山东农业大学学报(自然科学版),2026,57(1):166-178 VOL.57 NO.1 2026
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2026.01.017

基于对比学习的类别不均衡番茄叶片病虫害图像-文本检索方法研究

祝浩冉,芦 旭,张 亮*
山东农业大学信息科学与工程学院,山东 泰安 271018

摘 要: 番茄作为重要的经济作物,其病虫害防治对保障农业生产效益至关重要。番茄叶片病虫害图像能够直观展示病害形态特征和分布情况,但存在因分辨率有限、遮挡等情况导致细节丢失的缺点,而文本能够提供详细的发病症状描述、病因分析、防治策略,进而可以弥补图像的不足。然而,图像和文本两种模态存在语义鸿沟,使得利用图像精确检索相应文本描述困难,利用文本描述精确检索相应图像同样极具挑战性。此外,实际番茄叶片病虫害数据常存在类别不均衡问题,导致模型对主流类别过拟合、对稀有类别欠拟合。为解决上述问题,本文提出了一种基于对比学习的番茄叶片病虫害图像-文本跨模态检索方法并构建了首个中文番茄病虫害图像-文本数据集,其中模拟了实际生产中类别不均衡现象(如最大类别样本数为最小类别的近9倍),帮助研究更贴近实际生产中的挑战。为实现图像与文本模态的精确对齐,设计了一种基于三元组的对比学习方法,引入双曲空间建模层次语义关系,以拉近同类特征距离并拉远异类特征距离。提出自适应分类损失函数,动态调节模型对不同类别的关注度,有效缓解类别不均衡对检索性能的影响。此外,为提取局部病斑特征,设计了一种基于预训练编码器的参数冻结迁移学习模块,通过冻结编码器的参数来提取细粒度语义特征,以避免从头训练导致的额外训练成本。在实验部分,我们将所提出方法与先进的检索方法CCA、DSCMR、SCH、DDBH和DScPH进行对比,在图像检索文本任务上分别提升了28.68%、9.58%、3.38%、1.76%、1.03%,在文本检索图像任务上分别提升了35.71%、6.19%、0.94%、1.05%、0.54%。此外,为验证所提方法在图像编码器架构选择上的有效性,使用不同的图像特征编码器VGG16、MobileNet V2、CLIP-ViT-B/32进行对比,平均性能分别提升了12.965%、1.45%、1.005%。

关键词: 番茄病虫害识别;图像-文本检索;深度学习;对比学习
中图法分类号: TP391.4 **文献标识码:** A **文章编号:** 1000-2324(2026)01-0166-13

Research on Image-Text Retrieval Method for Tomato Leaf Diseases and Pests with Class Imbalance Based on Contrastive Learning

ZHU Hao-ran, LU Xu, ZHANG Liang*
College of Information Science and Engineering/Shandong Agricultural University, Tai'an 271018,China

Abstract: As an important cash crop, effective pest and disease control in tomatoes plays a crucial role in ensuring agricultural production efficiency. Images of tomato leaf pests and diseases can visually display the morphological characteristics and distribution patterns of the conditions, but they have drawbacks such as the loss of detail due to limited resolution, occlusion, and other factors. In contrast, text can provide detailed descriptions of disease symptoms, etiological analysis, and prevention strategies, thereby compensating for the deficiencies of images. However, a semantic gap exists between images and text, which makes it difficult to accurately retrieve corresponding text descriptions from images and vice versa. In addition, real-world tomato leaf disease and pest data often exhibit class imbalance, which causes models to overfit to dominant classes and underfit to rare classes. To address the above issues, this study proposes a contrastive learning-based image-text cross-modal retrieval method for tomato leaf disease and pest, and constructs the first Chinese tomato disease and pest image-text dataset. The dataset simulates the class imbalance observed in practical production (e.g., the number of samples in the largest class is nearly 9 times that in the smallest class), ensuring the research better reflects real-world challenges. To achieve accurate alignment between image and text modalities, this study designs a triplet-based contrastive learning method, which introduces hyperbolic space to model hierarchical semantic relationships, thereby pulling features of the same class closer while pushing features of different classes apart. Additionally, this study proposes an

收稿日期: 2025-03-25 **修回日期:** 2025-12-28
基金项目: 国家自然科学基金青年基金项目(62202281)
第1作者简介: 祝浩冉(2000-),男,硕士研究生,研究方向:多媒体信息检索,农业大数据与深度学习。E-mail:z1416161782@163.com
***通讯作者:** Author for correspondence. E-mail:zliang@sdau.edu.cn

adaptive classification loss function to dynamically adjust the model's attention to different classes, effectively mitigating the impact of class imbalance on retrieval performance. Furthermore, to extract local lesion features, this study designs a parameter-freezing transfer learning module based on pre-trained encoders. The module extracts fine-grained semantic features by freezing the encoder parameters, thereby avoiding additional training costs associated with training from scratch. In the experiment, this study compares the proposed method with advanced retrieval methods (CCA, DSCMR, SCH, DDBH, and DScPH). The proposed method achieves improvements of 28.68%, 9.58%, 3.38%, 1.76%, and 1.03% in the image-to-text retrieval tasks, and 35.71%, 6.19%, 0.94%, 1.05%, and 0.54% in the text-to-image retrieval tasks, respectively. In addition, to verify the effectiveness of the proposed method in selecting image encoder architectures, this study conducts comparisons using different image feature encoders (VGG16, MobileNet V2, and CLIP-ViT-B/32), with the average performance improvements of 12.965%, 1.45%, and 1.005%, respectively.

Keywords: Tomato disease and pest identification; image-text retrieval; deep learning; contrastive learning

番茄富含维生素和番茄红素等抗氧化物质，不仅具有增强免疫力、延缓衰老的营养价值，其含有的番茄红素还被证实具有抗癌和预防心血管疾病的功效。然而，番茄种植过程中易受多种病虫害侵袭，这不仅严重影响产量和品质，更会给种植户带来巨大经济损失，因此，病虫害防治直接关系到农业生产的整体效益。传统的番茄病虫害监测方法主要依赖于农业专家的经验判断和人工田间观察，存在效率低、成本高、主观性强等固有缺陷，随着农业规模的扩大和病虫害种类的增多，传统方法已难以满足现代农业的发展需求。近年来，计算机视觉技术在农业病虫害监测领域展现出巨大潜力[1-3]。番茄叶片病虫害图像作为一种直观的方式，能够有效展示病害的形态特征和分布情况。然而，图像数据通常受限于分辨率、叶片遮挡等因素，容易丢失关键细节，限制了其对病害细节的全面呈现。相对而言，文本数据能够提供详细的发病症状、病因分析及防治策略，可以弥补图像数据在语义信息表达上的不足。图像-文本跨模态检索方法[4-6]将图像和文本映射到统一空间中，进行图像和文本互相检索，充分利用两种模态的互补优势，有效弥补单模态信息不完整、特征表达有限及检索精度不足的问题。番茄病虫害图像-文本跨模态检索方法通过融合图像和文本信息，不仅能提升病虫害的识别准确性，还能提供病虫害的生物学特征、发病原因和防治措施等详细描述。图 1 展示了番茄病虫害图像-文本跨模态检索的主要流程，其核心在于通过将跨模态信息在统一空间的对齐，实现高效的检索。

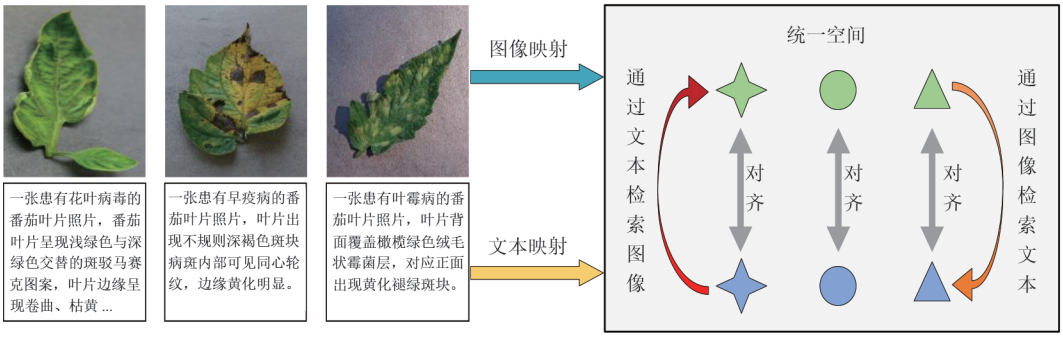

图 1 番茄病虫害图像-文本跨模态检索流程图

Fig. 1 Workflow of cross-modal image-text retrieval for tomato diseases and pests

然而，当前番茄病虫害图像-文本检索研究[7-9]仍然面临着一系列挑战：(1)数据方面，现有的数据集[10]主要以英文文本描述为主，缺乏高质量的中文图像文本数据集，这限制了国内相关研究和应用的发展。(2)图像-文本对齐方面，图像和文本两种模态在信息表达的方式、结构和语义层次上存在很大不同，导致难以对齐，且现有方法[11-15]在模态间语义关联建模上仍存在不足，未能充分利用标签这一显式语义信息。(3)特征提取方面，现有的基于深度学习的图像编码器[16]倾向于提取全局特征，难以有效捕捉番茄叶片病害的局部异常特征。(4)在实际生产中，番茄病虫害发病情况常出现类别不均衡问题[17,18]，由于其在自然分布上存在发生频率差异，导致不同种类

样本数量严重不均。本文构建数据集时模拟了这种实际生产中的分布情况，帮助研究更加贴近实际生产中的挑战，例如，黄化曲叶病毒图像多达3 202张，番茄花叶病毒仅有373张，最大与最小类别样本比例接近9∶1。现有的方法对此问题缺乏有效处理，如果只是简单的增加训练轮次，会导致部分类别过拟合，而其他类别则难以学习，从而影响检索精度和模型泛化能力；如果采用数据增强方法，通过旋转、裁剪、缩放等方式生成新的样本，虽增加了样本数量，但不足以捕捉样本多样性和复杂性。

为解决上述问题，本文提出一种基于对比学习的类别不均衡番茄叶片病虫害图像-文本检索方法。主要贡献如下：(1)针对目前缺乏中文番茄病虫害图像文本数据集的问题，本文构建首个高质量的中文番茄叶片病虫害图像-文本数据集，其中模拟了实际生产中番茄病虫害样本类别不均衡问题，帮助研究更加贴近生产中的挑战，并在后续模型设计中有效解决此问题。(2)针对图像-文本特征难以对齐问题，本文设计了一种基于三元组[19]的对比学习方法，充分利用标签显式语义特征，拉近同类特征距离，拉远异类特征距离，实现模态内和模态间的语义对齐。(3)针对类别不均衡问题，使用多任务学习架构[20]，设计自适应分类损失函数，动态调整模型对各类别的关注度，对难以学习的类别赋予更大的权重，提高其判别性。(4)针对局部病斑特征提取不足的问题，本文设计了一种基于预训练编码器的参数冻结迁移学习模块[21,22]，其中采用ResNet-50[23]图像编码器和CLIP文本编码器，冻结编码器的参数，实现局部特征的快速提取，有效捕捉病斑区域的细微差异，相较于从头开始训练特征提取模型更具优势。通过大量的实验对所提方法进行验证，充分证明其在番茄病虫害图像-文本检索中的优势。

1 数据获取与中文番茄病虫害图像-文本数据集构建

1.1 数据获取

本文构建了一个高质量番茄叶片病虫害图像-文本中文数据集。首先，通过网络爬虫技术从百度获取图像，并进行人工筛选以保证图像质量与标签准确性。经过筛选与整理，最终数据集包含10种类别的番茄叶片病虫害高质量图片，具体类别包括：番茄斑枯病、番茄花叶病毒、细菌性斑点病、黄化曲叶病毒、靶斑病、早疫病、晚疫病、叶霉病、双斑叶螨以及健康叶片。数据集共包含16 000张图片，且所有图片均为.jpg格式储存，表1展示了自建数据集的类别名称和各类别样本数目。

表1 数据集类别信息
Table 1 Dataset category information

类别序号 Category number	类别名称 Category name	样本数(图像-文本对) Number of samples (image-text pairs)
0	番茄斑枯病	1 771
1	番茄花叶病毒	373
2	细菌性斑点病	2 127
3	黄化曲叶病毒	3 202
4	靶斑病	1 404
5	健康	1 591
6	早疫病	1 000
7	叶霉病	952
8	双斑叶螨	1 671
9	晚疫病	1 909

如表1所示，本数据集模拟了类别不均衡问题，以真实反映番茄病虫害在实际生产中的分布情况，帮助研究更加贴近实际生产中的挑战。例如，“黄化曲叶病毒”类别包含3 202对图像-文本样本，而“番茄花叶病毒”仅有373对，最大与最小类别之间样本数相差近9倍。这种不均衡现象主要由病害自然发生频率差异、图像采集难度不一致及公开资源获取偏差等因素造成。然而，现有方法对此问题缺乏有效处理，如果只是简单的增加训练轮次，会导致部分类别过拟合，而其他类别则难以学习，从而影响检索精度和模型泛化能力；如果采用数据增强方法，通过旋转、裁剪、缩放等方式生成新的样本，虽增加了样本数量，但不足以捕捉样本多样性和复杂性。为此，本文设计了自适应分类损失函数，通过动态调整不同类别的权重，引导模型增强对稀有类别的关注，有效缓解类别不均衡对模型性能的不利影响。

1.2 中文番茄病虫害图像-文本数据集构建

本研究结合人工智能生成与专家筛选的文

本标注策略方法为每张图片生成了与其病虫害类别和属性特征(如病斑形态、颜色特征等)相关的中文文本描述,图2展示了部分类别的图像和文本样本示例。为便于模型读取与使用,数据集采用结构化方式存储:图片按类别名称存储在相应的文件夹中,同时将图片的相对路径、文本描述以及所属类别的one-hot标签存储在JSON文件中,模型通过读取JSON文件获取样本信息。

图2 番茄病虫害图像与文本样本示意图

Fig. 2 Example image-text pairs of tomato diseases and pests

为了更好地训练模型,本文参考跨模态检索[24]常用的比例划分数据集,将每个类别随机选择200个图像文本对,共计2 000个样本作为查询集对,其余14 000个图像文本对作为检索集,在检索集中,随机选择5 000个图像文本对作为训练集。模型的性能通过查询集在检索集上的检索精度进行评估,使用查询集中的图像(文本)作为查询输入,在检索集中检索与之匹配的文本(图像)。为了避免偶然性,模型的训练与测试重复进行10次,最终结果取10次结果的平均值。

2　研究方法

2.1　模型结构与符号定义

在本节中,对本文使用的符号进行定义,定义 $X^I=\{x_i^I\}_{i=1}^n$ 表示图像数据集,$X^T=\{x_i^T\}_{i=1}^n$ 表示文本数据集,n 表示样本总数,图像数据集的第 i 个样本与文本数据集的第 i 个样本表示同一对象。输入的数据集 $D=\{X,Y\}$,其中 $X=\{X^I,X^T\}$ 表示图像文本对,$Y\in\{0,1\}^c$ 表示样本类别信息,c 表示类别个数,每个样本只属于一个类别,当第 i 个样本属于第 j 类,可以表示为 $y_{i,j}=1$,其余元素为0。图3展示了本方法的整体网络结构,如图所示对输入的数据进行编码,通过各自编码器,特征提取后得到图像文本对 $\hat{X}=\{\hat{X}^I,\hat{X}^T\}$,将特征信息输入统一空间学习,通过映射层得到图像文本对低维特征 $H=\{H^I,H^T\}$,在自适应类别预测中,低维特征通过分类头,会得到图像预测标签 $\hat{y}^I$ 和文本预测标签 $\hat{y}^T$。

2.2　基于预训练编码器的参数冻结迁移学习模块

2.2.1 文本编码器　为了更好地提取中文番茄病虫害文本信息,本文采用CLIP文本编码器的架构。由于CLIP原始模型基于英文训练,为适配中文任务,我们进一步引入BERT-Base-Chinese分词器。CLIP文本编码器基于多层Transformer编码器堆叠而成,其核心思想是通过自注意力机制(Self-Attention)捕捉文本中的全局依赖关系,从而生成高质量的语义表示。每一层Transformer编码器都包含多头自注意力机制和前馈神经网络,并通过残差连接和层归一化来稳定训练过程,图4展示了本方法文本编码器的网络结构示意图。

具体而言,为了更有效地处理中文文本信息,本方法采用 BERT-Base-Chinese 分词器。该分词器能够将中文文本按单个汉字进行分割,并生成对应的 token,从而将文本信息映射到高维词向量空间中。具体而言,对于文本数据集中某一文本信息 $x_i^T=[w_1,w_2,\cdots,w_m]$,假设其包含 m

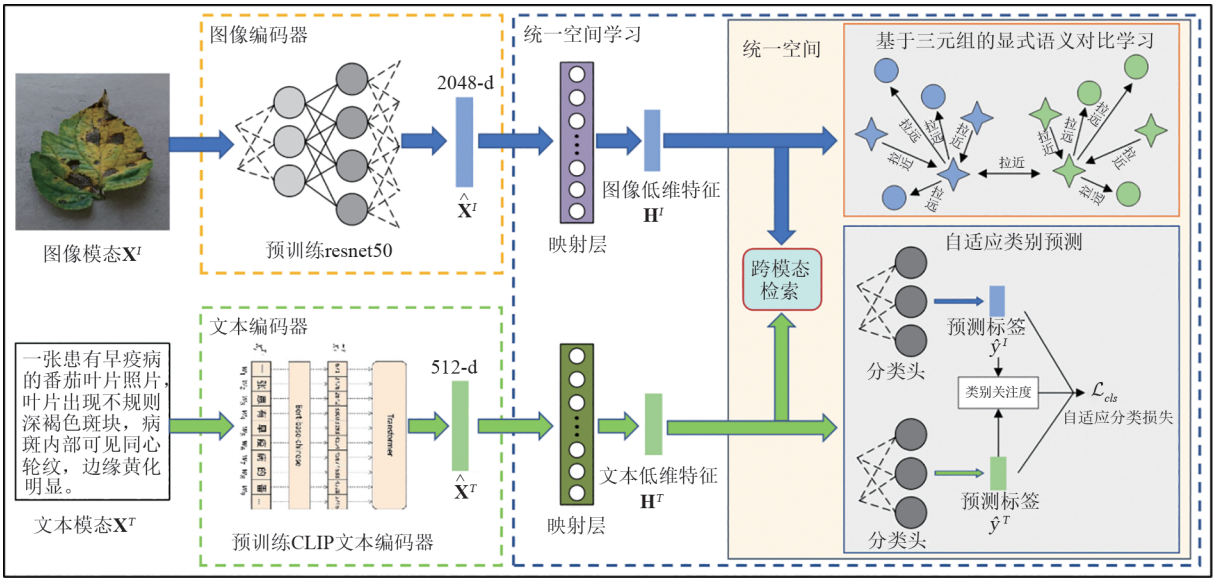

图3 本方法的整体网络结构图

Fig. 3 Overall network architecture of the proposed method

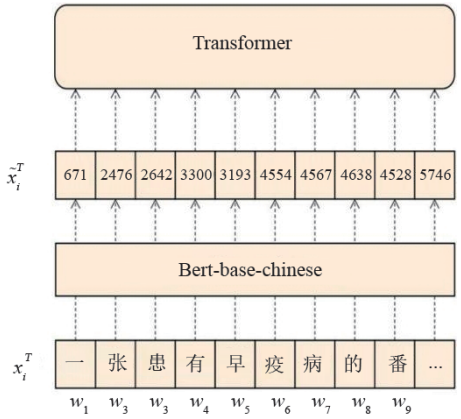

图4 文本编码器结构图

Fig. 4 Architecture of the text encoder

个汉字，使用分词器将文本逐词分割成 m 个 token，将其映射到对应的 token ID 中，将文本 x_i^T 逐词转换为对应的词向量 $\tilde{x}_i^T=[\tilde{w}_1,\tilde{w}_2,\cdots,\tilde{w}_m]$，然后得到的词向量经过多层 Transformer 编码器处理，Transformer 编码器层由多头注意力机制和前馈神经网络(Feed-Forward Network)组成，中间有残差连接和层归一化(Layer Normalization)，本节以一层为例介绍 Transformer 编码器的架构。多头注意力机制作为 Transformer 的核心组件，通过计算注意力分数，实现文本特征的注意。首先，计算查询 Q、键 K 和值 V 矩阵：

$$Q=\tilde{X}^T W^Q \tag{1}$$

$$K=\tilde{X}^T W^K \tag{2}$$

$$V=\tilde{X}^T W^V \tag{3}$$

其中，W^Q、W^K 和 W^V 是可学习的参数矩阵。然后计算单头的注意力得分：

$$Attention(Q,K,V)=softmax\left(\frac{QK^\top}{\sqrt{d_K}}\right) \tag{4}$$

其中，d_K 是键的维度。将多个注意力头的结果拼接在一起，得到多头注意力：

$$MultiHead(Q,K,V)=Concat(\text{head}_1,\cdots,\text{head}_h)W^M \tag{5}$$

其中，$\text{head}_i=\text{Attention}(QW_i^Q,KW_i^K,VW_i^V)$，$W_i^Q$、$W_i^K$、$W_i^V$ 和 W^M 均为可学习的权重矩阵，h 为注意力头的个数。最后，经过前馈神经网络(FFN)、残差连接和层归一化(LN)得到 512 维度的文本特征 $\hat{x}_i^T$。

$$Z=LN\left(\tilde{X}^T+\text{MultiHead}\left(\tilde{X}^T,\tilde{X}^T,\tilde{X}^T\right)\right) \tag{6}$$

$$\hat{X}_i^T=LN\left(Z+\text{FFN}(Z)\right) \tag{7}$$

2.2.2 图像编码器 在番茄病虫害图像-文本检索任务中，图像中往往包含大量如病斑、虫害痕迹等细粒度的局部特征信息，而这些细节正是实现精准检索的关键。考虑到本任务对图像细粒度语义表达能力的高度依赖，本文未直接采用 CLIP 原始结构中的 Vision Transformer(ViT)作为图像编码器，而是选择将其替换为具备更强局部感知能力的 ResNet-50。ViT 通过将图像划分为固定大小的 patch 并进行全局建模，能够捕捉

图像的整体语义信息,但在捕捉局部细节方面存在一定的不足。尤其是在病虫害图像中,局部病灶区域通常较小,ViT的特征提取方式可能导致其对这类细粒度结构的表征不够敏感。相较之下,ResNet-50作为一种典型的卷积神经网络(CNN),在结构设计上更适合于局部信息的提取。其多层卷积操作与逐层下采样机制使模型能够在不同尺度上学习图像的空间局部特征,并通过局部感受野和参数共享机制有效聚合周围区域的信息,从而增强对局部变化的表达能力。因此,结合番茄病虫害图像的特点和检索任务的需求,采用ResNet-50替代ViT能够更好地捕捉关键局部特征,增强模型对图像细节的敏感性,进而提升图像-文本检索的整体精度。基于上述分析,本文将ResNet-50用作图像编码器,以实现更具针对性的特征提取与语义匹配。

ResNet-50的网络结构如图5所示:

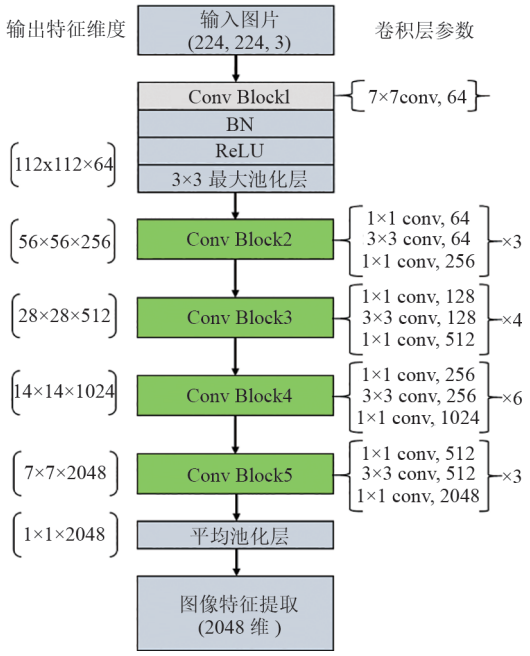

图5 图像编码器结构图

Fig. 5 Architecture of the image encoder

2.3　基于三元组的对比学习

统一空间学习通过将高维的特征映射到低维的二进制特征中,因其运算速度快、存储成本低的优势,而被广泛应用于图像-文本检索任务中。现有的方法大多基于对比学习框架,将图像与文本映射到同一特征空间,以实现跨模态语义对齐。然而,这些方法通常仅依赖样本级别的匹配关系,忽略了标签所携带的显式语义信息。为了应对细粒度图像-文本检索任务的需求,本方法提出了一种基于标签级显式语义信息的对比学习损失函数。该损失函数通过显式地建模类别级别的语义一致性,实现了同一类别下多模态特征的细粒度对齐,从而提升了检索精度与语义一致性。

首先,将各模态的高维特征信息通过映射层,得到低维特征H。

$$H^I = MapLayer\left(\hat{X}^I, \theta_I\right) \tag{8}$$

$$H^T = MapLayer\left(\hat{X}^T, \theta_T\right) \tag{9}$$

其中,映射层由全连接组成,θ_I和θ_T表示为可学习参数。双曲空间因为其几何性质能够自然地捕捉数据中的层次关系,为了挖掘深层语义结构,本方法将低维特征使用指数映射到双曲空间中,具体操作如下:

$$\bar{b}^{(*)} = Exp_p\left(h^{(*)}\right), s.t.* \in \{I, T\} \tag{10}$$

映射至此空间,可以在保持低维特征紧凑性的同时,进一步捕捉数据间的层次化语义关系。

在双曲空间中,样本间的距离可以表示为:

$$d\left(\bar{b}_i, \bar{b}_j\right) = arccosh\left(1+2\frac{\|\bar{b}_i - \bar{b}_j\|^2}{\left(1-\left\|\bar{b}_i\right\|^2\right)\left(1-\left\|\bar{b}_j\right\|^2\right)}\right) \tag{11}$$

为了充分利用双曲空间的层次化关系,本文使用三元组损失来学习样本之间的相对距离关系,实现细粒度的跨模态对齐。为了捕获标签级显式语义中的深层次细粒度信息,本文将属于同一类别的样本定义为正样本,不属于同一类别的样本定义为负样本。对于给定的三元组$(\bar{b}_i, \bar{b}_j, \bar{b}_k)$,其中$\bar{b}_i$表示锚点,$\bar{b}_j$表示正样本,$\bar{b}_k$表示负样本,三元组损失的目标是拉近锚点与正样本的距离,推远锚点与负样本的距离,通过这个损失可以实现模态间和模态内的同时对齐,损失可以表示为:

$$\mathcal{L}_c^{(*)} = L\left(\bar{b}_i^{(*)}, \bar{b}_j^{(*)}, \bar{b}_k^{(*)}\right) = max\left(0, d\left(\bar{b}_i^{(*)}, \bar{b}_j^{(*)}\right) - d\left(\bar{b}_i^{(*)}, \bar{b}_k^{(*)}\right) + \beta\right), \tag{12}$$

$$s.t.* \in I, T$$

$$\mathcal{L}_c = \mathcal{L}_c^{(I)} + \mathcal{L}_c^{(T)} \tag{13}$$

其中β是边距参数。

2.4　自适应类别预测

在低维特征的生成过程中,由于样本数量和

质量的不均衡，不同类别的检索精度可能存在较大差异，尤其是样本数量较少或语义边界模糊的类别更容易被忽略。为此，本文设计了一种自适应学习策略，旨在更充分地挖掘各类别的细粒度特征表示。其核心思想在于动态调整模型对各类别的关注度，实现各类别的自适应学习，确保难以区分的类别得到充分训练，避免已经充分学习的类别不会因为过多的训练导致过拟合。

具体而言，本文采用多任务学习架构，将得到的多模态低维特征分别输入到各自的模态特定分类头中，得到预测类别$\hat{y}$，并通过自适应权重调整机制优化训练过程。

$$\hat{y}^I = Softmax\left(W_1 h^I + b_1\right) \tag{14}$$

$$\hat{y}^T = Softmax\left(W_2 h^T + b_2\right) \tag{15}$$

其中，W_1和W_2是权重矩阵，b_1和b_2是偏置向量。

为了动态调整模型对各类别的关注度，本文使用交叉熵损失作为分类损失来衡量模型对于类别学习的好坏。

$$\mathcal{L}_{cls}^c = -\frac{1}{2N_c}\sum_{i=1}^{2N_c}\left[y\log\left(\hat{y}_i^{(*)}\right) + \left(1-y\right)\log\left(1-\hat{y}_i^{(*)}\right)\right], \tag{16}$$

$$s.t.* \in I,T$$

其中，N_c表示类别c的样本数，$\mathcal{L}_{cls}^c$表示属于类别c的所有图文样本计算得到的分类损失。本文通过对图像文本两个模态同时计算每个类别的分类损失。对于分类损失较大的类别给予一个更高的权重，使得模型更好地学习这个类别信息，类别c的关注度权重w_c

通过指数加权方法计算得到：

$$w_c = \frac{\exp\left(\alpha \cdot \mathcal{L}_{cls}^c\right)}{\sum_{j=1}^{C}\exp\left(\alpha \cdot \mathcal{L}_{cls}^j\right)} \tag{17}$$

其中，α是关注度控制超参数。通过引入类别关注度权重，使模型更加关注难以学习的类别信息，从而提高整体的分类性能，总分类损失可以表示为：

$$\mathcal{L}_{cls} = \sum_{j=1}^{C} w_j \cdot \mathcal{L}_{cls}^j \tag{18}$$

通过这种自适应方法，分类损失较大的类别会被赋予更高的权重，从而引导模型将更多的注意力集中在这些相对难以学习的类别上。随着训练的进行，各类别的分类误差会动态变化，权重也会相应调整，以实现自适应的学习过程。这种动态调整机制不仅能够有效提升模型对难以学习类别的识别能力，还能避免对已充分学习类别的过度训练，从而在整体上提高模型的泛化能力和鲁棒性。

统一空间学习模块的总损失可以表示为：

$$\mathcal{L}_h = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_c \tag{19}$$

其中，λ_1表示超参数。

3 结果与分析

3.1 模型性能与评价指标

对于模型检索性能的评价，本文使用平均精度均值（mAP）进行衡量，mAP综合考虑了检索结果的排序和相关性，能够有效反映模型在图像-文本检索任务中的性能，其计算公式如下：

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q} AP_q \tag{20}$$

其中，Q为查询样本，AP_q为第q个样本的平均精度。

此外，本文还采用准确度（Accuracy）作为评价指标，用以衡量模型在多任务学习中的分类性能，其计算公式如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

其中，TP表示预测为正且实际为正的样本数，TN表示预测为负且实际为负的样本数，FP表示预测为正但实际为负的样本数，FN表示预测为负但实际为正的样本数。

3.2 实验细节

本文通过图像编码器提取的图像特征维度为2048维，文本编码器提取的文本特征为512维。所有的实验都是在单块NVIDIA GeForce RTX 3090 GPU和2.10 GHz Intel（R）Xeon（R）Gold 5318Y CPU上完成的，实验环境包括cuda 11.7、python 3.7.13和pytorch 1.7.1。为提高训练效率，本文在图像编码器和文本编码器模块使用预训练权重，并在训练阶段保持这两个编码器参数不变，仅对统一空间学习模块进行训练。该模块训练100个轮次，每批次包含128对图像-文本样本对。训练过程中，超参数λ_1设置为2，β设置为0.5。所有实验均基于本文构建的番茄叶片病虫害图像-文本数据集进行，实验中涉及的i2t表示通过图像检索文本，t2i表示通过文本检索图像。

3.3　对比方法

本研究设计了两种对比实验方案,分别从图像-文本检索方法和编码器架构两个维度对所提模型的性能进行系统评估。在图像-文本检索方法对比方面,本研究选取了传统图像-文本检索方法中的典型相关分析(Canonical Correlation Analysis, CCA),以及四种先进的深度图像-文本检索方法:深度监督图像-文本检索(Deep Supervised Cross-modal Retrieval, DSCMR)、语义通道哈希(Semantic Channel Hashing, SCH)、深度判别边界哈希(Deep Discriminative Boundary Hashing, DDBH)和深度语义一致性惩罚哈希(Deep Semantic-consistent Penalizing Hashing, DScPH)作为对比基线。通过与这些方法的对比,全面评估本文方法在跨模态检索任务中的有效性与优势。在编码器架构对比方面,本研究系统地比较了多种主流编码器组合,包括VGG16、MobileNet v2 以及 CLIP 模型中的 ViT 作为图像编码器,同时选用 MiniLM 作为文本编码器进行对比实验,以全面评估不同编码器架构对模型性能的影响。

3.4　对比实验结果分析

3.4.1 图像-文本检索方法对比　为了验证本方法在图像-文本检索方面的有效性,本模型在番茄叶片病虫害图像-文本数据集上,与三种先进的图像-文本检索方法进行比较,为确保实验的公平性和可比性,所有对比方法统一使用 ResNet-50 图像编码器和 Transformer 文本编码器,表 2 展示了使用不同模型在通过图像检索文本和通过文本检索图像的平均精度均值对比结果。

为便于定量分析,表中对最优结果进行了加粗标注,次优结果则以下划线标示。实验结果表明,在图像检索文本和文本检索图像两个任务上,本方法在平均精度均值上均显著优于对比方法。传统方法CCA的性能明显低于基于深度学习的图像-文本方法,这证实了深度学习方法在特征表示学习方面具有更强的关系挖掘能力。与先进的检索方法 DSCMR、SCH、DDBH 和 DScPH 相比,本方法展现出显著的性能优势,在通过图像检索文本上精度分别提升 9.58%、3.38%、1.76% 和 1.03%,这说明本方法能够在低维空间中保留更丰富的特征信息,从而验证了本方法的有效性。

表 2 不同模型的检索平均精度均值对比结果
Table 2 Comparison results of mean average precision (MAP) for different models

模型 Model	i2t	t2i
CCA	0.641 6	0.593 2
DSCMR	0.832 6	0.888 4
SCH	0.894 6	0.940 9
DDBH	0.910 8	0.939 8
DScPH	0.918 1	0.944 9
OUR	0.928 4	0.950 3

3.4.2 图像编码器对比　为了验证本方法在图像编码器架构选择上的有效性,本研究在自建数据集上与多种图像编码器进行对比实验,为了控制唯一变量,统一空间学习网络保持不变,每次只替换一个图像编码器。本研究采用定量与定性相结合的方法评估图像编码器性能:通过计算平均精度均值(mAP)进行定量分析,同时基于多任务学习架构获得的预测标签分布绘制混淆矩阵,实现定性评估。

表 3 展示了使用不同的图像特征编码器的图像检索文本、文本检索图像的平均精度均值对比结果。实验结果充分表明,本方法采用的 ResNet-50 图像编码器和 Transformer 文本编码器的效果是最好的。在图像编码器对比方面,VGG16 效果最差,其在细粒度特征提取方面表现欠佳,难以准确区分病虫害的细微视觉差异。MobileNet V2 作为一种轻量级网络,虽然在计算效率方面表现出色,但不能充分提取细粒度特征信息。CLIP-ViT-B/32 虽然采用了视觉 Transformer 架构,能更好地对齐语义信息,但在处理局部细节特征时仍存在一定局限。

图 6 展示了不同图像编码器的混淆矩阵,通过可视化的形式,直观展现使用各种不同图像编

表 3 不同图像编码器的平均精度均值对比结果
Table 3 Comparison results of average precision means of different image encoders

模型 Model	i2t	t2i
VGG16	0.807 1	0.812 3
MobileNet V2	0.909 2	0.940 5
CLIP-ViT-B/32	0.917 8	0.940 8
OUR	0.928 4	0.950 3

码器的预测类别分布，每个混淆矩阵中横轴是预测的类别，纵轴是真实的类别，矩阵中对角线位置的数值越大、颜色越深，说明模型在该类上的分类准确率越高。通过混淆矩阵可以观察到，VGG16 分类性能最差，对角线分布松散，尤其是第 6 类误分类严重，MobileNet V2 表现有明显提升，大部分对角线数值较高，误差分布较少，CLIP 整体表现稳定，多数类别正确率较高，但个别类有较小的误判，本方法表现最佳，对角线数值最大，准确率最高。

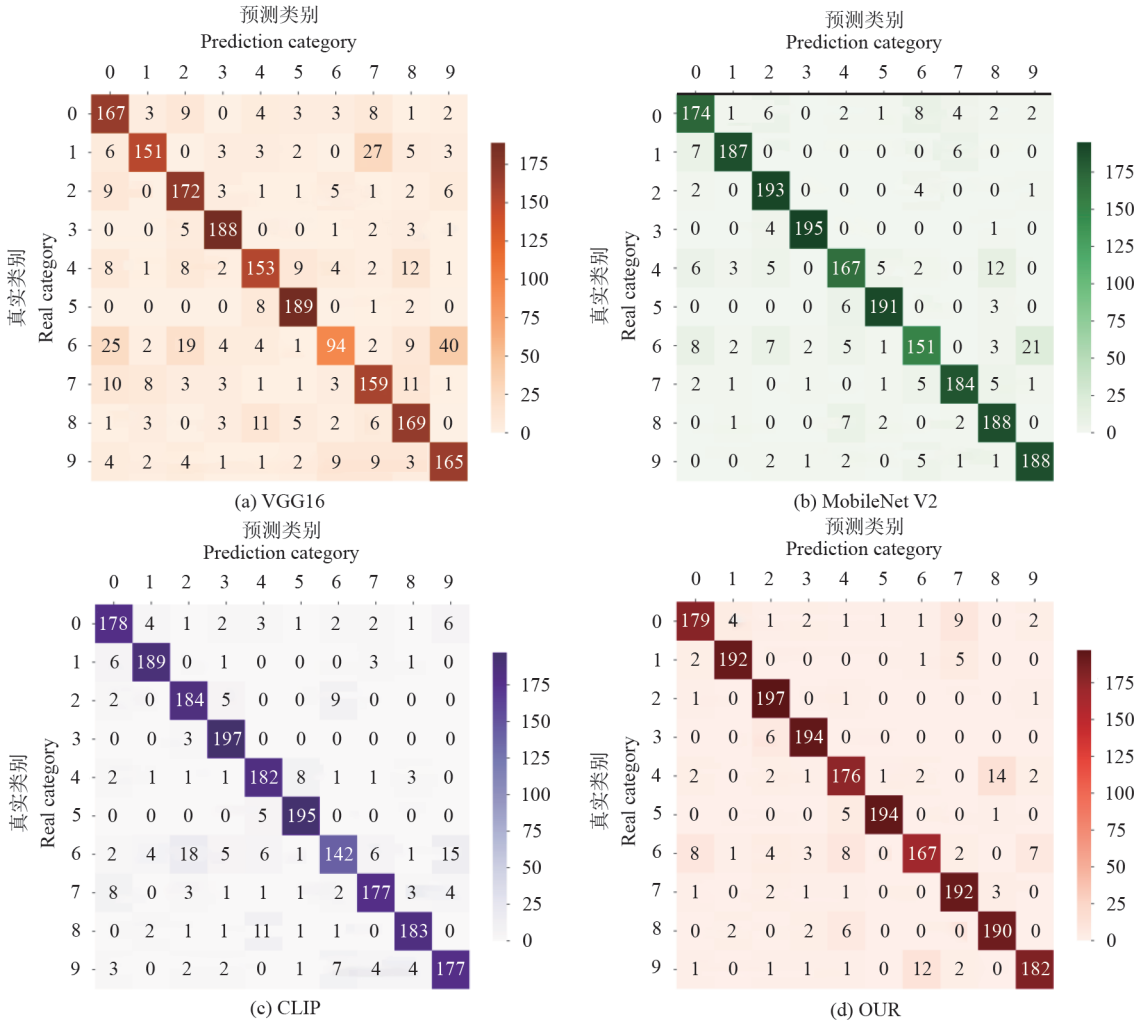

图 6 不同图像编码器分类结果的混淆矩阵

Fig. 6 Confusion matrix of different image encoder classification results

3.4.3 文本编码器对比　为了验证本方法在文本编码器架构选择上的有效性，通过对比不同的文本编码器，采用和图像编码器架构对比一样的实验设置。

表 4 展示了使用不同文本编码器的图像检索文本、文本检索图像的平均精度均值对比结果。在文本编码器对比方面，MiniLM 作为 Transformer 的轻量级改进版本，其具有效率高、参数量少的优势，但不利于对性能要求高的应用场景。

表 4 不同文本编码器的平均精度均值对比结果

Table 4 Comparison results of average precision means of different text encoders

模型 Model	i2t	t2i
MiniLM	0.893 9	0.889 0
OUR	0.928 4	0.950 3

图 7 展示了不同文本编码器的混淆矩阵，通过可视化的形式，直观展现使用各种不同文本编码器的预测类别分布，每个混淆矩阵中横轴是预测的类别，纵轴是真实的类别，矩阵中对角线位

置的数值越大、颜色越深,说明模型在该类上的分类准确率越高。通过混淆矩阵可以观察到,MiniLM整体表现良好,但个别类略有混淆,精确度较低,本方法各类别的准确率更高。

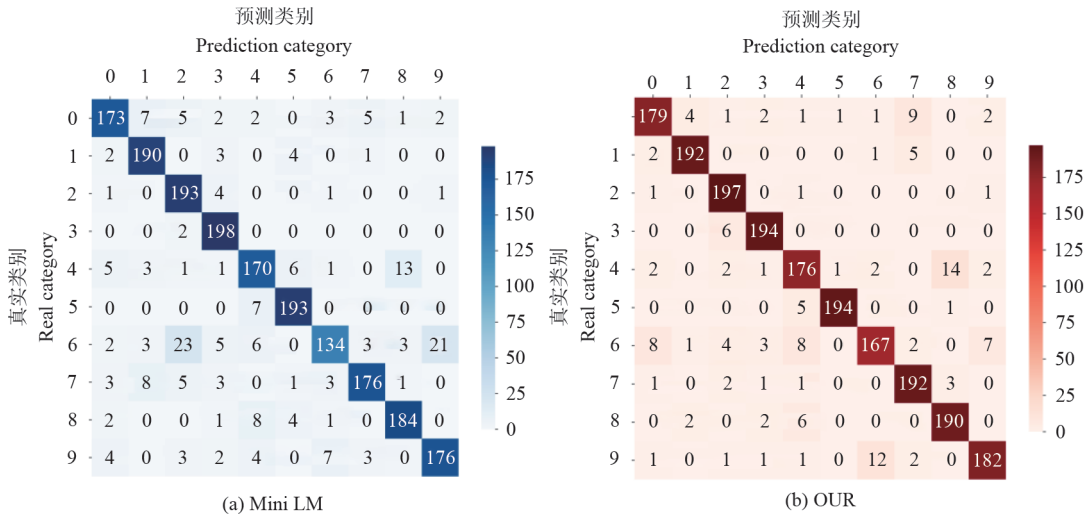

图7 不同文本编码器分类结果的混淆矩阵

Fig. 7 Confusion matrix of different text encoder classification results

3.5　消融实验结果分析

为了进一步验证本方法每个模块的有效性,本文分别对图像编码器、文本编码器和统一空间学习模块中的三元组损失和自适应分类损失进行消融实验。本节在自建数据集上进行图像-文本检索,表5展示了消融实验结果:

表5 消融实验

Table 5 Ablation experiments

视觉编码器 Visual encoder	文本编码器 Text encoder	三元组损失 Triplet loss	双曲空间 Hyperbolic space	自适应分类损失 Adaptive classification loss	i2t	t2i
	√	√	√	√	0.807 1	0.812 3
√		√	√	√	0.893 9	0.889 0
√	√		√	√	0.906 3	0.919 7
√	√	√		√	0.923 9	0.943 5
√	√	√	√		0.922 1	0.939 6
√	√	√	√	√	0.928 4	0.950 3

实验结果表明,将图像编码器替换为VGG16,检索性能显著下降,说明使用ResNet-50网络可以提取更多图像细粒度特征,从而精准地识别病虫害局部细节信息。通过将文本编码器替换为MiniLM,效果有所下降,说明Transformer编码器具有更深度的语义理解能力,适用于细粒度的文本判别。此外,三元组损失能充分借助标签显式语义信息,通过对比学习的方式对齐同类别间的特征信息,使得生成的低维特征更具判别性,更好地实现图像-文本检索。本研究通过将统一空间学习映射到汉明空间中,实现双曲空间映射的消融,可以观察到,精度出现了下降,说明双曲空间更有利于挖掘深层语义结构,其几何特性能够更好地增强检索任务的判别能力。在对自适应分类损失的消融实验中,本研究将其替换成普通交叉熵损失,实验结果显示检索性能有所下降,说明本研究设计的自适应分类损失通过自适应的学习策略,重点关注难以区分的类别样本,从而优化了模型的类别判别能力,进一步提高图像-文本检索的准确度。

为了进一步验证本方法在各类别上的性能表现,本研究基于自建番茄病害数据集进行了多种方法对比实验,各类别准确率对比结果如图8所示。实验结果表明,本方法在多数类别上的预测结果优于其他方法,并表现出更高的稳定性。值得注意的是,本方法与语义通道哈希(SCH)方

法使用相同的图像文本编码器，仅在统一空间学习模块不同，在个别类别上，本方法与 SCH 相比，具有更高的类别预测准确率，且预测结果波动更小。说明本方法设计的自适应分类损失函数在缓解类别不均衡问题方面具有积极效果。与 CLIP 相比，本方法使用 ResNet-50 网络，可以提取到更多的图像局部特征，有效捕捉病斑区域的细微差异。整体而言，本方法在不同类别上的准确率分布更加均衡，进一步验证了模型的泛化能力和鲁棒性。

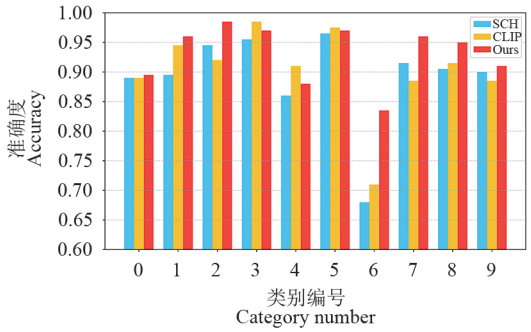

图 8 各图像-文本检索方法在不同类别上的准确率对比图

Fig. 8 Comparison of retrieval accuracy across categories for different image-text retrieval methods

3.6 可视化

图 9 和 10 展示了本方法在自建数据集上的检索结果可视化效果。图 9 展示了本方法通过图像检索文本信息的结果，本节选取了前 5 个检索结果，并使用绿色边框表示正确的检索结果，红色边框代表错误的检索结果。图 10 展示了本方法通过文本检索相应图像的可视化结果，进一步验证了本方法在跨模态检索任务中的有效性。实验结果表明，所提方法能够准确捕捉图像与文本之间的语义关联，实现了高质量的跨模态检索。

图 9 图像检索文本任务中的前五个文本检索结果示意图

Fig. 9 Top-5 retrieved texts by image queries

3.7 收敛性分析

为了研究模型的收敛性，对统一空间学习网络进行收敛性分析，图 11 展示了本方法在自建数据集上，随着轮次的增加，模型的损失值和 mAP 值变化曲线。从图中可以清晰地观察到，模型的损失逐渐下降并趋向于稳定，通过图像检

图 10 文本检索图像任务中的前五个图像检索结果示意图

Fig. 10 Top-5 retrieved images by text queries

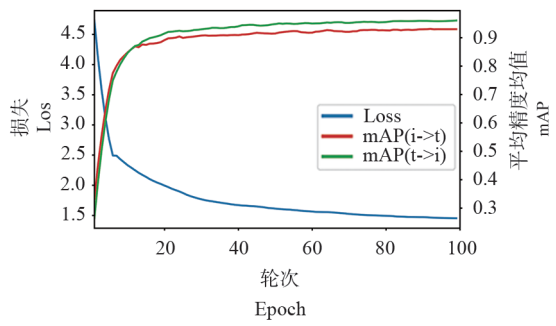

图 11 本方法训练过程的损失与 mAP 收敛曲线图

Fig. 11 Training loss and mAP convergence curves of the proposed method

索文本 mAP 值和通过文本检索图像 mAP 值逐渐上升并趋向于稳定。这一结果表明,模型在训练过程中实现了有效收敛,验证了所提方法的稳定性和可靠性。

3.8　超参数分析

为了研究超参数对模型性能的影响,本节在自建数据集上对超参数 λ_1 和 β 的敏感性进行了实验分析。图 12 展示了在不同超参数取值下,模型 mAP 的变化趋势。图中可以观察到,当 λ_1 取 2 和 β 取 0.5 时,模型的性能达到了最优,验证了所选超参数的有效性与合理性。

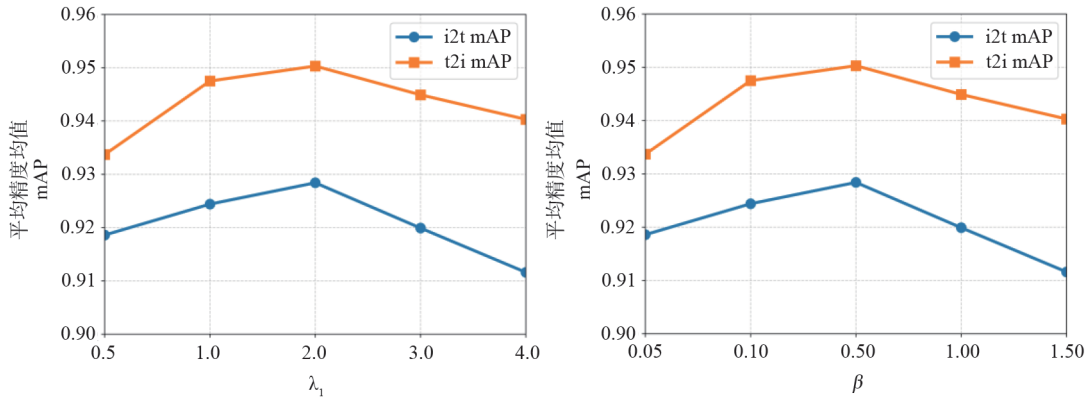

图 12 不同超参数设置对模型性能的影响图

Fig. 12 Performance variation under different hyperparameter settings

4　结论

(1)针对目前缺乏中文番茄病虫害图像文本数据集的问题,本文构建一个高质量番茄叶片病虫害图像-文本中文数据集,其中包含 10 种类别的番茄病虫害图像文本对。

(2)针对特征提取不足的问题,本文使用ResNet-50图像编码器和CLIP文本编码器,并且提出直接使用编码器的预训练权重不进行额外训练这种迁移学习的方法,实现局部特征的快速提取,有效捕捉病斑区域的细微差异。

(3)针对图像-文本特征难以进行对齐的问题,本文使用基于三元组的对比学习,使同类别的特征信息尽可能地接近,不同类别的特征信息尽可能地远离。针对类别不均衡问题,设计自适应分类损失函数,对难以学习的类别赋予更大的权重,提高其判别性。

最后,进行大量实验,通过定量分析与定性分析相结合,充分验证本方法的有效性。

参考文献

[1] 郭小清,范涛杰,舒 欣.基于改进Multi-Scale AlexNet的番茄叶部病害图像识别[J].农业工程学报,2019,35(13):162-169.

[2] 柴 帅,李壮举.基于迁移学习的番茄病虫害检测[J].计算机工程与设计,2019,40(06):1701-1705.

[3] 陶兆胜,石鑫宇,王 勇,等.基于改进YOLOv5s的番茄叶片病害检测方法[J].沈阳农业大学学报,2023,54(06):712-721.

[4] 刘立波,赵斐斐.融合注意力机制的枸杞虫害图文跨模态检索方法[J].农业机械学报,2022,53(02):299-308.

[5] 宋文韬,姜茹月,舒 欣.基于零样本学习的枸杞虫害识别[J].江苏农业学报,2024,40(02):320-330.

[6] Dai G, Fan J, Dewi C. ITF-WPI: Image and text based cross-modal feature fusion model for wolfberry pest recognition[J]. Computers and Electronics in Agriculture, 2023, 212: 108129.

[7] Xu J, Zhou H, Hu Y, et al. High-accuracy tomato leaf disease image-text retrieval method utilizing LAFANet[J]. Plants, 2024, 13(9): 1176.

[8] Feng X, Zhao C, Wang C, et al. A vegetable leaf disease identification model based on image-text cross-modal feature fusion[J]. Frontiers in Plant Science, 2022, 13: 918940.

[9] 杨 筝,程云志,常开心,等.基于细粒度特征增强交互网络的植物病虫害识别[J].河南大学学报(自然科学版),2024,54(06):722-729.

[10] 陈 磊,刘立波,王晓丽.2020年宁夏枸杞虫害图文跨模态检索数据集[J].中国科学数据(中英文网络版),2022,7(03):149-156.

[11] 聂 葳,叶成炜,杨家慧,等.基于CHINESE-CLIP跨模态图像文本检索研究[J].电子制作,2024,32(22):61-66.

[12] 张北辰,李 亮,查正军,等.基于跨模态对比学习的视觉问答主动学习方法[J].计算机学报,2022,45(08):1730-1745.

[13] 周 坤,徐黎明,郑伯川,等.自适应高效深度跨模态增量哈希检索算法[J].计算机工程与应用,2023,59(02):85-93.

[14] 邢嘉璐,刘建平,周国民,等.基于参数高效微调的跨模态枸杞虫害识别模型D-PAG[J].农业大数据学报,2024,6(04):509-521.

[15] Wei Y, Liu X, An D, et al. Citrus diseases and pests image-text retrieval based on multi-modal transformer [C]. 2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS). IEEE, 2023: 66-70.

[16] 黄 溪,王先兵,林 海,等.基于自集成视觉Transformer的图像检索[J].武汉大学学报(工学版),2024,57(12):1795-1802.

[17] 姜 飞,叶 炜,李兆星,等.基于特征交互的样本不均衡的玉米病害检测方法[J].华南农业大学学报,2025,46(03):399-406.

[18] 孙宝刚,何国斌.不均衡少标签样本下基于语义自动编码网络的高光谱图像分类[J].红外技术,2025,47(04):429-436.

[19] 郑宗生,霍志俊,高 萌,等.基于类中心优化辅助三元组损失的遥感图像检索[J].计算机工程,2025,51(05):305-313.

[20] 代佳洋,周 栋.基于多任务学习的跨语言信息检索方法研究[J].广西师范大学学报(自然科学版),2022,40(06):69-81.

[21] 尚皓玺,郭小燕,朱恒宇.基于迁移学习与GhostNet模型的农业害虫图像识别研究[J].软件导刊,2022,21(11):137-143.

[22] 王丽妍.基于迁移学习的智慧农业病虫害图像识别方法[J].农业工程技术,2024,44(02):120-121.

[23] 万 鹏,赵竣威,朱 明,等.基于改进Res Net50模型的大宗淡水鱼种类识别方法[J].农业工程学报,2021,37(12):159-168.

[24] Sun C, Latapie H, Liu G, et al. Deep normalized cross-modal hashing with bi-direction relation reasoning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 4941-4949.