern Recognition and Artificial Intelligence, 2024, 37(5): 459-468.
[18] FENG X G, ZHAO C J, WANG C S, et al. A vegetable leaf disease identification model based on image-text cross-modal feature fusion[J]. Frontiers in Plant Science, 2022, 13: 918940.
[19] CAO Y 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.
[20] LIU W J, WU G Q, WANG H, et al. Cross-modal data fusion *via* vision-language model for crop disease recognition[J]. Sensors, 2025, 25(13): 4096.
[21] DENG S M, ZHU J L, HU Y, et al. Tomato leaf disease identification framework FCMNet based on multimodal fusion[J]. Plants, 2025, 14(15): 2329.
[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. arXiv:1706.03762, 2017.
[23] ERICSSON L, GOUK H, LOY C C, et al. Self-supervised representation learning: Introduction, advances, and challenges[J]. IEEE Signal Processing Magazine, 2022, 39(3): 42-62.
[24] BERAHMAND K, DANESHFAR F, SALEHI E S, et al. Autoencoders and their applications in machine learning: A survey[J]. Artificial Intelligence Review, 2024, 57(2): 28.
[25] HUGHES D P, SALATHE M. An open access repository of images on plant health to enable the development of mobile disease diagnostics[EB/OL]. arXiv: 1511.08060, 2015.
[26] PENN STATE EXTENSION. Pests and diseases[EB/OL]. [2025-12-16]. https://extension.psu.edu/pests-and-diseases.
[27] THE AMERICAN PHYTOPATHOLOGICAL SOCIETY. Common names of plant diseases[EB/OL]. [2025-12-16]. https://www.apsnet.org/edcenter/resources/commonnames/Pages/default.aspx.
[28] ZHANG T Y, KISHORE V, WU F, et al. BERTScore: Evaluating text generation with BERT[EB/OL]. arXiv: 1904.09675, 2019.
[29] SAI A B, MOHANKUMAR A K, KHAPRA M M. A survey of evaluation metrics used for NLG systems[J]. ACM Computing Surveys, 2023, 55(2):1-39.
[30] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB/OL]// arXiv:1711.05101, 2017.
[31] LOSHCHILOV I, HUTTER F. SGDR: Stochastic gradient descent with restarts[EB/OL]. arXiv: 1608.03983, 2016.
[32] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, New Jersey, USA: IEEE, 2016: 770-778.
[33] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, New Jersey, USA: IEEE, 2022: 9992-10002.
[34] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks *via* gradient-based localization[J]. International Journal of Computer Vision, 2020, 128 (2): 336-359.
[35] 王科平, 左鑫浩, 杨艺, 等. 基于伪全局Swin Transformer的遥感图像识别算法[J]. 模式识别与人工智能, 2023, 36(9): 818-831.
WANG K P, ZUO X H, YANG Y, et al. Remote sensing image recognition algorithm based on pseudo global swin transformer [J]. Pattern Recognition and Artificial Intelligence, 2023, 36(9): 818-831.
[36] 朱杰, 陈黎飞. 核密度估计的聚类算法[J]. 模式识别与人工智能, 2017, 30(5): 439-447.
ZHU J, CHEN L F. Clustering algorithm with kernel density estimation[J]. Pattern Recognition and Artificial Intelligence, 2017, 30(5): 439-447.

Self-Supervised Adaptive Multimodal Feature Fusion Recognition of Crop Diseases and Pests

YE Penglin[1,2], MIN Chao[1,2,3*], GOU Liangjie[2,3], WANG Pengcheng[1,2], HUANG Xiaopeng[1,2], LI Xin[1,2], MENG Yuping[4]

(1. *School of Science*, *Southwest Petroleum University*, *Chengdu* 610500, *China*; 2. *Institute of Artificial Intelligence*, *Southwest Petroleum University*, *Chengdu* 610500, *China*; 3. *National Key Laboratory of Oil and Gas Reservoir Geology and Development Engineering*, *Southwest Petroleum University*, *Chengdu* 610500, *China*; 4. *Information Center*, *Sinopec Zhongyuan Oilfield Company*, *Puyang* 457001, *China*)

Abstract: [Objective] Crop diseases and pests are significant factors restricting global agricultural production. Traditional intelligent recognition technologies predominantly rely on single-modal image data processed by convolutional neural networks (CNNs) or Transformers. However, in complex natural environments, these methods often suffer from insufficient information utilization and lim-

ited robustness due to the lack of semantic guidance. Although emerging multimodal approaches like CLIP have introduced textual information, they typically rely on shallow feature alignment in the embedding space without achieving deep semantic interaction or effective feature fusion. Furthermore, the asymmetry between the quantity of image samples and text labels during training poses a challenge for effective cross-modal learning. In this study, a self-supervised adaptive multimodal feature fusion recognition (SAFusion-CLIP) method is proposed, aiming to significantly enhance classification accuracy and model generalization in fine-grained diseases and pests recognition tasks. **[Methods]** A comprehensive recognition framework was constructed, integrating four key components to achieve deep fusion of visual and textual features. First, prompt engineering was conducted by utilizing large language models (LLMs) combined with authoritative agricultural guides to transform simple category labels into fine-grained pathological semantic descriptions. These descriptions encapsulated morphological details, color gradients, and texture features, with quality verified by BERTScore and ROUGE-L metrics. Second, a cross-modal balanced alignment module was designed to resolve the problem of sample asymmetry between image batches and fixed text labels. This module employed a dot-product attention mechanism to calculate the correlation between image and text projections, applying Softmax normalization to dynamically align image features with their corresponding textual representations. Third, an adaptive fusion mechanism was employed to achieve deep semantic interaction. A gating unit based on the Sigmoid function was designed to calculate a gate value, which dynamically allocated weights to image and text features, allowing the model to adaptively integrate complementary information from both modalities. Finally, a self-supervised feature reconstruction task was introduced to enhance the robustness of feature representation. A simple decoder was utilized to reconstruct the original image and text embeddings from the fused features, and the model was optimized using a composite objective function combining image-text contrastive loss, mean squared error reconstruction loss, and weighted cross-entropy classification loss. **[Results and Discussions]** Extensive experiments were conducted on the standard PlantVillage dataset, which includes 39 categories covering 14 crop species. The proposed SAFusion-CLIP model achieved a classification accuracy of 99.67%, with precision, recall, and F_1-Score all exceeding 99.00%. Comparative analysis demonstrated that the proposed method significantly outperformed mainstream single-modal and baseline multimodal models, ResNet50 (96.51%), Swin-Transformer (97.48%), and baseline CLIP (98.23%), respectively. Visualization analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) indicated that, unlike single-modal models which were susceptible to background noise or non-specific physical damage, the SAFusion-CLIP model focused more precisely on core lesion areas, effectively suppressing background interference. Furthermore, ablation studies confirmed the effectiveness of the proposed modules, showing that the combination of the self-supervised architecture and the adaptive fusion mechanism resulted in a 2.46 percentage points accuracy improvement over the baseline, validating the necessity of deep feature interaction and reconstruction tasks. **[Conclusions]** By fusing textual semantics with visual features, the SAFusion-CLIP method effectively overcame the limitations of single-modal recognition. The adaptive fusion mechanism ensured deep interaction between modalities, while the self-supervised reconstruction task significantly enhanced the robustness of feature representation. The experimental results verified that this data-driven approach significantly improves accuracy and generalization capabilities in fine-grained crop disease classification tasks, providing a new and effective solution for precision agricultural prevention and control.

Key words: diseases and pests recognition; multimodal fusion; adaptive feature fusion; self-supervised learning

Foundation items: National Natural Science Foundation of China (52574048); Sichuan Science and Technology Program (2025NSFTD0016)

First author: YE Penglin, E-mail: yplin0907@outlook.com
***Corresponding author:** MIN Chao, E-mail: minchaosc@126.com