

多肉植物识别数据集构建与应用研究

孙俊

广东松山职业技术学院, 广东 韶关 512000

DOI: 10.61369/ETR.2026230044

摘 要 : 现实常见的多肉植物不但品种繁杂, 而且形态相近, 并且外观易受外界环境等因素影响而发生变异, 通过人工识别方式低效且精度差, 现有深度学习模型存在优质数据不足而泛化能力弱等问题。本文针对多肉植物数据瓶颈问题, 构建 SPRD 多肉植物数据集, 通过采集、预处理、标注、质检与数据增强全流程, 提出三级标注方案, 用 StyleGAN3 解决罕见样本稀缺问题。将 SPRD 用于三类主流深度学习模型测试, 结果显示识别准确率得到显著提升; 生成式增强技术使罕见品种准确率平均提升 15.7%。

关 键 词 : 多肉植物; 数据集构建; 数据标注; 生成式数据增强; 深度学习; 图像识别

Research on Construction and Application of Succulent Plant Recognition Dataset

Sun Jun

Guangdong Songshan Polytechnic, Shaoguan, Guangdong 512000

Abstract : Common succulent plants in daily life feature a wide variety of species and similar appearances. Their morphology is also prone to variation due to external environmental factors. Manual recognition is inefficient and inaccurate, while existing deep learning models suffer from weak generalization caused by the lack of high-quality data. Aiming at the data bottleneck of succulent plant recognition, this paper constructs the SPRD succulent plant dataset. Following the whole process of data collection, preprocessing, annotation, quality inspection and data augmentation, a three-level annotation scheme is proposed. StyleGAN3 is adopted to address the scarcity of rare samples. The SPRD dataset is applied to tests on three mainstream deep learning models. The experimental results show that the recognition accuracy is greatly improved, and the generative augmentation technology raises the average accuracy of rare species by 15.7%.

Keywords : succulent plants; dataset construction; data annotation; generative data augmentation; deep learning; image recognition

引言

多肉植物凭借造型别致、观赏价值高、养护成本低、兼具空气净化功能等特点, 被广泛应用于家庭园艺、室内装饰、城市绿化与景观设计, 已逐渐成为大众主流观赏植物品类^[1-2]。据中国花卉协会《中国多肉植物产业发展报告(2023)》统计, 我国多肉植物市场年均增速超 20%, 栽培品种突破 800 种, 从业人员超 100 万人, 产业规模持续扩容^[3]。

随着计算机视觉与深度学习技术发展, 植物图像识别已成为农业信息化、园林智能化的研究热点, 市面上相继涌现多种通用植物识别软件工具, 但现有工具对多肉植物的识别精度与泛化能力仍无法满足实际落地需求^{[4][7]}。核心瓶颈在于缺乏高质量、大规模、多场景的标准化标注数据集, 现有小型研究数据集普遍存在品种覆盖少、样本数量不足、应用场景单一等缺陷, 难以还原真实应用场景^{[5][8]}。同时, 数据集标注精度参差不齐、无统一标注规范与共享机制, 严重制约了多肉智能识别技术的产业化应用^[6-7]。

基于以上因素, 本文围绕多肉植物识别数据集(以下简称: SPRD 数据集)构建与应用开展系统研究, 结合生成式数据增强技术突破小样本瓶颈, 在主流深度学习模型中验证数据集的适配效果, 为多肉植物智能化识别提供完整技术方案与数据支撑。

项目信息: 本文系韶关市科技计划项目-支持科技工作者项目《从图像到知识图谱: 多肉植物图像数据集的构建和信息提取的分析研究》(课题编号: 230326138036609)研究成果。

作者简介: 孙俊(1972.06-), 男, 教授, 主要研究方向图像识别, 职业教育, Email: 47325694@qq.com

一、多肉植物识别数据集构建

（一）数据集需求分析与方案设计

1. 需求分析

结合多肉植物的形态特征、生长规律及实际应用场景，结合深度学习模型训练的需求，明确SPRD数据集构建的目标与标准：一是品种覆盖全面性，满足不同应用场景的识别需求；二是样本多样性，需覆盖不同生长周期、光照条件、栽培环境、形态特征的图像，提升模型的泛化能力；三是标注精细化，需实现品种级标注与关键形态特征标注，为模型的精细化识别与特征提取提供支撑；四是数据真实性，以实拍图像为主，合成样本辅助，确保数据贴近实际应用场景，避免合成样本失真对模型训练结果的影响；五是可扩展性，便于后续新增多肉品种、补充样本数据，适应多肉植物识别技术的发展需求。

2. 方案设计

基于上述需求分析，本文制定“多渠道采集→标准化预处理→三级标注→多维度评估→生成式增强”的系统化数据集构建流程：首先通过实地拍摄、公开数据集筛选、合作机构获取三种渠道采集原始数据，确保数据来源广泛、品种覆盖全面；在预处理环节，通过图像清洗、增强、尺寸统一与归一化等操作，提升数据质量，丰富样本多样性；通过“机器初标—人工复核—交叉验证”的三级标注方案，完成品种级与特征级的精细化标注，确保标注精度；然后从样本规模、多样性、标注精度、真实性四个维度构建质量评估指标体系，对初始数据集进行全面评估，针对存在的问题开展针对性优化；最后引入StyleGAN3生成式数据增强技术，补充罕见品种样本，解决小样本数据瓶颈，最终形成标准化的SPRD多肉植物识别数据集。

（二）数据采集

为确保数据来源的广泛性与样本的多样性，数据采集规范如下：一是样本数量规范，每个常见品种采集 ≥ 300 张原始图像，每个罕见品种采集 ≥ 200 张原始图像，确保样本数量满足模型训练需求；二是场景覆盖规范，每种品种在不同生长周期、光照条件、栽培环境、形态状态下的采集样本数 ≥ 50 张，确保样本分布均衡，能够真实模拟实际应用场景；三是图像质量规范，拍摄时保证多肉植株主体清晰，无明显无关背景干扰，避免反光、模糊、失真等问题；四是参数记录规范，同步记录每张图像的拍摄参数与样本信息，确保数据的可追溯性。

（三）数据预处理

为解决采集的原始图像存在重复、模糊、尺寸不一、亮度不均、背景杂乱等系列问题，本文的预处理流程严格遵循“清洗—增强—归一化”的流程。采用Perceptual Hash（感知哈希）算法剔除重复与模糊的图像，避免样本冗余；采用拉普拉斯算子计算图像清晰度，剔除清晰度值低于50的模糊图像。通过人工审核进行无效样本筛选，剔除主体占比不足、严重遮挡、图像失真的无效样本。通过“传统图像增强+深度学习增强”相结合的方式，提升样本多样性以及模型的泛化能力与鲁棒性。通过采用StyleGAN3生成式模型生成合成样本，补充罕见品种样本数量，

解决小样本数据瓶颈。采用双线性插值法缩放，以及像素值归一化，消除像素值差异对模型训练的影响，避免因像素值过大导致模型训练不稳定。

（四）数据标注

数据标注工作选用开源标注工具，标注格式采用VOC2007标准，生成XML格式标注文件，标注文件包含样本路径、品种类别、特征位置等关键信息，可直接适配主流深度学习框架的输入需求；同时，配套生成JSON格式标注文件，为后续数据管理、特征分析以及模型优化提供便利。

（五）数据集质量评估与优化

本文构建样本规模、多样性、标注精度、真实性四个核心维度的数据集质量评估指标体系，如表1所示。

表1 SPRD数据集质量评估指标体系

评估维度	核心指标	评估标准
样本规模	品种样本数、总样本数	常见品种 ≥ 350 张/种，罕见品种 ≥ 250 张/种，总样本数 ≥ 18000 张
多样性	场景覆盖度、形态特征标注覆盖率	光照/生长周期/形态场景覆盖度 $\geq 95\%$ ，形态特征标注覆盖率 $\geq 80\%$
标注精度	品种标注准确率、特征标注准确率	品种标注准确率 $\geq 98\%$ ，特征标注准确率 $\geq 95\%$
真实性	合成样本占比、实拍样本占比	合成样本占比 $\leq 10\%$ ，实拍样本占比 $\geq 90\%$

按照上述评估指标体系，对初始SPRD数据集进行全面评估，针对评估中发现的问题，对数据集进行针对性优化：采用StyleGAN3生成式模型生成罕见品种合成图像，满足样本规模要求。补充极端光照条件（如强光直射、弱光昏暗）与特殊形态状态（如重度轻微、部分遮挡）的样本，提升场景覆盖度以满足多样性要求。对标注准确率未达标的样本进行重新复核，修正错误标注，确保标注精度符合评估标准。

（六）生成式数据增强实现

采用StyleGAN3生成式模型开展数据增强，通过生成高质量合成样本，补充罕见品种样本数量，破解小样本学习瓶颈。

1. 模型训练

选取15种罕见品种的实拍样本作为模型训练集，构建StyleGAN3生成式模型，模型输入分辨率设置为 224×224 像素，与预处理后的样本分辨率保持一致；模型学习率设置为0.0002，训练轮次设置为10000轮，采用Adam优化器进行参数优化，损失函数采用对抗损失与感知损失相结合的方式，确保生成样本的真实性与细节还原度。

2. 样本生成与筛选

模型训练完成后，为每种罕见品种生成100张合成图像，生成过程中控制生成样本的多样性，避免出现重复样本。为确保合成样本的质量，采用SSIM（结构相似性指数）对生成样本与实拍样本进行相似度评估，剔除与实拍图像差异过大、细节失真的样本。

3. 增强效果验证

为验证生成式数据增强的效果，选取5种典型罕见品种，对比

增强前后样本数量与模型识别效果。实验采用 EfficientNet-B0 模型，分别基于增强前、增强后的样本进行训练，测试模型的识别准确率，验证结果如下表 2 所示：

表 2 罕见品种生成式数据增强效果验证表

罕见品种	增强前 样本数 (张)	增强后 样本数 (张)	增强前 准确率 (%)	增强后 准确率 (%)	提升幅度 (%)
龟甲龙	220	300	72.5	91.3	18.8
断崖女王	220	300	75.8	92.6	16.8
弹簧草	220	300	78.2	93.5	15.3
玉扇	220	300	80.1	94.2	14.1
万象	220	300	79.5	93.8	14.3
平均	220	300	78.3	94.0	15.7

二、多肉植物识别数据集应用

本文选取 3 类主流深度学习模型搭建标准化深度学习实验环境，验证 SPRD 数据集的有效性、实用性与通用性，基于 SPRD 数据集开展模型训练与测试，对比分析模型识别性能，同时拓展数据集的应用场景，验证其在实际应用中的价值。

(一) 实验环境搭建

实验环境分为硬件环境与软件环境两部分，均采用当前深度学习研究的主流配置，确保实验的稳定性与高效性。选取 3 类不同级别、不同结构的主流深度学习模型开展实验，覆盖轻量级、中量级、高性能三种模型类型：CNN（基础卷积神经网络）、ResNet50（残差网络），EfficientNet-B0（高效卷积神经网络）。3 种模型的训练参数统一设置，具体参数如下：Batch Size 为 64，学习率为 0.001，优化器采用 Adam 优化器（ $\beta_1=0.9$ ， $\beta_2=0.999$ ， $\epsilon=1e-8$ ），损失函数为交叉熵损失（Cross-Entropy Loss）；训练轮次为 50 轮，学习率衰减策略采用余弦退火衰减，避免模型训练过程中出现过拟合；数据集按照“训练集：验证集：测试集=6:2:2”的比例进行划分；采用早停法（Patience=10），防止模型过拟合，确保模型的泛化能力。

(二) 实验结果与分析

实验通过分析模型的准确率、精确率、召回率、F1 值等核心性能指标，全面验证 SPRD 数据集的有效性 with 实用性。

1. 模型识别性能对比

为验证 SPRD 数据集的质量优势，将 SPRD 数据集与公开植物数据集（PlantNet 多肉子集）分别同 3 种模型进行训练与测试，对比两种数据集训练下模型的核心性能指标，实验结果如下表 3 所示：

表 3 不同数据集下模型识别性能对比表

模型	数据集	准确率 (%)	精确率 (%)	召回率 (%)	F1 值 (%)
CNN	SPRD	89.5	88.2	87.6	87.9
CNN	PlantNet	76.3	74.1	73.5	73.8
ResNet50	SPRD	95.8	95.2	94.9	95.0

ResNet50	PlantNet	82.7	80.5	79.8	80.1
EfficientNet-B0	SPRD	98.2	97.9	97.7	97.8
EfficientNet-B0	PlantNet	86.4	84.2	83.5	83.8

由表 3 可知，基于 SPRD 数据集训练的 3 种模型，其准确率、精确率、召回率、F1 值均显著优于基于 PlantNet 多肉子集训练的模型。分析其原因，主要是 SPRD 数据集品种覆盖全面、样本多样性丰富、标注精度高，能够为模型训练提供更充足、更准确的特征信息。实验结果充分证明了 SPRD 数据集的高质量与有效性。

2. 相似品种识别效果分析

实验选取 5 组形态高度相似的多肉品种，测试 EfficientNet-B0 模型在 SPRD 数据集上的识别准确率，同时分析错误识别的类型，实验结果如下表 4 所示：

表 4 相似品种识别效果分析表

相似品种组	识别准确率 (%)	错误类型及占比 (%)
吉娃娃 vs 桃蛋	96.5	颜色误判 (2.1)、叶形误判 (1.4)
仙人球 vs 仙人掌	97.2	株型误判 (1.8)、细节纹理误判 (1.0)
玉露 vs 寿	95.8	叶尖形态误判 (2.5)、窗面纹理误判 (1.7)
熊童子 vs 熊爪	96.9	爪尖数量误判 (1.9)、叶色误判 (1.2)
紫珍珠 vs 雨燕座	95.5	颜色渐变误判 (2.3)、株型紧凑度误判 (2.2)
平均	96.4	—

由表 4 可知，5 组相似品种的平均识别准确率达 96.4%。错误识别类型主要集中在颜色、叶形、纹理、株型等细微特征。实验结果表明，SPRD 数据集的形态特征精细化标注能够有效帮助模型提取相似品种的细微差异特征，提升模型对相似品种的区分能力，验证了精细化标注的有效性。

3. 罕见品种识别效果分析

为进一步验证生成式数据增强技术的有效性，对 15 种罕见品种进行对比分析，实验采用 EfficientNet-B0 模型，基于增强前、后的样本进行训练与测试，统计每种罕见品种的识别准确率及提升幅度，部分典型品种的实验结果如下表 5 所示：

表 5 罕见品种增强前后识别准确率对比表

罕见品种	增强前准确率 (%)	增强后准确率 (%)	提升幅度 (%)
龟甲龙	72.5	91.3	18.8
断崖女王	75.8	92.6	16.8
弹簧草	78.2	93.5	15.3
玉扇	80.1	94.2	14.1
万象	79.5	93.8	14.3
平均	78.3	94.0	15.7

由表 5 可知，生成式数据增强后，15 种罕见品种的识别准确率均得到显著提升，平均提升幅度达 15.7 个百分点。实验结果验证了生成式数据增强技术能够有效破解罕见品种小样本学习困境，提升模型对罕见品种的识别性能。

三、结论

本文构建了包含50个品种、18600张图像的SPRD多肉植物识别数据集，通过标准化构建流程，采用StyleGAN3解决罕见品种小样本问题。实验表明，该数据集可使EfficientNetB0识别准

确率达98.2%，罕见品种准确率平均提升15.7%，可为多肉智能识别、园艺管理、品种选育等提供高质量数据支撑。进一步研究可通过扩大品种覆盖、补充多季节与病虫害样本，优化标注体系与生成模型，构建更大规模、更高精度的多肉专用数据集。

参考文献

[1] 杨伽飞, 郝转. 多肉植物的特点及应用价值研究进展 [J]. 现代园艺, 2020(19):112-114.
[2] 胡莹冰, 沈守云. 多肉植物的景观应用及发展趋势 [J]. 广东农业科学, 2013, 40(12):46-48.
[3] 贝哲斯咨询. 全球及中国多肉植物行业: 2023年市场供需现状与发展前景分析报告 [R]. 长沙: 湖南贝哲斯信息咨询有限公司, 2023.
[4] 王永财, 万华伟, 高吉喜. 基于深度学习的我国北方常见天然草地植物识别 [J]. 生物多样性, 2024, 32(7):56-65.
[5] 袁敏, 李翔, 张磊. 细粒度植物种子识别数据集构建与应用 [J]. 农业大数据学报, 2024, 6(2):45-53.
[6] 陈艺荃, 罗远华, 吴建设. 景天科多肉植物品种资源评价体系构建及应用 [J]. 福建农业学报, 2024, 39(3):320-329.
[7] 宋琰, 杨兰海, 杜忠友. 深度学习在多肉植物图像识别中的应用进展 [J]. 计算机工程与应用, 2023, 59(18):1-8.
[8] 李昊, 孙文博. 小样本植物图像数据增强方法研究 [J]. 智能计算机与应用, 2023, 13(5):78-83.