

大豆室内考种图像数据集

www.sciengine.com/CSD

ISSN 2096-2223

CN 11-6035/N



文献 CSTR:

32001.14.11-6035.csd.2023.0104.zh



文献 DOI:

10.11922/11-6035.csd.2023.0104.zh

数据 DOI:

10.57760/sciencedb.08286

文献分类: 生物学

收稿日期: 2023-05-15

开放同评: 2023-05-16

录用日期: 2023-11-08

发表日期: 2024-08-21

张哲^{1,2,3}, 饶元^{1,2,3*}, 束雅莉^{1,2,3}, 陈浩然⁴, 朱尚尚⁴, 王晓波⁴,金秀^{1,2,3}, 王丰仪^{1,2,3}, 李佳佳⁴, 徐文强^{1,2,3}, 吴康磊^{1,2,3}, 王安然^{1,2,3}

1. 安徽农业大学信息与人工智能学院, 合肥 230036
2. 农业农村部农业传感器重点实验室, 合肥 230036
3. 智慧农业技术与装备安徽省重点实验室, 合肥 230036
4. 安徽农业大学农学院, 合肥 230036

摘要: 选育优质高产品种是提升大豆品质和产量的重要保障。传统大豆考种方法存在费时费力、数据误差大等问题, 距大豆表型数据的一站式高通量获取尚有不小的差距。因此, 构建大豆室内考种图像数据集, 对于开展大豆植株表型信息的高通量准确获取方法研究, 实现大豆的自动智能考种尤为重要。本数据集包含原始图像数据、标注文件数据、测试图像数据三部分, 涵盖了典型大豆品种的离体豆荚和植株, 其中离体豆荚包括无重叠离体豆荚和重叠离体豆荚, 大豆植株包括单分枝型、双分枝型和复杂分枝型; 标注文件数据包含了大豆植株主茎的实例分割标注、主茎节的检测框标注、在体豆荚的检测框标注、离体豆荚的实例分割和检测框标注, 其中在体豆荚和离体豆荚的检测框标注采用了正框和旋转框两种标注方式, 共计 4.2 GB。通过本数据集进行训练和验证的模型在无重叠离体豆荚、重叠豆荚以及大豆植株形态等各项指标上均表现出了良好的检测和分割效果。本数据集可为大豆植株的离体豆荚检测、在体豆荚检测与植株形态分析的目标检测和实例分割模型方法研究提供宝贵的基础图像数据资源, 并且对于促进大豆关键考种信息的一站式高通量获取等智能化考种方法研究也具有重要价值。

关键词: 大豆; 智能考种; 目标检测; 实例分割

数据库(集)基本信息简介

数据库(集)名称	大豆室内考种图像数据集
数据通信作者	饶元 (raoyuan@ahau.edu.cn)
数据作者	张哲、饶元、束雅莉、陈浩然、朱尚尚、王晓波、金秀、王丰仪、李佳佳、徐文强、吴康磊、王安然
数据时间范围	2022年9月至11月
地理区域	安徽合肥
数据量	4.2 GB
数据格式	*.JPG, *.json, *.xml
数据服务系统网址	https://doi.org/10.57760/sciencedb.08286

* 论文通信作者

饶元: raoyuan@ahau.edu.cn

基金项目	国家自然科学基金(32371993); 安徽省重点研究和开发计划面上攻关项目(2023n06020057, 202204c06020022); 安徽省高校自然科学研究重大项目(2022AH040125)。
数据库(集)组成	本数据集由3部分数据组成, 大豆植株数据、离体豆荚数据、测试图像数据。 (1) 大豆植株数据包含4个文件夹: bean-height、bean-stalk、bean-pod、bean-image。前3个文件夹各包含了2个子文件夹 Annotations-VOC 和 Annotations-COCO, 分别存放大豆植株主茎、主茎节、在体豆荚的 VOC 格式标注文件和 COCO 格式标注文件, bean-image 存放了大豆植株可见光图像, 以 JPG 格式保存; (2) 离体豆荚数据包含2个文件夹: nocover、cover, 这2个文件夹各包含了3个子文件夹 Annotations-VOC、Annotations-COCO、pod-image, 分别存放离体豆荚的 VOC 格式标注文件、COCO 格式标注文件和离体豆荚可见光图像, 其中离体豆荚图像以 JPG 格式保存; (3) 测试图像数据包含两个文件夹: test-soybean、test-pod, 分别存放用于测试的大豆植株可见光图像和离体豆荚可见光图像, 均以 JPG 格式保存。

引言

大豆作为中国重要的粮食和油料作物之一, 其高度依赖进口的现状已经给国家粮油安全带来了显著的威胁^[1-2]。因此, 选育国产优质高产品种已成为提升我国大豆质量、摆脱国外种源依赖的关键任务。近年来, 中国正致力于培育高产大豆新品种^[3], 在寻求增产的过程中, 除了通过各种方法培育优质植株外, 还需要对大豆植株性状进行测量与分析, 以评估不同品种的表型信息^[4-5]。对大豆表型信息进行详细的观察和记录(即考种)是全面评价新品种的一个重要步骤。选育大豆品种的过程中需要对其各类表型信息进行全方位实验, 比如株高、主茎节数、单株荚数等。大豆产量主要受三个因素的影响: 单株荚数、每荚粒数和籽粒大小。其中, 单株荚数是最关键的参数, 传统的大豆豆荚计数主要是依靠人工。由于大豆形态各异, 且豆荚易被植株上的枝条和其他豆荚遮挡, 导致人工计数不仅容易出错, 且耗时耗力。

随着智能化与信息化技术不断在农业生产中的应用, 农业生产领域正逐渐向智慧农业领域转变^[6-8]。为了提高作物考种效率, 减轻人力工作量, 研究人员开展了一系列的探索。运用机器视觉技术和图像处理技术对农作物图像的研究一直是农作物自动化考种领域的研究热点^[9-12]。陈姝^[13]为解决在特征提取网络中因连续下采样导致大豆豆荚特征丢失的问题, 提出了一种基于 D-YOLOV3 算法的策略来精确计量大豆单株的荚数。Xiang S 等^[14]提出了一种基于 YOLO X 框架的方法, 引入了卷积块注意模块, 在不降低推理速度的情况下实现了豆荚的准确识别和计数。Li Y 等^[15]结合了基于密度估计和卷积神经网络的技术, 实现了从单一视角对单个大豆豆荚图像中的种子数进行精确估计。Yang S 等^[16]针对豆荚在图像中重叠密集的情况下, 提出了一种基于人工合成离体豆荚数据集的迁移学习在植株上豆荚的原位分割方法, 实现了植株上豆荚的高精度分割。

然而, 受限于当前大豆植株图像数据多样性不足、背景复杂等情况, 不同品种大豆在颜色和形态上的差异导致其在目标检测和分割算法中的识别准确性上存在显著差别^[17-20]。为了实现大豆的高效、智能化考种, 并满足目标检测和分割算法的准确率要求, 构建一个专门针对大豆室内考种的图像数据集显得尤为重要。这不仅有助于研究大豆植株关键表型信息的高通量准确获取方法, 也是实现大豆自动智能考种的关键一步。

因此,本文提出了一个大豆室内考种图像数据集。本数据集包括原始图像数据、标注文件数据和测试图像数据三部分,涵盖了典型大豆品种的离体豆荚和植株。本数据集旨在为大豆植株的离体豆荚、在体豆荚及植株形态等表型信息分析的目标检测和分割模型方法研究提供宝贵的基础图像数据资源,同时也为推进大豆关键表型信息的一站式高通量获取等智能化考种方法研究提供重要支撑。

1 数据采集和处理方法

本研究中大豆数据的获取过程具体可分为两个步骤:数据采集和数据标注。

1.1 数据采集

本研究中选种及收获均在安徽农业大学宿州皖北综合试验站大豆育种基地完成,选取的大豆品种均为国内大豆主产区主栽品种,大豆植株的采集集中于收获期(2022年9月份)进行。

为建立大豆室内考种图像数据集,本研究在自主搭建的大豆图像表型检测平台上对成熟期的大豆植株以及离体豆荚进行了数据采集。该检测平台配备了佳能 6D Mark II 型相机和佳能 EF 24–105 mm 1:4 L IS USM 型变焦镜头,相机被垂直固定于距离放置样本的采集平台 2000 mm 处。在拍摄过程中,图像的分辨率设置为 5472×3648 像素,拍摄时的镜头焦段为 55 mm,光圈值为 f 11,曝光时间为 0.005 s,ISO 为 100,曝光补偿为 0。

图像采集过程中,大豆植株被平铺放置在检测台上,银色直尺作为标定物,黑色吸光布作为背景,以降低背景复杂度对实验带来的干扰。相机与植株的垂直距离保持在 80 cm 左右,拍摄角度垂直于植株表面,确保植株表面受光均匀,以保证图像采集质量。大豆植株图像采集完成后,人工剥离植株上的豆荚,并将剥落后的豆荚平铺在检测平台的遮光布上,在距离豆荚 80 cm 左右的垂直距离进行拍摄,分别拍摄了无重叠离体豆荚图像和重叠离体豆荚图像。

1.2 数据标注

在本研究中,为了满足不同的标注需求并确保标注的完整性和准确性,选用了 LabelImg、LabelMe 和 rolabelImg 三种标注软件。LabelImg 提供了直观的用户界面,适合常规的矩形框标注。因此,在本数据集中,使用 LabelImg 对在体豆荚、离体豆荚和主茎节进行正框标注。LabelMe 则以其对多边形等多种标注形式的支持,适用于较为复杂的实例分割标注任务,故在本数据集中,使用 LabelMe 对大豆主茎进行多边形标注。rolabelImg 则由于具备旋转边界框功能,有助于处理需要识别对象方向的标注任务,所以被用于在体豆荚和离体豆荚的旋转框标注。考虑到与现代目标检测和图像分割算法的兼容性,本研究采用了广泛被认可的 PASCAL VOC 和 MS COCO 两种标注格式。

2 数据样本描述

本数据集的文件组织形式如表 1 所示,共包括大豆植株数据、离体豆荚数据、测试图像数据三部分数据。其中,大豆植株数据存放了用于本次实验的大豆植株图像及大豆植株主茎、主茎节数、在体豆荚的 VOC 格式标注文件和 COCO 格式标注文件;离体豆荚数据存放了用于本次实验的离体豆荚图像及离体豆荚的 VOC 格式标注文件、COCO 格式标注文件;测试图像数据存放了可用于测试的大豆植株图像和离体豆荚图像,均以 JPG 格式保存。

表 1 数据集文件组织形式

Table 1 Dataset file organization

文件夹	子文件夹				文件内容/说明
大豆植株数据	RGB/*.JPG				大豆植株图像目录
	bean_stalk	Annotations-VOC		*.xml	PASCAL VOC 格式标注文件
		Annotations-COCO		*.json	MS COCO 格式标注文件
	bean_height	Annotations-VOC		*.xml	PASCAL VOC 格式标注文件
		Annotations-COCO		*.json	MS COCO 格式标注文件
	bean_pod	Annotations-VOC	nomal	*.xml	PASCAL VOC 格式标注文件
			rotation	*.xml	PASCAL VOC 格式标注文件
		Annotations-COCO	nomal	*.json	MS COCO 格式标注文件
			rotation	*.json	MS COCO 格式标注文件
	bean_image	*.JPG			大豆植株图像目录
离体豆荚数据	nocover	Annotations-VOC	nomal	*.xml	PASCAL VOC 格式标注文件
			rotation	*.xml	PASCAL VOC 格式标注文件
		Annotations-COCO	nomal	*.json	MS COCO 格式标注文件
			rotation	*.json	MS COCO 格式标注文件
		pod_image	*.JPG		无重叠离体豆荚图像目录
	cover	Annotations-VOC	nomal	*.xml	PASCAL VOC 格式标注文件
			rotation	*.xml	PASCAL VOC 格式标注文件
		Annotations-COCO	nomal	*.json	PASCAL VOC 格式标注文件
			rotation	*.json	PASCAL VOC 格式标注文件
		pod_image	*.JPG		重叠离体豆荚图像目录
测试图像数据	test_soybean	category_1	*.JPG		大豆植株品种 1 测试图像目录
		category_2	*.JPG		大豆植株品种 2 测试图像目录
		...	...		...
		category_n	*.JPG		大豆植株品种 n 测试图像目录
	test_pod	nocover	*.JPG		无重叠离体豆荚测试图像目录
		cover	*.JPG		重叠离体豆荚测试图像目录

2.1 图像数据样本

数据集中大豆植株和离体豆荚包含有 RGB 图像数据，分辨率为 5472×3648 像素，8 位深，每张 RGB 图像约 1.3 MB。图像文件名按序对应从 1.JPG 开始依次递增至 N.JPG 编号，其中 N 为对应图像数据的总数量。

数据集中包含了单分枝大豆植株、双分枝大豆植株、复杂分枝大豆植株等多种类别的大豆植株和不同重叠程度下的离体豆荚，以保证图像数据集的多样性。将豆荚从植株上人工脱落后，由于在视觉上豆荚之间无显著差异，故在本数据集中对于离体豆荚的分类仅分为无重叠离体豆荚和重叠离

体豆荚两类。图 1-2 显示了离体豆荚及大豆植株图像的示例：包括无重叠离体豆荚、不同重叠情况下的离体豆荚和单分枝大豆植株、双分枝大豆植株、复杂分枝大豆植株。

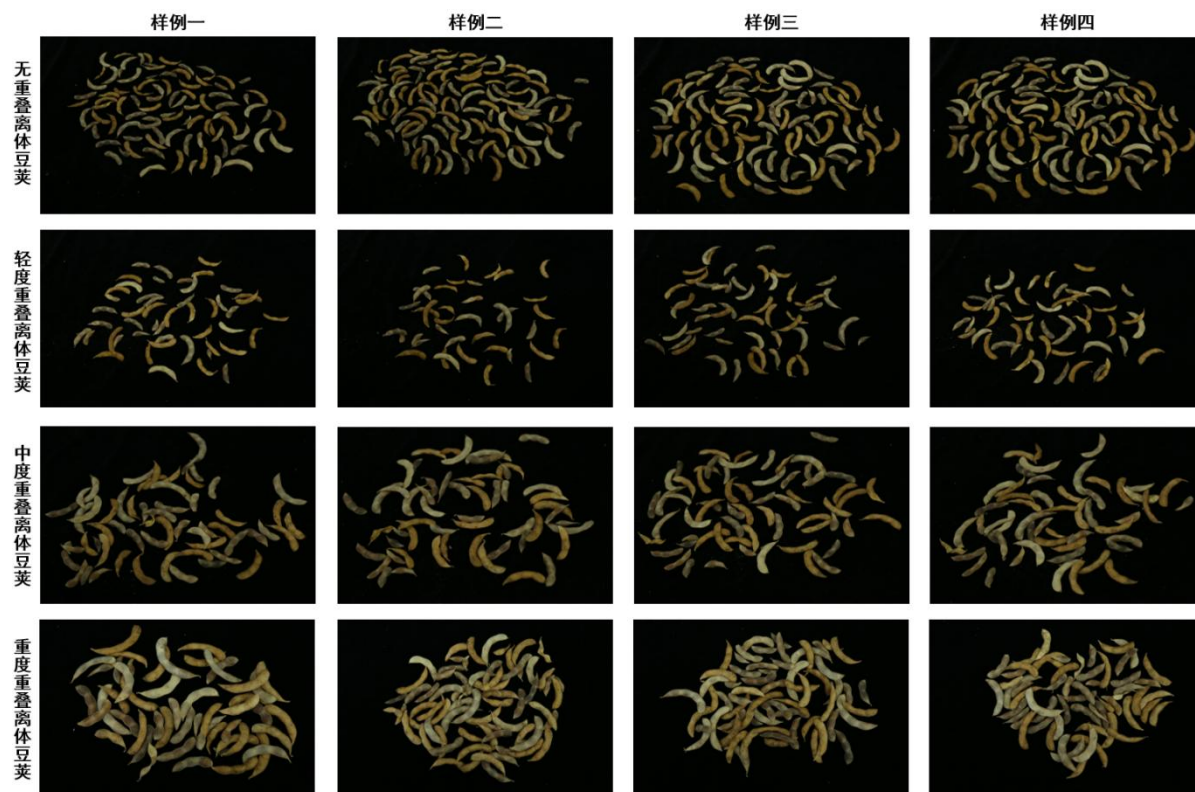


图 1 离体豆荚图像示例

Figure 1 Representative images of in-vitro pods



图 2 大豆植株图像示例

Figure 2 Representative images of soybean plants

2.2 标注数据样本

本研究基于大豆考种的因素与特点，在收集并整理好拍摄的图像后，需要对所得到的图像进行数据标注。本研究使用的数据标注软件为 LabelImg、LabelMe 和 roLabelImg。每种标注方法都生成 PASCAL VOC 格式标注文件和 MS COCO 格式两种格式的标注文件。其中，PASCAL VOC 格式的标注结果以.xml 格式保存，MS COCO 格式的标注结果以 json 格式保存。数据集中图像样本可通过旋转、放缩、错切、平移、高斯模糊、亮度调整、对比度调整、饱和度调整等手段进一步增加数据量。

(1) 离体豆荚标注方面。对离体豆荚分别进行检测框标注和实例分割标注，标注时需要将豆荚进行分类标注为 zero、one、two、three、four，分别对应 0 粒荚、1 粒荚、2 粒荚、3 粒荚、4 粒荚。由于 5 粒荚罕见且数量较少，非大豆植株考种必须选择，通常育种专家更关注 0 粒荚到 4 粒荚数量与识别准确率，并作为考种的主要依据。对离体豆荚进行检测框标注时，使用 LabelImg 软件打开离体豆荚图像，通过矩形框对图像中的豆荚进行检测框标注。每个标注完成的矩形框会生成对应的 XML 文件，该文件包含矩形框的中心坐标及边框的长、宽信息。对离体豆荚进行实例分割标注时，使用 LabelMe 软件打开离体豆荚图像，沿着每颗豆荚的边缘进行标注。标注完成后会生成对应的 JSON 文件，该文件包含每个豆荚的边界多边形 (mask) 以及 labels。同时，为提高离体豆荚标注样本的多样性，本数据集对不同重叠情况下的豆荚分别采用不同的标注策略。对于无重叠豆荚的样本，每个豆荚作为一个独立的分割样本；对于豆荚间存在重叠现象的样本，视野可见的各部分都作为独立的分割样本，且为确保算法能够正确识别来自同一豆荚的不同样本，在 LabelMe 软件中为同一豆荚的每个独立样本分配了相同的 ID。图 3 显示了无重叠离体豆荚和重叠离体豆荚图像的标注示例：包括离体豆荚的实例分割标注、正框标注、旋转框标注。

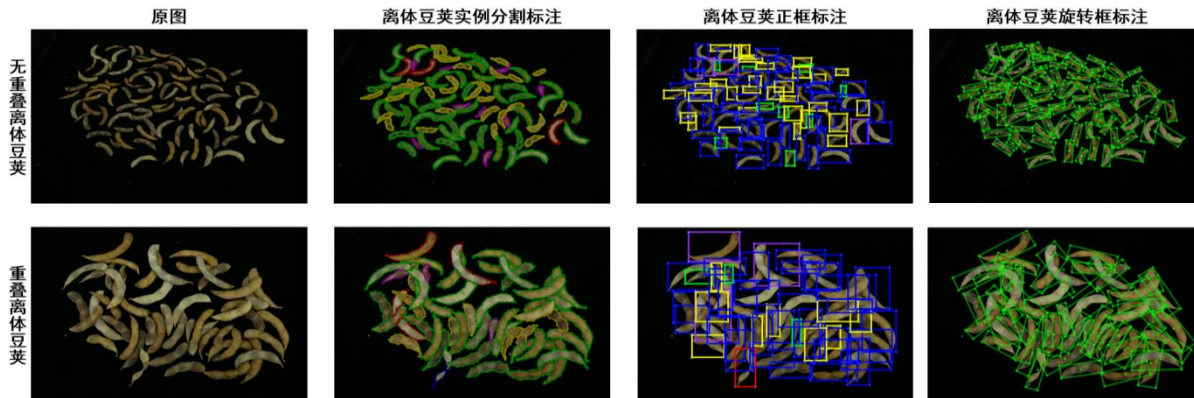


图 3 离体豆荚标注文件示例

Figure 3 Representative images of the annotated in-vitro pods

(2) 大豆植株标注方面。为满足大豆考种及目标检测和实例分割模型算法的研究需求，本数据集对大豆植株的主茎节进行检测框标注、对主茎进行实例分割标注、对在体豆荚进行检测框和旋转框标注。使用 LabelImg 软件打开大豆植株图像，通过矩形框对图像中的豆荚、主茎节进行检测框标注，每个标注完成的矩形框会生成对应的 XML 文件，该文件包含矩形框的中心坐标及边框的长、宽信息。标注主茎节点时，使用最小边界框在子叶节到主茎顶端的范围标注大豆豆荚与大豆植株的相连部分，主茎节标注类别为 stalk。标注豆荚时（正框标注），由于大豆形态各异且易被植株上的枝条和其他豆荚遮挡，遮挡较严重难以识别豆粒数的豆荚不予标注。此外，每个大豆豆荚的豆粒数存在差异，且存在豆荚里无豆粒的情况。因此在标注时需要将豆荚进行分类标注为 zero、one、two、

three、four，对应 0 粒荚、1 粒荚、2 粒荚、3 粒荚、4 粒荚，分类标准同（1）。标注植株主茎时，使用 LabelMe 软件对大豆植株图像从子叶节沿着目标边缘到植株生长点进行标注，主茎标注类别为 *caulis*。使用 rolabelImg 软件对豆荚进行旋转框标注时，根据豆荚位置将矩形框旋转一定角度对豆荚进行标注，标注完成后生成对应的 XML 文件，该文件包含旋转框的中心坐标以及该框的长、宽和旋转角度信息。图 4 显示了不同种类大豆植株图像的标注示例：包括大豆植株的主茎节检测框标注、主茎实例分割标注、在体豆荚的正框标注和旋转框标注。

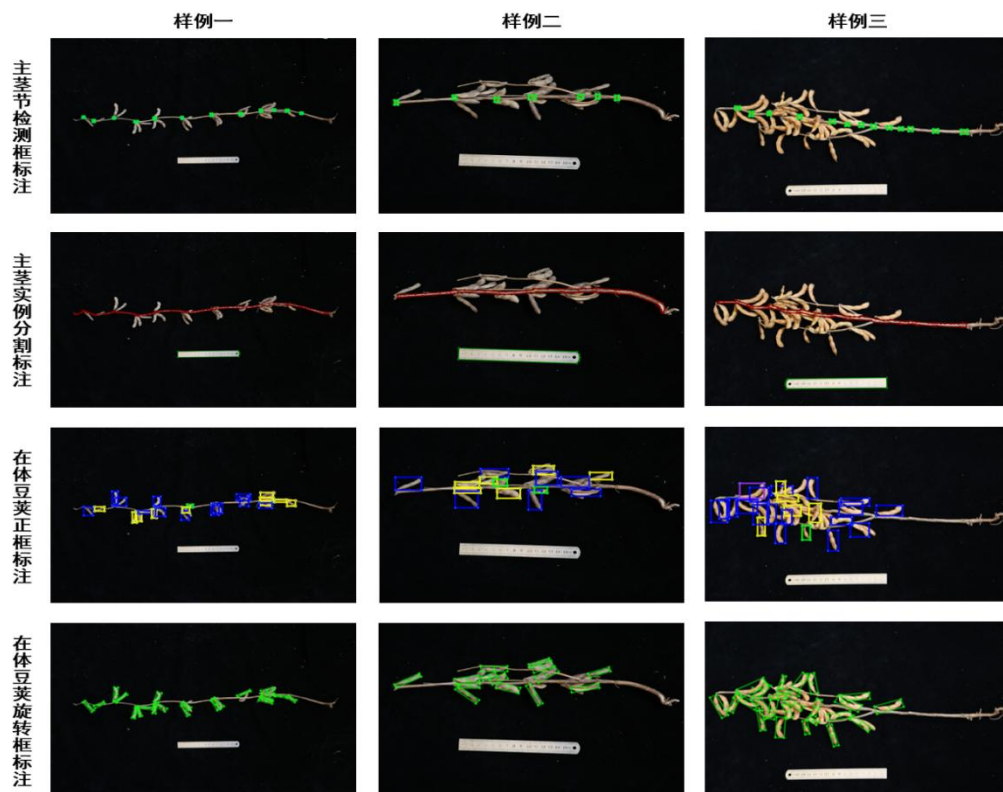


图 4 大豆植株标注文件示例

Figure 4 Representative annotation images of soybean plants

3 数据质量控制和评估

在数据质量控制方面，本研究对于大豆种类的选择基于多样性原则，同时遵循以下几点要求：

（1）采集的大豆均为国内大豆主产区中主栽品种；（2）采集大豆植株包括单分枝型、双分枝型和复杂分枝型；（3）采集数据时尽量剔除一些不符合要求的大豆植株。

在数据质量评估方面，本研究中数据集里的全部图像均是利用标准高清相机在特定的实验室环境下采集得到，数据处理环节严格按照上述流程进行，未经任何加工处理，并在后期对采集的图像进行人工筛选，剔除了一些不符合要求的文件，保证了数据来源的质量和可靠性。标注文件均为人工依照相应格式进行标注。数据集中的图像数据及标注文件经过多次核查检验，以确保数据的质量。

目前，目标检测和实例分割是深度学习中较为热门的研究领域，YOLOv5s 是目标检测领域的经典模型之一，在灵活性与速度上要强于其他检测模型，在模型的快速部署上具有显著优势^[21-23]。Deeplabv3+网络采用的结构是编码器-解码器结构，增强了反向推导能力，降低了网络的延迟。相比传统的分割模型，Deeplabv3+在分割精度和速度上都有显著的提升^[24-25]。因此，使用 YOLOv5s 和

DeepLabv3+模型对本数据集进行验证。模型的分类结果由 TP（真实的正样本数量）、FP（虚假的正样本数量）、TN（真实的负样本数量）、FN（虚假的负样本数量）4 种情况来构成混淆矩阵。常用的衡量算法精度的评价指标如下：

（1）Recall（召回率）：预测正确的正样本在所有正样本中所占的比例。Recall 计算如公式（1）所示。

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

（2）Precision（精确率）：在全部预测为正的结果中，被预测正确的正样本所占的比例。Precision 计算如公式（2）所示。

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

（3）AP（平均精度）：对每一类目标，计算在不同阈值下的 Precision 和 Recall，绘制 PR 曲线，并通过积分计算该曲线与坐标轴围成的面积。AP 的取值范围为 0 到 1，数值越高说明模型的检测精度越高。

（4）mAP_{0.5}：当 Intersection over Union（IoU）阈值为 0.5 时，计算每个类别的 AP 值，并对所有类别的 AP 取平均。mAP_{0.5} 计算如公式（3）所示，式中 N 为类别总数，AP_i 为第 i 类的平均精度。

$$mAP_{0.5} = \frac{1}{N} \sum_{i=1}^N AP_i \quad (3)$$

在图像实例分割中，一般采用平均交并比（mean Intersection over Union, mIoU）和平均像素精度（Mean Pixel Accuracy, MPA）作为模型性能的评价标准。公式（4）–（5）中：k 表示图像分割标签标记的类别总数，P_{ii} 表示为实际分割为 i 类、预测也为 i 类的像素数量，P_{ij} 表示为实际分割为 i 类、预测为 j 类的像素数量，P_{ji} 表示为实际分割为 j 类、预测为 i 类的像素数量。

（4）mIoU（平均交并比）：计算分割图像实际标签与模型预测标签的交集和并集之比，以此来反映模型预测结果与原始图像标签的重复程度。mIoU 计算如公式（4）所示。

$$mIoU = \frac{I}{k+I} \frac{\sum_{i=0}^k P_{ii}}{\sum_{j=0}^k P_{ij} + \sum_{j=0}^k P_{ji} - \sum_{i=0}^k P_{ii}} \quad (4)$$

（5）MPA（平均像素精度）：指每类正确分类的像素占该类所有像素的比例，并对图像中每个类求平均值。MPA 计算如公式（5）所示。

$$MPA = \frac{I}{k+I} \frac{\sum_{i=0}^k P_{ii}}{\sum_{i=0}^k \sum_{j=0}^k P_{ij}} \quad (5)$$

数据集训练的模型检测量化与可视化效果如表 2 和图 5-6 所示。对离体豆荚分别进行了无重叠和重叠两种情况的检测，包括了正框和旋转框检测；对大豆整株分别进行了在体豆荚、主茎节、主茎的检测，并对在体豆荚进行了正框和旋转框检测。经大豆室内考种数据集训练后的模型在无重叠

离体豆荚图像的 mAP_{0.5} 分别为 0.9447（无重叠离体豆荚正框）、0.9938（无重叠离体豆荚旋转框），在重叠离体豆荚图像的 mAP_{0.5} 分别为 0.8010（重叠离体豆荚正框）、0.8626（重叠离体豆荚旋转框），在大豆植株图像的 mAP_{0.5} 分别为 0.6318（主茎节）、0.8631（在体豆荚正框）、0.8727（在体豆荚旋转框），mIoU 和 MPA 分别为 0.6836（主茎）、0.7962（主茎）。

表 2 数据集的目标检测模型训练与验证性能定量分析

Table 2 Quantitative analysis of the training and validation performance for object detection models in the dataset

检测指标		Recall	Precision	mAP _{0.5}	mIoU	MPA
离体豆荚	无重叠离体豆荚（正框）	0.8781	0.9524	0.9447	-	-
	无重叠离体豆荚（旋转框）	0.9986	0.9871	0.9938	-	-
	重叠离体豆荚（正框）	0.7450	0.8097	0.8010	-	-
	重叠离体豆荚（旋转框）	0.8156	0.8516	0.8626	-	-
大豆植株	在体豆荚（正框）	0.8162	0.8718	0.8631	-	-
	在体豆荚（旋转框）	0.8337	0.8636	0.8727	-	-
	主茎节	0.6388	0.7398	0.6318	-	-
	主茎	0.7926	0.8366	-	0.6836	0.7962

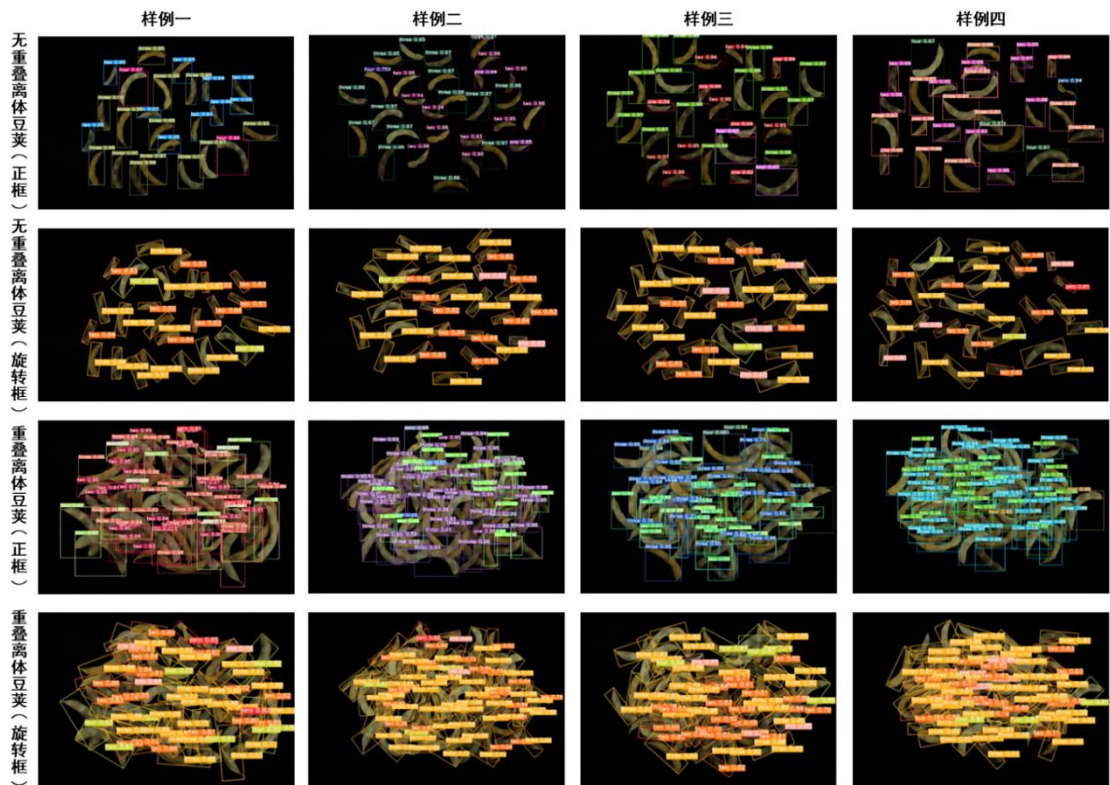


图 5 离体豆荚检测可视化结果

Figure 5 Visualization results of in-vitro pod detection

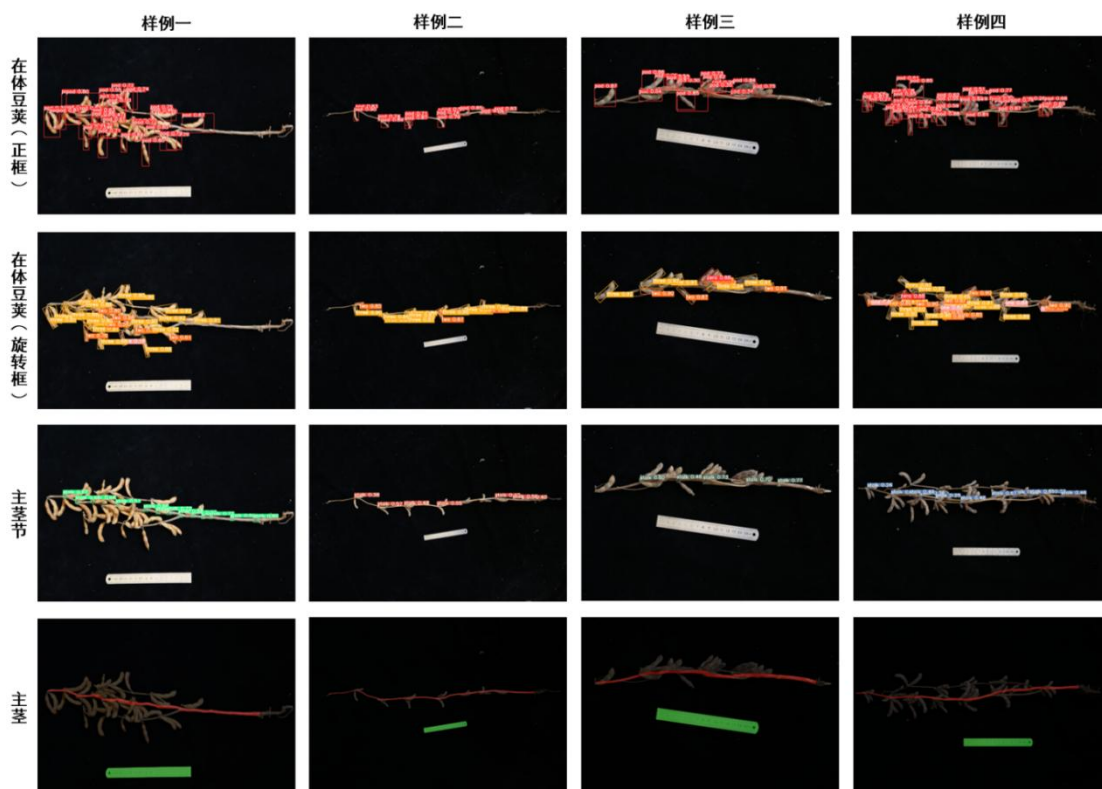


图6 大豆植株检测可视化结果

Figure 6 Visualization results of soybean plant detection

由表2和图5、图6可知，当豆荚重叠或被枝干遮挡时，检测精度相较于无遮挡时会有明显的下降或漏检。针对此类难题，采用文献^[26]提出的 Re-YOLOv5 和检测区域搜索算法的大豆植株表型参数获取方法，通过引入圆形平滑标签技术实现旋转目标检测，减少传统方法检测区域冗余信息，从而提高大豆豆荚的检测精度。

上述实验结果表明，通过本数据集进行训练和验证的模型在无重叠离体豆荚、重叠豆荚以及大豆植株形态等各项指标上均表现出了良好的检测能力，但不同指标间仍然存在一定的差异。特别是在重叠离体豆荚和大豆植株形态信息的检测上，模型仍有较大的提升空间。这充分说明本数据集可作为实际大豆智能化考种研究的优质基础图像数据资源，能够支持开展大豆室内考种新型目标检测和实例分割模型的研究，促进大豆关键考种信息一站式高通量获取方法研发。

4 数据价值

本数据集面向大豆智能化考种方法研究需求，收集和整理了多种大豆植株和离体豆荚图像数据，包含原始图像数据、标注文件数据、测试图像数据3部分，旨在扩充当前大豆关键考种信息方面的数据样本，推动大豆智能化考种研究及应用。大豆品种多样，品种间的株高、主茎节数、单株荚数等往往存在着较大差异，本数据集中收集到的多种豆荚有利于提高模型算法的泛化能力。不同重叠度的离体豆荚图像为探索检测、泛化能力更强的豆荚检测模型提供了训练和测试数据，且为基于迁移学习的在体豆荚目标检测和实例分割提供了基础图像数据资源。同时，大豆植株图像数据为开展大豆植株表型信息的一站式高通量准确获取方法研究，实现大豆的自动智能考种提供了数据支撑。

因此,本数据集可为大豆的离体豆荚检测、在体豆荚检测与植株形态分析的目标检测和实例分割模型方法研究提供宝贵的基础图像数据资源,有利于大豆考种信息的定量化分析和标准化表型数据库的建立,促进大豆关键考种信息的标准化、高通量、低成本获取,提高种质资源鉴定效率,以期为大豆育种提供辅助。

5 数据使用方法和建议

本数据集收集并整理了多种大豆植株和离体豆荚图像数据,并对豆荚、主茎节和主茎进行了标注,为大豆考种表型信息的分析提供了丰富的基础数据资源。在使用本数据集之前,需使用批量替换工具,将 XML 文件中的文件路径修改为使用者计算机中的绝对路径。同时,为确保模型在未见过的数据上能够展现良好的性能,建议以 8:2 的比例划分训练集和测试集。本数据集包含的所有数据均可通过 Python 的 PIL、OpenCV、json、ElementTree 等模块,或主流深度学习框架的自带模块进行读取。通过采用适当的机器学习或深度学习模型,使用者可以充分利用本数据集来进行豆荚、主茎节和主茎的识别与分类新方法、新模型研究,从而提高大豆种质资源鉴定效率。

数据作者分工职责

张哲(2000—),男,安徽省亳州市人,学士,研究方向是人工智能。主要承担工作:数据采集、数据清洗、数据整理、数据标注、论文撰写。

饶元(1982—),男,安徽省合肥市人,博士,教授,研究方向是计算机应用技术。主要承担工作:提出数据采集与处理方法。

束雅莉(1997—),女,安徽省六安市人,硕士,研究方向是机器视觉。主要承担工作:数据采集、论文修改。

陈浩然(1999—),男,安徽省合肥市人,学士,研究方向是作物遗传育种。主要承担工作:数据采集、数据标注。

朱尚尚(1996—),男,安徽省宿州市人,硕士,研究方向是作物遗传育种。主要承担工作:数据采集。

王晓波(1982—),男,安徽省滁州市人,博士,教授,研究方向是大豆育种。主要承担工作:提出数据采集方法。

金秀(1983—),男,安徽省合肥市人,博士,副教授,研究方向是人工智能。主要承担工作:数据采集。

王丰仪(1998—),女,安徽省宿州市人,硕士,研究方向是机器视觉。主要承担工作:数据采集、论文修改。

李佳佳(1988—),女,安徽省阜阳市人,博士,副教授,研究方向是大豆育种。主要承担工作:提出数据标注方法。

徐文强(2002—),男,安徽省安庆市人,本科在读,研究方向是机器视觉。主要承担工作:数据标注。

吴康磊(2003—),男,安徽省蚌埠市人,本科在读,研究方向是机器视觉。主要承担工作:数据标注。

王安然（2003—），男，安徽省马鞍山市人，本科在读，研究方向是机器视觉。主要承担工作：数据标注。

参考文献

- [1] 栾健, 张斌, 胡钰. 中国大豆产业的发展态势、政策演进与趋势展望[J]. 农业展望, 2022, 18(8): 35–41. DOI: 10.3969/j.issn.1673-3908.2022.08.006. [LUAN J, ZHANG B, HU Y. Development trend, policy evolution and trend prospect of China's soybean industry[J]. Agricultural Outlook, 2022, 18(8): 35–41. DOI: 10.3969/j.issn.1673-3908.2022.08.006.]
- [2] 汤碧, 李妙晨. 后疫情时代我国大豆进口稳定性及产业发展研究[J]. 农业经济问题, 2022, 43(10): 123–132. DOI: 10.13246/j.cnki.iae.2022.10.007. [TANG B, LI M C. China's soybean import stability and industrial development in the post-covid era[J]. Issues in Agricultural Economy, 2022, 43(10): 123–132. DOI: 10.13246/j.cnki.iae.2022.10.007.]
- [3] 盖钧镒. 发展我国大豆遗传改良事业解决国内大豆供给问题[J]. 中国工程科学, 2003, 5(5): 1–6. DOI: 10.3969/j.issn.1009-1742.2003.05.001. [GAI J Y. Expanding and enhancing the research allocation on soybean breeding and genetics for the establishment of market supply based on domestic production[J]. Strategic Study of CAE, 2003, 5(5): 1–6. DOI: 10.3969/j.issn.1009-1742.2003.05.001.]
- [4] BAEK J, LEE E, KIM N, et al. High throughput phenotyping for various traits on soybean seeds using image analysis[J]. Sensors, 2020, 20(1): 248. DOI: 10.3390/s20010248.
- [5] SILVA E V V, BRUZI A T, SILVA F D S, et al. Genotypic and phenotypic parameters associated with early maturity in soybean[J]. Pesquisa Agropecuária Brasileira, 2022, 57. DOI: 10.1590/s1678-3921.pab2022.v57.02545.
- [6] KUMAR R, KUMAR P, ALJUHANI A, et al. Deep learning and smart contract-assisted secure data sharing for IoT-based intelligent agriculture[J]. IEEE Intelligent Systems, 2023, 38(4): 42–51. DOI: 10.1109/MIS.2022.3201553.
- [7] 康孟珍, 王秀娟, 华净, 等. 平行农业: 迈向智慧农业的智能技术[J]. 智能科学与技术学报, 2019, 1(2): 107–117. DOI: 10.11959/j.issn.2096-6652.201904. [KANG M Z, WANG X J, HUA J, et al. Parallel agriculture: intelligent technology toward smart agriculture[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(2): 107–117. DOI: 10.11959/j.issn.2096-6652.201904.]
- [8] 陈桂芬, 李静, 陈航, 等. 大数据时代人工智能技术在农业领域的研究进展[J]. 吉林农业大学学报, 2018, 40(4): 502–510. DOI: 10.13327/j.jjlau.2018.4566. [CHEN G F, LI J, CHEN H, et al. A survey of researches on artificial intelligence in the field of agriculture in big data era[J]. Journal of Jilin Agricultural University, 2018, 40(4): 502–510. DOI: 10.13327/j.jjlau.2018.4566.]
- [9] MAIMAITIJIAN M, SAGAN V, SIDIKE P, et al. Soybean yield prediction from UAV using multimodal data fusion and deep learning[J]. Remote Sensing of Environment, 2020, 237: 111599. DOI: 10.1016/j.rse.2019.111599.
- [10] KAMILARIS A, PRENAFETA-BOLDÚ F X. Deep learning in agriculture: a survey[J]. Computers and Electronics in Agriculture, 2018, 147: 70–90. DOI: 10.1016/j.compag.2018.02.016.
- [11] MAMAT N, OTHMAN M F, ABDOLGHAFOR R, et al. Advanced technology in agriculture

industry by implementing image annotation technique and deep learning approach: a review[J]. Agriculture, 2022, 12(7): 1033. DOI: 10.3390/agriculture12071033.

[12] ZHAO W, YAMADA W, LI T X, et al. Augmenting crop detection for precision agriculture with deep visual transfer learning—a case study of bale detection[J]. Remote Sensing, 2020, 13(1): 23. DOI: 10.3390/rs13010023.

[13] 陈姝. 基于深度目标检测的大豆单株荚数自动测量方法与表型测量系统[D]. 南京: 南京农业大学, 2019. [CHEN S. Automatic Measurement Method of Soybean pod Number per Plant based on Deep Object Detection and Soybean Phenotype Measurement System[D]. Nanjing: Nanjing Agricultural University, 2019.]

[14] XIANG S, WANG S Y, XU M, et al. YOLO POD: a fast and accurate multi-task model for dense Soybean Pod counting[J]. Plant Methods, 2023, 19(1): 8. DOI: 10.1186/s13007-023-00985-4.

[15] LI Y, JIA J D, ZHANG L, et al. Soybean seed counting based on pod image using two-column convolution neural network[J]. IEEE Access, 2019, 7: 64177–64185. DOI: 10.1109/ACCESS.2019.2916931.

[16] YANG S, ZHENG L H, CHEN X, et al. Transfer learning from synthetic In-vitro soybean pods dataset for In-situ segmentation of on-branch soybean pods[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, LA, USA. IEEE, 2022: 1665–1674. DOI: 10.1109/CVPRW56347.2022.00173.

[17] AHMAD A, SARASWAT D, AGGARWAL V, et al. Performance of deep learning models for classifying and detecting common weeds in corn and soybean production systems[J]. Computers and Electronics in Agriculture, 2021, 184: 106081. DOI: 10.1016/j.compag.2021.106081.

[18] PRATAMA M T, KIM S, OZAWA S, et al. Deep learning-based object detection for crop monitoring in soybean fields[C]//2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK. IEEE, 2020: 1–7. DOI: 10.1109/IJCNN48605.2020.9207400.

[19] YANG S, ZHENG L H, YANG H J, et al. A synthetic datasets based instance segmentation network for High-throughput soybean pods phenotype investigation[J]. Expert Systems With Applications, 2022, 192: 116403. DOI: 10.1016/j.eswa.2021.116403.

[20] LI S, YAN Z Z, GUO Y X, et al. SPM-IS: an auto-algorithm to acquire a mature soybean p henotype based on instance segmentation[J]. The Crop Journal, 2022, 10(5): 1412–1423. DOI: 10.1016/j.cj.2021.05.014.

[21] 王丰仪, 饶元, 罗庆, 等. 毛桃多模态图像目标检测数据集[J/OL]. 中国科学数据, 2022, 7(4). (2022-12-02). DOI: 10.11922/11-6035.csd.2022.0055.zh. [WANG F Y, RAO Y, LUO Q, et al. A dataset of multi-modal peach images for object detection[J/OL]. China Scientific Data, 2022, 7(4). (2022-12-02). DOI: 10.11922/11-6035.csd.2022.0055.zh.]

[22] LI Y P, RAO Y A, JIN X, et al. YOLOv5s-FP: a novel method for In-field pear detection using a transformer encoder and multi-scale collaboration perception[J]. Sensors, 2022, 23(1): 30. DOI: 10.3390/s23010030.

[23] ZHANG Z, JIN X, RAO Y, et al. DSBEAN: An innovative framework for intelligent soybean breeding phenotype analysis based on various main stem structures and deep learning methods[J]. Computers and

Electronics in Agriculture, 2024: 109135. DOI: 10.1016/j.compag.2024.109135.

[24] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. European Conference on Computer Vision. Cham: Springer, 2018: 833-851.10.1007/978-3-030-01234-2_49.

[25] WAN T, JIN X, RAO Y, et al. SoybeanTracer: An In-Field scene property-based framework for high-throughput soybean canopy coverage extraction and evaluation[J]. Computers and Electronics in Agriculture, 2024, 220: 108869. DOI: 10.1016/j.compag.2024.108869.

[26] 郭希岳, 李劲松, 郑立华, 等. 利用 Re-YOLOv5 和检测区域搜索算法获取大豆植株表型参数[J]. 农业工程学报, 2022, 38(15): 186–194. DOI: 10.11975/j.issn.1002-6819.2022.15.020. [GUO X Y, LI J S, ZHENG L H, et al. Acquiring soybean phenotypic parameters using re-YOLOv5 and area search algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(15): 186–194. DOI: 10.11975/j.issn.1002-6819.2022.15.020.]

论文引用格式

张哲, 饶元, 束雅莉, 等. 大豆室内考种图像数据集[J/OL]. 中国科学数据, 2025, 10(1). (2024-08-21). DOI: 10.11922/11-6035.csd.2023.0104.zh.

数据引用格式

张哲, 饶元, 束雅莉, 等. 大豆室内考种图像数据集[DS/OL]. Science Data Bank, 2023. (2023-05-16). DOI: 10.57760/sciencedb.08286.

A dataset of images for indoor soybean survey

**ZHANG Zhe^{1,2,3}, RAO Yuan^{1,2,3*}, SHU Yali^{1,2,3}, CHEN Haoran⁴, ZHU Shangshang⁴,
WANG Xiaobo⁴, JIN Xiu^{1,2,3}, WANG Fengyi^{1,2,3}, LI Jiajia⁴,
XU Wenqiang^{1,2,3}, WU Kanglei^{1,2,3}, WANG Anran^{1,2,3}**

1. College of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, P. R. China

2. Key Laboratory of Agricultural Sensors, Ministry of Agriculture and Rural Affairs, Hefei 230036, P. R. China

3. Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment, Hefei 230036, P. R. China

4. College of Agronomy, Anhui Agricultural University, Hefei 230036, P. R. China

*Email: raoyuan@ahau.edu.cn

Abstract: Breeding superior varieties is an important guarantee for enhancing the quality and yield of soybeans. However, traditional soybean phenotyping methods are often time-consuming, labor-intensive, and prone to data, which is a far cry from the one-stop high-throughput acquisition of soybean phenotypic information. Therefore, building a dataset of images for indoor soybean survey is very important for

developing methods for high-throughput and accurate phenotypic data acquisition and for further paving the way for automated and intelligent soybean survey. This dataset consists of three parts: original image data, annotation file data, and test image data. It features in-vitro pods and plants of typical soybean varieties, in which the former includes non-overlapping in-vitro pods and overlapping ones, and the latter includes plants of single branch, double branches and complex branches. The annotation file data includes instance segmentation labeling of main stems in soybean plants, detection box labeling of main stem nodes, detection box labeling of in-body pods, and the labeling of instance segmentation and detection box of the in-vitro pods. Moreover, the horizontal and rotation boxes are employed regarding the detection box labeling of soybean plant pods, and in-vitro pods, with a total amount of 4.2 GB. The models trained and validated on this dataset have demonstrated excellent detection and segmentation performance across various metrics, effectively handling non-overlapping and overlapping in-vitro pods, as well as soybean plant structures. This dataset can provide foundational image data for in-vitro pod detection, in-body pod detection and plant morphological analysis of object detection and instance segmentation models. Furthermore, it is of great value for advancing the research on intelligent soybean survey for the one-stop high-throughput acquisition of soybean key information.

Keywords: soybean; intelligent soybean survey; object detection; instance segmentation

Dataset Profile

Title	A dataset of images for indoor soybean survey
Data corresponding author	RAO Yuan (raoyuan@ahau.edu.cn)
Data authors	ZHANG Zhe, RAO Yuan, SHU Yali, CHEN Haoran, ZHU Shangshang, WANG Xiaobo, JIN Xiu, WANG Fengyi, LI Jiajia, XU Wenqiang, WU Kanglei, WANG Anran
Time range	Sep. 2022—Nov. 2022
Geographical scope	Hefei City, Anhui Province
Data volume	4.2 GB
Data format	*.JPG, *.txt, *.xml
Data service system	< https://doi.org/10.57760/sciencedb.08286 >
Sources of funding	The National Natural Science Foundation of China (32371993); the Key Research and Development Plan of Anhui Province (2023n06020057, 202204c06020022); the Natural Science Research Key Project of Anhui Provincial University (2022AH040125).
Dataset composition	This dataset consists of three parts of data: soybean plant data, in-vitro pod data and test image data. (1) The soybean plant data are organized into four folders: bean-height, bean-stalk, bean-pod, and bean-image. The first three folders each include two subfolders, namely Annotations-VOC and Annotations-COCO, which respectively store the VOC format annotation files and COCO format annotation files of soybean plant main stems, main stem nodes, and in-body pods. The bean-image folder stores visible image of soybean plant, saved in JPG format.

	(2) The in-vitro pod data are organized into two folders: nocover and cover. These two folders each include three subfolders, namely Annotations-VOC, Annotations-COCO, and pod-image. which respectively store VOC format annotation files, COCO format annotation files and visible images of in-vitro pod, where the in-vitro pod images are saved in JPG format. (3) The test image data are organized into two folders: test-soybean and test-pod, which store the visible light images of soybean plants used for testing and the visible light images of in-vitro pod respectively, both saved in JPG format.
--	--