scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Annotated 3D Point Cloud Dataset of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform

Alexander Galba[1], Jan Masner[1]✉, Jana Kholová[1,2], Serkan Kartal[3], Michal Stočes[1], Vojtěch Mikeš[1], Pavel Šimek[1], Štěpánka Prokopová[1], René Fiala[1], Thorsten Karrer[4] & András Tóth[4]

This data descriptor presents novel, annotated 3D point cloud plant scans generated by a high-throughput phenotyping platform (LeasyScan, ICRISAT, India). It focuses on broad-leaf legume species (mungbean, common bean, cowpea, and lima bean). The dataset, generated by PlantEye(R) F600 technology, captures multispectral 3D scans of plant canopies. It includes 223 scans, providing detailed organ-level segmentation annotations for embryonic leaves, leaves, petioles, stems, and whole plants. The dataset fills a critical gap in plant phenomics research by offering a base of annotated data to support AI model development efforts in 3D computer vision. Data preprocessing, annotation procedures, and potential applications in crop research disciplines are further discussed. The dataset, preprocessing code, annotations, and a MIAPPE-compliant data sheet are also presented via the GitHub repository for further updates and expansion.

Background & Summary

Background. There is an increasing demand from plant-related research disciplines (e.g., crop breeding, gene banks, plant biologists, etc.), which require access to specific plant characters in large numbers of plants and with high precision and throughput. This has become possible with the development of sensor-based technologies (i.e., plant phenomics). These technologies typically generate vast amounts of data. However, the digital signal generated by the sensors requires data-processing algorithms to infer the desired plant features. Many of these algorithms are AI-based and require specific data inputs and pre-treatment (e.g., annotation) that are time- and resource-consuming to generate. To advance the development of plant traits inference algorithms in support of plant biology disciplines, there is a need to share the relevant datasets with the broad scientific community.

Related datasets. Not many public datasets provide annotated data in the form of 3D point clouds. A comprehensive list of public repositories can be found in the Papers with Code portal[1]. Another list is provided by Zifeng *et al.*[2]. Those datasets are mainly either LiDAR data generated by autonomous vehicles (complemented by 2D RGB image) or full indoor and outdoor scenes. There are only a few articles providing 3D point cloud plant scans (not available in standard repositories), such as soybean[3,4], rose[5], strawberry[6], or maize and tomato[7]. The mentioned datasets provide high-quality scans mostly generated by high-precision systems that do not allow high-throughput essays. To the best of our knowledge, no such an annotated dataset from a high throughput phenotyping platform as presented hereby is available publicly.

Provided plant species. In the presented dataset, we focus on broad-leaf legume species that have relatively simple canopy structures compared to other crop species. Namely, we provide the following species:

[1]Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences, Kamýcká 129, Prague, 165 00, Czech Republic. [2]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, 502 324, Telangana, India. [3]Department of Computer Engineering, Çukurova University, 01380, Adana, Türkiye. [4]Phenospex B. V., Jan Campertstraat 11, 6416 SG, Heerlen, The Netherlands. ✉e-mail: masner@pef.czu.cz

Name	Count
Total number of annotated scans	223
common bean	50
cowpea	45
lima bean	58
mungbean	71
Scans with all plants annotated using organs	141
Scans containing plants unannotated using organs	85
Scans containing some unannotated plants	3
Annotated classes	5
Annotated objects (all classes)	3 712
Annotated objects (Embryonic leaf)	1287
Annotated objects (Leaf)	1224
Annotated objects (Petiole)	814
Annotated objects (Stem)	88
Annotated objects (Plant)	299

Table 1. Summary of the dataset, including counts of annotated scans, species, plants, and classes. There are 223 scans (files) in total. Each scan contains 1–12 plants. Some plants could not be annotated using organ-level classes due to, e.g., wind distortion or overlapping. Instead, those plants were labeled by the Plant class (85 scans, 299 plants). 141 scans contain all plants annotated by organ-level classes.

Fig. 1 The LeasyScan[9] high-throughput phenotyping platform used to gather the data. The picture shows the dual position (2 complementary, partially overlapping scanners capture the same area). The mounted scanners are moving over the field to capture the data (~2 500 m^2 area in 90 minutes).

mungbean (*Vigna radiata* L.), common bean (*Phaseolus vulgaris* L.), cowpea (*Vigna unguiculata* L.), and lima bean (*Phaseolus lunatus* L.). These have been generated as a part of crop improvement efforts at the International Crops Research Institute for Semi-arid Tropics (ICRISAT) and are dry-land grain legume crops – an important source of food and nutrition in semi-arid tropical agricultural systems.

Dataset summary. In summary, we provide annotated high-throughput plant scans in the form of 3D point clouds (*.PCD format). The counts of scans, species, objects, etc., are provided in Table 1. The dataset can be used for research not only in the field of plant phenomics but also generally in the development of 3D computer vision AI models that are currently far less developed than traditional 2D computer vision. The dataset is available at Figshare[8]. The provided dataset is annotated using the Segments.ai platform and can be easily re-imported into this software. All the code and data are also available as the GitHub (https://github.com/kit-pef-czu-cz/3d-point-cloud-dataset-plants) repository, which will be continuously updated with newly annotated data.

Methods

Technical equipment. The presented data were generated using a commercially available PlantEye technology (F600), which is a unique plant phenotyping sensor that combines a 3D scanner with multispectral imaging developed by Phenospex B. V. (PlantEye F600 multispectral 3D scanner for plants - PHENOSPEX). At the ICRISAT field (located in Hyderabad, India), during the data collection, the F600 scanners were mounted in a dual position (2 complementary, partially overlapping scanners capture the same area, illustrated in Fig. 1) and are set to cover the total cropped area of ~2 500 m^2 in 90 minutes. Details of the LeasyScan platform design can be found in[9].

The scanner captures plants' digital reflection in the form of two multi-spectral 3D point clouds where each point contains information on:

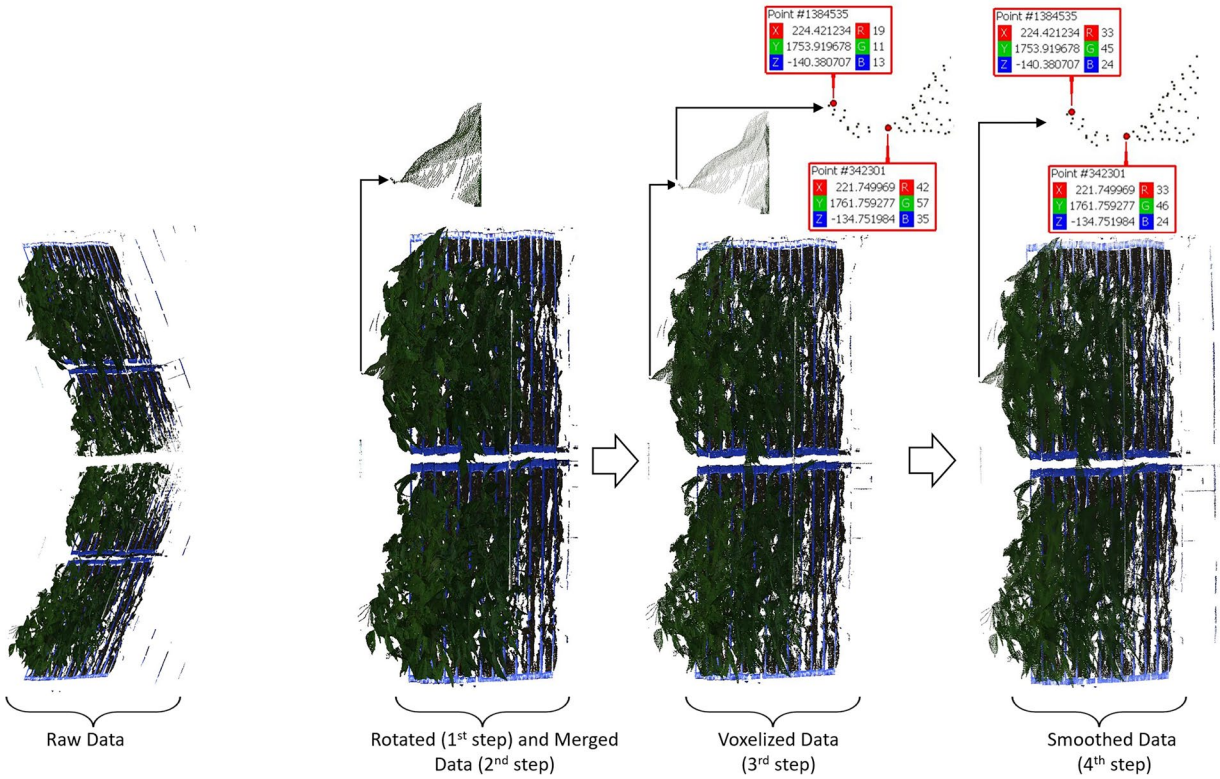

Fig. 2 Pre-processing steps of the raw scan files to extract only plant data for individual microplots.

- x, y, z coordinates in space
- Reflectance in Red, Green, Blue, and Near-Infrared spectra
- Reflectance of the 3D Laser (940 nm)

The 3D model of the plant is stored and pre-processed in proprietary software (HortControl) in an open *.PLY format. The files are accessible through a standard Breeding API (BrAPI) interface[10]. The 3D model of the plant can be used to directly measure or infer a range of plant parameters related to plant morphology and functions; at the moment, the inference algorithms are mostly limited to statistical-based prediction models.

Experiments. The hereby reported data comes from three experiments conducted in 2022 and 2023. Briefly, a single plant genotype was planted in one experimental unit ("microplot") consisting of a PVC tray (blue ones in Fig. 1) of 64 × 40 × 42.5 (length × width × height) cm containing ~50–60 kg of homogenized *Vertisols* collected from the ICRISAT farm. 12 seeds were planted in each tray and later thinned to 1–8 plants per tray, maintained throughout the vegetative growth phase. Plants were maintained up to ~ 35 days after planting, and the 3D point cloud data was obtained throughout the plant growth, typically twice a day. In each experiment, there were a multiple replications of each crop and genotype. At the LeasyScan platform, 24 microplots are grouped under a "barcode" area recognized by the scanning system and provided as a single raw scan. Each microplot has its identification based on a position within the barcode area (0_0 to 1_11).

Data preprocessing. The raw data obtained from the platform for each barcode is represented by the two scans (files) that are rotated to each other (illustrated on the left side of Fig. 2). The raw data are then pre-processed to extract only plant data for individual microplots. The initial step involves rotating the data to align flatly on the x-plane (see left side of Fig. 2). Both scans are merged into a single file in the second step. This merging process increases the point cloud density in the overlapping areas scanned by both scanners. Therefore, the third step involves a voxelization process[11] to rearrange the points in space uniformly. During the scanning process, certain point cloud data may be considered outlier values, where the color values differ significantly from the others. A smoothing process (4th step) was applied to unify these outliers in some point cloud data. In this process, each point data takes the average color value of the N nearest point data.

In the last step, we use a custom AI-based segmentation algorithm to separate plant data from background data such as soil and trays (details are the subject of another publication – please see the GitHub repository for details). The plant data are cut based on the fixed coordinates of each tray in the fifth step (Fig. 3). This step produces the input data for the annotation process.

The pre-processing was implemented independently to skip the coordinate-based cropping of soil data performed by the Phena pipeline in the HortControl software that operates the scanners. The current cropping is based on coordinates and, in some cases, cuts the plant data below the tray edge. We additionally implemented

Fig. 3 Last pre-processing step. On the left side, the zoomed-in plant data shows a single tray separated from the whole scan on the right side. Similarly, the data for all trays are separated from each other and saved as individual files for annotation.

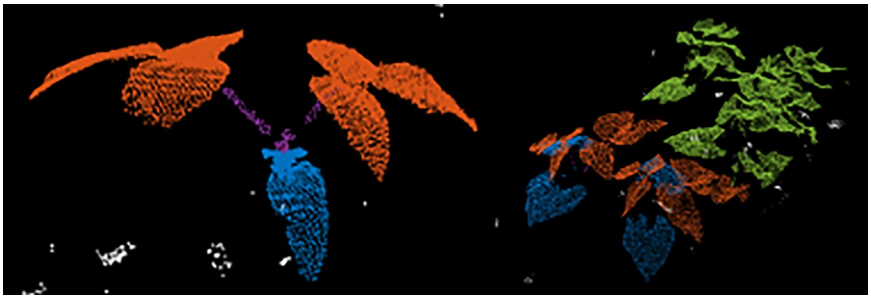

Fig. 4 Examples of annotated scans (orange color – Leaf, blue – Embryonic leaf, violet – Petiole, green - Plant). On the right side is a sample of a lower-quality scan on which it is impossible to recognize plant organs or individual plants.

the smoothing step, which is not part of the Phena pipeline. The pre-processing source code is provided to help work with the raw data.

Data annotation. The data were annotated using the online platform Segments.ai (https://segments.ai) under an academic license. Initial efforts included simultaneous drawing of cuboids (object detection) and segmentation for plant organs and whole plants. It was motivated by doing all possible annotations at once. However, this approach was too time-consuming. The Initial annotation of a single microplot took an average of two hours. It was also apparent that the segmentation drawing was less difficult than drawing cuboids. Annotation was, therefore, restricted to plant organs only. This reduced the time needed to annotate a file to an average of 30 minutes. *Plant* class segmentation annotations for all plants or cuboids for object detection (plants and organs) can be algorithmically generated using the segmentation annotations (see example code in the Usage Notes section).

There are 5 annotated classes within the dataset, specifically: *Embryonic leaf* (the juvenile leaves that are already present in the seed embryo and which have different morphology from other leaves), *Leaf, Petiole* (Leaf petiole), *Stem*, and *Plant*. The overlapping or distorted plants due to environmental conditions (e.g., wind) were additionally annotated using the *Plant* class (see right side of Fig. 4). Those unrecognizable plants naturally appear within the scans and cannot be avoided. In the provided dataset, each plant is either fully annotated by plant organs, annotated using the *Plant* class only, or unannotated. There are no partially annotated plants, only partially annotated scans that include unannotated plants (3 scans).

Data Records

The dataset has been deposited in Figshare[8]. All the shared data is structured into the following directories:

- Readme.md
  - Basic documentation for the dataset. Serves also as a description of the initial GitHub Repository.
- Data
  - Generated cuboid annotations
    - A folder that contains generated cuboids in .txt files using KITTI annotations format.
  - Point clouds
    - A folder containing all 3D point cloud files in .pcd format. The file naming convention is described in Table 2.
  - Segments-ai annotations.json
    - A file that contains segmentation annotations (organ-level mostly), where each point has an assigned class. The file is in the format from the Segments.ai platform (see *Segments-ai annotation format.md* for format description).
  - Segments-ai annotation format.md
    - A file that contains a description of the segments.ai annotation format.

Column name	Content description
Specie	Name of the plant specie that the file contains.
Exp. num.	Number of experiment, under which the scan was obtained at ICRISAT.
Bar code	Identification of a section within the experiment (position in the LeasyScan platform).
Tray	Identification of the tray within the section.
Date time	Timestamp of the scan in format *YYYYMMDDTHHMMSS*. The T is a divider.
Full-Part-Organs	"Full" determines, thar all organs of all plants in the scan were fully annotated. "Part", otherwise, means that the scan contains plant(s) where it is not possible to recognize their organs.
Full-Part-Plants	"Full" determines, whether all plants in the scan were annotated at least using the Plant class. Otherwise "Part". Part means, that in the scan, there are two or more plants that overlap so they cannot be distinguished from each other.
File name	Name of the file in the provided dataset. The name consist of the following columns, divided by dash ("-"): Exp. Num., Bar code, Tray, Date time.
Obj ID X	Multiple columns named Obj. ID X contains IDs of objects (annotated classes) that belongs to one plant.

Table 2. Description of the columns in annotation data.csv file that contains annotation records to assign annotated objects to individual plants.

	Average Precision		**F-Score @ best threshold**		**R^2**	**RMSE**
IoU	***0.3***	***0.5***	***0.3***	***0.5***	**count-only**	
Mean	0.701	0.317	0.759	0.478	0.788	2.153
Median	0.723	0.336	0.773	0.503	0.799	2.000
Best	0.796	0.389	0.817	0.554	0.905	1.491
Worst	0.546	0.162	0.659	0.323	0.601	2.944
Std. Dev.	0.071	0.071	0.046	0.068	0.085	0.450
Range	0.250	0.226	0.158	0.231	0.304	1.453
Var. Coeff.	10.1%	22.4%	6.1%	14.3%	10.8%	20.9%

Table 3. Summary of the results of the SECOND model for all outer and inner cross-validation combinations.

- Annotation data.csv (and .xlsx)
  - Annotations for plant organs to track their assignment to individual plants. A CSV file containing associations of annotated objects and individual plants in scan files. A single line in the file represents an individual plant and its organs. Table 2 provides a description of each column.
  - Raw data.zip
    - Contains raw data from the scanner. There are always two files (each from a single scanner) for each bar code.
  - MIAPPE_data.xlsx
    - Contains MIAPPE-compliant data sheet including mapping to the *Annotation data.csv* file.
- Code
  - Preprocessing from raw data
  - Cuboids generation
    - The folder contains an example code for generating cuboids for object detection for whole plants, together with the organs in the KITTI format. The folder contains an example annotation, an input point cloud file, and an output.
- Baseline evaluation
  - This folder contains full code and results for baseline evaluation using the SECOND and PointRCNN models with instructions on how to install, run, and reproduce the results.

Technical Validation

Plant scanning. The high-throughput phenotyping platform in ICRISAT was originally conceptualized in 2012 to detect key crop adaptations to environmental constraints (e.g., drought, heat, salinity) at the scale relevant to assist crop improvement programs[9]. In 2022, the platform was upgraded with the PlantEye F600 scanners, which has been used to generate data in the presented work. The LeasyScan fully automates the phenotyping process and creates insights into plant growth or changes in health for applications where detailed information or high numbers of plants are required, as referred in, e.g.,[12–14]. The PlantEye is built with high-quality standards to operate in any environment, such as growth chambers, labs, greenhouses, and fields. The technology is patented and widely used by scientists globally. For details about the scanner technology, refer to Phenospex website.

Annotation process. In order to validate the annotated data and minimize human errors, we created a protocol that included a double-checking process. Firstly, we trained every annotator and provided a detailed manual. Each file was assigned to a certain annotator. Another one was assigned to check the annotation first. The

	Average Precision		F-Score @ best threshold		R^2	RMSE
IoU	*0.3*	*0.5*	*0.3*	*0.5*	count-only	
Mean	0.544	0.258	0.709	0.480	0.718	3.615
Median	0.554	0.269	0.717	0.491	0.710	3.826
Best	0.692	0.366	0.809	0.588	0.847	2.384
Worst	0.385	0.158	0.595	0.374	0.580	4.835
Std. Dev.	0.076	0.059	0.054	0.060	0.065	0.703
Range	0.308	0.208	0.214	0.214	0.267	2.451
Var. Coeff.	14.0%	22.7%	7.6%	12.5%	9.0%	19.5%

Table 4. Summary of the results of the PointRCNN model for all outer and inner cross-validation combinations.

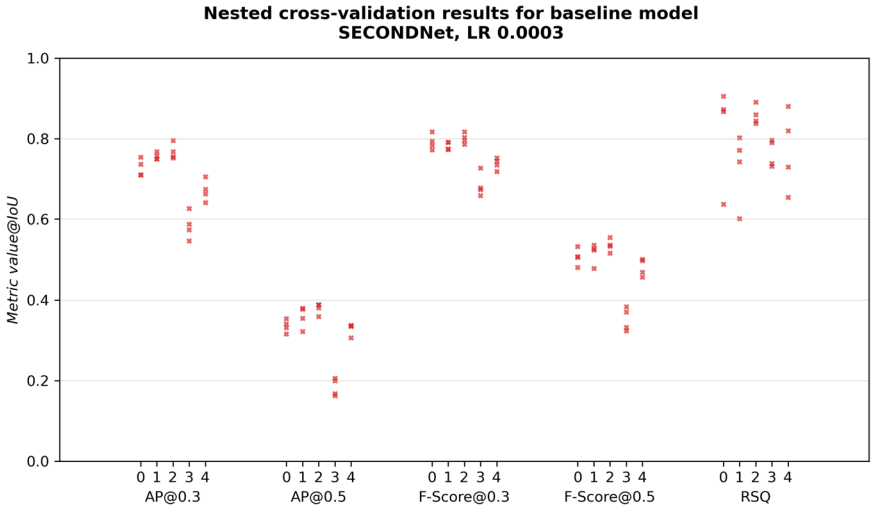

Fig. 5 Results visualization of the SECOND model (voxelization) evaluation for selected metrics. The X-axis shows different metrics for different outer test sets (0,1,2,3,4). Inner cross-validation combinations (train-validation) are represented by the dots. All results and their values can be found in the dataset[8] in the "Baseline Evaluation/Baseline Results" folder.

scan could either be returned to re-annotate or marked as checked. The checked scans were afterward checked again (marked as re-annotate or double-checked) by an expert or senior (experienced) annotator.

Baseline evaluation on object detection models. We conducted baseline experiments to assess the utility and applicability of the presented dataset using two standard object detection architectures: SECOND[15], which operates on voxel grids, and PointRCNN[16], which processes raw points. The codebase utilized the OpenPCDet library (https://github.com/open-mmlab/OpenPCDet) with minor modifications tailored to our dataset.

The dataset was randomly shuffled and partitioned into five-fold splits, each comprising 20%, enabling cross-validation. This resulted in training, validation, and test subsets in a 60:20:20 ratio. Models were trained using nested cross-validation (each fold is rotated as a test set; for each one, the remaining four are rotated as a validation set), ensuring a thorough evaluation and reducing biases related to fold selection, particularly beneficial given the relatively small size of the presented dataset[8].

Each model underwent training for up to 300 epochs, with an early stopping mechanism (patience of 100 epochs and warm-up of 25 epochs) based on the Average Precision (AP) metric at an Intersection-over-Union (IoU) threshold of 0.3. Default hyperparameters were used, with minor adjustments specific to dataset characteristics, including changes to the learning rate, detailed in the Baseline *evaluation/Code/OpenPCDet/tools/cfgs/README.md* file in the dataset repository[8].

Evaluation metrics included AP, F-Score, Precision, Recall, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2, RSQ), providing a comprehensive view of model performance. Calculation methodologies for these metrics are explained in the supplementary Metrics notes. Results are summarized in Table 3 (SECOND) and Table 4 (PointRCNN), presenting descriptive statistics from each inner cross-validation combination evaluated across five different test sets. To illustrate the distribution across folds, selected metrics are visualized in Fig. 5 (SECOND) and Fig. 6 (PointRCNN). Comprehensive results for all metrics can be found in the dataset repository[8] under the *Baseline evaluation/Baseline results* folder.

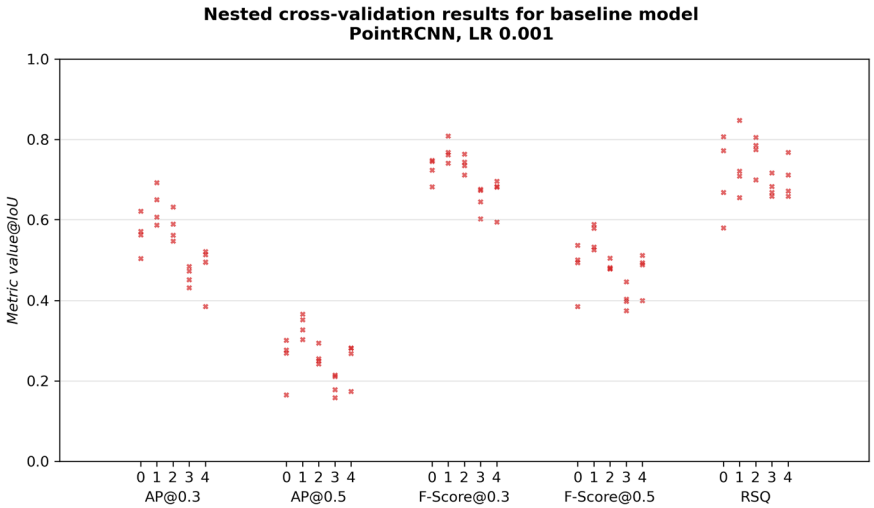

Fig. 6 Results visualization of the PointRCNN model (raw points) evaluation for selected metrics. The X-axis shows different metrics for different outer test sets (0,1,2,3,4). Inner cross-validation combinations (train-validation) are represented by the dots. All results and their values can be found in the dataset[8] in the "Baseline Evaluation/Baseline Results" folder.

Two main limitations were identified in regard to the data splitting and training procedures. First, Figs. 5, 6 highlight significant variations across different test sets. Employing a more sophisticated splitting strategy, considering plant numbers, scan size, or species, might yield more balanced results. Second, additional hyperparameter tuning could further enhance the models' performance.

Code availability

Together with the dataset[8], we first provide a sample code for preprocessing raw scan files to the format that is used for annotation. We also provide code for the automatic generation of cuboids for object detection tasks. Both codes take one file as input and output the result as another file. Third, we provide code for the model evaluations. The code is available in the Code directory.

Received: 31 January 2025; Accepted: 24 September 2025;
Published online: 10 November 2025

References

1. Machine Learning Datasets | Papers With Code. at https://paperswithcode.com/datasets?mod=point-cloud (2024).
2. Ding, Z. *et al*. Recent Advances and Perspectives in Deep Learning Techniques for 3D Point Cloud Data Processing. *Robotics 2023, Vol. 12, Page 100* **12**, 100 (2023).
3. Luo, L. *et al*. Eff-3DPSeg: 3D Organ-Level Plant Shoot Segmentation Using Annotation-Efficient Deep Learning. *Plant Phenomics* **5** (2023).
4. Sun, Y. *et al*. Soybean-MVS: Annotated Three-Dimensional Model Dataset of Whole Growth Period Soybeans for 3D Plant Organ Segmentation. *Agriculture 2023, Vol. 13, Page 1321* **13**, 1321 (2023).
5. Dutagaci, H., Rasti, P., Galopin, G. & Rousseau, D. ROSE-X: An annotated data set for evaluation of 3D plant organ segmentation methods. *Plant Methods* **16**, 1–14 (2020).
6. James, K. M. F., Heiwolt, K., Sargent, D. J. & Cielniak, G. Lincoln's Annotated Spatio-Temporal Strawberry Dataset (LAST-Straw). at https://arxiv.org/abs/2403.00566v1 (2024).
7. Schunck, D. *et al*. Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLoS One* **16**, e0256340 (2021).
8. Galba, A. *et al*. Annotated 3D Point Cloud Dataset of Broad-Leaf Legumes Captured by High-Throughput Phenotyping Platform. at https://doi.org/10.6084/m9.figshare.28270742 (2025).
9. Vadez, V. *et al*. LeasyScan: A novel concept combining 3D imaging and lysimetry for high-throughput phenotyping of traits controlling plant water budget. *J Exp Bot*. https://doi.org/10.1093/jxb/erv251 (2015).
10. Selby, P. *et al*. BrAPI—an application programming interface for plant breeding applications. *Bioinformatics* **35**, 4147–4155 (2019).
11. Aleksandrov, M., Zlatanova, S. & Heslop, D. J. Voxelisation Algorithms and Data Structures: A Review. *Sensors (Basel)* **21**, 8241 (2021).
12. Sivasakthi, K. *et al*. Plant vigour QTLs co-map with an earlier reported QTL hotspot for drought tolerance while water saving QTLs map in other regions of the chickpea genome. *BMC Plant Biol* **18** (2018).
13. Tharanya, M. *et al*. Quantitative trait loci (QTLs) for water use and crop production traits co-locate with major QTL for tolerance to water deficit in a fine-mapping population of pearl millet (Pennisetum glaucum L. R.Br. *Theor Appl Genet* **131**, 1509–1529 (2018).
14. Sivasakthi, K. *et al*. Functional dissection of the chickpea (Cicer arietinum l.) stay-green phenotype associated with molecular variation at an ortholog of mendel's i gene for cotyledon color: Implications for crop production and carotenoid biofortification. *Int J Mol Sci* **20** (2019).
15. Yan, Y., Mao, Y. & Li, B. Second: Sparsely embedded convolutional detection. *Sensors (Switzerland)* **18** (2018).
16. Shi, S., Wang, X. & Li, H. PointRCNN: 3D object proposal generation and detection from point cloud. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2019-June** (2019).

Acknowledgements
The results and knowledge included herein have been obtained owing to support from the following grants: Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. 2023B0005 (Oborově zaměřené datové modely pro podporu iniciativy Open Science a principu FAIR); Ministry of Agriculture of the Czech Republic, grant number QK23020058 (Precision agriculture and digitization in the Czech Republic). We also want to thank the Segments.ai platform for the university license that was provided. Additional acknowledgments go to Anbazhagan Krithika, Sunita Choudhary, Baddam Rekha, and their students from ICRISAT for help with the initial data annotation protocol testing and the first round of annotations.

Author contributions
Alexander Galba – paper writing, code management, annotations generation; Jan Masner – conceptualization, paper writing, annotation management; Jana Kholová – conceptualization, data acquisition, paper revision; Serkan Kartal – data pre-processing; Michal Stočes – data management; Vojtěch Mikeš – data annotation; Pavel Šimek – paper revision, data management; Štěpánka Prokopová – data annotation; René Fiala – AI models development; Thorsten Karrer – data acquisition, pre-processing; András Tóth – data acquisition, pre-processing.

Competing interests
We hereby declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06049-7.

Correspondence and requests for materials should be addressed to J.M.