From Simulation to Field: A Ground Truth-Free Approach for 3D Orchard Monitoring

Harold Murcia, Simon Lacroix

From Simulation to Field: A Ground Truth-Free Approach for 3D Orchard Monitoring

H. Murcia Moreno[1, 2] and S. Lacroix[1]
[1]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France
[2]Facultad de Ingeniería, Universidad de Ibagué, Colombia
hfmurciamo@laas.fr

Abstract
A lack of annotated data and model transfer challenges limits accurate structural analysis of 3D orchards in field conditions. This study presents a scalable framework that trains deep learning models on synthetic LiDAR data. It applies them to unlabelled real-world point clouds acquired with an unmanned ground vehicle in an apple orchard. By integrating supervised classification, contrastive learning, and clustering, the framework segments trees, generates structural maps, and detects anomalies, all without relying on field ground-truth annotations. These results contribute to flexible orchard monitoring and structural analysis in precision agriculture.

Keywords: LiDAR point clouds, tree characterization, contrastive learning

Introduction
Precision agriculture has increasingly integrated advanced technologies to improve crop management and phenotyping. Light Detection and Ranging (LiDAR) has emerged as a key tool for acquiring 3D data, providing enhanced precision in tree mapping and structural analysis for orchard environments. Recent advances in LiDAR-based point cloud processing have shown promising results for agricultural phenotyping. Nevertheless, significant challenges remain in achieving accurate and robust models, particularly for large-scale or heterogeneous vegetation scenarios. This study identifies two primary obstacles: firstly, the generation of real-world databases that capture sufficient variability, hindering the development of robust and generalizable models for diverse orchard scenarios. Secondly, LiDAR-based analyses face a dependency on large annotated datasets, the production of which is labor-intensive and restricts their practical application in agricultural applications (Jin et al., 2021). This dependency underlines the need to develop autonomous data tagging methods to overcome scalability limitations in field phenotyping (Xu et al., 2022).

In recent years, several innovative solutions have emerged for efficient tag learning, such as active learning, semi-supervised learning, weakly supervised learning, self-supervised learning, and unsupervised clustering (Li et al., 2023). These methods aim to reduce reliance on manual annotations and improve the adaptability to large datasets. Despite these advancements, their effectiveness is often constrained by the need for representative training data, particularly in diverse and variable field conditions. For example, while active and semi-supervised learning approaches reduce dependence on labeled datasets, they still struggle with generalization in heterogeneous environments. Similarly, loosely supervised methods, which leverage coarse labels, encounter difficulties in resolving fine-grained phenotyping tasks, such as individual tree characterization in complex environmental scenarios (Singh et al., 2021).
To address the above challenges, this study presents a LiDAR-based framework that combines different techniques and adapts them to the typical needs of an orchard

analysis. These include the identification of ground, invasive plants, and trees; the separation of individual trees; and the generation of specific features to analyze each one individually. To overcome the difficulty of generating diverse real-world databases, the proposed framework employs synthetic data to replicate varied and complex orchard scenarios, facilitating model training and seamless transfer to real-world data without requiring retraining. The framework integrates multiple learning strategies to enhance the analysis of 3D orchard environments. It leverages existing knowledge to identify key elements such as trees and ground while organizing data into meaningful categories. By focusing on both shared and unique features of individual trees, the approach improves its ability to recognize patterns and differences within the orchard. The validity of these patterns can be assessed by clustering trees based on their similarities, ensuring the extracted features accurately reflect meaningful patterns and distinctions within the orchard. Together, these techniques take raw orchard data in a three-dimensional format and transform it into a detailed description of each tree, highlighting its individual features. The proposed approach demonstrates three key properties: (1) a self-driven extraction of unique features by comparing trees, (2) the capability to perform flexible multi-scale orchard analysis based on 3D point clouds, and (3) successful knowledge transfer from simulated to real-world data.

Materials and methods

The proposed framework integrates both simulated and real LiDAR data for detailed analysis of individual trees through a multi-stage process. As shown in Figure 1, the process starts with data generation, followed by 3D scene interpretation and isolation of individual trees. This leads to a contrastive learning phase for anomaly detection and tree clustering based on proximity. The framework concludes with the application of trained models to real-world data to validate their performance.

Data generation — The 3D orchard reconstruction models were simulated using the Discrete Anisotropic Radiative Transfer (DART) software (Gastellu-Etchegorry et al., 2004). In these simulations, a virtual 3D LiDAR system was configured to replicate the characteristics of the real device, including its resolution, field of view, sampling rate, wavelength, and height above the ground. The process started by moving the Mobile Laser Scanner (MLS) in an environment with a set of 22 simulated trees, labeled in 13 groups based on a predefined user-defined taxonomy. Each simulated point in the orchard is represented as $p_i^{syn} = (x_i, s_i)$, where $x_i = (x_i, y_i, z_i) \in R^3$ denotes the spatial coordinates of the point in the world reference frame, and $s_i \in S$ represents the semantic class of the point. The simulated dataset is represented as a matrix $P_{syn} \in R^{nxd}$, where n is the total number of simulated points and d is the dimensionality of each data point (including spatial coordinates and class label). If a point p_i^{syn} belongs to the category $S = tree$, an additional value $g_i \in G_{syn}$ is assigned, where $G_{syn} = \{0,1\ldots,12\}$ represents specific tree group attributes. To enhance the learning process, data augmentation was performed via voxelization with a fixed voxel size v_s. Voxel centers served as seed points, and for each center, the n_{NN} k-nearest neighbors were selected from the original dataset. Field data was collected using an Ouster OS0-128 LiDAR (2048x128 beams @10 Hz) mounted at 1.65 m above ground on an Unmanned Ground Vehicle (UGV). The UGV, remotely operated, navigated through three 50-m rows of an apple orchard (approximately 12 trees per row) in Toulouse, France, over 8 minutes in August 2024.

Data fusion from GPS+RTK and an IMU were used for position estimation. LiDAR scans and pose data were recorded in ROS bag format, with the robot's pose represented by geographic coordinates and quaternion orientation.

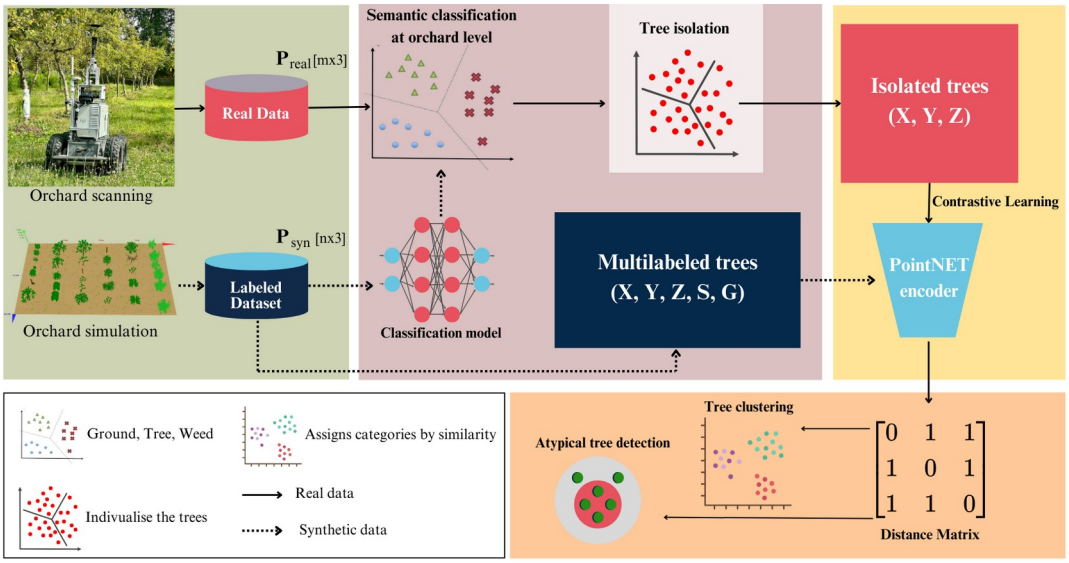


Figure 1. Multi-stage LiDAR framework for individual tree analysis with pre-trained models.

An extended Kalman filter was employed to fuse the sensor data, providing an estimate of the robot's pose T_t at each timestamp t. This pose is represented as a transformation matrix $T_t \in SE(3)$, which defines the relationship between the LiDAR sensor reference frame F_{Lidar} and the absolute reference frame F_{World}. For precise alignment, iterative closest point and pose graph optimization refined the transformations. Post-processing included down-sampling, orchard area delimitation, and noise filtering with a standard deviation multiplier of 2 and 10 nearest neighbors, parameters fine-tuned through multiple trials to maximize noise reduction. The final output was a 3D reconstruction matrix $P_{real} \in R^{mxd}$, where m is the number of points number of points collected from the real-world LiDAR scans and d, is the dimensionality of each data point, similar to the simulated dataset.

3D scene analysis — Supervised classification was conducted using synthetic data subsets generated for training. The RandLA-Net model (Hu et al., 2021), a state-of-the-art neural architecture designed for efficient semantic segmentation of large-scale point clouds, was employed. The network consists of five layers, each applying random sampling and a novel local feature aggregation module to preserve geometric details while reducing computational costs. Subsampling rates of 4 ,4 ,4 ,4 and 2 were applied, progressively increasing output dimensions to 16, 64, 128, 256 and 512. The Adam optimizer was used with an initial learning rate of 0.001, reduced using a cosine annealing scheduler. The training was executed on a GPU-enabled machine, with a batch size of 8, containing n_{NN} points per batch, during 50 epochs.

Isolation of individual trees — Once the classification models were established, the orchard-scale model was applied to the point cloud obtained in real field conditions, producing three spatial maps: soil, weeds, and trees. The data labeled as trees were

extracted, and each tree was isolated using an algorithm based on Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). A voxelisation step with a 5 cm resolution was applied to reduce the computational cost, followed by clustering with a minimum cluster size of 200 voxels and at least 200 voxels per neighborhood. Sparse clusters were discarded, resulting in N_{trees} clusters, each representing a potential individual tree. Original point data within the voxels were then retrieved by finding the points occupied for each voxel.

Triplet-based contrastive learning — The proposed contrastive learning framework follows a supervised approach to learn 128-dimensional feature descriptors $f \in R^{128}$ from simulated point cloud data $p_i^{real} = (x_i, s_i)$. This model builds upon the foundational PointNet architecture (Qi et al., 2017), but introduces key enhancements to process both spatial and additional point descriptors. It leverages permutation invariance, local and global feature aggregation, and a Spatial Transformer Network (STN) to robustly align the spatial (XYZ) coordinates. Furthermore, additional descriptors are concatenated after the STN transformation, and a sequence of convolutional and fully connected layers with increased capacity and regularization via dropout is used to extract robust and discriminative global features. To standardize input dimensions and to compatibilise the use of the GPU, each synthetic tree point cloud was randomly subsampled to [$n_{\square} \times 3$], ensuring consistency across instances. The model was trained using synthetic tree data, with labels s_i. During training, a triplet loss function was employed, designed to maximize inter-class separation while minimizing intra-class distances in the feature space, ensuring discriminative descriptors (Wu et al., 2017). The triplet loss is defined as:

$$L = \frac{1}{N_{triplet}} \sum_{i=1}^{N_{triplet}} max\left(0, \lambda_P . d_E\left(w_i^A, w_i^P\right) - \lambda_N . d_E\left(w_i^A, w_i^N\right) + \alpha\right)$$

Where:

- $N_{triplet}$ is total number of triplets in the batch
- w_i^A, w_i^P, w_i^N are the embeddings of the anchor, positive, and negative samples, respectively.
- $d_E(.,.)$ denotes the Euclidean distance between embeddings
- α is the margin that enforces a minimum difference between positive and negative distances

In this framework, triplets (A,P,N) are selected based on tree groups G, ensuring that the anchor and positive samples belong to the same group, while the negative sample comes from a different group. This encourages the formation of compact clusters in the latent space that are separable for clustering tasks. Furthermore, weighted losses $\lambda = (\lambda_P, \lambda_N)$ are applied to the positive and negative distances to balance their contributions according to class distributions.

Outlier Analysis and Tree Grouping — Starting from the pre-trained PointNet encoder, a distance matrix $D \in R^{N_{trees} \times N_{trees}}$ was computed to represent the pairwise Euclidean distances between point clouds corresponding to individual trees. The diagonal elements

of D were set to zero to ensure self-similarity does not contribute to the analysis. For each tree, the mean distance to all other trees was calculated, resulting in a vector of average distances $\mu \in R^{N_{trees} \times 1}$, where the i-th element is defined as:

$$\mu_i = \frac{1}{N_{trees} - 1} \sum_{i=1,\, j \neq i}^{N_{trees}} D_{ij}$$

To identify outliers, an Isolation Forest algorithm was applied to μ, detecting trees with unusually high average distances. After outlier removal, agglomerative clustering grouped the data, with the optimal cluster count identified via the graph Laplacian's eigenvalues and the elbow method for silhouette scoring

Results

The synthetic point cloud $P^{syn}[n \times 5]$, comprising 22.2 million points over 700 m², was voxelised with a v_s of 0.25 m to enhance learning. For each voxel, 30,000 nearest points were selected, resulting in 18,936 training subsets, and 6,313 each for validation and testing. Real data $P^{real}[m \times 4]$ comprised approximately 28 million points covering 728 m². The testing process was carried out independently for each stage. A reserved 20% subset of the synthetic dataset was utilised for a priori classification evaluation. The total testing time was 3.35 hours. As summarised in Table 1, the classification stage demonstrated high performance, achieving over 99% accuracy across the Ground, Tree, and Grass|Weed classes. Specifically, it attained an overall accuracy (OA) of 99.99%, a mean Intersection over Union (mIoU) of 99.93%, and a mean accuracy (mAcc) of 99.97%. These results underscore the robustness and reliability of the proposed classification framework.

Table 1: Confusion matrix for classification at the orchard scales with simulated data

True/Predicted Class	Ground	Tree	Grass|Weed
Ground	99.999%	0.032%	0.022%
Tree	0.267%	99.976%	0.020%
Grass|Weed	0.085%	0.053%	99.945%

The reconstructed point cloud shown in Figure 2a, subdivided into segments of 30,000 points like the simulation point clouds, was input to the pre-trained model for the classification stage, producing labeled spatial coordinate data. Field data showed some differences from the simulation environment, such as the presence of larger weeds and poles aligned with the trees, which in the case of the real experiments can be large bushes or actual poles. Despite these variations, the color-coded results in Figure 2b visually indicate effective classification, distinguishing key elements like ground, vegetation, and tree structures. In this way, three spatial maps of the orchard are generated: soil model, weed or low grass, and potential trees. Continuing the framework, the points for which a tree class prediction was obtained were used to feed the tree isolation stage. As illustrated in Figure 2c, 49 clusters with between 38,000 and 55,000 points were detected.

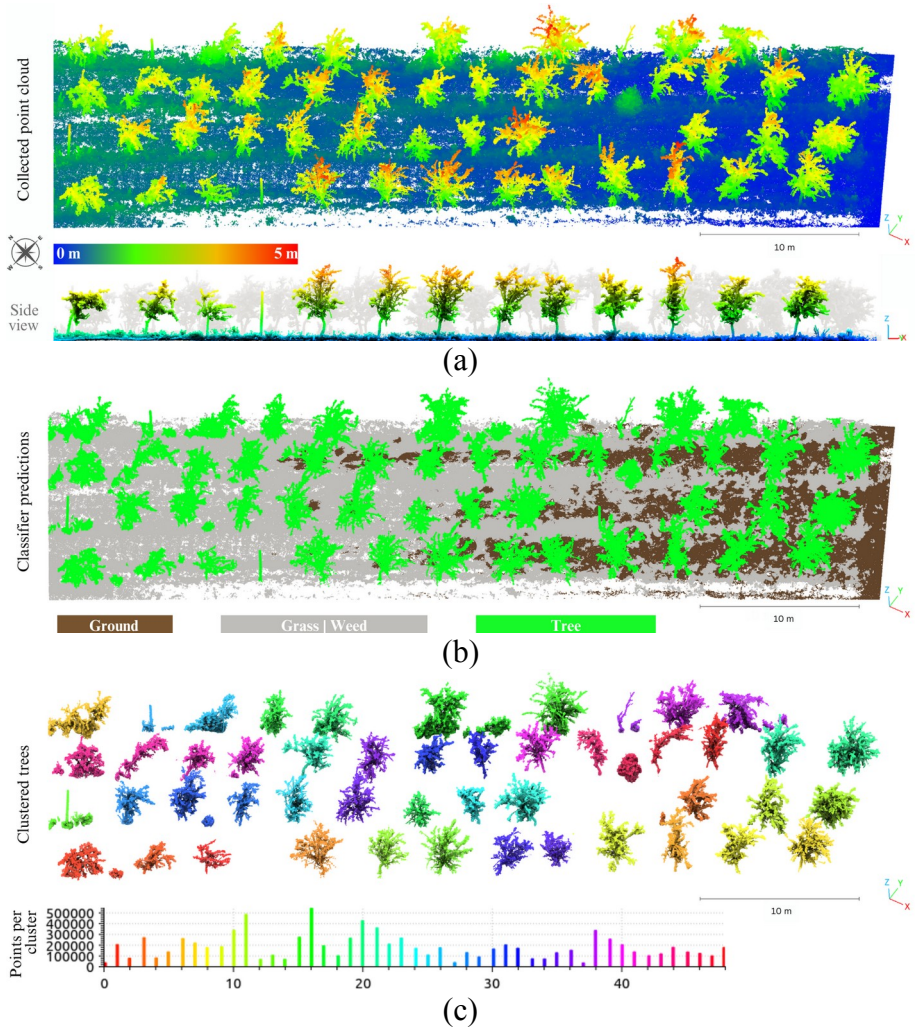
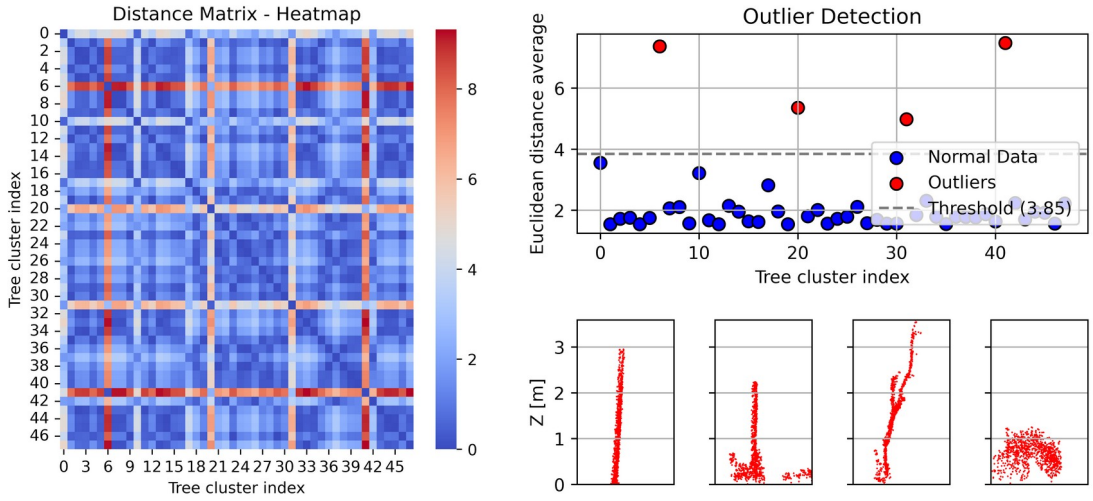


Figure 2. Bird's-eye views of the reconstructed point cloud: (a) colored by height; (b) predicted classes at orchard scale (brown represents the ground, grey indicates grass or weeds, and green denotes trees); and (c) 49 filtered clusters with their respective histogram.

To reduce computational cost and standardize the number of points per tree, a subsampling process was applied. After training, the model's performance was evaluated on the original labeled synthetic data using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Once validated, the model was applied to real-world tree data. On synthetic data, the model achieved an ARI of 0.78 and NMI of 0.95, confirming the accurate clustering of the tree groups. The similarity matrix (Figure 3a), derived from the pre-trained PointNet encoder, reveals clear clustering patterns among tree point clouds, reflecting structural and spatial relationships within the orchard. The Isolation Forest algorithm, applied to the average distance vector μ, identified clusters 6, 20, 31, and 41 as outliers (Figure 3b). These clusters deviate from typical patterns, potentially representing malformed trees, prediction errors in dense vegetation, or misclassified non-tree objects.



(a) (b)

Figure 3. Analysis of tree clusters based on: (a) Similarity matrix showing pairwise distances between tree clusters; (b) Outlier detection using Isolation forest.

Figure 4 presents the clustering results of individual tree point clouds, visualized in 2D projections of the X and Z axes. Each plot represents a tree, color-coded by its assigned cluster label. This visualization highlights the structural and morphological similarities within clusters, as well as the variability across clusters. The clustering was optimized using a Silhouette Score, which evaluates the cohesion and separation of clusters. The optimal number of clusters, determined to be 9 based on the Silhouette Score, reflects moderate cluster separation. This indicates a moderate level of cluster compactness and separation, reflecting the complex and variable nature of tree structures in orchard environments.

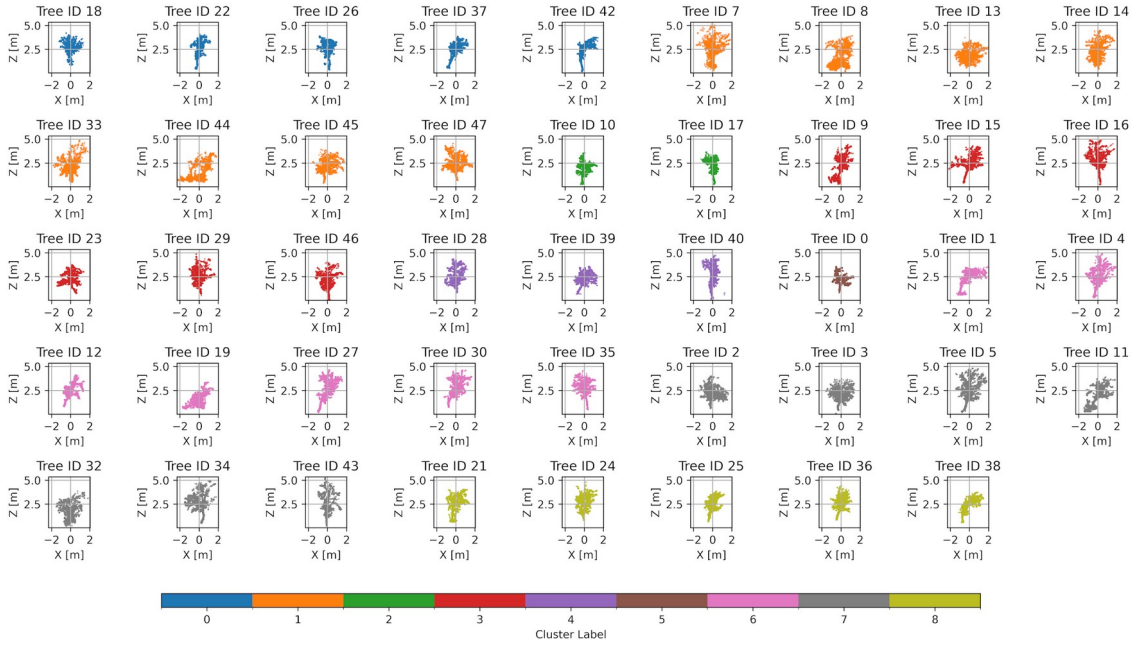


Figure 4. Clustered 3D point clouds of individual trees, color-coded by cluster assignment.

Conclusions

This study introduces a scalable framework for 3D orchard monitoring, training models on synthetic data for application to real-world scenarios. The framework successfully integrates supervised and contrastive learning, along with clustering methods, to enable precise tree segmentation, anomaly detection, and individual tree characterization. Validation through morphology-based clustering demonstrates its ability to extract meaningful features without the need for annotated field data. The proposed pipeline addresses limitations in precision agriculture by transferring pre-trained models from simulation to field conditions, producing terrain, vegetation, and tree maps, and supporting large-scale structural monitoring and anomaly detection in orchard environments.

Future work will focus on incorporating additional LiDAR-derived information, such as multiple echoes and intensity data, to enhance the framework's accuracy, understanding, and comprehensive evaluation.

Acknowledgements

This research was supported by LAAS-CNRS, Toulouse, France, in collaboration with the Universidad de Ibagué, Colombia. Field experiments were carried out at the Lycée Agricole de Toulouse apple orchard. The authors gratefully acknowledges financial support from “Fundación para el Futuro de Colombia” Colfuturo.

References

Gastellu-Etchegorry, J.P., Martin, E. and Gascon, F., 2004. DART: a 3D model for simulating satellite images and studying surface radiation budget. International journal of remote sensing, 25(1), pp.73-96.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N. and Markham, A., 2021. Learning semantic segmentation of large-scale point clouds with random sampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(11), pp.8338-8354.

Jin, S., Sun, X., Wu, F., Su, Y., Li, Y., Song, S., Xu, K., Ma, Q., Baret, F., Jiang, D. and Ding, Y., 2021. Lidar sheds new light on plant phenomics for plant breeding and management: Recent advances and future prospects. ISPRS Journal of Photogrammetry and Remote Sensing, 171, pp.202-223.

Li, J., Chen, D., Qi, X., Li, Z., Huang, Y., Morris, D. and Tan, X., 2023. Label-efficient learning in agriculture: A comprehensive review. Computers and Electronics in Agriculture, 215, p.108412.

Qi, C.R., Su, H., Mo, K. and Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 652-660).

Singh, A., Jones, S., Ganapathysubramanian, B., Sarkar, S., Mueller, D., Sandhu, K. and Nagasubramanian, K., 2021. Challenges and opportunities in machine-augmented plant stress phenotyping. *Trends in Plant Science*, *26*(1), pp.53-69.

Wu, C.Y., Manmatha, R., Smola, A.J. and Krahenbuhl, P., 2017. Sampling matters in deep embedding learning. In Proceedings of the IEEE international conference on computer vision (pp. 2840-2848).

Xu, R. and Li, C., 2022. A review of high-throughput field phenotyping systems: focusing on ground robots. Plant Phenomics.