

• 数据挖掘与循证医学 •

HerbVLM: 一种基于图文提示增强与视觉去混淆的中药材鉴别模型

叶 诚¹, 陈天雪², 于哲轩², 赵希曦³, 虞 立^{2*}, 金 伟^{1*}

1. 浙江中医药大学中医药人工智能学院, 浙江 杭州 310053

2. 浙江中医药大学基础医学院, 浙江 杭州 310053

3. 浙江中医药大学中医药科学院, 浙江 杭州 310053

摘要: **目的** 构建一种面向中药材鉴别的视觉-文本融合模型 HerbVLM, 提升中药材智能识别在相似类别区分与跨场景泛化中的性能。**方法** 基于视觉-语言预训练双编码器框架, 在文本侧引入中药知识增强提示模块 (traditional Chinese medicine knowledge-enhanced prompt module, TKPM), 在图像侧构建分层适配与去混淆模块 (hierarchical image adaptation and de-confusion module, HIA), 并结合动态融合与联合损失进行优化。在 Chinese-Medicine-163 开源数据集上开展对比与消融实验, 并在自采“新浙八味”外部数据集上进行跨域与小样本适配验证。**结果** 在测试集上, HerbVLM 的首位预测准确率 (top-1 accuracy, Top-1 Acc)、宏平均精确率 (macro precision, P_{macro})、宏平均召回率 (macro recall, R_{macro})、宏平均 F1 值 (macro F1-score, $F1_{macro}$) 分别达到 98.69%、98.70%、98.71% 和 98.70%, 优于残差网络 (residual network, ResNet)、上下文优化 (context optimization, CoOp)、条件上下文优化 (conditional context optimization, CoCoOp)、免训练 CLIP 适配器 (training-free CLIP-Adapter, Tip-Adapter)、多模态提示学习 (multi-modal prompt learning, MaPLe)、中文视觉-语言对比预训练模型 (contrastive vision-language pretraining in Chinese, Chinese-CLIP) 等对比方法。消融实验显示, 文本增强、图像适配、去混淆约束和动态融合均带来稳定增益。在自采浙八味数据集外部测试中, 已知类别灵芝、覆盆子、铁皮石斛、前胡识别率分别为 96.00%、100.00%、100.00% 和 100.00%; 未知类别在 1-shot、2-shot、5-shot 条件下平均准确率分别约为 84.00%、91.54% 和 98.20%, 显示出较好的小样本迁移能力。**结论** HerbVLM 通过图文知识多模态协同建模显著提升了中药材细粒度识别精度与跨域泛化性能, 可为中药材教学、质检辅助等场景提供方法支持。

关键词: 中药材鉴别; 视觉-文本融合; 多模态; 提示学习; 小样本学习; 跨域泛化**中图分类号:** TP18; R282.5 **文献标志码:** A **文章编号:** 0253-2670(2026)19-7722-17**DOI:** 10.7501/j.issn.0253-2670.2026.19.018

HerbVLM: A traditional Chinese medicinal materials identification model based on vision-text prompt enhancement and visual deconfusion

YE Cheng¹, CHEN Tianxue², YU Zhexuan², ZHAO Xixi³, YU Li², JIN Wei¹

1. School of Artificial Intelligence for Traditional Chinese Medicine, Zhejiang Chinese Medical University, Hangzhou 310053, China

2. School of Basic Medical Sciences, Zhejiang Chinese Medical University, Hangzhou 310053, China

3. Academy of Chinese Medical Sciences, Zhejiang Chinese Medical University, Hangzhou 310053, China

Abstract: **Objective** To construct HerbVLM, a vision-text fusion model for the identification of traditional Chinese medicinal materials, aiming to improve intelligent recognition performance in distinguishing visually similar categories and generalizing across scenarios. **Methods** Based on a vision-language pretrained dual-encoder framework, a traditional Chinese medicine knowledge-enhanced prompt module (TKPM) was introduced on the text side, while a hierarchical image adaptation and de-confusion module

收稿日期: 2026-06-05**基金项目:** 浙江省自然科学基金项目/优青项目 (LZYQ25H270001); 浙江省自然科学基金项目/探索一般 (LY24H270007); 浙江省中医药管理局中医药健康服务研究计划项目 (2026ZF41); 2026 年浙江省大学生科技创新活动计划 (新苗人才计划) 项目 (2026R410A003)**作者简介:** 叶 诚 (2005—), 本科生, 浙江省大学生新苗人才计划项目负责人, 主要从事图像算法研究。E-mail: yecheng_yyg@163.com***通信作者:** 虞 立, 高级实验师、高级工程师, 硕士生导师, 主要从事中医药信息学及其药理研究。E-mail: yuli9119@126.com

金 伟, 讲师, 主要从事深度学习、图像处理等人工智能算法研究。E-mail: jinwei@zcmu.edu.cn

(HIA) was constructed on the image side. Dynamic fusion and joint loss optimization were further incorporated. Comparative and ablation experiments were conducted on the open-source Chinese-Medicine-163 dataset, and cross-domain and few-shot adaptation validation was performed on a self-collected external dataset of the “new Zhebawei”. **Results** On the test set, HerbVLM achieved top-1 accuracy (Top-1 Acc), macro precision (P_{macro}), macro recall (R_{macro}), and macro F1-score ($F1_{\text{macro}}$) of 98.69%, 98.70%, 98.71%, and 98.70%, respectively, outperforming comparative methods including residual network (ResNet), context optimization (CoOp), conditional context optimization (CoCoOp), training-free CLIP-Adapter (Tip-Adapter), multi-modal prompt learning (MaPLe), and contrastive vision-language pretraining in Chinese (Chinese-CLIP). Ablation experiments showed that text enhancement, image adaptation, de-confusion constraints, and dynamic fusion all brought stable performance gains. In external testing on the self-collected Zhejiang Eight Flavors dataset, the recognition rates for known categories Lingzhi (*Ganoderma*), Fupenzi (*Rubi Fructus*), Tiepishi (*Dendrobii Officinalis Caulis*), and Qianhu (*Peucedani Radix*) were 96.00%, 100.00%, 100.00%, and 100.00%, respectively. For unknown categories, the average accuracies under 1-shot, 2-shot, and 5-shot settings were approximately 84.00%, 91.54%, and 98.20%, respectively, demonstrating favorable few-shot transfer capability. **Conclusion** HerbVLM significantly improves the fine-grained recognition accuracy and cross-domain generalization performance of traditional Chinese medicinal materials through multimodal collaborative modeling of image-text knowledge, and can provide methodological support for traditional Chinese medicinal material teaching, quality inspection assistance.

Key words: traditional Chinese medicinal materials identification; vision-text fusion; multimodality; prompt learning; few-shot learning; cross-domain generalization.

深度学习技术已广泛应用于中药材智能识别,并在饮片分类、药材图像识别及相似药材区分等任务中取得了长足进展^[1-3]。这些研究多依赖卷积神经网络、注意力机制或特征度量学习等视觉模型,通过自动提取药材的形态、颜色和纹理等图像特征,提高中药材识别的效率与准确率^[4]。然而,这类方法总体上仍属于单模态视觉识别范畴,模型判断主要依赖图像信息,对药材名称、性状描述、来源属性及功效特征等文本语义知识利用不足。当面对外观相似、类间差异细微或样本数量有限的中药材类别时,仅依赖视觉特征容易产生误判,模型的泛化能力与领域适应能力仍有进一步提升空间^[5-6]。这表明,中药材智能识别在本质上属于典型的细粒度识别问题,即类别间差异小而类内变化大。

近年来,多模态学习尤其是视觉-文本融合方法,在图像分类^[7]、细粒度识别^[8]、医学影像分析^[9-10]、智能问答与跨模态检索^[11]等领域取得了显著成效。依托视觉-语言预训练模型的跨模态表征能力,模型能够同时建模图像内容与文本语义信息,实现不同模态间的特征对齐与知识互补,因此在类别区分细微、样本规模有限及场景变化复杂等任务中表现出更强的识别性能与泛化能力。对于中药材智能鉴别而言,药材图像所呈现的形态、色泽与纹理等视觉线索与药材名称、性状描述、来源属性及功效特征等文本知识之间具有天然互补关系^[12-13]。将视觉信息与文本语义协同建模不仅契合中药材实际鉴别逻辑,也有助于缓解单模态视觉识别中混淆误判与

泛化不足的问题。基于此,本研究继承了 logits de confusion (LDC) 模型^[14]的多层视觉特征适配和对数几率(logits)去混淆的基本思想,新增中药文本模块增强、面向中药材任务的提示词(prompt)构造、样本级融合解释,构建基于视觉-语言预训练模型的多模态识别框架 HerbVLM (herb vision-language model),同时引入外部中药材验证方案,提升中药材细粒度识别的准确性、鲁棒性与跨场景适应能力。

1 材料和方法

1.1 数据集

1.1.1 图像数据集 为验证 HerbVLM 在中药材图像识别任务中的有效性,采用 Chinese-Medicine-163 开源数据集^[15]开展实验。该数据集包含 163 个中药材类别,共 266 767 张图像,其中训练集 256 767 张、测试集 10 000 张。数据覆盖根、茎、叶、花、果实及切片等多种药材形态,具有较强的类内变化和跨类相似性,能够较好反映真实场景下中药材细粒度识别的难点。数据集原始目录包含训练集与测试集。本研究保留原始测试集作为最终测试集,并在原始训练集内部按照类别分层随机划分出 20%作为验证集,其余 80%作为训练集。验证集和测试集保持完整,用于模型选择与最终性能评估。考虑到该数据集中药材种类众多,按形态多样性与类别覆盖从中选取 12 种类别作为展示图像(图 1),用于展示不同药材在形态、纹理和色泽上的典型差异。

为验证 HerbVLM 在真实场景下的泛化性能,



图 1 代表性中药材图像示例及样本数量

Fig. 1 Image examples and sample quantities of representative traditional Chinese medicinal materials (TCMM)

本研究进一步构建了自采集“新浙八味”中药材图像数据集(图 2),并将其作为外部测试集进行评估。该数据集共包含 8 类药材,分别为铁皮石斛、衢枳壳、乌药、三叶青、覆盆子、前胡、灵芝和西红花;每类对应 5 个独立采购批次,共 40 个采购批次。采购日期跨度为 2023 年 1 月至 2026 年 6 月,分别购自方回春堂、桐君堂、胡庆余堂、叶重德堂和张同泰等中药零售药店或其他流通渠道,产地覆盖浙江衢州、浙江台州、浙江金华、浙江丽水、浙江杭州、浙江温州、浙江建德、浙江淳安、浙江桐庐及进口来源等。各药材的批号和产地概况见表 1。药材经浙江中医药大学生命科学院张宇燕教授鉴定,铁皮石斛为兰科植物铁皮石斛 *Dendrobium officinale* Kimura et Migo 的干燥茎,衢枳壳为芸香科植物酸橙 *Citrus aurantium* L.及其栽培变种的干

燥未成熟果实,乌药为樟科植物乌药 *Lindera aggregata* (Sims) Kosterm.的干燥块根,三叶青为葡萄科植物三叶崖爬藤 *Tetrastigma hemsleyanum* Diels et Gilg 的干燥块根,覆盆子为蔷薇科植物华东覆盆子 *Rubus chingii* Hu 的干燥未成熟果实,前胡为伞形科植物白花前胡 *Peucedanum praeruptorum* Dunn 的干燥根,灵芝为多孔菌科真菌赤芝 *Ganoderma lucidum* (Leyss. ex Fr.) Karst.的子实体,西红花为鸢尾科植物番红花 *Crocus sativus* L.的干燥柱头,均符合《中国药典》2025 年版相关规范。

药材的图像采集设备为 HUAWEI nova 12 活力版手机,原始图像分辨率为 $3\ 072 \times 3\ 072$ 像素,并保留可交换图像文件格式(EXIF)基础信息。采集过程中以单份药材样本为对象,在统一设备条件下获取药材整体外观与局部纹理图像,同时允许样本

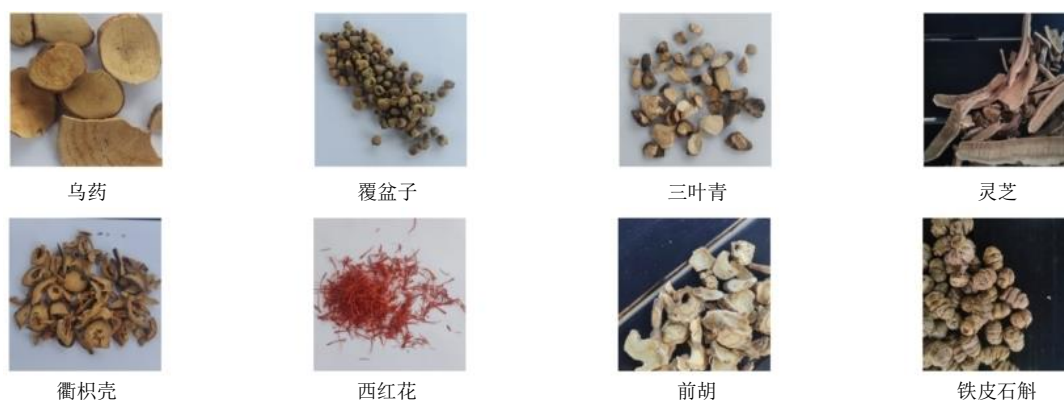


图 2 新浙八味图像示例

Fig. 2 Image examples of new ZhebaWei (eight genuine medicinal materials originated from Zhejiang)

表 1 新浙八味各药材的批号和产地信息

Table 1 Batch numbers and origin information of each medicinal materials among new Zhebwawei

药材名称	批号	产地	药材名称	批号	产地
铁皮石斛	wl23030601	浙江乐清	三叶青	wl24011201	浙江丽水
	wl25011503	浙江天台		wl24101102	浙江金华
	wl25032201	浙江温州		wl25051801	浙江丽水
	wl25111201	浙江建德		wl25121201	浙江台州
	wl26032302	浙江义乌		wl26051501	浙江衢州
衢枳壳	wl24011601	浙江衢州	覆盆子	wl24051101	浙江杭州
	wl24070501	浙江衢州		wl24121101	浙江临安
	wl25032201	浙江衢州		wl25040701	浙江衢州
	wl25110502	浙江衢州		wl25101001	浙江丽水
	wl26021101	浙江衢州		wl26032202	浙江金华
乌药	wl24071101	浙江天台	前胡	wl25012302	浙江杭州
	wl24120701	浙江天台		wl25041801	浙江淳安
	wl25022001	江西德清		wl25092601	安徽广德
	wl25052401	江西婺源		wl25110701	贵州毕节
	wl25120101	湖南衡阳		wl26021301	浙江桐庐
灵芝	wl23111201	浙江衢州	西红花	wl22031901	进口
	wl24030801	吉林长白山		wl23022201	进口
	wl24092701	浙江金华		wl24051101	浙江丽水
	wl25011903	浙江丽水		wl25030901	浙江衢州
	wl25071901	安徽大别山		wl26052901	浙江建德

摆放姿态、取景角度和拍摄距离在小范围内变化，以尽可能保留实际流通场景下中药材外观性状的自然差异。

外部数据实验包含 2 类设置，分别是已知类别直接迁移测试（不更新模型权重）和未知类别 few-shot 原型适配测试（1-shot、2-shot、5-shot），进行 10 次重复采样，统计均值与标准差，以评估模型在开放场景下的稳定迁移能力。由于该外部数据集在成像条件、背景分布及样本来源等方面与训练集存在明显差异，因此可作为跨域测试数据用于考察模型在新场景下的识别能力与泛化水平。

1.1.2 文本数据集 除图像数据外，进一步构建了面向中药材类别语义建模的类别级文本数据，用于支持视觉-文本联合识别。本研究采用类别级提示文本而非样本级文本，主要基于 2 方面考虑。其一，视觉语言分类领域的经典方法通常以类别名称构建文本提示，并通过图像特征与类别文本特征之间的相似度实现分类，对比语言-图像预训练模型（contrastive language-image pre-training, CLIP）^[16]、上下文优化（context optimization, CoOp）^[17]和条件上下文优化（conditional context optimization,

CoCoOp）^[18]均遵循这一基本范式。其二，已有研究表明，更丰富的类别描述性 prompts 有助于提升开放词表图像分类性能^[19]，因此本研究进一步在类别级提示中加入拼音、别名及形态描述，以增强中药材类别的语义可分性。需要说明的是，样本级文本并非不可行，但其通常需要额外的逐样本文本标注或自动生成机制；同时，图文数据中可能存在上下文与背景偏置，因此本研究在当前分类设定下优先采用更稳定的类别级文本原型。

具体而言，对于 Chinese-Medicine-163 开源数据集集中的 163 个药材类别，本研究为每个类别建立 1 条文本元信息记录，包含类别名称、拼音、英文别名以及形态学描述等字段。在此基础上，采用多模板提示策略对每个类别生成多种文本描述，包括类别名称模板、干燥药材模板、中文描述模板、拼音模板、英文别名模板以及形态模板等。最终，每个类别生成 7 条提示文本，共形成 1 141 条类别级提示语句。这些提示文本经 CLIP 文本编码器映射到共享语义空间，并聚合为类别文本原型，以用于与图像特征进行相似度匹配和类别判别。文本字段设置如表 2 所示。

表 2 文本字段设置
Table 2 Text field settings

字段名	示例	含义	作用
classname	人参切片	药材类别名称	作为类别语义中心
pinyin	ren shen qie pian	药材类别拼音	缓解中文标签在预训练语义空间中的稀疏问题
english_alias	Ren Shen Qie Pian	类别英文别名或拼音式英文写法	提供英文语义入口
morphology_en	slices	英文形态描述	强化药材外观结构语义
morphology_zh	切片	中文形态描述	补充中文领域知识表达

文本元信息如下所示。

提示文本

a photo of {classname}, a Chinese medicinal herb.
a photo of dried {classname}, a traditional Chinese medicine herb.
一张中药材{classname}的照片。
a photo of {pinyin}, a Chinese medicinal herb name in pinyin.
a photo of {english_alias}, a Chinese medicinal herb.
a close-up photo of dried {morphology_en} of {classname}.
中药材{classname}, 其形态为{morphology_zh}。

1.2 模型

1.2.1 HerbVLM 模型 HerbVLM 整体框架基于 CLIP 视觉-语言双编码器构建, 在保留预训练跨模态先验优势的基础上, 针对中药材细粒度识别任务中的 3 类核心问题进行改进: (1) 针对类别名称与鉴别知识利用不足的问题, 引入中药材知识增强提示模块, 以增强文本语义表征; (2) 针对外观相近类别的细节区分困难问题, 引入多层视觉特征建模与类别去混淆模块, 以增强局部形态与纹理判别能力; (3) 针对不同样本难度差异较大的问题, 引入自适应融合机制, 以动态平衡不同判别分支的贡献。整体框架如图 3 所示。

具体而言, 模型在文本层引入中药材知识增强提示模块 (traditional Chinese medicine knowledge-enhanced prompt module, TKPM), 对类别文本进行多模板扩展与领域先验注入, 将类别名称、拼音及形态描述联合编码为更具区分性的文本特征。在图像层面, 以 LDC 主干思想为基础, 在冻结主干编码器的条件下, 通过多层适配特征融合模块 (Herb-MAF) 提取和聚合层次化视觉信息, 以增强对药材纹理、切片形态及边缘结构的表达能力。为缓解中

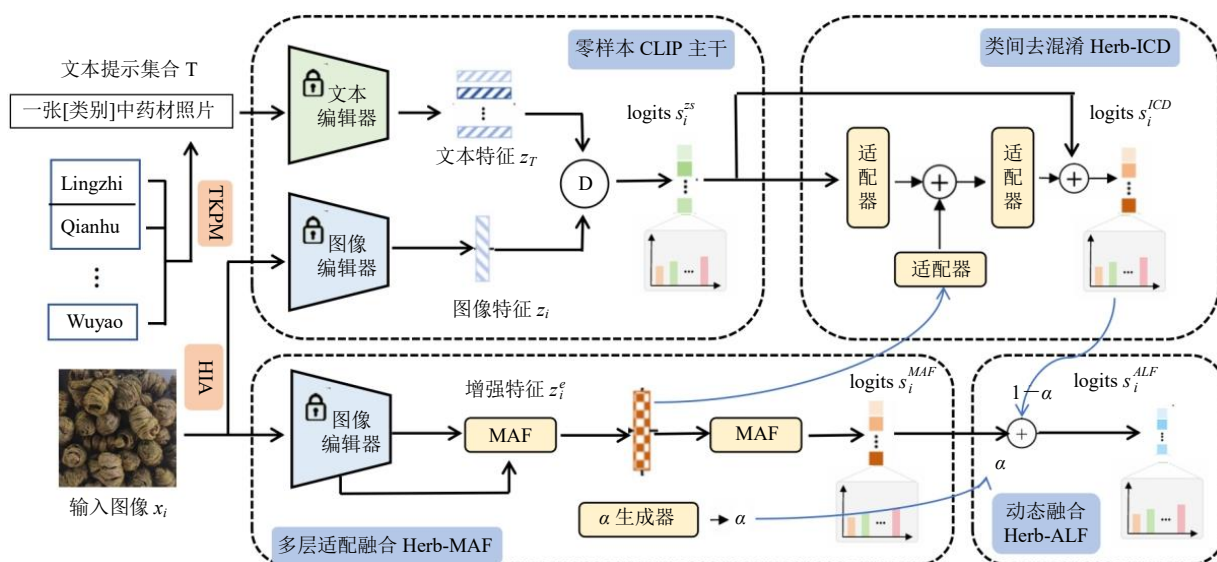


图 3 HerbVLM 整体框架结构

Fig. 3 Overall framework of HerbVLM

药材类别间外观近似导致的误判问题, 模型进一步在 logits 空间引入类别去混淆模块 (Herb-ICD), 对易混类别关系进行残差校正, 降低类别重叠区域的决策歧义。模型通过自适应 logits 融合模块 (Herb-ALF) 对不同分支输出进行动态加权, 综合 CLIP 先

验稳定性与任务分支判别性, 得到最终预测结果。

HerbVLM 的前向过程可分为 4 步。首先, 输入图像经 CLIP 图像编码器得到全局视觉表示, 类别文本经 TKPM 与文本编码器得到增强语义表示, 并通过点积匹配获得基础 logits。其次, Herb-MAF 从

多层视觉特征中构建任务相关表示，经轻量判别头输出补充分支 logits。再次，Herb-ICD 对类别间混淆模式进行学习式修正，强化近邻类别边界。最后，Herb-ALF 根据样本级置信信息自适应融合各分支 logits，输出最终分类结果。

与仅依赖固定提示或单一路径判别的方法相比，HerbVLM 同时在语义描述能力与类别边界建模能力上进行优化，能够更好适应中药材“类间差异细微、类内形态变化较大”的识别场景。为保证训

练稳定性，本研究采用联合损失对主分类目标与分支一致性进行协同约束，从而在保持预训练知识迁移能力的同时提升下游任务判别性能。

1.2.2 TKPM 模块（文本层） 为提升视觉-语言模型在中药材智能鉴别任务中的文本表征能力，本研究在 HerbVLM 模型文本侧设计了 TKPM。该模块按照“输入-提示词构造-分析与编码-提示权重学习-加权聚合输出”的过程，逐步生成具有判别性的类别文本原型。TKPM 模块具体流程如图 4 所示。

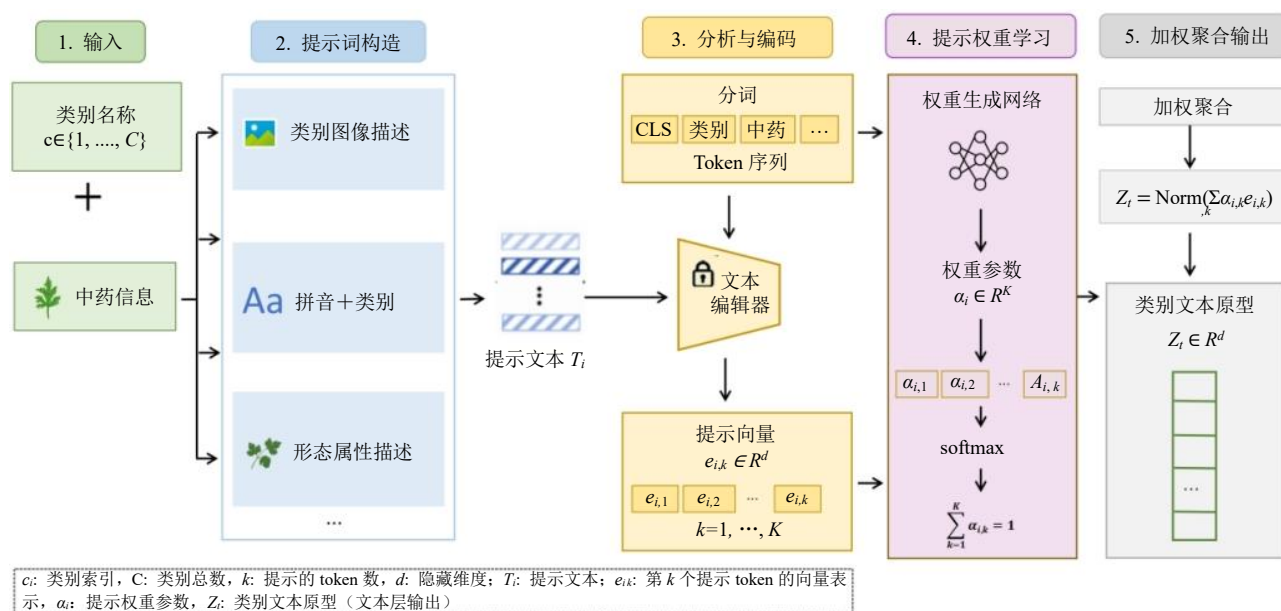


图 4 TKPM 模块流程

Fig. 4 TKPM module process

在输入部分，模块以类别名称作为基础输入，并引入与该类别相关的中药信息作为语义补充。类别名称提供类别身份标识，而中药信息则用于补充更细粒度的语义内容，如拼音、类别别名以及可用于区分类别的形态属性描述等。该设计旨在缓解单一类别名称语义不足的问题，使文本端能够更贴近中药材实际鉴别过程中“名称和性状”的认知方式。

在提示词构造部分，模块围绕同一类别构造多条提示文本 T_i 。这些提示文本并不完全相同，而是从不同语义角度对该类别进行描述，包括类别图像描述、拼音+类别、形态属性描述等形式。通过多模板提示构造，模型能够从不同语言视角表达同一药材类别，从而增强类别语义的丰富性与区分性。

在分析与编码部分，各提示文本经分词后形成 token 序列，并输入冻结的文本编码器进行编码，得到对应的提示向量 $e_{i,k}$ 。其中， i 表示类别索引， $k=$

$1, \dots, K$ 表示该类别下第 k 条提示文本， K 为提示条数。该过程实现了从离散文本描述到统一语义向量空间的映射，为后续跨模态对齐提供文本特征基础。

此外，考虑到不同提示文本对类别判别的贡献并不相同，本研究在图 4 第 4 部分引入提示权重学习机制。模块通过权重生成网络为同一类别下各提示向量学习对应权重参数 $\alpha_{i,k}$ ，并经 softmax 归一化，使其满足

$$\sum_{k=1}^K \alpha_{i,k} = 1$$

该过程使模型能够自动强调更具判别力的提示描述，降低冗余或弱区分提示对类别表示的干扰。模块依据学习得到的权重对同一类别下的多个提示向量进行加权聚合，得到类别文本原型 Z_i 。

$$z_{i,i} = \text{Norm}(\sum_k \alpha_{i,k} e_{i,k})$$

其中, $Norm(\cdot)$ 表示归一化操作, $e_{i,k}$ 为第 k 个提示 token 的向量表示。所有类别文本原型共同构成类别文本原型矩阵 $Z_t = [z_{t,1}, \dots, z_{t,n}]$, 并用于后续与图像特征进行相似度计算和视觉-文本对齐。

从功能上看, TKPM 为 HerbVLM 提供了稳定的“语义边界先验”。在中药材类别外观相近而语义描述差异较显著的场景下, 文本侧先验能够有效辅助后续图像分支进行类别判别与去混淆。在样本数量较少或成像条件波动较大的场景中, 文本信息还可作为相对稳定的知识补充, 提高模型的鲁棒性。

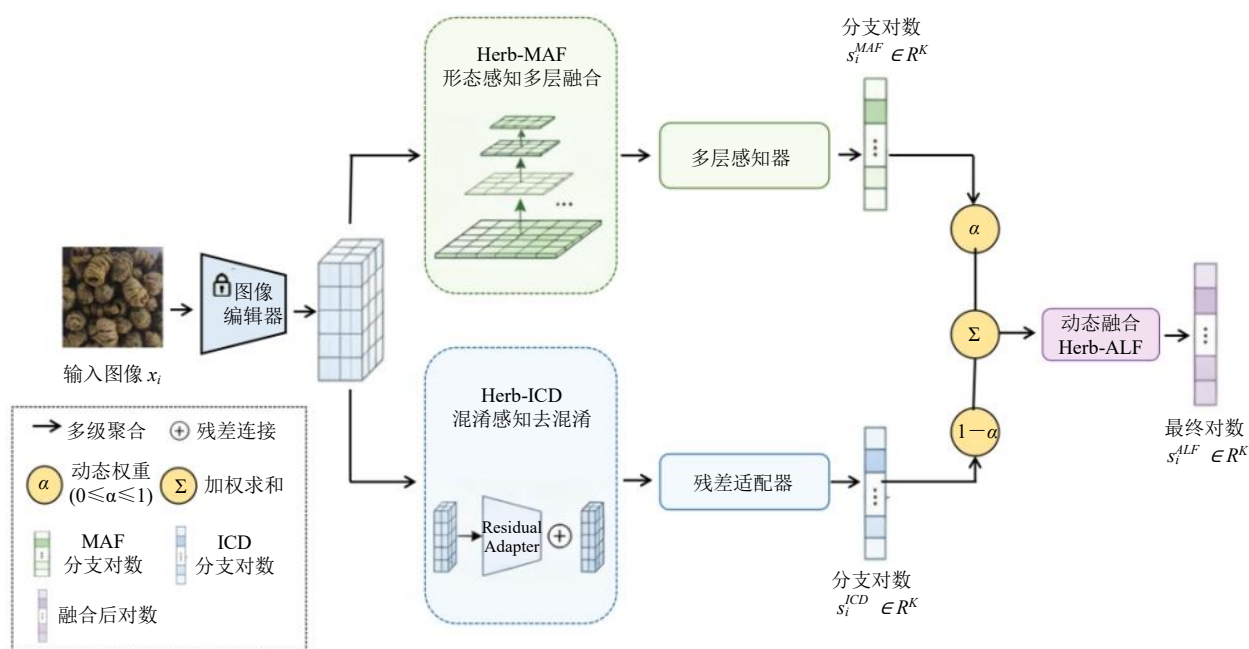


图 5 HIA 结构

Fig. 5 HIA structure

首先, Herb-MAF (形态感知多层融合) 从视觉编码器的多层特征中提取纹理与形态信息, 并通过轻量适配器和 MLP 进行融合, 得到任务相关 logits s_i^{MAF} 。该分支重点强化切片边缘、表面纹理、干燥结构等中药材关键视觉线索, 弥补仅依赖全局特征时的细节缺失。对于外观整体相近但局部纹理或断面差异较明显的类别, Herb-MAF 能够提供更细粒度的判别依据。其次, Herb-ICD (混淆感知去混淆) 在 logits 空间建模易混类别关系。针对中药材中存在的相似类别对, ICD 通过残差式适配结构对类别响应进行校正, 得到去混淆 logits s_i^{ICD} , 从而抑制相近类别间的误激活, 提高决策边界清晰度。该模块的引入主要用于解决“模型已经看到了特征, 但边界仍然不够清晰”的问题。Herb-ALF (自适应融合) 对分支输出进行动态加权融合。

1.2.3 分层适配与去混淆模块 (hierarchical image adaptation and de-confusion module, HIA) 模块 (图像层) 在图像端, HerbVLM 在 LDC 框架基础上构建 HIA, 由 Herb-MAF、Herb-ICD 和 Herb-ALF 3 个子模块组成。该模块主要针对中药材识别中的 2 个关键难点: 一是局部细节信息丰富但容易被全局特征淹没, 二是相似类别之间的决策边界模糊。因此, HIA 的目标不是简单增加模型复杂度, 而是提升细节建模能力与类别边界判别能力, 具体结构如图 5 所示。

该机制使模型能够根据样本难度在“细节判别能力”(MAF)与“边界去混淆能力”(ICD)之间自适应平衡。对于形态典型、纹理清晰的样本, 模型可更多依赖判别性较强的特征分支。对于近缘易混或成像条件较复杂的样本, 则可适当增加去混淆分支的贡献, 以提高识别稳定性。

综合而言, HIA 在视觉端形成了一套完整的技术链路, 先通过多层建模提取细粒度图像细节, 再对易混淆类别进行特征解耦区分, 最终以自适应机制完成多分支决策融合。与 TKPM 提供的文本语义先验互补后, HerbVLM 能够在中药材智能识别任务中同时提升语义区分能力与视觉判别能力。

1.2.4 损失函数 为同时保证中药材识别任务中的分类准确性、跨模态语义一致性以及易混类别区分能力, 本研究采用由分类损失、语义对齐损失和

混淆感知损失组成的联合优化目标。设输入图像为 x_i ，其真实类别标签为 $y_i \in \{1, \dots, C\}$ ，其中 C 表示类别总数。对于样本 x_i ，模型产生 4 类输出：零样本先验分支输出 s_i^{ZS} 、Herb-MAF 判别分支输出 s_i^{MAF} 、Herb-ICD 去混淆分支输出 s_i^{ICD} 、Herb-ALF 最终融合输出 s_i^{ALF} 。其中， s_i^{ZS} 、 s_i^{MAF} 、 s_i^{ICD} 和 s_i^{ALF} 分别承担跨模态语义先验、任务判别、易混校正和最终决策融合的功能。

设图像 x_i 经视觉编码器得到归一化图像特征 z_i ，类别文本原型矩阵为 $Z_t = [z_{t,1}, \dots, z_{t,n}]$ ， γ 表示可学习温度系数，则有：

$$s_i^{ZS} = \gamma z_{i,n} Z_t$$

在视觉端，经 Herb-MAF 模块获得任务相关融合特征 h_i ，在此基础上，判别分支输出定义为：

$$s_i^{MAF} = f_{mlp}(h_i)$$

其中 $f_{mlp}(\cdot)$ 是判别分支分类映射函数。随后，Herb-ICD 在保留先验分支语义信息的基础上，对类别响应进行残差式校正，得到去混淆分支输出。

$$s_i^{ICD} = s_i^{ZS} + g(s_i^{ZS}, h_i)$$

其中， $g(\cdot)$ 表示去混淆残差校正函数。最终，Herb-ALF 根据样本特征生成动态融合系数 $\alpha_i \in [0, 1]$ ，并对判别分支与去混淆分支进行加权融合。

$$s_i^{ALF} = \alpha_i s_i^{MAF} + (1 - \alpha_i) s_i^{ICD}$$

为保证各分支及最终输出均具备稳定的类别判别能力，本研究对 Herb-MAF、Herb-ICD 和 Herb-ALF 分支分别施加交叉熵监督，其定义如下。

$$L_{ce}^{MAF} = CE(s_i^{MAF}, y_i)$$

$$L_{ce}^{ICD} = CE(s_i^{ICD}, y_i)$$

$$L_{ce}^{ALF} = CE(s_i^{ALF}, y_i)$$

$CE(\cdot, \cdot)$ 表示交叉熵损失函数

为避免领域判别分支在训练过程中偏离视觉-文本预训练模型原有的跨模态语义结构，本研究进一步引入语义对齐损失，对 Herb-MAF 和 Herb-ICD 分支分别施加与零样本先验分支的一致性约束。具体定义为

$$L_{align}^{MAF} = \|s_i^{MAF} - s_i^{ZS}\|_1$$

$$L_{align}^{ICD} = \|s_i^{ICD} - s_i^{ZS}\|_1$$

L_{align}^{MAF} 为判别分支与先验分支之间的语义对齐损失， L_{align}^{ICD} 为去混淆分支与先验分支之间的语义对齐损失

由于中药材识别属于典型细粒度分类任务，不同类别之间常存在较强外观相似性，仅依赖标准分类损失难以充分拉开易混类别间的决策边界。为此，本研究进一步引入混淆感知损失 (confusion-aware

loss, CA-Loss)，对最易混淆的负类施加间隔约束。

设样本 x_i 的真实类别在最终融合输出中的得分为 s_i^+ ，从零样本先验分支 s_i^{ZS} 中筛选得到的易混负类集合为 μ_i ，其中第 j 个易混负类在最终融合输出中的得分记为 s_{ij}^- ，其对应权重为 ω_{ij} ，间隔超参数为 $m > 0$ ，则混淆感知损失定义为

$$L_{CA} = \sum_{j \in \mu_i} \omega_{ij} \max(0, m - (s_i^+ - s_{ij}^-))$$

最终训练目标定义为：

$$L = \lambda_1 L_{ce}^{MAF} + \lambda_2 L_{ce}^{ICD} + \lambda_3 L_{ce}^{ALF} + \lambda_4 L_{align}^{MAF} + \lambda_5 L_{align}^{ICD} + \lambda_6 L_{CA}$$

1.3 评估指标

为全面验证本研究提出的 HerbVLM 模型在中药材细粒度识别任务中的有效性，选取典型视觉-语言分类方法作为对比，包括：CLIP (Zero-shot) [16]、CoOp [17]、CoCoOp [18]、CLIP 适配器 (Clip-Adapter) [20]、免训练 CLIP 适配器 (training-free CLIP-Adapter, Tip-Adapter) [21]、多模态提示学习 (multi-modal prompt learning, MaPLe) [22]、LDC [14]、开源对比语言图像预训练模型 (open contrastive language-image pre-training model, OpenClip) [23]、中文视觉-语言对比预训练模型 (contrastive vision-language pretraining in Chinese, Chinese-CLIP) [24] 以及本文方法 HerbVLM。所有模型均在统一数据集、相同数据划分、相同骨干网络设置与统一硬件环境下进行训练与测试，以保证对比实验结果的公平性与可靠性。

针对分类性能评估，采用首位预测准确率 (top-1 accuracy, Top-1 Acc)、精确率 (precision, P)、召回率 (recall, R) 和 F1 分数 (F1-score, F1) 作为主要指标，并报告宏平均 (macro) 结果。共设有 C 个类别，测试样本总数为 N ，则总体准确率定义为

$$Acc = \frac{1}{N} \sum_{i=1}^N 1(\hat{y}_i = y_i)$$

$\hat{y}_i$ 和 y_i 分别表示第 i 个样本的预测标签与真实标签

对于第 c 类，精确率 (P_c) 与召回率 (R_c) 定义为

$$P_c = \frac{TP_c}{TP_c + FP_c}$$

$$R_c = \frac{TP_c}{TP_c + FN_c}$$

其中， TP_c 、 FP_c 、 FN_c 分别表示第 c 类的真正例、假正例和假负例。进一步，F1-score ($F1_c$) 定义为

$$F1_c = \frac{2 \cdot P_c R_c}{P_c + R_c}$$

为避免类别不平衡对结果造成偏置,本研究采用宏平均指标 (P_{macro} 、 R_{macro} 、 $F1_{\text{macro}}$),公式如下所示。

$$P_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C P_c$$

$$R_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C R_c$$

$$F1_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C F1_c$$

除总体准确率外,特别关注 P_{macro} 、 R_{macro} 和 $F1_{\text{macro}}$ 的变化情况,以判断模型性能提升是否仅集中于少数优势类别,还是在多类别层面具有较好的均衡性。由于中药材识别任务中存在较多细粒度相似类别,宏平均指标更能反映模型在整体类别层面的识别质量。

1.4 训练设置

本研究在 Linux (Ubuntu 20.04.5 LTS) 环境下完成,硬件平台为 NVIDIA RTX A5000 GPU (24

GB 显存) 双卡配置,实验采用单卡运行。软件配置包括 Python 3.9.23 与 PyTorch 2.0.1 (2.0.1+cu117, CUDA Runtime 11.7 驱动侧 CUDA Version 12.4)。实验以 CLIP 模型为核心,在 Chinese-Medicine-163 数据集上开展中药材识别任务;输入分辨率统一为 224×224 ,主实验训练轮数设为 20, Batch Size 设为 64。优化器采用 AdamW,并配合余弦退火学习率调度。其余超参数沿用当前 CLIP 实现的默认配置。

2 结果与分析

2.1 模型训练结果分析

本研究对 HerbVLM 进行了 20 轮训练,并采用 Top-1 Acc、 P_{macro} 、 R_{macro} 和 $F1_{\text{macro}}$ 作为主要评估指标。结果表明模型具备良好的收敛性与稳定性,如图 6 所示。训练前期主指标提升较快,后期逐渐趋于平稳;训练损失持续下降,说明联合优化目标能够有效引导模型学习。整体来看,本研究所构建的多模块协同框架未出现明显训练震荡,能够在验证集上获得稳定的性能提升。

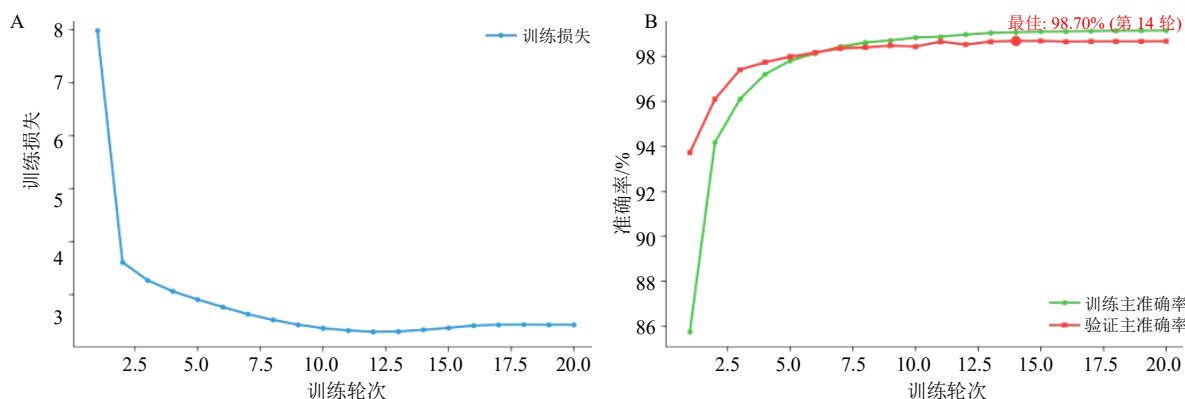


图 6 HerbVLM 的训练损失 (A) 和主指标变化曲线 (B)

Fig. 6 Training loss (A) and main indicator change curve (B) of HerbVLM

如图 7 所示,模型在验证集上的 P_{macro} 、 R_{macro} 和 $F1_{\text{macro}}$ 随训练推进持续提升,并在后期保持稳定,说明模型不仅总体识别精度较高,而且在类别层面具有较好的均衡性。该结果表明,模型性能提升并非仅依赖少数易识别类别,而是在多类别中均表现出较稳定的识别能力。

如图 8 所示,从不同输出分支的验证结果看,ZS (CLIP Zero-shot) 先验分支的准确率始终处于较低水平,说明通用视觉-语言零样本先验在中药材细粒度识别任务中存在明显领域偏移;相比之下,Herb-MAF 分支和 Herb-ICD 分支在训练过程中均表现出更强的任务适配能力,而 Herb-ALF 输出进

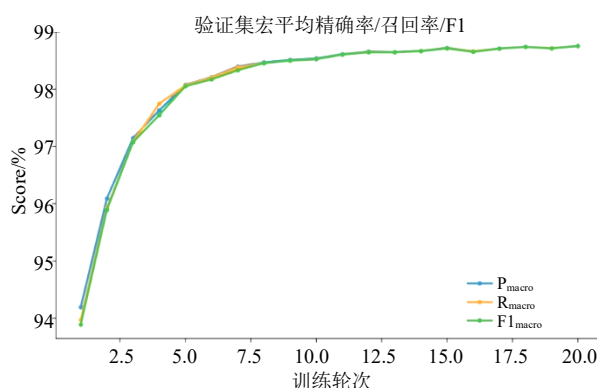
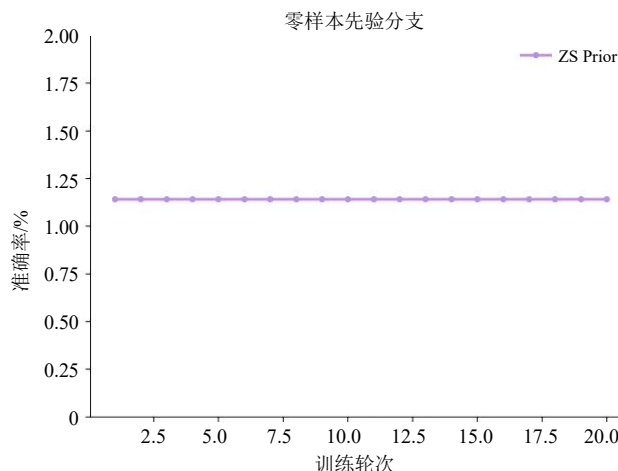


图 7 验证集宏平均精确率、召回率、F1 的过程曲线

Fig. 7 Process curves of P_{macro} 、 R_{macro} 、 $F1_{\text{macro}}$ on validation set

一步取得了稳定且最优的综合表现。该结果说明，仅依赖通用零样本语义对齐难以满足中药材细粒度识别需求，而面向领域任务的视觉适配、去混淆校正与动态融合设计具有必要性。

如图 9 所示，各损失项在训练过程中总体呈下降趋势，说明本研究构建的联合优化目标能够有效促进模型收敛。其中，分类损失分别作用于 Herb-



MAF、Herb-ICD 和 Herb-ALF 输出分支，而对齐项用于约束任务分支与 ZS 先验分支之间的表示一致性；CA-Loss 在训练后期逐步减小，说明模型对易混类别的判别边界不断得到优化。结合整体训练损失的持续下降与后期稳定，可以认为本研究所设计的优化目标在提升总体分类性能和强化相似类别区分能力 2 方面均具有积极作用。

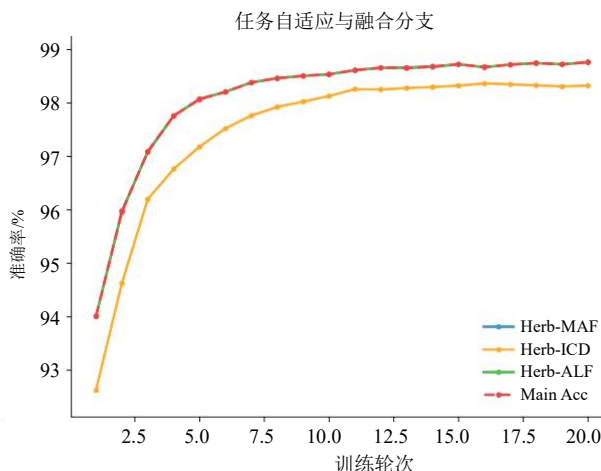


图 8 不同分支验证集准确率对比

Fig. 8 Comparison of accuracy of validation set across different branches

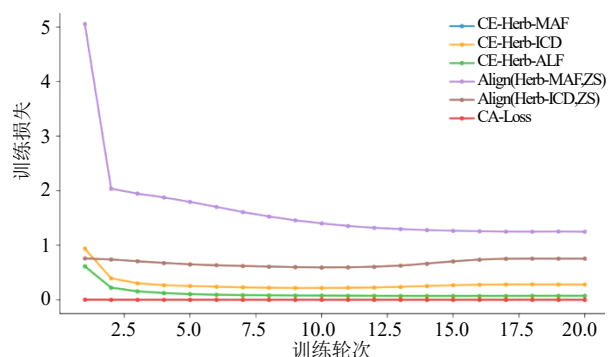


图 9 HerbVLM 各损失项收敛过程

Fig. 9 Convergence process of each loss term of HerbVLM

从训练过程可以进一步看出，本研究所提出的“文本提示增强、视觉端自适应学习、动态融合”设计具有明确的任务针对性。首先，原始 CLIP 先验分支的 $F1_{macro}$ 始终处于极低水平，说明通用视觉-文本对齐能力难以直接适用于细粒度中药材识别任务，单纯依赖原始类别文本与预训练先验难以形成有效判别。其次，在引入领域增强文本原型和任务相关视觉适配后，模型验证集 Top-1 Acc 与 $F1_{macro}$ 随训练逐步提升并最终稳定在较高水平，表明模型能够在中药材语义知识约束和视觉细粒度特征建

模的共同作用下学习到更具判别性的类别表示。进一步从不同分支的性能变化可见，任务适配分支和去混淆分支优于原始零样本先验分支，而融合分支在训练后期保持了稳定且接近最优的综合表现，说明固定的单一路径判别难以兼顾先验稳定性与任务判别性，采用样本级动态融合是合理且必要的。需要说明的是，此次训练过程结果主要从优化与收敛角度对上述设计提供动态证据，而各模块的独立贡献将在后文消融实验中进一步验证。

2.2 测试集识别结果

如表 3 所示，进一步从各类别识别结果来看，163 个类别中有 116 个类别的识别准确率达到 99.0% 及以上，且仅有 4 个类别低于 95.0%，表明本方法对大多数中药材类别均具有较好的识别稳定性和泛化性能。对于三七、丹参、人参、决明子、火麻仁、灵芝、石斛、穿心莲、金钱草、黄精和龙骨等类别，模型均取得了 100% 的识别准确率，说明当药材类别具有较为稳定的外观形态或较明显的判别特征时，HerbVLM 能够有效完成准确识别。

与此同时，部分外观相近或细粒度差异较小的类别仍存在一定识别困难。例如，仙鹤草、枳实、天冬、板蓝根等类别的识别准确率相对略低，但整

表 3 HerbVLM 模型对每种中药材的识别准确率
Table 3 Recognition accuracy of HerbVLM model for each type of TCMMs

中药材	准确率/%	中药材	准确率/%	中药材	准确率/%	中药材	准确率/%	中药材	准确率/%
三七	100.0	大青叶	97.9	火麻仁	100.0	红花	100.0	贯众	100.0
丹参	100.0	天冬	95.5	灵芝	100.0	红蔻	100.0	赤石脂	100.0
乌梅	100.0	天南星	96.9	炮姜	100.0	细辛	100.0	赤芍	100.0
五加皮	98.8	天葵子	100.0	牛膝	96.8	络石藤	100.0	路路通	100.0
五味子	100.0	天麻块	100.0	牡丹皮	100.0	续断	98.3	辛夷	100.0
五灵脂	98.1	天麻片	98.4	牡蛎	96.6	罗汉果	100.0	远志	96.3
陈皮	100.0	女贞子	100.0	玉竹条	100.0	羌活	98.6	连翘	100.0
青蒿	100.0	姜黄	100.0	玉竹片	100.0	肉桂	98.3	通草	100.0
仙鹤草	94.5	射干	100.0	珍珠母	100.0	肉苁蓉根	100.0	郁金	100.0
何首乌	100.0	小茴香	100.0	甘草	98.7	肉苁蓉片	100.0	酸枣仁	100.0
佛手	100.0	山楂	100.0	白头翁	98.2	肉豆蔻	100.0	野菊花	100.0
佩兰	100.0	山茱萸	98.3	白扁豆	100.0	艾叶	100.0	金钱草	100.0
侧柏叶	100.0	山药	98.4	白术	100.0	苍术	98.3	金银花	100.0
僵蚕	98.6	巴戟天	100.0	白矾	100.0	苦参	100.0	钩藤	98.2
党参	100.0	干姜	98.2	白芍	100.0	茯苓	98.6	防风	98.2
全蝎	100.0	木香	100.0	麦冬	100.0	茵陈	98.4	阿胶	100.0
决明子	100.0	杜仲	100.0	白茅根	100.0	荆芥	100.0	附子	97.0
刺楸	100.0	板蓝根	95.0	白蔻	100.0	草寇	100.0	人参	100.0
前胡	96.2	枳壳条	100.0	百合	96.4	草果	100.0	人参切片	100.0
香附	100.0	枳壳片	100.0	百部	98.5	荔枝核	100.0	首乌藤块	100.0
鳖甲	100.0	枳实	94.6	益母草	100.0	莱菔子	98.2	首乌藤片	97.3
升麻	100.0	枸杞子	100.0	知母	97.0	莲子心	100.0	北沙参块	100.0
半夏	98.3	柏子仁	100.0	石斛	100.0	菟丝子	98.2	北沙参条	100.0
厚朴	98.5	柴胡	96.6	石膏	100.0	葛根	100.0	鸡内金	95.5
合欢皮	100.0	桃仁	100.0	石菖蒲	100.0	蒲公英	100.0	鸡血藤	95.7
土鳖虫	100.0	桑椹	100.0	砂仁	100.0	蒲黄	100.0	白花蛇舌草	100.0
地榆	96.6	桑螵蛸	100.0	神曲	100.0	薏苡仁	97.1	麦芽	60.0
地骨皮	100.0	槐花	100.0	穿山甲	100.0	虎杖	96.7	水红花子	100.0
地龙	100.0	水牛角	100.0	穿心莲	100.0	虫草	100.0	紫花地丁	100.0
墨旱莲	100.0	黄柏	100.0	竹茹	100.0	蛇床子	98.5	黄芩	100.0
夏枯草	100.0	沉香	100.0	黄精	100.0	蝉蜕	100.0	龙骨	100.0
大腹皮	100.0	泽兰	100.0	紫草	100.0	覆盆子	98.0		
大血藤	96.7	浙贝母	100.0	紫菀	100.0	谷芽	49.4		

体仍保持在 95%左右，反映出模型在复杂类别上的性能虽有波动，但总体仍处于较高水平。谷芽和麦芽 2 类药材识别效果明显偏低，其中谷芽的识别准确率仅为 49.4%、麦芽为 60.0%，为所有类别中识别难度最大的 2 类。预测结果显示，这 2 类错误主要表现为彼此互判，其中谷芽误判为麦芽 20 例，麦芽误判为谷芽 39 例，提示模型虽已将其定位为近邻类别，但对二者细粒度边界的刻画仍不充分。其原因可能与 2 类药材在颗粒形态、色泽及发芽状态

等方面高度相似有关；同时，当前类别级文本提示对二者的性状区分信息仍较有限，导致视觉与文本两端均难以形成稳定的判别约束。该结果说明，针对高度相似药材类别的识别仍是后续优化的重要方向。

2.3 对比试验

为验证 HerbVLM 在中药材识别任务中的有效性，本研究将其与 ResNet50^[25]、ViT-B/16^[26]、CLIP(Zero-shot)^[16]、CoOp^[17]、CoCoOp^[18]、Clip-

Adapter^[20]、Tip-Adapter^[21]、MaPLe^[22]、OpenClip^[23]、Chinese-CLIP^[24]及 LDC^[14]进行对比。评价指标采用 Top-1 Acc、P_{macro}、R_{macro} 和 F1_{macro}，结果见表 4 所示。

表 4 不同模型对比实验

Table 4 Comparative experiments of different models				
模型	Top-1 Acc/%	P _{macro} /%	R _{macro} /%	F1 _{macro} /%
CLIP (Zero-shot)	1.22	1.35	1.21	0.81
CoOp	90.13	90.35	89.60	87.51
CoCoOp	92.03	90.39	90.77	89.04
ResNet50	92.37	91.52	91.01	90.33
Clip-Adapter	92.40	91.77	91.03	90.58
ViT-B/16	92.53	91.79	92.32	92.01
Tip-Adapter	92.61	92.07	92.50	92.01
MaPLe	93.80	93.13	93.32	92.07
LDC	93.83	93.30	93.01	92.80
OpenClip	94.49	94.03	94.12	94.55
Chinese-CLIP	95.63	94.85	94.92	94.61
HerbVLM(ours)	98.69	98.70	98.71	98.70

从现有实测结果可见，各类引入提示学习或适配机制的方法均能在不同程度上改善性能，表明针对领域任务进行有监督适配具有必要性。进一步比较可见，HerbVLM 在 Chinese-Medicine-163 数据集上 Top-1 Acc、P_{macro}、R_{macro} 和 F1_{macro} 分别达到 98.69%、98.70%、98.71%和 98.70%，整体优于其他对比方法，说明本研究提出的文本知识增强、图像去混淆与动态融合策略能够有效提升中药材细粒度识别性能。与 Chinese-CLIP、OpenClip 和 MaPLe 等视觉-语言

方法相比，HerbVLM 的提升并不只体现为整体准确率升高，更体现为宏平均指标的同步改善，说明该方法在多数类别层面均具备较好的识别一致性。

与 LDC 的比较更能体现本方法的实际增益。LDC 作为强基线，已经通过参数高效适配改善了视觉-语言模型在领域任务中的表现；而 HerbVLM 在此基础上进一步引入文本知识增强、类别去混淆与动态融合机制，说明性能提升并非仅来自“更多参数”或“更复杂结构”，而是来自针对中药材识别难点进行的有针对性设计。也就是说，HerbVLM 的改进路径更符合中药材智能鉴别任务的实际需求。

整体来看，HerbVLM 在保持高准确率的同时，宏平均指标更均衡，说明模型不仅在整体分类上有效，也在类别间保持了较好的识别一致性。该结果验证了本研究在文本端（中药材提示增强）与图像端（任务适配与融合）联合优化的必要性与有效性。

2.4 消融实验

为验证本文各改进模块对中药材识别性能的独立贡献及协同作用，在相同训练、验证和测试划分下开展消融实验。各模型均先基于验证集选择最优权重，再在测试集上报告 Top-1 Acc、P_{macro}、R_{macro} 和 F1_{macro}，结果见表 5。需要说明的是，表 5 中的 LDC 为原始基线模型，已包含其原始版本的 MAF、ICD 和 ALF；其中 Herb-MAF、Herb-ICD 和 Herb-ALF 表示是否采用本研究提出的改进版本，×并不表示原始 LDC 中不存在对应模块。

由表 5 可见，在 LDC 基础上，无论引入 TKPM，

表 5 HerbVLM 模型消融实验结果

Table 5 Ablation experiment results of HerbVLM model									
模型	TKPM	Herb-MAF	Herb-ICD	Herb-ALF	CA-Loss	Top-1 Acc/%	P _{macro} /%	R _{macro} /%	F1 _{macro} /%
LDC	×	×	×	×	×	93.83	93.30	93.01	92.80
LDC+TKPM	√	×	×	×	×	94.50	94.92	94.70	94.39
LDC+Herb-MAF	×	√	×	×	×	94.33	94.07	94.82	94.16
LDC+Herb-ICD	×	×	√	×	×	94.52	94.15	94.03	94.29
LDC+Herb-ALF	×	×	×	√	×	94.53	94.30	94.01	94.80
LDC+Herb-MAF+ Herb-ICD+Herb- ALF	×	√	√	√	×	95.33	95.50	95.81	94.67
LDC + TKPM + Herb-MAF+Herb- ICD+Herb-ALF	√	√	√	√	×	97.19	97.20	97.20	97.16
HerbVLM	√	√	√	√	√	98.69	98.70	98.71	98.70

×表示该列对应的模块/损失函数未使用（未加入该组实验）；√表示该列对应的模块/损失函数使用了（加入该组实验）。
× means the module/loss function corresponding to this column is not used (not included in this group of experiments); √ means the module/loss function corresponding to this column is used (included in this group of experiments).

还是分别采用 Herb-MAF、Herb-ICD 和 Herb-ALF，模型性能均较基线有所提升，说明文本提示增强与视觉端改进均能够有效改善中药材细粒度识别效果。其中，TKPM 的引入说明类别级文本知识能够补充中药材名称与性状语义，增强类别表征的可分性；Herb-MAF、Herb-ICD 和 Herb-ALF 的增益则表明，针对局部纹理建模、相似类别去混淆以及分支自适应融合的优化，对于提升中药材识别性能均具有实际意义。进一步地，多个视觉端改进模块联合后性能继续提高，在此基础上叠加 TKPM 后取得更明显增益，说明文本语义增强与视觉判别优化之间具有较好的互补性。最终，加入 CA-Loss 后模型获得最优结果，表明该损失约束有助于增强多分支协同学习，进一步提升整体识别性能与类别判别一致性。

综上，消融实验表明 HerbVLM 的各组成模块均对性能提升具有独立价值，且多模块联合可形成稳定的协同增益。其中，图像层适配与动态融合是提升性能上限的关键因素，而文本增强与去混淆约束进一步优化了类别级别的识别均衡性与鲁棒性。这一结果与本研究假设基本一致，即中药材智能鉴别的性能提升需要同时依赖语义知识补充、视觉细节建模和相似类别边界优化。

2.5 自采浙八味数据集上的外部测试与泛化分析

为评估 HerbVLM 模型在外部数据上的泛化能力，本研究在不改变模型权重的条件下，直接将训练好的模型迁移到浙八味数据集上进行测试。具体地，首先将外部数据划分为基础已知类别（预训练集上已训练过的类别）与新增类别 2 部分，其中基础类别用于检验模型的跨数据来源迁移能力，新增类别用于模拟实际应用中新增药材类别的快速接入场景。对于新增类别，本研究采用图像级随机采样策略构建支持（support）集与查询（query）集：在每个新增类别中随机选取 K 张图像作为 support 样本，其余图像作为该类别的 query 样本。因此，1-shot、2-shot 和 5-shot 分别表示每个新增类别提供 1 张、2 张和 5 张标注图像用于类别原型构建。在评估阶段，query 集不仅包含新增类别中未参与 support 构建的剩余图像，同时保留基础已知类别图像，以考察模型在类别扩展后对已知类与新类的联合识别能力。

在直接迁移测试中，对与原训练标签空间可对应的类别进行评估，包括灵芝、覆盆子、铁皮石斛

和前胡。其中，覆盆子、铁皮石斛与前胡的识别准确率均达到 100.00%，灵芝为 96.00%。结果表明，HerbVLM 对部分形态稳定类别具有较强的跨域鲁棒性。这说明模型在公开数据集上学到的表征并非完全依赖特定背景或特定拍摄分布，而具有一定场景迁移能力。

对于新增药材，即乌药、三叶青、衢枳壳、西红花，将其纳入 few-shot 类别，并在少样本适配实验中采用 10 次重复实验统计均值与标准差，如表 6 和图 10 所示。结果显示，模型在 1-shot、2-shot 和 5-shot 设置下的新类平均准确率分别为 $(84.00 \pm 6.96)\%$ 、 $(91.54 \pm 3.72)\%$ 和 $(98.20 \pm 1.89)\%$ ，呈现稳定上升趋势。以“乌药”为例，其准确率由 1-shot 的 $(73.57 \pm 10.71)\%$ 提升至 5-shot 的 $(97.00 \pm 3.00)\%$ 。这说明 HerbVLM 在仅需极少标注样本的条件下即可较快建立新类别判别能力，提示该方法在新药材类扩展方面具有一定应用潜力。

表 6 未知类别少样本学习准确率

Table 6 Accuracy of few-shot learning for unknown classes

样本	准确率/%			
	乌药	三叶青	衢枳壳	西红花
1-shot	73.57	90.00	83.57	88.86
2-shot	86.92	94.62	91.23	93.39
5-shot	97.00	97.00	100.00	98.80

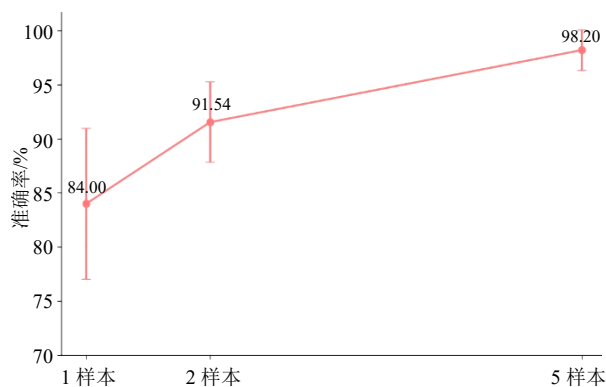


图 10 未知类别少样本学习适配准确率

Fig. 10 Adaptation accuracy for unknown-class few-shot learning

为评估新类别接入对基础类别识别性能的影响，在外部 few-shot 适配实验中进一步比较了新增类别接入前后的基础类别准确率。首先在仅包含基础类别标签空间的条件下，对基础类别 query 样本进行评估，得到基线准确率 baseline_old_acc ；随后，

在加入新增类别原型后,于“基础类别+新增类别”的联合标签空间中再次对同一基础类别 query 样本进行判别,得到 acc_old 。二者差值 $\Delta_{\text{old}} = \text{acc}_{\text{old}} - \text{baseline}_{\text{old}}$ 用于衡量新增类别接入对基础类别性能的影响。实验结果如图 11 所示。

结果表明,新增类别接入后基础类别准确率分别下降 0.62%、1.25%和 2.92%,但总体仍保持在 95%以上。这说明本方法在进行少样本新类别扩展时,基础类别识别能力会受到一定干扰,干扰出现的根本原因是新类加入后在同一特征空间中与基础类产生了边界竞争,而中药材本身又具有较强细

粒度相似性。但整体保持性能较好,体现出一定的类别扩展可行性。

总体而言,HerbVLM 不仅在封闭数据集内具有较高识别精度,而且在新场景和新类别条件下也表现出一定适应能力。这种能力对于中药材智能鉴别的实际应用具有重要意义,因为真实应用场景往往难以严格受限于固定类别和固定拍摄条件。

3 系统设计与实现

3.1 系统设计

面向中药材智能鉴别的实际应用需求,本研究在模型研究基础上实现了可交互的中药材智能鉴别系统。系统设计遵循以下目标:1) 识别准确且结果便于判读。基于 HerbVLM 模型输出 Top-3 候选类别及其置信度,并给出类别来源(基础类别或自定义类别)与辅助判读信息,以提高结果的可用性。2) 支持增量扩展。针对新药材类别,允许用户通过少量样本进行快速学习,以降低新类别接入成本。3) 流程完整、操作简便。构建登录、识别、学习和状态查看的完整流程,便于非算法用户使用。4) 安全可控。采用账号认证与会话管理机制,保证系统访问安全,并对学习数据规模设置边界约束,以保障服务稳定运行。总体架构如图 12 所示。

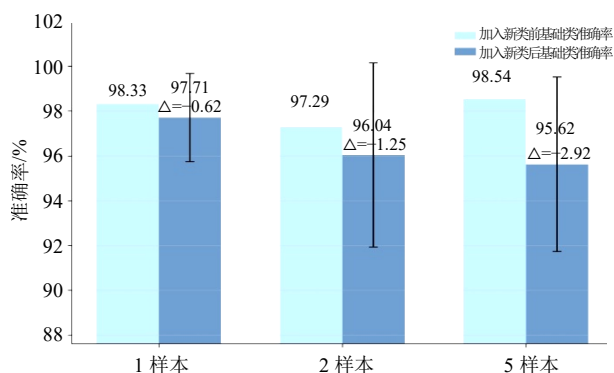


图 11 新增类别接入前后基础类别准确率对比

Fig. 11 Comparison of base category accuracy before and after new category access

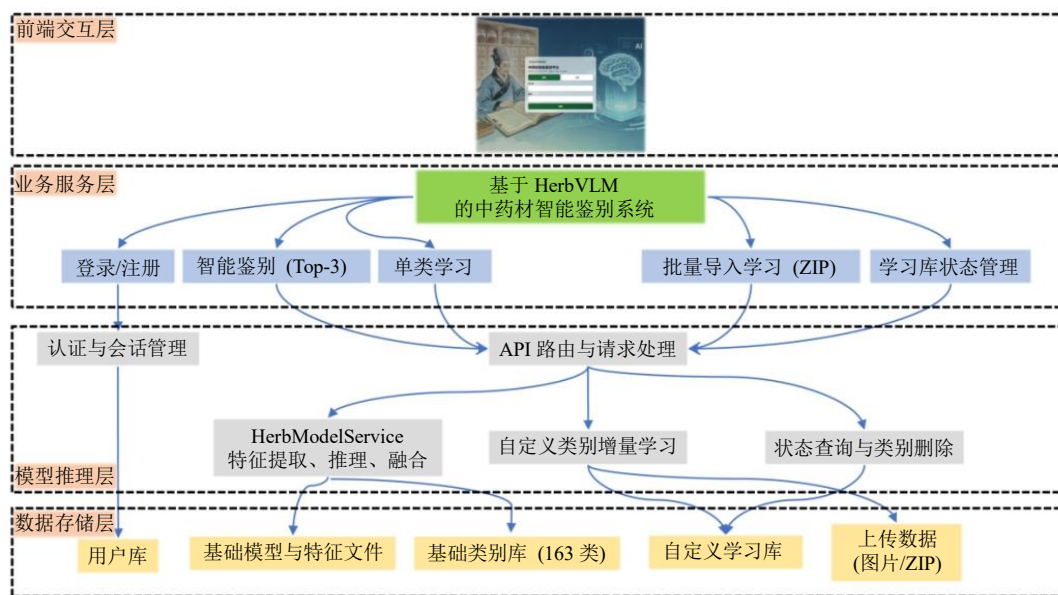


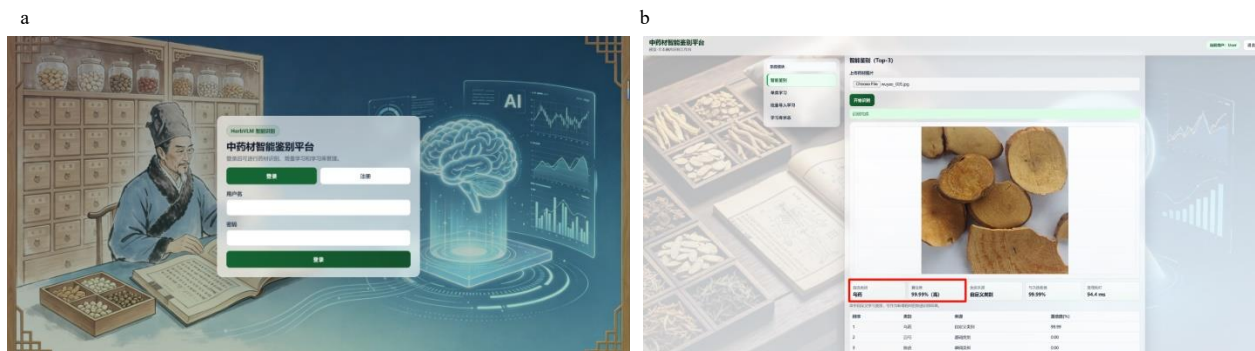
图 12 系统总体架构

Fig. 12 Overall system architecture

系统采用“前端交互层-业务服务层-模型推理层-数据存储层”的分层架构。前端交互层提供登录

注册、图像上传、结果展示、样本学习与学习库管理等界面。业务服务层负责请求路由、参数校验、

权限控制与流程编排。模型推理层调用 HerbVLM 模型完成特征提取与类别判别。数据存储层用于保存用户信息与自定义类别原型。其中,基础识别类别来自预训练体系,自定义类别则通过在线增量学习加入推理候选集合,从而实现基础类别识别与场景扩展识别的统一。



a-登录/注册入口; b-智能鉴别界面。
a-login/regisration entry; b-intelligent identification interface.

图 13 系统局部界面

Fig. 13 Partial system interface

在智能鉴别流程中,用户上传药材图像后,系统完成预处理与特征匹配,返回 Top-3 候选类别及其置信度,同时给出推理耗时,便于用户快速判断结果可靠性。对于不在现有类别中的样本,用户可通过增量学习扩展类别库:单类学习支持输入类别名并上传 1~5 张样本图,系统提取特征后写入自定义学习库;批量导入学习支持上传按类别组织的 ZIP 数据集,系统自动解析并批量完成学习。学习状态模块用于展示模型名称、基础类别数量、自定义类别数量、各类样本数与更新时间,并支持按类删除,便于后续维护。

从应用效果看,系统在保证基础识别能力的同时,可通过少量样本快速适配新类别。对同一测试样本,在完成“少样本学习”后,系统可将其识别为新增类别且相关置信度提升,说明该方案具备较好的工程可用性与扩展性。

3.3 系统测试

为验证系统可用性与稳定性,本研究开展功能测试、性能测试与泛化验证。功能测试覆盖注册登录、图像识别、单类学习、ZIP 批量学习、学习库查询与删除等核心流程,结果显示各模块交互正常,异常输入可被有效拦截。性能测试表明,在连续请求条件下系统能够稳定返回识别结果,推理耗时处于可接受范围。泛化验证采用“基础类识别、

3.2 系统实现

本系统以中药材图像识别为核心搭建完整业务闭环,整体划分智能鉴别、单类样本增量训练、批量样本导入学习、学习状态管理 4 大功能模块,同时配套账号登录与注册通道,整体界面布局如图 13 所示。

新类别少样本学习、再次识别”的流程,结果显示系统可在保留基础识别能力的同时快速适配新类别,满足实际应用中的增量扩展需求。

4 讨论与结论

本研究围绕中药材细粒度识别中“类间相似、类内差异大、文本知识利用不足”等问题,提出了视觉-文本融合框架 HerbVLM,并在开源数据集与自采浙八味外部测试集上进行系统验证。结果表明,HerbVLM 在 Chinese-Medicine-163 测试集上的 Top-1 Acc、 P_{macro} 、 R_{macro} 和 $F1_{macro}$ 分别达到 98.69%、98.70%、98.71%和 98.70%;在浙八味外部测试中,已知类别灵芝、覆盆子、铁皮石斛和前胡的识别准确率分别为 96.00%、100.00%、100.00%和 100.00%,外部新增类别在 1-shot、2-shot 和 5-shot 条件下平均准确率分别为 84.00%、91.54%和 98.20%。上述结果表明,HerbVLM 在中药材细粒度识别、跨域泛化和少样本类别扩展方面具有较好的应用潜力。

目前,中药材智能识别研究多以图像特征提取和分类模型优化为主要思路,通常将中药材识别视为一般视觉分类任务^[27]。例如,韩勇等^[28]构建了包含 50 种中药材饮片的数据集,并通过优化 DenseNet-201 网络实现了较高识别准确率;Xu 等^[29]提出引入注意力机制的金字塔网络模型,以提升中草药图像识别效果;Sun 等^[30-31]则通过 softmax 损失函数与

三重损失函数相结合的方式,对中药材识别网络进行优化。尽管上述研究在识别精度方面取得了一定进展,但多数方法仍主要依赖视觉信息,对药材名称、性状描述等文本知识利用相对不足。与已有研究相比,本研究从视觉-语言协同建模角度出发,将类别级文本提示与视觉细粒度特征共同引入识别过程,并针对中药材任务特点对视觉适配、类间去混淆和动态融合进行改进,构建了多模态识别模型 HerbVLM。实验结果表明,该设计并非简单增加模型复杂度,而是较好契合了中药材识别中“相似类别难区分、近邻边界易重叠”的实际需求。

通过实验结果分析,中药材图像识别的难点并不只在于总体准确率的提升,更在于对细粒度“图像鉴别边界”的刻画。这里的类别边界并非一般物体分类中的轮廓边界,而是由药材形态、纹理、色泽、切面结构及局部器官特征共同构成。当某些类别具有较稳定且明显的视觉性状时,模型通常能够获得较高识别率;而对于外观高度接近、局部差异较弱的类别,即使模型能够将其定位于近邻类别范围,仍可能在最终判别时发生混淆。本研究中,谷芽和麦芽即为典型困难类别。二者在颗粒形态、色泽分布、发芽程度及堆叠方式等方面相似度较高,导致图像端可分辨线索有限;同时,当前类别级文本提示对二者的区分信息仍相对不足,难以形成足够强的语义约束。提示对于高度相似药材类别的识别,后续仍需结合更细化的性状知识、局部关键区域建模及困难样本约束策略加以改进。

值得指出的是,现有中药材识别研究多基于封闭数据集开展性能比较,对开放场景下的泛化与迁移能力关注相对不足^[32-35]。本研究进一步引入浙八味外部测试与少样本增量实验,考察模型在新域和外部新增类别条件下的适应性。结果显示,模型在外部已知类别上保持较高识别水平;对于未参与当前训练的外部新增类别,仅借助少量标注样本即可实现有效迁移,且性能随 shot 数增加而稳定提升,表明 HerbVLM 具有一定的“基础识别+快速扩展”应用潜力。此外,本研究还实现了配套网页系统,支持登录管理、Top-3 智能鉴别、单类学习与批量导入,使研究由算法验证进一步延伸至可交互应用验证。

本研究仍有进一步完善空间。例如,自采数据规模与采集条件仍可继续扩展,后续可纳入更多产地、批次及拍摄域差异,以进一步提升模型的跨域

稳健性;同时,针对真伪鉴别、炮制等级判别等更细粒度任务,仍可结合更丰富的中药材性状知识、专家经验规则及多源信息开展深入研究。总体来看,HerbVLM 为中药材智能鉴别提供了一种兼顾识别精度、跨域泛化与类别扩展能力的可行方法,也为后续中药材教学、质检辅助及真实场景应用提供了参考。本研究模型推理代码已开源,详见 <https://github.com/Yyg314/herbvml>。

利益冲突 所有作者均声明不存在利益冲突

参考文献

- [1] Tian D C, Zhou C, Wang Y, *et al.* NB-TCM-CHM: Image dataset of the Chinese herbal medicine fruits and its application in classification through deep learning [J]. *Data Brief*, 2024, 54: 110405.
- [2] 张谊, 万华, 涂淑琴. 基于计算机视觉的中药饮片分类技术综述与案例研究 [J]. *计算机应用*, 2022, 42(10): 3224-3234.
- [3] 孙兴, 李杨, 毛天驰, 等. 改进 YOLOv8 在中药饮片自动识别标注中的研究与应用 [J]. *中草药*, 2025, 56(1): 216-226.
- [4] Miao J F, Huang Y N, Wang Z S, *et al.* Image recognition of traditional Chinese medicine based on deep learning [J]. *Front Bioeng Biotechnol*, 2023, 11: 1199803.
- [5] Ji H H, Liu X J, Wang L, *et al.* Image recognition of Chinese herbal medicine using adaptive gamma correction based on convolutional neural network [C] // 2024 IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS). Kaifeng, China: IEEE, 2024: 1428-1433.
- [6] Wang H, Chen Z, Huo J, *et al.* Research on micro-morphological feature recognition method of traditional Chinese medicinal herbs based on deep convolutional neural network [J]. *J Combinat Mathemat Combinat Comput*, 2025, 127B: 7759-7776.
- [7] 王双双, 吴纯洁, 艾莉, 等. 自动识别技术在中药饮片行业的展望 [J]. *中药与临床*, 2012, 3(3): 52-55.
- [8] Datsi T, Aznag K, BenAli B A, *et al.* A short survey on multimodal data fusion in image classification [C] // 2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST). Marrakesh, Morocco: IEEE, 2024: 1-4.
- [9] Zhang T B, Xu H R, Wang Z Y, *et al.* Deep learning for subtype recognition of Yang deficiency tongue images in traditional Chinese medicine [J]. *Digital Chin Med*, 2026, 9(2): 197-210.
- [10] Lu X, Feng Y, Jia X D, *et al.* Multi-label tongue image classification via label semantics-driven feature interaction

- enhancement [J]. *Comput Biol Chem*, 2026, 120: 108735.
- [11] Chen J Y, Cai Z Y, Liu Z H, *et al.* ShizhenGPT: Towards multimodal LLMs for traditional Chinese medicine [EB/OL]. *arXiv* [2025-08-20]. <https://arxiv.org/abs/2508.14706>
- [12] Li Q, Zhao F, Zhao L L, *et al.* A system of multimodal image-text retrieval based on pre-trained models fusion [J]. *Concurr Comput Pract Exp*, 2025, 37(3): e8345.
- [13] Shen Q, Ruan C G, Yang L. Multi-granularity knowledge embedding for multi-label herbal medicine recognition [J]. *Neurocomputing*, 2025, 649: 130707.
- [14] Li S, Liu F, Hao Z H, *et al.* Logits DeConfusion with CLIP for few-shot learning [C] // 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025: 25411-25421.
- [15] 飞桨 AI Studio. 智慧岐黄: 163 种中药材深度图像辨识库 [EB/OL]. [2026-02-15]. <https://aistudio.baidu.com/datasetdetail/246739>.
- [16] Radford A, Kim J W, Hallacy C, *et al.* Learning transferable visual models from natural language supervision [EB/OL]. *arXiv* [2021-02-26]. <https://arxiv.org/abs/2103.00020>.
- [17] Zhou K Y, Yang J K, Loy C C, *et al.* Learning to prompt for vision-language models [J]. *Int J Comput Vis*, 2022, 130(9): 2337-2348.
- [18] Zhou K Y, Yang J K, Loy C C, *et al.* Conditional prompt learning for vision-language models [C] // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 16795-16804.
- [19] Pratt S, Covert I, Liu R, *et al.* What does a *Platypus* Look like? Generating customized prompts for zero-shot image classification [C] // 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2024: 15645-15655.
- [20] Gao P, Geng S J, Zhang R R, *et al.* CLIP-adapter: Better vision-language models with feature adapters [J]. *Int J Comput Vis*, 2024, 132(2): 581-595.
- [21] Zhang R R, Zhang W, Fang R Y, *et al.* Tip-adapter: Training-free adaption of CLIP for few-shot classification [C] // Computer Vision—ECCV 2022. Cham: Springer, 2022: 493-510.
- [22] Khattak M U, Rasheed H, Maaz M, *et al.* MaPLe: Multi-modal prompt learning [C] // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 19113-19122.
- [23] Cherti M, Beaumont R, Wightman R, *et al.* Reproducible scaling laws for contrastive language-image learning [C] // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 2818-2829.
- [24] Yang A, Pan J S, Lin J Y, *et al.* Chinese CLIP: Contrastive vision-language pretraining in Chinese [EB/OL]. *arXiv* [2022-11-02]. <https://arxiv.org/abs/2211.01335>
- [25] 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). Las Vegas, NV, USA: IEEE, 2016: 770-778.
- [26] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. *arXiv* [2020-10-22]. <https://arxiv.org/abs/2010.11929>.
- [27] Zeng J, Wu H, He M. Image classification combined with faster R-CNN for the peak detection of complex components and their metabolites in untargeted LC-HRMS data [J]. *Anal Chim Acta*, 2023, 1238: 340189.
- [28] 韩勇, 兰杰, 郭瑞瑶, 等. 基于深度学习的中药材饮片识别 [J]. *西北农业学报*, 2023, 32(11): 1859-1867.
- [29] Xu Y X, Wen G H, Hu Y, *et al.* Multiple attentional pyramid networks for Chinese herbal recognition [J]. *Pattern Recognit*, 2021, 110: 107558.
- [30] Sun X, Qian H N. Chinese herbal medicine image recognition and retrieval by convolutional neural network [J]. *PLoS One*, 2016, 11(6): e0156327.
- [31] Sun X, Qian H N, Xiong Y L, *et al.* Deep learning-enabled mobile application for efficient and robust herb image recognition [J]. *Sci Rep*, 2022, 12: 6579.
- [32] Kawadkar K. Comparative analysis of vision transformers and convolutional neural networks for medical image classification [EB/OL]. *arXiv* [2025-07-24]. <https://arxiv.org/abs/2507.21156>.
- [33] Jung D H, Kim H Y, Won J H, *et al.* Development of a classification model for *Cynanchum wilfordii* and *Cynanchum auriculatum* using convolutional neural network and local interpretable model-agnostic explanation technology [J]. *Front Plant Sci*, 2023, 14: 1169709.
- [34] Dai W B, Ma Y X, Fan Y, *et al.* A multi-scale feature extraction algorithm for Chinese herbal medicine image classification [J]. *Appl Sci*, 2025, 15(8): 4271.
- [35] Pan D P, Guo Y L, Fan Y F, *et al.* Development and application of traditional Chinese medicine using AI machine learning and deep learning strategies [J]. *Am J Chin Med*, 2024, 52(3): 605-623.