www.csdata.org

ISSN 2096-2223
CN 11-6035/N

2020–2021 年新疆荒漠植物深度学习识别图像数据集

王亚鹏[1]，李全胜[1]，古丽米拉·克孜尔别克[1]，闫燊[2,3,4]，刘婷婷[2,3,4]，孙伟[2,3,4*]，曹姗姗[2,3,4*]

1. 新疆农业大学，乌鲁木齐　830052
2. 中国农业科学院农业信息研究所，北京　100081
3. 国家农业科学数据中心，北京　100081
4. 中国农业科学院国家南繁研究院，海南三亚　572024

摘要： 荒漠植物类型的机器视觉自动识别可为防风固沙、生态系统价值评估、植被恢复与重建等研究提供技术支持，减少对植物专家鉴定的依赖。目前荒漠植物的机器判别模型研究主要依靠标准化处理过的高质量植物标本图片，缺少复杂自然条件下获取的荒漠植物图像。本数据集提供了可用于深度学习图像分类模型训练的新疆典型荒漠植物图像，包含不同季节、自然背景和光照条件下获取的 15550 张新疆荒漠植物数码相机图片，涵盖 19 种典型荒漠植物类型，碱蓬图像数量最少，沙蒿最多，分别为 465 张和 1240 张，中位数为 800，已满足主流深度学习模型训练需要。本数据集可为荒漠植物图像分割、目标检测、自动识别等研究提供基础数据。

关键词： 荒漠植物；识别；标准图库；训练样本

文献 CSTR：
32001.14.11-6035.nasdc.2021.0050.zh
文献 DOI：
10.11922/11-6035.nasdc.2021.0050.zh
数据 DOI：
10.12205/A0007.20211123.32.is.1918
文献分类：农学

收稿日期：2021-11-01
开放同评：2021-11-29
录用日期：2022-03-21
发表日期：2023-02-17

数据库（集）基本信息简介

数据库（集）名称	2020–2021 年新疆荒漠植物深度学习识别图像数据集
数据作者	孙伟
数据通信作者	孙伟（sunwei2@caas.cn）；曹姗姗（caoshanshan@caas.cn）
数据时间范围	2020-2021年
地理区域	新疆准噶尔盆地
数据量	8.54 GB
数据格式	*.JPG
数据服务系统网址	http://dx.doi.org/10.12205/A0007.20211123.32.is.1918
基金项目	国家自然科学基金（32060321、41807005）；新疆维吾尔自治区2020年度创新环境（人才、基地）建设专项（PT2012）
数据库（集）组成	数据集包含新疆准噶尔盆地荒漠植物的图像15550张，涵盖19种荒漠植物，包括碱蓬、蛇麻黄、沙拐枣、沙漠绢蒿、花花柴、白刺、费尔干猪毛菜、盐节木、叉毛蓬、骆驼刺、沙蓬、梭梭、类碱蓬、琵琶柴、角果藜、怪柳、骆驼绒、三芒草及沙蒿。

* 论文通信作者
孙伟：sunwei2@caas.cn
曹姗姗：caoshanshan@cass.cn

引　言

荒漠植被虽然稀疏、生物量低下，但由这些适应荒漠生境的植物构成的荒漠植被，既维持着荒漠区域能量与物质的循环过程，又防止了风蚀、流沙和进一步荒漠化，并为人类提供木材、药材、燃料、饲料、肥料以及其他以及其它副产品，是荒漠生态系统的核心，具有重要的生态和经济意义[1]。野外调查人员经常不能准确、快速地获取荒漠植物的学名、科属及性状等信息，给植物鉴定和保护研究带来了一定的困难。因此，植物准确及快速识别是植物识别和保护过程中不可或缺的任务。

传统上，植物学家将荒漠植物的不同特征（视觉和非视觉）作为识别因子。然而，由于自然图像的复杂性、荒漠植物物种的多样性、物种间的相似性以及大规模的外观变异性，从不可控的自然图像中识别荒漠植物是一个具有挑战性的问题。根据视觉特征准确地识别荒漠植物种类需要相当多的专业知识和长期的实践经验，甚至需要掌握相应的关键技术才能有效地完成分类，这对普通大众来说几乎是不可能的，甚至对专家来说也是富有挑战性的。近年来，伴随着大量标记数据集的出现以及图形处理器（GraphicalProcessingUnit，GPU）计算能力的发展，深度学习[2]方法被广泛应用于计算机视觉领域，识别效果显著，同时也解放了人力。

在计算机辅助植物识别中，常以叶片图像作为研究对象，基于机器视觉的叶片图像识别已成为主流方法[3]，如 Cope 等[4]采用 Gabor 滤波器提取纹理特征对 32 种叶片进行识别，识别率为 85.16%。另有通过花朵图像进行植物识别，如袁培森等[5]采集了大量菊花图像数据进行标注和分类，研究构建了一个 6 层的卷积神经网络，实现了菊花品种智能识别。通过对当下植物识别技术的分析研究，发现已有植物图像识别局限于单一背景或实验室环境，且研究对象单一，不包含植物的整体特征，不适用于真实场景。

本研究基于荒漠植物准确及快速识别的任务，考虑到结合深度学习方法需要大量的标记数据集，在新疆噶尔盆地采集了包含猪毛菜、琵琶柴、类碱蓬等 19 种荒漠植物图像资源，建立了一个可以为深度学习建模提供训练样本和测试样本的荒漠植物识别图像数据集，填补了当前我国荒漠植物相关数据资源的空白。同时，数据集中的荒漠植物图像包含复杂背景，例如光照、土壤、沙砾等，且拍摄的是荒漠植物的整体图像，非植物的单一特征，使训练出的模型具有泛化能力，更适用于真实场景。

自 2011 年起，CLEF Initiative 实验室[6]每年组织基于图像分析的植物信息识别竞赛，可见植物图像识别已成为植物分类学和计算机视觉领域的一个跨学科研究热点。本数据集可填补在机器视觉领域荒漠植物数据资源的空缺，具有较高的实用性和价值性，为植物图像分类、目标检测及图像分割等研究提供基础数据。

1　数据采集和处理方法

项目组于 2020 年 5 月至 2021 年 9 月在新疆准噶尔盆地进行荒漠植物生态照片采集。图像采集及处理工作由从事荒漠植物图像识别研究的专业技术人员按照如下标准操作流程进行。

（1）根据《中国植物志》[7]、《新疆植物志》[8]及植物通（http://1.zhiwutong.com/）确定新疆现有的荒漠植物种类以及准噶尔盆地现有的荒漠植物种类，确定待采集的荒漠植物种类，预计每种荒漠植物的图像数量在 800 张左右，保证数据采集量。图像采集设备为 Canon EOS 90D 型数码单反相

机，配备佳能 EF-S 18-200mmf/3.5-5.6 IS 镜头与佳能 EF 100mmf/2.8L IS USM 微距镜头，以及索尼 A105 数码相机。

（2）拟定如下数据采集标准，进行数据采集。使用上述采集设备拍摄荒漠植物图像，保证照片质量。拍摄图像时采用 2400×1600 分辨率，从多种角度、多种光线下进行拍摄，所拍摄的荒漠植物占据图像的中央主要位置。分多季度前往准噶尔盆地进行野外实地考察，拍摄不同季节下的荒漠植物图像，充分采集荒漠植物图像。

（3）荒漠植物种类的鉴定、筛选及构建数据集。数据采集完成后需要对荒漠植物物种鉴定，过程中听取、整合各类专家意见，保证荒漠植物信息的准确性。删除质量差、无效以及无法鉴定的植物图像，构建新疆荒漠植物深度学习识别图像数据集。

2 数据样本描述

调查得到的新疆准噶尔盆地荒漠植物图像数据集，数据集按照荒漠植物名称建立文件夹，共有 19 个文件夹，每张图像代表一个数据样本。数据集中的部分样本示例如图 1 所示。

（a）琵琶柴 *Reaumuria soongorica* (Pall.) Maxim

（b）类碱篷 *Suaeda glauca* (Bunge) Bunge

（c）叉毛篷 *Petrosimonia sibirica* (Pall.) Bunge

（d）角果藜 *Ceratocarpus arenarius* Linn

图 1 荒漠植物图像数据集中的样本示例

Figure 1 Dataset sample of desert plant images

荒漠植物图像的统计数据如图 2 所示，19 种荒漠植物的图像数量在 465–1240 张之间，中位数为 800，已满足主流深度学习模型训练需要。

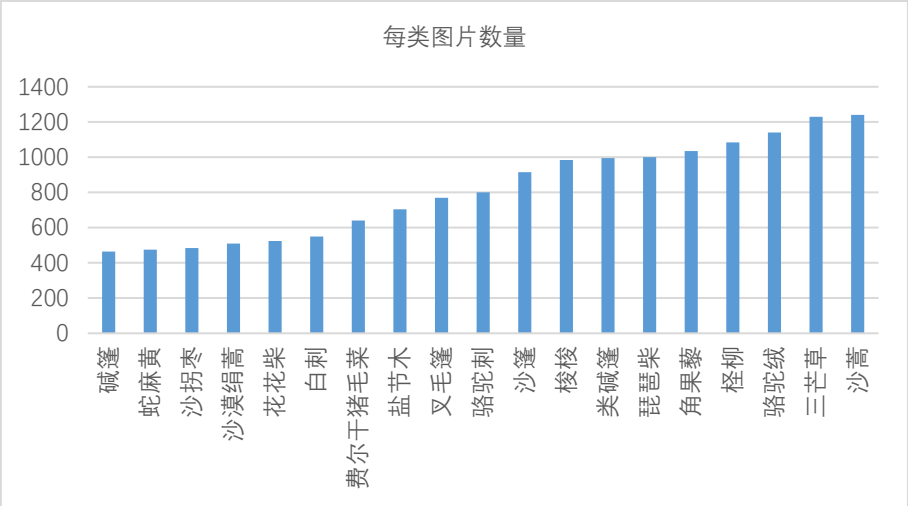

图 2　荒漠植物图像数据分布

Figure 2　Distribution of desert plant image data

3　数据质量控制和评估

本研究中，图像采集和处理工作由专业技术人员按照上述标准操作流程进行。为保证数据质量要求，采取以下措施：（1）调查现场拍摄图像并记录物种名称，确保植物种类鉴定有凭可依。对于鉴定模糊的物种，根据专家意见对该物种信息进一步修改和完善。（2）图像采集时需注意多角度、多光线下拍摄，且包含植物整体图像和部分细节图像，保证拍摄的荒漠植物占据图像的中央主要位置，保证数据采集量，达到数据采集标准。（3）对采集的所有图像进行人工筛选，剔除不符合要求的图像，保证本数据集中荒漠植物图像数据的质量和可靠性。

4　数据价值

本数据集通过野外调查，获得真实、原始的荒漠植物图像数据资源。与现有荒漠植物图谱相比，本数据集中的每种荒漠植物有几百乃至上千张图像，随着本数据集的不断更新发展，今后将建设成为国内标准的荒漠植物图像识别研究数据资源，为相关研究人员提供统一的训练与测试数据。通过深度学习方法建立相关的计算机视觉识别模型，推动荒漠植物图像识别研究的发展。

致　谢

感谢新疆农业大学草业学院张鲜花副教授、安沙舟教授对荒漠植物图像采集、鉴定与分类的大力支持！

数据作者分工职责

王亚鹏（1996—），男，河南商丘人，硕士，研究生，研究方向为农林信息化。主要承担工作：数据汇总整理及论文撰写。

李全胜（1979—），男，新疆库尔勒市人，硕士，讲师，研究方向为农业信息化。主要承担工作：数据采集及汇总整理。

古丽米拉·克孜尔别克（1970—），女，新疆乌鲁木齐市人，硕士，副教授，研究方向为农业信息化、计算机应用。主要承担工作：数据汇总整理。

闫桑（1987—），男，山东省菏泽市人，博士，助理研究员，研究方向为农业大数据。主要承担工作：数据汇总整理。

刘婷婷（1985—），女，北京人，硕士，助理研究员，研究方向为科学数据管理。主要承担工作：数据汇总整理。

孙伟（1978—），男，山东海阳人，博士，副研究员，研究方向为农林时空信息智能分析。主要承担工作：总体方案设计与组织实施。

曹姗姗（1984—），女，黑龙江哈尔滨人，博士，副研究员，研究方向为农林时空信息智能分析。主要承担工作：总体方案设计与组织实施。

参考文献

[1] 尹林克. 中国温带荒漠区的植物多样性及其易地保护[J]. 生物多样性, 1997(01): 40-48. [YIN L K. Species Catalogue of China: a remarkable achievement in the field of biodiversity science in China[J]. Biodiversity Science, 1997(01): 40-48.]

[2] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[3] 李化乐. 基于植物图像特征的识别研究[D]. 兰州大学, 2017. [LI H L. Study on plant identification based on image features[D]. Lanzhou University, 2017.]

[4] COPE J S, REMAGNINO P, BARMAN S, et al. Plant texture classification using gabor co-occurrences[C]//International Symposium on Visual Computing. Springer, Berlin, Heidelberg, 2010: 669-677.

[5] 袁培森, 黎薇, 任守纲, 等. 基于卷积神经网络的菊花花型和品种识别[J]. 农业工程学报, 2018, 34(05):152-158. [YUAN P S, LI W, REN S G, et al. Recognition for flower type and variety of chrysanthemum with convolutional neural network[J]. Transactions of the Chinese Society of Agricultural Engineering, 2018, 34(05):152-158.]

[6] CLEF Initiative[EB/OL]. 2017. (2017-07-25). http://www.clef-initiative.eu.

[7] 中国科学院植物研究所系统与进化植物学国家重点实验室.《中国植物志》网络版[EB/OL]. 2019-08-01. http://www.iplant.cn/frps. [State Key Laboratory of systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences. Flora of China online edition[EB/OL]. 2019- 08-01. http://www.iplant.cn/frps.]

[8] 安争夕. 新疆植物志[M]. 新疆科技卫生出版社, 1999. [AN Z X. Flora of Xinjiang[M]. Xinjiang science and Technology Health Publishing House, 1999.]

论文引用格式

王亚鹏, 李全胜, 古丽米拉·克孜尔别克, 等. 2020–2021 新疆荒漠植物深度学习识别图像数据集[J/OL]. 中国科学数据, 2023, 8(1). (2023-02-17). DOI: 10.11922/11-6035.nasdc.2021.0050.zh.

数据引用格式

孙伟. 2020–2021 新疆荒漠植物深度学习识别图像数据集[DS/OL]. 中国农业科学院农业信息研究所. 国家农业科学数据中心, 2021. (2021-11-24). DOI: 10.12205/A0007.20211123.32.is.1918.

A dataset of desert plant images for deep learning recognition in Xinjiang in 2020–2021

WANG Yapeng[1], LI Quansheng[1], Gulimila KEZIERBIEKE[1], YAN Shen[2,3,4], LIU Tingting[2,3,4], SUN Wei[2,3,4*], CAO Shanshan[2,3,4*]

1.Xinjiang Agricultural University, Urumqi 830052, P.R.China

2.Agricultural Information Institute of CAAS, Beijing 100081, P.R.China

3.National Agriculture Science Data Center, Beijing 100081, P.R.China

4.National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya 572024, P.R. China

*Email: sunwei02@caas.cn; caoshanshan@caas.cn

Abstract: Automatic recognition of desert plant types by machine vision can support the research on wind prevention and sand fixation, ecosystem value assessment, vegetation restoration and reconstruction, and reduce the dependence on plant expert identification. At present, the research on the machine discrimination model of desert plants mainly relies on the standardized high-quality plant specimen images, lacking the desert plant images obtained under complex natural conditions. This dataset provides typical desert plant images of Xinjiang that can be used for the model training of deep learning image classification, including 15,550 digital camera images of desert plants in Xinjiang obtained under different seasons, natural backgrounds and lighting conditions, and covering 19 typical desert plant types. *Suaeda salsa* has the smallest number of images and *Artemisia desertorum* has the biggest, 465 and 1,240 respectively, with a median of 800, which has met the training needs of mainstream deep learning model. This dataset can provide basic data for desert plant image segmentation, target detection and automatic recognition.

Keywords: desert plants; identification; standard image dataset; training sample

Dataset Profile

Title	A dataset of desert plant images for deep learning recognition in Xinjiang in 2020–2021
Data corresponding author	SUN Wei (sunwei2@caas.cn); CAO Shanshan (caoshanshan@caas.cn)
Data authors	SUN Wei
Time range	2020–2021
Geographical scope	Junggar Basin, Xinjiang
Data volume	8.54 GB
Data format	*.JPG
Data service system	<http://dx.doi.org/10.12205/A0007.20211123.32.is.1918>
Sources of funding	National Natural Science Foundation of China (32060321, 41807005); Special Project for the Construction of Innovation Environment (Talent, Base) in Xinjiang Uygur Autonomous Region in 2020 (PT2012).
Dataset composition	The dataset contains 15,550 images of desert plants in Junggar basin, Xinjiang, covering 19 kinds of desert plants, namely *Suaeda glauca*, *Ephedra distachya*, *Calligonum mongolicum*, *Seriphidium santolinum*, *Karelinia caspia*, *Nitraria tangutorum*, *Salsola ferganica*, *Halocnemum strobilaceum*, *Petrosimonia sibirica*, *Alhagi sparsifolia*, *Agriophyllum squarrosum*, *Haloxylon ammodendron*, *Suaeda salsa*, *Reaumuria soongorica*, *Ceratocarpus arenarius*, *Tamarix taklamakanensis*, *Peganum harmala*, *Aristida adscensionis*, and *Artemisia desertorum*.