

# 基于层次标注和自适应预处理的多源农业病害图像数据集构建

胡婷<sup>1</sup>, 孙晓海<sup>2</sup>, 宋海龙<sup>2</sup>, 廖昌义<sup>2</sup>, 王福德<sup>2,3</sup>

(1. 长春工业大学 新闻与传播学院, 长春 130012;

2. 吉林海诚科技有限公司, 长春 130119; 3. 吉林农业大学 信息技术学院, 长春 130118)

**摘要:** 针对农业病害图像数据集存在多样性和图像质量欠佳的问题, 提出一种基于层次标注和自适应预处理的多源农业病害图像数据集构建方法. 首先, 利用智能手机、专业相机和无人机等设备从不同地区、作物种类及生长阶段采集图像, 以确保数据的多样性. 其次, 构建层次标注体系, 涵盖农业病害类型、程度和部位3个层次, 使用 LabelImg 和 LabelMe 等工具进行标注, 并经专家审核. 最后, 应用自适应预处理方法, 包括自动裁剪、归一化、去噪和增强, 根据图像特征调整参数以提升质量. 实验采用基于 ResNet-50 架构的卷积神经网络 (CNN) 模型进行验证, 结果表明, 层次标注和自适应预处理方法显著提升了数据集的质量和模型性能, 模型在准确率、召回率和  $F_1$  分数上分别达 92.5%, 91.8% 和 92.1%, 优于其他数据集训练结果.

**关键词:** 农业病害图像; 数据集构建; 层次标注; 自适应预处理; 多源数据

**中图分类号:** TP391 **文献标志码:** A **文章编号:** 1671-5489(2025)03-0815-07

## Construction of Multisource Agricultural Disease Image Dataset Based on Hierarchical Annotation and Adaptive Preprocessing

HU Ting<sup>1</sup>, SUN Xiaohai<sup>2</sup>, SONG Hailong<sup>2</sup>, LIAO Changyi<sup>2</sup>, WANG Fude<sup>2,3</sup>

(1. School of Journalism and Communication, Changchun University of Technology, Changchun 130012, China;

2. Jilin Haicheng Technology Co., Ltd., Changchun 130119, China;

3. College of Information Technology, Jilin Agricultural University, Changchun 130118, China)

**Abstract:** Aiming at the problems of diversity and poor image quality in agricultural disease image datasets, we proposed a multisource agricultural disease image dataset construction method based on hierarchical annotation and adaptive preprocessing. Firstly, images were collected from different regions, crop types, and growth stages by using devices such as smartphones, professional cameras, and drones to ensure data diversity. Secondly, we constructed a hierarchical annotation system that covered three levels of agricultural disease type, severity, and location, we used tools such as LabelImg and LabelMe for annotation, and requested expert review. Finally, we applied adaptive preprocessing methods, including automatic cropping, normalization, denoising and enhancement, to adjust parameters based on image features to improve quality. The experiment used a convolutional

收稿日期: 2024-01-12.

**第一作者简介:** 胡婷(1981—), 女, 汉族, 硕士, 讲师, 从事计算机应用的研究, E-mail: huting@ccut.edu.cn. **通信作者简介:** 王福德(1990—), 男, 汉族, 硕士, 中级工程师, 从事人工智能的研究, E-mail: 562324919@qq.com.

**基金项目:** 吉林省科学技术厅项目(批准号: 20230401092YY).

neural network (CNN) model based on the ResNet-50 architecture for validation, and the results show that hierarchical annotation and adaptive preprocessing methods significantly improve the quality of the dataset and model performance, the model achieves accuracy, recall, and  $F_1$  score of 92.5%, 91.8%, and 92.1%, respectively, which are better than the training results of other datasets.

**Keywords:** agricultural disease image; dataset construction; hierarchical annotation; adaptive preprocessing; multisource data

农业病害是影响农作物产量和质量的重要因素之一<sup>[1-2]</sup>。真菌病害可导致作物叶片枯萎,影响光合作用效率,进而降低作物产量;病毒病害则可能改变作物的生长发育过程,导致果实畸形或品质下降<sup>[3]</sup>。因此,准确监测和识别农业病害并及时采取有效防治措施至关重要。不仅能减少农药的使用,降低环境污染,还能保障农作物的质量和产量,起到降本增效的作用<sup>[4]</sup>。

近年来,随着深度学习技术的快速发展,尤其是卷积神经网络(CNN)<sup>[5]</sup>在图像识别领域的成功应用,使农业病害图像识别技术也取得了显著进展。CNN通过多层的卷积和池化操作,能自动学习图像的层次结构特征,无需人工设计复杂的特征提取器<sup>[6]</sup>。这种自底向上的特征学习方式,使CNN在处理图像数据时有显著优势<sup>[7]</sup>。例如,在图像识别任务中,CNN能自动提取图像的边缘、纹理、形状等特征<sup>[8]</sup>,从而实现了对图像内容的准确分类和识别。但CNN模型的性能高度依赖于训练数据的质量和数量,高质量的数据集能提供丰富的特征信息,帮助模型更好地泛化到新的、未识别的数据上。数据集的质量不仅取决于图像的清晰度和标注的准确性,还取决于数据的多样性和覆盖范围<sup>[9]</sup>。一个多样化的数据集包含不同光照、角度、背景和尺度下的图像,这样训练出的模型才能更好地适应各种实际情况<sup>[10]</sup>。此外,模型训练过程中的参数调整和优化也是提升CNN性能的关键,包括选择合适的损失函数衡量预测结果与实际标签之间的差异,选择合适的优化算法调整网络权重,以及调整学习率、批处理大小等超参数控制训练过程<sup>[11]</sup>。这些因素共同决定了模型训练的效果和最终性能。

构建多源农业病害图像数据集具有重要意义。首先,多源数据能涵盖不同地区、不同农作物种类、不同生长阶段的病害图像,从而提高数据集的多样性和代表性。这种多样性的提升有助于增强模型对不同环境和病害类型的泛化能力,使其能在多种实际应用中表现出色。例如,不同地区的气候条件和土壤类型可能导致农业病害的表现形式存在差异,通过多源数据采集,可确保数据集涵盖这些差异,从而提高模型的适应性。其次,层次标注体系的构建能更精准地描述农业病害特征,为模型提供更丰富的语义信息。将病害类型、病害程度、病害部位等作为不同层次的标注内容,形成一个结构化的标注体系,可以显著提升数据集的质量。这种结构化的标注方式不仅能提高模型的识别精度,还能为后续的研究提供更详细的信息。此外,自适应预处理方法的应用能根据图像的特征和质量自适应地调整处理参数,进一步提升数据质量<sup>[12]</sup>。可通过自动裁剪去除无关背景信息,利用归一化调整图像的亮度和对比度,利用去噪减少图像中的噪声干扰,这些方法能显著提高图像的质量,从而提升模型的性能。

## 1 研究方法

### 1.1 数据来源与采集

多源农业病害图像数据的采集需涉及多个维度,以确保数据集的多样性和代表性。本文数据来源包括不同地区、不同作物种类以及不同生长阶段的农业病害图像。数据采集覆盖多个农业产区,涉及小麦、水稻、玉米等多种主要农作物,以及番茄、苹果、茶叶等经济作物。这些图像涵盖了从苗期到成熟期的各生长阶段,以反映农业病害在不同发育时期的特征。

采集设备的选择对图像质量有重要影响。本文采用多种采集设备,包括智能手机、专业相机和无人机。智能手机因其便捷性和普及性,被广泛用于现场快速采集图像;专业单反相机因其高分辨率和图像质量,适用于需要精细图像的场景;无人机平台则用于获取大范围农田的图像,以评估病害的分布情况。图1为不同设备采集的图像数据示例。



图1 不同设备采集的图像示例

Fig.1 Examples of images collected by different devices

为确保采集的图像清晰、完整且具有代表性,本文采取了一系列质量控制措施.首先,采集人员接受专业的培训,以确保他们能准确识别病害症状,并在合适的光照和角度条件下进行拍摄;其次,采集过程中严格遵循标准化的操作流程,包括固定拍摄距离、角度和光照条件等参数;再次,采集的图像经过初步筛选,剔除模糊、遮挡严重或信息不完整的图像;最后,所有采集的图像均需经过专家审核,以确保其准确性和科学性.

### 1.2 层次标注体系构建

层次标注是一种结构化的标注方法,旨在更精准地描述病害特征<sup>[13]</sup>.本文的层次标注体系包括3个主要层次:病害类型、病害程度和病害部位.病害类型标注用于区分不同的病害种类,如真菌病害、细菌病害和病毒病害等;病害程度标注则根据病害的严重程度进行分类,如轻微、中度和重度;病害部位标注指明病害在植物上的具体位置,如叶片、茎干或果实.这种多层次的标注方式能为模型提供更丰富的语义信息,有助于提高病害的识别精度.

标注工具的选择对标注效率和准确性至关重要.本文采用 LabelImg 和 LabelMe 等标注工具,这些工具支持多种标注格式,能满足不同层次标注的需求.标注流程包括标注人员培训、标注任务分配、标注过程监控以及标注结果的审核.标注人员在开始标注前需接受系统培训,以确保他们能准确理解标注标准和操作流程.标注过程中,通过定期检查标注结果,及时发现并纠正错误.最终,所有标注结果都需经过专家审核,以确保标注的准确性和一致性.

### 1.3 自适应预处理方法

自适应预处理是提升图像质量、减少噪声干扰和增强病害特征的关键步骤.本文采用多种自适应预处理方法,包括自动裁剪、归一化、去噪和增强等<sup>[14]</sup>.自动裁剪方法根据图像中的病害区域自动调整裁剪范围,以去除无关背景信息,突出病害特征.归一化处理将图像的像素值调整到统一范围,减少光照条件对图像的影响.去噪方法利用先进的图像处理算法去除图像中的噪声,提高图像的清晰度<sup>[15]</sup>.增强方法则通过调整图像的对数度、亮度和饱和度等参数,进一步突出病害特征.

这些自适应预处理方法能根据图像的特征和质量自适应地调整处理参数.例如,对光照不均匀的图像,增强方法会自动调整亮度和对数度,以改善图像质量;对噪声较多的图像,去噪方法会自动选择合适的算法进行处理<sup>[16]</sup>.这种自适应性处理方法不仅提高了图像预处理的效率,还确保了预处理后的图像能更好地服务于后续的病害识别任务.

自适应预处理在提升图像质量方面具有显著优势.通过去除无关背景信息和噪声,预处理后的图像能更清晰地展示病害特征,从而提高模型的识别性能.此外,自适应预处理还能减少不同采集条件下的图像差异,增强数据集的一致性,为模型训练提供更高质量的数据支持.

## 2 数据集构建与分析

### 2.1 数据集构建流程

本文构建的多源农业病害图像数据集流程如图2所示,涵盖了数据采集、层次标注、自适应预处理、数据筛选和整合等关键环节,以确保数据集的高质量和高可用性.

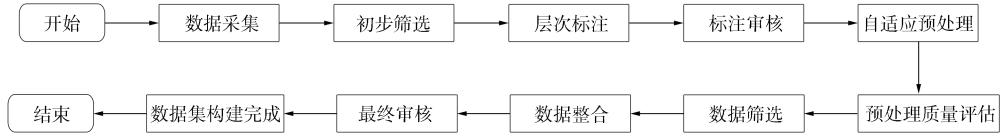


图 2 数据集构建流程

Fig. 2 Flow chart of dataset construction

**数据采集：**数据采集是构建数据集的基础环节。本文从不同地区、不同作物种类和不同生长阶段采集农业病害图像，以确保数据的多样性。采集设备包括智能手机、专业相机和无人机等，以适应不同的采集场景和需求。采集过程中，严格遵循标准化操作流程，确保图像清晰、完整且具有代表性。

**初步筛选：**采集的图像先经过初步筛选，剔除模糊、遮挡严重或信息不完整的图像。该过程通过自动化算法和人工审核相结合的方式完成，以确保筛选后的图像质量符合后续处理的要求。

**层次标注：**层次标注是数据集构建中的关键环节。标注内容包括病害类型、病害程度和病害部位 3 个层次。标注工具采用 LabelImg 和 LabelMe 等专业软件，标注员在标注前需进行系统培训，以确保标注的准确性和一致性。标注完成后，所有标注结果均需经过专家审核，以确保标注质量。

**自适应预处理：**自适应预处理方法包括自动裁剪、归一化、去噪和增强等。这些方法能根据图像的特征和质量自适应地调整处理参数，提升图像质量，减少噪声干扰，增强病害特征。预处理后的图像质量通过自动化评估工具进行评估，以确保预处理效果符合要求。

**数据筛选和整合：**预处理后的图像再次经过筛选，去除不符合要求的图像。筛选后的图像按病害类型、作物种类和采集设备等维度进行整合，形成结构化的数据集。整合后的数据集经过最终审核，以确保数据集的完整性和一致性。图 3 为数据集的部分图像。

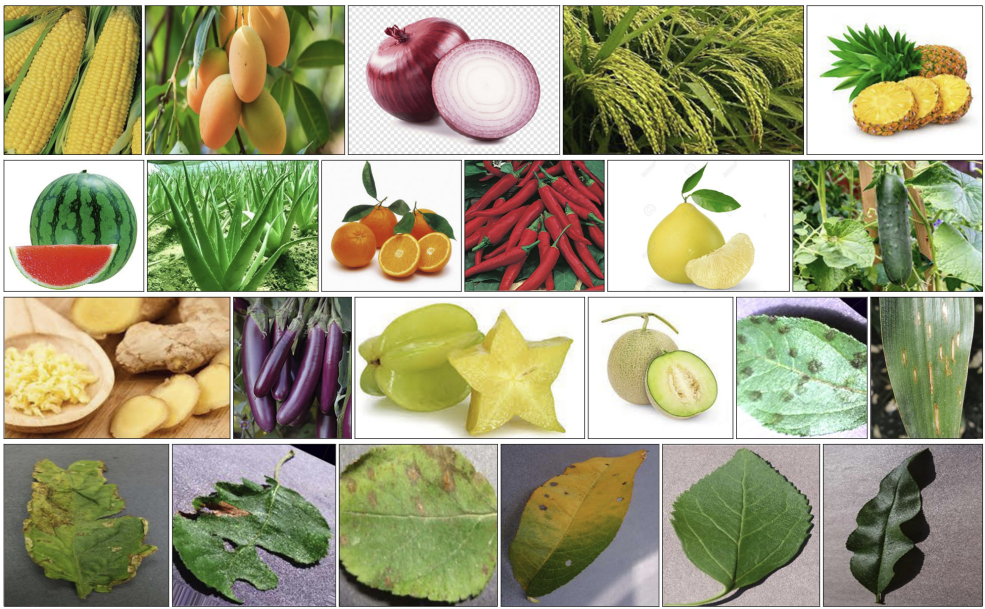


图 3 数据集部分图像

Fig. 3 Partial images of dataset

## 2.2 数据集特征分析

本文构建的多源农业病害图像数据集具有丰富的特征，涵盖了多种病害类型、不同作物种类和不同生长阶段的图像。数据集的主要特征包括病害类型分布、图像数量、图像分辨率和标注信息等，其中：病害类型包括真菌病害(40%)、细菌病害(30%)、病毒病害(20%)、其他(10%)；图像数量总计 10 000 张，训练集 7 000 张，验证集 1 500 张，测试集 1 500 张；图像分辨率主要集中在  $640 \times 480$ ， $1\ 280 \times 720$  和  $1\ 920 \times 1\ 080$ ；标注信息中病害类型包括白粉病、稻瘟病、炭疽病、斑点病、枯萎病、线虫病害、黄化病、线虫病；病害程度分为轻微、中度和重度；病害部位包括果实、叶片、表皮、茎干；作

物种类包括小麦、水稻、玉米、棉花、番茄、苹果、香蕉、茶叶、草莓、辣椒、橙子、芒果、生姜、茄子、芦荟、柚子、黄瓜、哈密瓜、杨桃、洋葱、西瓜、菠萝、猕猴桃;采集设备为手机(30%)、相机(40%)、无人机(30%)。

本文构建的多源农业病害图像数据集涵盖了丰富的病害类型和作物种类,图像数量充足且分辨率多样,标注信息详细。与现有的农业病害图像数据集相比,本文数据集有显著的多样性和代表性。例如,与 Plant Village 数据集相比,本文数据集不仅涵盖了更多的病害类型和作物种类,还通过层次标注提供了更丰富的语义信息。此外,自适应预处理方法的应用进一步提升了图像质量,增强了数据集的可用性。

### 3 实验与验证

#### 3.1 实验设计

为验证本文构建的多源农业病害图像数据集的有效性和优越性,选择 CNN 作为验证工具。CNN 因其在图像识别任务中的卓越性能而被广泛应用于农业病害图像识别领域。本文选用的 CNN 模型结构基于经典的 ResNet-50 架构,该架构在处理大规模图像数据集时性能优异,能有效提取图像的层次结构特征,同时具有良好的泛化能力。

将数据集按 70%,15%,15% 划分为训练集、验证集和测试集。这种划分方式旨在确保模型在训练过程中能充分学习数据集的特征,同时通过验证集和测试集对模型的性能进行准确评估。数据划分采用随机抽样的方法,以保证每个子集都具有代表性。

#### 3.2 模型训练与评估

模型训练过程包括损失函数的选择、优化算法的使用以及超参数的调整。损失函数选择交叉熵损失函数,该函数在分类任务中广泛应用,能有效衡量模型预测值与真实标签之间的差异。优化算法采用 Adam 优化器,其自适应学习率调整机制能加速模型的收敛速度,提高训练效率。超参数调整包括学习率、批处理大小和训练轮次等,通过在验证集上进行多次实验,确定最优的超参数设置。

模型评估采用准确率、召回率和  $F_1$  分数等指标。准确率衡量模型正确预测的样本比例,召回率衡量模型能识别出的正样本比例, $F_1$  分数则是准确率和召回率的调和平均值,可综合反映模型的性能。实验结果如图 4 和图 5 所示。

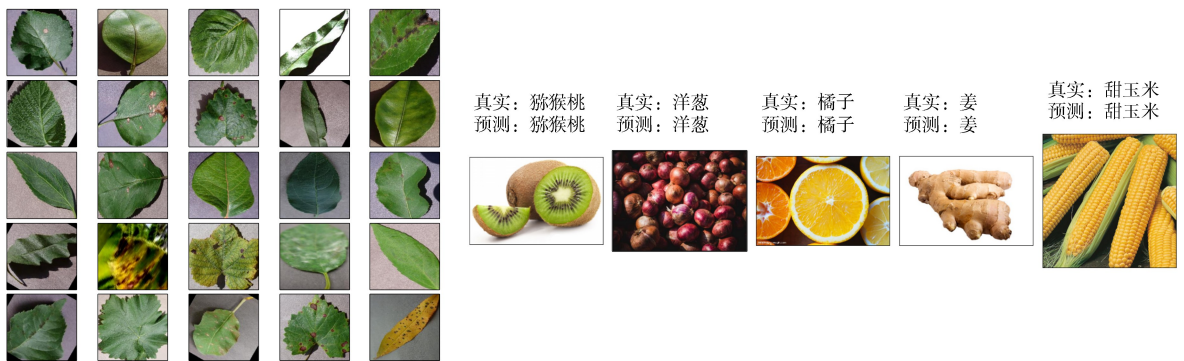


图 4 检测结果

Fig. 4 Detection results

通过在测试集上对模型进行评估,本文模型在准确率、召回率和  $F_1$  分数上均取得了优异成绩,分别为 92.5%,91.8% 和 92.1%。与使用其他数据集训练的结果相比,本文数据集训练的模型在性能上有显著优势,从而验证了本文数据集的有效性和优越性。

#### 3.3 结果分析

实验结果表明,层次标注和自适应预处理方法对模型性能的提升有显著贡献。层次标注通过提供更丰富的语义信息,使模型能更准确地识别不同类型的病害。例如,在处理复杂病害图像时,层次标注能显著提高模型的准确率。自适应预处理方法通过提升图像质量,减少了噪声干扰,增强了病害特

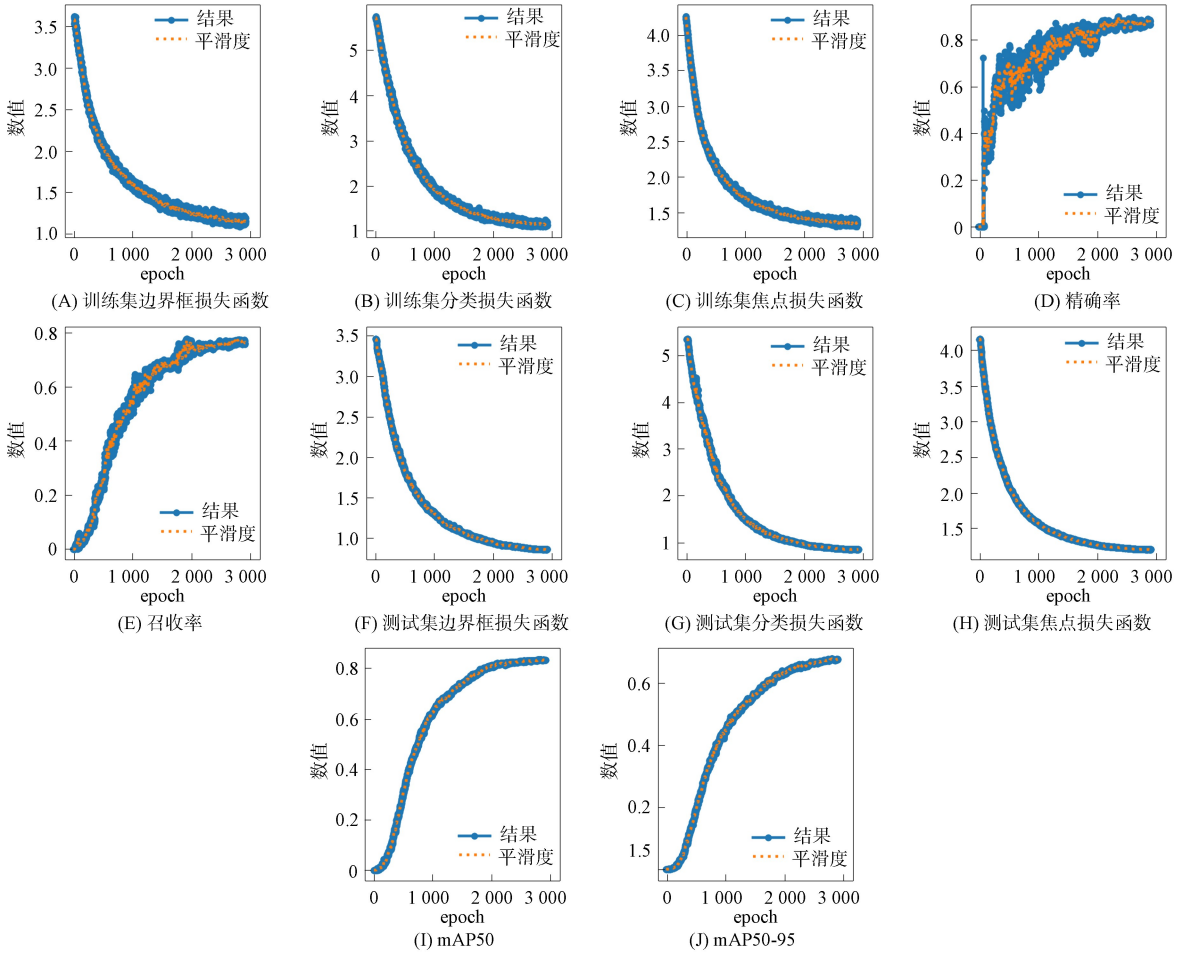


图 5 性能指标

Fig. 5 Performance indicators

征,从而提高了模型的鲁棒性.在不同图像质量条件下,自适应预处理方法均表现出色,特别是在处理低质量图像时,能显著提升模型的性能.

尽管模型在整体性能上较优,但在某些特定病害类型上仍存在不足.例如,对某些罕见农业病害的识别,模型的准确率较低,这可能是由于这些病害的图像数量较少,导致模型训练不足.此外,在极端光照条件下的图像识别中,模型的性能也有下降.

综上所述,针对农业病害图像数据集存在多样性和图像质量欠佳的问题,本文提出了一种基于层次标注和自适应预处理的多源农业病害图像数据集构建方法,并通过实验验证了本文构建的多源农业病害图像数据集的有效性和优越性.实验结果表明,层次标注和自适应预处理方法的引入显著提升了模型的性能,为农业病害识别领域提供了高质量的数据支持.

参 考 文 献

[ 1 ] 雷仲仁,郭予元,李世访.中国主要农作物有害生物名录[M].北京:中国农业科学技术出版社,2014:1-525. (LEI Z R, GUO Y Y, LI S F. Catalogue of Pests on Major Crops in China [M]. Beijing: China Agricultural Science and Technology Press, 2014: 1-525.)

[ 2 ] 吴钜文,陈红印.蔬菜害虫及其天敌昆虫名录[M].北京:中国农业科学技术出版社,2013:1-738. (WU J W, CHEN H Y. Catalog of Insect Pests and Their Natural Enemies of Vegetable Crops [M]. Beijing: Agricultural Science and Technology Press, 2013: 1-738.)

[ 3 ] 翟肇裕,曹益飞,徐焕良,等.农作物病虫害识别关键技术研究综述[J].农业机械学报,2021,52(7):1-18. (ZHAI Z Y, CAO Y F, XU H L, et al. Review of Key Techniques for Crop Disease and Pest Detection [J].

Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(7): 1-18.)

- [4] 何俊杰, 胡建君. 融合注意力机制的多特征艺术图像感知方法研究 [J]. 信息技术, 2024(11): 105-111. (HE J J, HU J J. Research on Multi-feature Art Image Perception Method Based on Attention Mechanism [J]. Information Technology, 2024(11): 105-111.)
- [5] 白黎明, 王道明, 赵玲. 人工智能技术在无牌行李匹配查找的应用探讨 [J]. 中国设备工程, 2024(22): 28-30. (BAI L M, WANG D M, ZHAO L. The Application of Artificial Intelligence Technology in Unlicensed Luggage Matching Search [J]. China Equipment Engineering, 2024(22): 28-30.)
- [6] 谭金鸿. 基于深度学习的人体异常行为识别 [D]. 成都: 电子科技大学, 2022. (TAN J H. Human Abnormal Behavior Recognition Based on Deep Learning [D]. Chengdu: University of Electronic Science and Technology of China, 2022.)
- [7] 屈伟洋, 俞扬. 多样性正则的神经网络训练方法探索 [J]. 南京大学学报(自然科学), 2017, 53(2): 340-349. (QU W Y, YU Y. Exploring Diversity Regularization in Neural Networks [J]. Journal of Nanjing University (Natural Sciences), 2017, 53(2): 340-349.)
- [8] 郭静博. 基于改进 CNN-Bi-LSTM 模型故障诊断与改进随机森林模型的湿法冶金流程评价研究 [J/OL]. 湿法冶金, (2023-12-01)[2024-11-26]. <https://link.cnki.net/urlid/11.3012.TF.20241126.1508.002>. (GUO J B. Research on Hydrometallurgical Process Evaluation Based on Improved CNN-Bi-LSTM Model Fault Diagnosis and Improved Random Forest Model [J/OL]. Hydrometallurgy, (2023-12-01)[2024-11-26]. <https://link.cnki.net/urlid/11.3012.TF.20241126.1508.002>.)
- [9] 佟忠正, 孙畅子. 基于 U-Net 网络的电力设备巡检图像增强模型及其自动控制研究 [J]. 自动化与仪表, 2024, 39(11): 79-82. (TONG Z Z, SUN Y Z. Research on Image Enhancement Model and Automatic Control of Power Equipment Inspection Based on U-Net Network [J]. Automation and Instrumentation, 2024, 39(11): 79-82.)
- [10] 隋浩然, 周晓航, 张宁. 基于 UGC 的产品改进: 属性提取和属性情感分类的方法与应用综述 [J]. 计算机科学, 2024, 51(增刊 2): 22-30. (SUI H R, ZHOU X H, ZHANG N. Product Improvement Based on UGC: Review of Methods and Applications of Attribute Extraction and Attribute Sentiment Classification [J]. Computer Science, 2024, 51(Suppl 2): 22-30.)
- [11] KLAWONN A, LANSER M, WEBER J. A Domain Decomposition-Based CNN-DNN Architecture for Model Parallel Training Applied to Image Recognition Problems [J]. SIAM Journal on Scientific Computing, 2024, 46(5): C557-C582.
- [12] 闵锋, 张雨薇, 刘焯晖, 等. 改进 YOLOv8 的轻量化水下生物检测模型 [J]. 计算机工程与应用, 2025, 61(6): 96-105. (MIN F, ZHANG Y W, LIU Y H, et al. To Improve the Lightweight Underwater Biological Detection Model of YOLOv8 [J]. Computer Engineering and Application, 2025, 61(6): 96-105.)
- [13] 吕彦朋, 赵颖彤, 苏晓明, 等. 基于改进 YOLOv7 的微生物细胞识别算法 [J]. 现代电子技术, 2025, 48(1): 47-54. (LV Y P, ZHAO Y T, SU X M, et al. Microbia Cell Recognition Algorithm Based on Improved YOLOv7 [J]. Modern Electronic Technology, 2025, 48(1): 47-54.)
- [14] 王畅, 李朝, 赵霞, 等. 基于图像和深度学习的驾驶人状态监测算法综述 [J]. 中国公路学报, 2025, 38(1): 324-347. (WANG C, LI C, ZHAO X, et al. An Overview of Driver Condition Monitoring Algorithms Based on Image and Deep Learning [J]. Chinese Journal of Highways, 2025, 38(1): 324-347.)
- [15] 马震环. 深度学习技术在图像语义分割中的应用研究 [D]. 成都: 中国科学院光电技术研究所, 2020. (MA Z H. Research on the Application of Deep Learning Technology in Image Semantic Segmentation [D]. Chengdu: Institute of Optoelectronic Technology, Chinese Academy of Sciences, 2020.)
- [16] 于成成, 郭芝源. 基于轻量级神经网络的人脸表情识别研究 [J]. 物联网技术, 2024, 14(8): 49-52. (YU C C, GUO Z Y. Research on Facial Expression Recognition Based on Lightweight Neural Network [J]. Internet of Things Technology, 2024, 14(8): 49-52.)

(责任编辑: 韩 啸)