ISPRS Journal of Photogrammetry and Remote Sensing 184 (2022) 243–263

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

PlantNet: A dual-function point cloud segmentation network for multiple plant species

Dawei Li [a,b,1], Guoliang Shi [c,1], Jinsheng Li [a,b], Yingliang Chen [a,b], Songyin Zhang [d], Shiyu Xiang [a,b], Shichao Jin [d,e,*]

[a] College of Information Sciences and Technology, Donghua University, Shanghai 201620, China
[b] Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, China
[c] Intel Asia-pacific Research and Development Ltd., Shanghai 200246, China
[d] Plant Phenomics Research Centre, Academy for Advanced Interdisciplinary Studies, Collaborative Innovation Centre for Modern Crop Production co-sponsored by Province and Ministry, Jiangsu Key Laboratory for Information Agriculture, Nanjing Agricultural University, Nanjing 210095, China
[e] Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, International Institute for Earth System Sciences, Nanjing University, Nanjing, Jiangsu 210023, China

ARTICLE INFO

Keywords:
Plant phenotyping
Point cloud
Semantic segmentation
Instance segmentation
Deep learning

ABSTRACT

The accurate plant organ segmentation is crucial and challenging to the quantification of plant architecture and selection of plant ideotype. The popularity of point cloud data and deep learning methods make plant organ segmentation a feasible and cutting-edge research. However, current plant organ segmentation methods are specially designed for only one species or variety, and they rarely perform semantic segmentation (stems and leaves) and instance segmentation (individual leaf) simultaneously. This study innovates a dual-function deep learning neural network (PlantNet) to realize semantic segmentation and instance segmentation of two dicotyledons and one monocotyledon from point clouds. The innovations of the PlantNet include a 3D Edge-Preserving Sampling (3DEPS) strategy for preprocessing input points, a Local Feature Extraction Operation (LFEO) module based on dynamic graph convolutions, and a semantic-instance Feature Fusion Module (FFM). The semantic segmentation results of tobacco, tomato, and sorghum in average *Precision*, *Recall*, *F1-score*, and *IoU* reached 92.49%, 92.04%, 92.13%, and 85.86%, respectively; and the instance segmentation results in the mean precision (*mPrec*), the mean recall (*mRec*), the mean coverage (*mCov*), *and* the mean weighted coverage (*mWCov*) reached 83.30%, 74.08%, 78.62%, and 84.38%, respectively. The PlantNet outperformed state-of-the-art deep learning networks including PointNet, PointNet++, SGPN, and ASIS, which achieved an average improvement of 5.56%, 3.58%, 4.78%, and 6.74% in *Precision, Recall, F1-score, IoU* on semantic segmentation, and an average improvement of 22.18%, 16.37%, 14.13%, and 13.35% in *mPrec*, *mRec*, *mCov*, and *mWCov* on instance segmentation. In addition, the effectiveness of 3DEPS, sub-modules, and the new loss function were verified separately by the ablation analysis, in which the removal of any of them can result in a segmentation performance decline of up to 2.0% on average quantitative measures. This study may contribute to the development of plant phenotype extraction, ideotype selection, and intelligent agriculture.

1. Introduction

Plant phenotyping studies a set of traits formed by the dynamic interaction between genes and the environment (Brown et al., 2014; Su et al., 2019). Traditional studies on phenotypic traits are concentrated on the plot and individual plant levels through manual measurements, while recent interdisciplinary studies of genomics and phenomics put forward higher requirements for high-throughput and high-precision phenotype acquisition, such as the organ-level leaf and stem traits that are related to ideal plant architecture (Jin et al., 2018b). Advances in image sensing and analysis technologies shed new light on the high-throughput phenotyping (Sun et al., 2021), which also boosts the high-precision phenotyping at organ levels. Therefore, how to achieve high-throughput and high-precision trait extraction at the organ level is

* Corresponding author.
E-mail address: jschaon@njau.edu.cn (S. Jin).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.isprsjprs.2022.01.007
Received 5 August 2021; Received in revised form 23 November 2021; Accepted 11 January 2022
Available online 17 January 2022

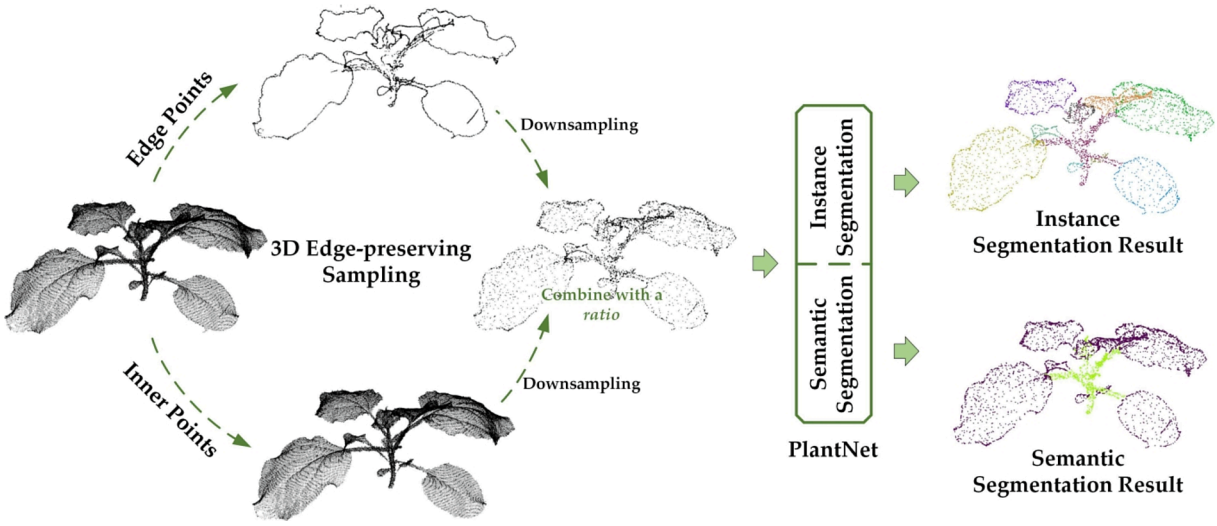


Fig. 1. Overview of the proposed method.

vital and is still in its infancy.

The prerequisite of high-throughput and high-precision phenotypic trait extraction is the automatic plant organ segmentation from accurate data. Optical imagery is the most widely used data for two-dimensional (2D) phenotype analysis. However, images are easily affected by illumination and lack of complete spatial information, restricting these 2D phenotypic analysis methods on simple rosettes plants (e.g., Arabidopsis, tobacco) or several monocotyledonous plants with fewer leaves (e.g., wheat, maize). By contrast, 3D models recorded by emerging 3D sensors, such as light detection and ranging (LiDAR) (Jin et al., 2021; Yuan et al., 2019), Structured Light (Nguyen et al., 2015; Yang et al., 2015), and Time-of-Flight camera (Vázquez-Arellano et al., 2018), can characterize the most important 3D depth and spatial information. These 3D data alleviate the data occlusion and overlapping problems and have been widely used for high-precision phenotyping in the forest (Koma et al., 2018; Livny et al., 2010), agriculture (Jin et al., 2018b; Su et al., 2018; Sun et al., 2017), and grass (Schulze-Brüninghoff et al., 2019), which shows great potential in crop organ segmentation and phenotyping.

Traditional methods for plant organ segmentation from 3D data, such as LiDAR point cloud, usually utilized hand-crafted features such as the octree algorithm (Duan et al., 2016), Difference of Normals (Li et al., 2017), and 3D skeleton (Zermas et al., 2017). These methods can handle several types of plants with simple structures through tedious and labor-intensive parameter tuning, but they still lack generalization ability on the segmentation of different crop species with diverse leaf shapes and canopy structures. Designing a universal 3D segmentation method for different varieties at multiple growth stages is the current research frontier of plant phenotyping.

Recent 3D-based deep learning methods show great potential in improving the generality and accuracy of organ segmentation benefits from not only the massive data and high-performance hardware but also the advances in neural network architectures (Guo et al., 2020). Some studies focus on the Multi-view CNNs (Boulch et al., 2018; Guerry et al., 2017; Kalogerakis et al., 2017), hoping to indirectly realize the understanding of 3D data by strengthening the connection between 2D and 3D via 2D CNNs (Jin et al., 2018a). The main difficulties of this method include that it is hard to determine the angle and the quantity of projection from a point cloud to a 2D image, and how to re-project the segmented models from 2D to 3D space. In addition, to generate regularized structures similar to images for point clouds, some studies focus on voxel-based 3D CNNs (Huang and You, 2016; Li et al., 2016; Maturana and Scherer, 2015; Wang and Posner, 2015; Wu et al., 2015). The point cloud is first divided into a large number of voxels and then a 3D convolution is used to achieve the direct segmentation on the point cloud (Jin et al., 2019). However, this methodology requires a large amount of computation. PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b) are among the earliest end-to-end deep learning networks that operate directly on points, which can simultaneously conduct object classification and point semantic segmentation. Since then, many researchers have adopted similar frameworks to improve their network performance by optimizing and improving the feature extraction modules (Jin et al., 2020; Li et al., 2021), such as the SGPN (Wang et al., 2018b) that works with the similarity of each pair of points in the feature space for instance and semantic segmentation. To better enhance the connection between local features of the point cloud, researchers have resorted to Recurrent Neural Networks (RNN) (Engelmann et al., 2017; Huang et al., 2018; Ye et al., 2018), Conditional Random Fields (CRFs) (McCormac et al., 2017; Pham et al., 2019a; Pham et al., 2019b; Wolf et al., 2015; Yang et al., 2017), and convolutions on local adjacent points (Ben-Shabat et al., 2017; Marulanda et al., 2018; Tatarchenko et al., 2018; Wang et al., 2018a; Xu et al., 2018), such as the Pointwise CNN (Hua et al., 2018), PointCNN (Li et al., 2018), PointConv (Wu et al., 2019) and Graph Neural Networks (Landrieu and Simonovsky, 2018; Qi et al., 2017c; Shen et al., 2018; Simonovsky and Komodakis, 2017; Wang et al., 2019b).

Despite recent advances in 3D sensing and deep learning networks, there are a few studies that successfully achieved plant organ segmentation from the point cloud using deep learning networks. Shi et al. (2019) applied image-based deep learning on plant point clouds constructed by a multi-view camera system to segment organs, and their method mapped 2D organ segmentation results back onto the point clouds with a voting mechanism. Jin et al. (2019) collected LiDAR point clouds of 3000 maize plants, and then proposed a Voxel-based Convolutional Neural Network (VCNN) to realize semantic classification and leaf instance segmentation. Despite these early efforts, current challenges mainly exist in three aspects. First, the lack of a well-labeled 3D plant dataset restricts further progress in this field (the 3D Plant Dataset Challenge). Second, it is difficult to simultaneously achieve high-accurate point-level organ semantic segmentation and instance segmentation (the Network Architecture Challenge); Third, the existing deep learning networks mainly focus on one single species and cannot be generalized to other species (Species Challenge). Therefore, this study aims to propose a dual-function point cloud segmentation network (PlantNet) that can work on several plant species (i.e., tobacco, tomato, and sorghum). The main contributions are as follows:

(i) A well-labeled point cloud dataset for plant stem-leaf semantic segmentation and leaf instance segmentation was built. The dataset

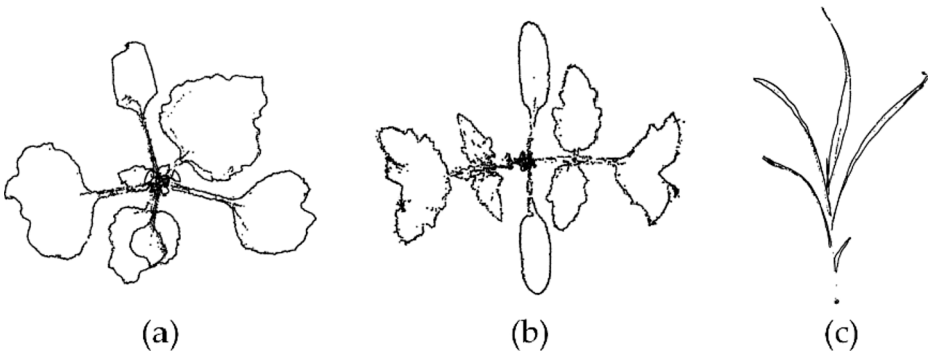


Fig. 2. The extracted edge points of three different plant point clouds by using the 3D Surface Boundary Filter (SBF). (a) is a tobacco plant. The SBF extracts 6171 edge points, which accounts for 7.1% of the total points. (b) is a tomato plant. The SBF extracts 4010 edge points, which accounts for 8.9% of the total points. (c) is a sorghum plant. The SBF extracts 6143 edge points, accounting for 9.1% of the total points.

contains 5460 LiDAR-scanned crops (including 1050 labeled tobaccos, 3120 tomatoes, and 1290 sorghums) with manual labels after data augmentation under several growth periods covering 20 days.

(ii) A 3D Edge-Preserving Sampling (3DEPS) module was proposed, which intentionally increases the proportion of edge points. 3DEPS not only reduces the computation burden by down-sampling raw points but also significantly improves the network performance by introducing more edge points than traditional sampling methods.

(iii) A point-based dual-function network for organ segmentation of different plant species was designed, and its effectiveness and generality were systematically evaluated. The network can simultaneously realize semantic segmentation for two classes—the stems and the leaves, and leaf instance segmentation. PlantNet is featured with a two-pathway architecture that integrates an encode-decoder structure, a Local Feature Extraction Operation (LFEO) module based on dynamic graph convolutions, a semantic-instance Feature Fusion Module (FFM), and a new comprehensive loss function. The effectiveness of PlantNet modules was separately verified. The generalization ability of pre-trained PlantNet was also proved by testing on several other species and two other types of point clouds.

2. Methods

2.1. Framework

The proposed method consists of two parts (Fig. 1). The first part is the down-sampling of the input point cloud, which fixes the input point number to a constant (e.g., 4096) to ensure reliable training of networks. Here, we novelly propose a 3D Edge-Preserving Sampling (3DEPS) strategy with inspiration from sketching to not only reduce the number of network parameters but also preserve the 3D structure of plants as much as possible. The 3DEPS down-sampling strategy focuses on preserving the edge information of point clouds and highlights the sharp contours of plant leaves and stems. The second part is the design of a dual-function point cloud segmentation network (PlantNet), which simultaneously realizes semantic segmentation (stems and leaves) and leaf instance segmentation from point clouds of different species at different growth stages.

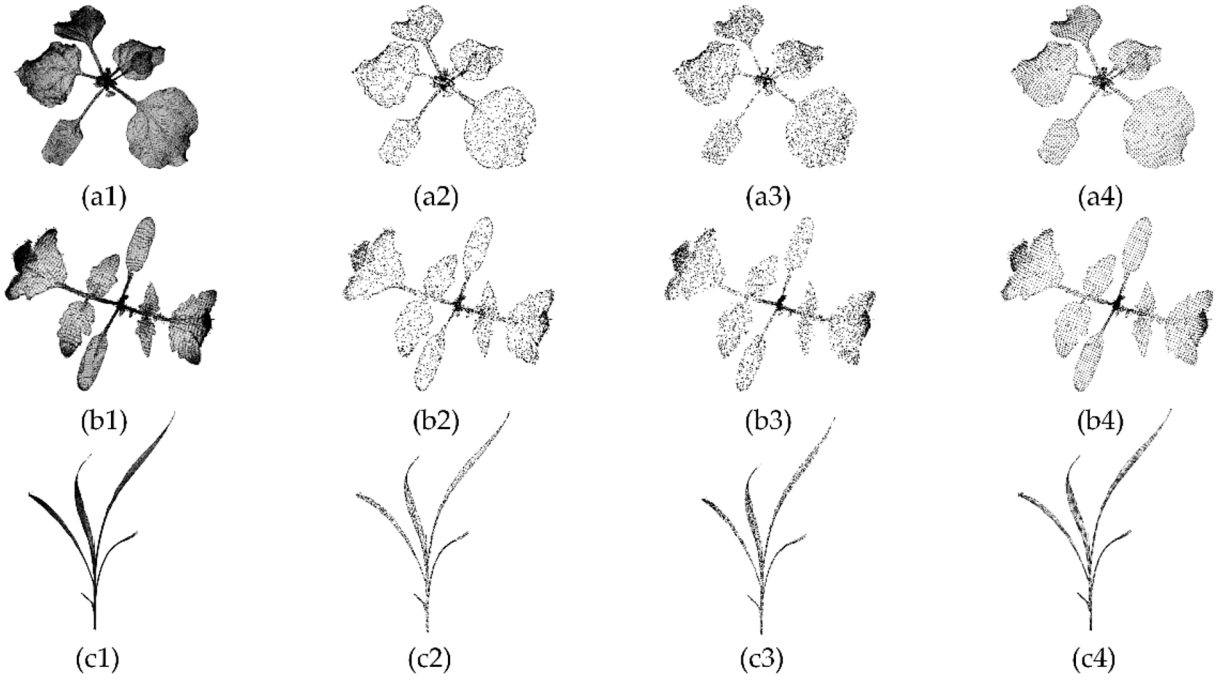


Fig. 3. a visual comparison of the three down-sampling methods on three plants. The first column shows the original point clouds of a tobacco, a tomato, and a sorghum plant, respectively. The second column shows the results after 3DEPS, and each down-sampled plant contains 4096 points. The third column shows the results of IRFPS, each of which contains 4096 points. The fourth column shows the results of VBS, whose output point numbers cannot be easily controlled and were all close to 4100 after multiple trials.

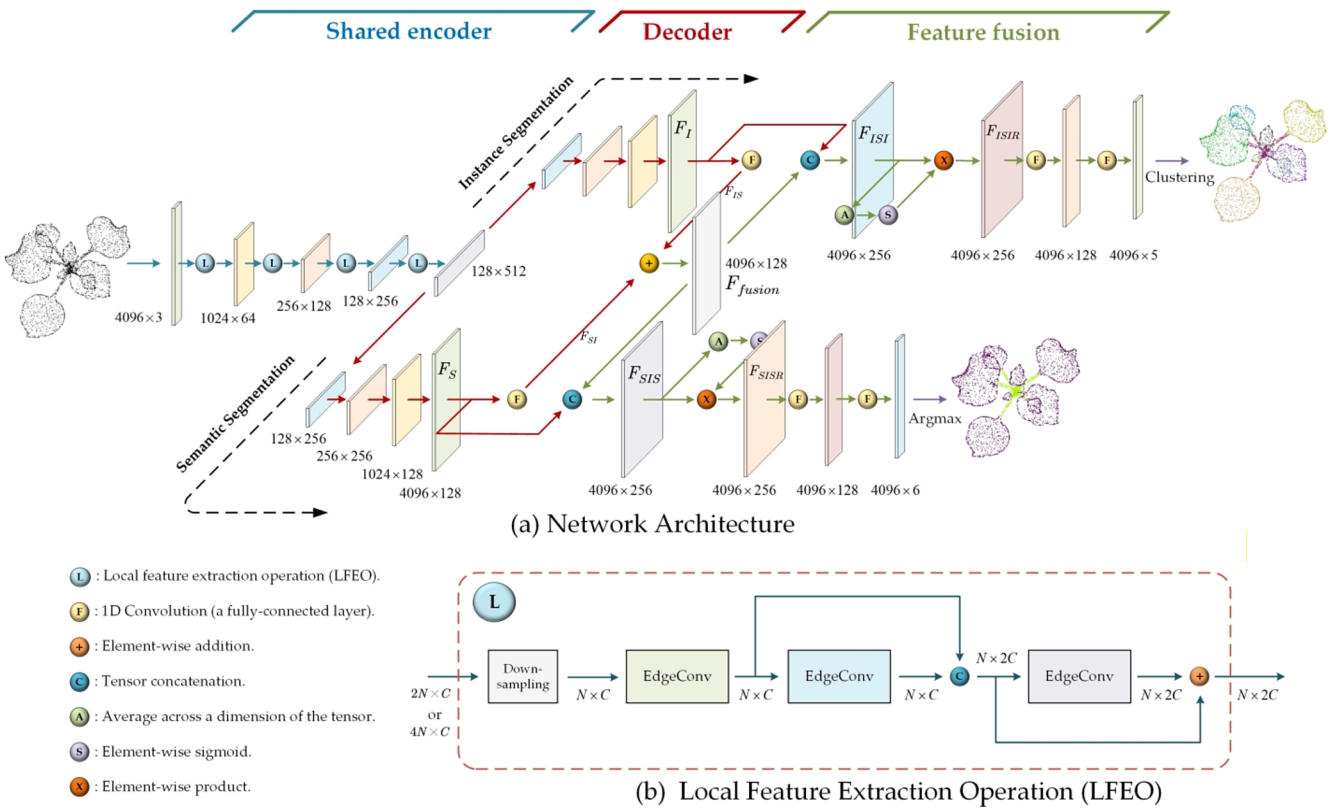


Fig. 4. Overview of PlantNet. (a) is the main structure of the network, which mainly includes three parts—a shared encoder in the front, a dual-pathway decoder in the middle, and a feature fusion module at the end. (b) is a clear demonstration of the Local Feature Extraction Operation (LFEO) used in the encoder. The input feature of LFEO has a dimension of $N \times C$, where N is the number of points, and C is the feature length. For the four LFEOs in a row, the inputs of a LFEO change dynamically with N becoming smaller while C becoming larger.

2.2. Point cloud down-sampling

Theoretically, deep learning networks can accept any input size, but increasing the number of input points without limit will bring serval problems. Particularly, the network parameters will increase with the number of points, harming the training speed and converging ability. Moreover, redundant input points bring no obvious improvement on the model accuracy but will cost more computation resources. Therefore, an effective down-sampling strategy that not only reduces the number of points but also preserves critical features is necessary and beneficial for the deep learning methods on point clouds.

Currently, there are two commonly used down-sampling methods: the Voxel-based Sampling (VBS) (Rusu and Cousins, 2011) and the Iterative Random Farthest Point Sampling (IRFPS) (Qi et al., 2017b). VBS can simply generate regular point clouds with an even density, but it has two main disadvantages: (i) The voxelization needs to be adjusted according to the density and the size of the point cloud, and the number of points after VBS sampling is uncertain; (ii) The computed gravity point is used to represent all points in a voxel, resulting in the local density being constant everywhere. By contrast, IRPFS has a low computational cost and can keep the local density of the point cloud, but it may easily lose details of sparse and small areas.

To maintain more details during down-sampling, we realized that the most effective way to describe a complex 3D object under limited resources was to outline the edge information of the object, which was inspired by sketch drawing. The extracted edge points by a Surface Boundary Filter (SBF) (Li et al., 2019) from point clouds of three different plant species showed that less than 10% of the total points can clearly characterize the overall structure of a plant, indicating that the edge points contain important global information of the point cloud (Fig. 2).

Therefore, this study proposes a new down-sampling strategy—3D Edge-Preserving Sampling (3DEPS) to maintain both the detail of the plant structure and the randomness of effective sampling. This new strategy includes two steps. First, the plant point cloud is divided into edge part A and non-edge part B through SBF. Second, the effective IRFPS method is separately applied to set A and set B, and the sampled points are then combined with a proportion to form the final point cloud. The 3DEPS can deliberately increase the proportion of edge points in the final point cloud to ensure the integrity of the global structure of plants. Moreover, the IRFPS in the second step introduces randomness, which is also helpful to augment the training dataset because we can generate various sampled point clouds (with tiny differences) from the same original object.

A visual comparison of the three kinds of down-sampling methods on three plants is given in Fig. 3. The IRFPS results have jagged contours (especially on leaves) and lose details of tips of several leaves (e.g., Fig. 3 (c3)). The number of points after VBS sampling cannot be easily controlled, and we tried multiple times to make the number of sampled points to be around 4100. By contrast, the 3DEPS method can easily control the output point number and deliberately lift the proportion of edge points in the sampled point cloud, which contributes to a perfect description of the overall contour of the three plants.

2.3. Network architecture

The PlantNet network features a dual-pathway structure, in which the semantic pathway solves the semantic segmentation task, and the instance pathway handles the instance segmentation (Fig. 4). Specifically, the PlantNet structure is composed of three unique parts: the first part is a shared encoder built on a series of EdgeConv—a dynamic graph convolution operation; The second part is a decoder structure that starts

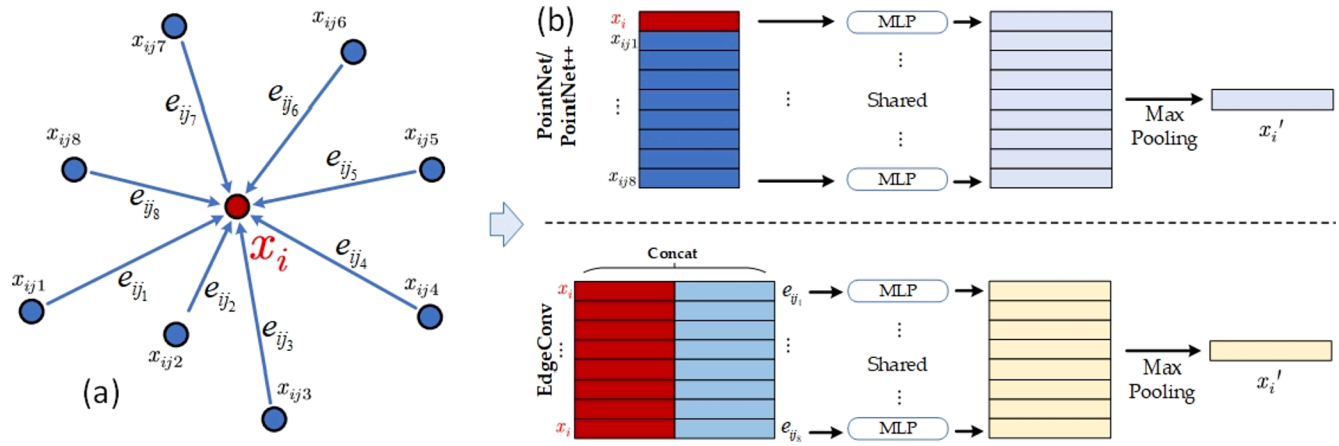


Fig. 5. Comparison between the PointNet and the EdgeConv operation on the extraction of local features.

with a dual-pathway structure, one pathway for extracting advanced semantic features and the other for instance features. The last part is a self-designed Feature Fusion Module (FFM) inspired by Jin et al. (2019), Wang et al. (2019a), and Zhao and Tao (2020). The FFM enables features to interactively merge and support each pathway. At the end of the network, the semantic flow obtains the predicted semantic label for each point through the Argmax operation, while the instance flow performs MeanShift clustering (Comaniciu and Meer, 2002) at the last feature layer to achieve instance segmentation.

2.3.1. Feature extraction of local points (The encoder and the decoder)

Extracting features of local points is the key to point-based deep learning, especially for feature decoder in this study. PointNet-series networks are the earliest architectures that extract local features directly from points. For an input set $\boldsymbol{P}_n = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^c$, c is the number of input feature channels of a point. PointNet (Qi et al., 2017a) uses Multi-layer Perceptron (MLP) networks to extract the pointwise high-dimensional features $\boldsymbol{F}_i \in \mathbb{R}^d$. The PointNet structure can be approximated by the mapping as follows:

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) = \mu\left\{ Concat\left[h(\boldsymbol{x}_1), \underset{i=1,...,n}{MAX}(h(\boldsymbol{x}_i)) \right], ..., Concat\left[h(\boldsymbol{x}_n), \underset{i=1,...,n}{MAX}(h(\boldsymbol{x}_i)) \right] \right\} \quad (1)$$

where μ and h represent MLPs. *MAX* is the max-pooling operation. To improve the lack of multi-scale local neighborhood information aggregation, PointNet++ (Qi et al., 2017b) proposed a hierarchical point set feature learning mechanism that is more effective in learning local structures. The key points are obtained via sampling the feature space, and the features of k-neighboring points of each key point are grouped at the local scale using the PointNet structure. However, it should be noted that most of the current pointwise deep learning networks still conduct feature extraction just by simple MLPs, and then obtain feature vectors through max-pooling (Fig. 5(b)), which is insufficient in the learning of local spatial information (e.g., local connectivity) of point clouds.

To improve the ability of local information extraction, this study adopts a graph-based EdgeConv operation to mine the internal connections between local points (Wang et al., 2019b). The graph structure can represent the local relationships between points efficiently. For each point and its surrounding k-nearest neighbors, a directed graph $G(V, E)$ is formed (Fig. 5(a)), of which $V = \{1, ..., n\}$ is the set of vertices and $E \subseteq V \times V$ represents the set of edges. The directed edge between two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_{ij}$ is defined as $\boldsymbol{e}_{ij} = \boldsymbol{x}_{ij} - \boldsymbol{x}_i$. The edge convolution function is defined as $h_{\boldsymbol{\Theta}}[Concat(\boldsymbol{x}_i, \boldsymbol{e}_{ij})]$, where $h(\cdot)$ represents an MLP based on parameters $\boldsymbol{\Theta}$ to be learned. When taking $\boldsymbol{x}_i$ as the query point, the edge information $\boldsymbol{x}_{ij} - \boldsymbol{x}_i$ is calculated according to its k-nearest neighbors. Finally, the new feature vector of point $\boldsymbol{x}_i$ extracted by EdgeConv is defined as follows:

$$\boldsymbol{x}'_i = MAX\left\{ \underset{j=i,j1,...,jk}{h_{\boldsymbol{\Theta}}} \left[Concat(\boldsymbol{x}_i, \boldsymbol{x}_{ij} - \boldsymbol{x}_i) \right] \right\}, \quad (2)$$

It can be seen from the above equation that the EdgeConv forms a combined feature on each vertex with its adjacent edges in the feature graph, and then uses a shared MLP for feature extraction. Finally, it aggregates a new local feature vector of the vertex point by max-pooling.

The encoder part of PlantNet applies 4 consecutive LFEOs for feature abstraction. Each LFEO mainly includes 2 stages (Fig. 4(b)): (i) down-sampling the current input point feature space to a quarter; (ii) three cascaded EdgeConvs are used for feature extraction of the point feature space after down-sampling. In LFEO, we use the *L2* distance to measure the k neighborhood of a point. Because the graph around the same point formed in the feature space will change in each EdgeConv operation, the calculation of the directed graph $G(V, E)$ is a dynamic process. The residual structure is also introduced into LFEO to alleviate the loss of low-level features and the disappearance of gradients. For example, the output of the first EdgeConv is connected to the output of the second EdgeConv and then becomes the input of the last EdgeConv. Moreover, the outputs of the last two EdgeConvs are combined with an element-wise addition, which combines features while avoiding the increase of network parameters.

The decoder of the network is divided into two pathways, one corresponds to the semantic segmentation, and the other is designed for the instance segmentation. The number of points is gradually restored from 128 to the original 4096 points after multiple up-sampling and MLPs. At both the decoder output of the semantic stream and instance flows, the length of each point feature vector is 128.

2.3.2. Fusion of semantic features and instance features

The third part of PlantNet (i.e., FFM) was designed to carry out effective information interaction between the instance feature map F_I and the semantic feature map F_S. In the feature space F_I, points that belong to the same instance are close to each other, and points that

belong to different instances will repel by each other. This distribution tendency has the potential to enhance the accuracy of semantic segmentation. In the semantic segmentation feature space F_S, the points with the same semantic label will be close to each other, and different semantic categories will repel from each other, which in turn is helpful to reduce the errors of instance segmentation. For F_I and F_S, 1D convolution operation (a single fully-connected layer) is used to extract the feature F_{IS} and F_{SI}, respectively. Then F_{IS} and F_{SI} are combined to generate F_{fusion}, which is then concatenated with F_{IS} and F_{SI} to obtain the characteristic layer F_{ISI} and F_{SIS}, respectively.

For the instance segmentation pathway, we introduce a spatial attention mechanism (Woo et al., 2018) to strengthen the learning of important features inside F_{ISI}. First, a feature vector with a length of 4096 is obtained by calculating the average value of each point from F_{ISI}. Then, the feature vector is transformed into a weight vector with all elements normalized in the range of (0, 1) by carrying out a sigmoid operation on each element. Finally, the feature graph F_{ISIR} is calculated by a point-by-point multiplication of the weight vector and the feature map F_{ISI}. After two 1D convolution operations, a feature embedding O_{ins} with a dimension of $4096 \times S$ is obtained. For all experiments in this study, S is set to 5. The MeanShift clustering (Comaniciu and Meer, 2002) is then applied to conduct instance segmentation form O_{ins}. The above process can be formulated as follows:

$$F_{ISI} = Concat(F_I, Conv1D(F_I) + Conv1D(F_S)), \tag{3}$$

$$F_{ISIR} = F_{ISI} \otimes Sigmoid(Average(F_{ISI})) \tag{4}$$

$$O_{ins} = Conv1D(Conv1D(F_{ISIR})) \tag{5}$$

For the semantic segmentation pathway, the processing on F_{SIS} is similar to the instance branch. We also calculate the weight vector of spatial attention for F_{SIS}, and the feature channels of F_{SIS} are multiplied by each feature channel to obtain the feature map F_{SISR}. Finally, two 1D convolution operations for F_{SISR} are carried out to obtain the semantic output feature O_{sem} of a dimension of $4096 \times C$, where C is the number of semantic categories. For example, if the training dataset used in this paper contains a total of 3 crops, and points of each crop can be divided into the leaf class and the stem class, then C is equal to 6. The predicted semantic label for each input point is then obtained by using Argmax operation on O_{sem} across the feature dimension. The above semantic process can be defined as follows:

$$F_{SIS} = Concat(F_S, Conv1D(F_I) + Conv1D(F_S)) \tag{6}$$

$$F_{SISR} = F_{SIS} \otimes Sigmoid(Average(F_{SIS})) \tag{7}$$

$$O_{sem} = Conv1D(Conv1D(F_{SISR})) \tag{8}$$

2.4. Loss functions

The design of the loss function plays an important role in the training of the network. For the semantic segmentation task, PlantNet uses the standard cross-entropy loss L_c. For the instance segmentation pathway, because the output of the instance segmentation branch is a feature embedding with a size of $4096 \times S$, the final instance prediction label is obtained by clustering. A comprehensive discriminative loss function (Eq. (9)) was designed to constrain the feature embedding enlightened by the class-agnostic instance embedding learning strategy proposed in Wang et al. (2019a).

$$Lf = \alpha \cdot L_{same} + \beta \cdot L_{diff} + \gamma \cdot L_{reg}, \tag{9}$$

where L_{same} pulls the points that belong to the same instance towards the mean of the instance. L_{diff} makes the points from different instances to repel from each other. L_{reg} is a regularization term to bring all clusters close to the feature origin to form a valid boundary for the whole feature space. Parameters α, β and γ are the weights assigned to the three types of sub-losses, and we fix $\alpha = \beta = 1$, and $\gamma = 0.001$ throughout this study. The exact forms of the three sub-losses are given below:

$$L_{same} = \frac{1}{|I|} \sum_{t=1}^{|I|} \frac{1}{N_t} \sum_{i=1}^{N_t} [\|\mu_t - e_i\| - \delta_s]_+^2, \tag{10}$$

$$L_{diff} = \frac{1}{|I|(|I|-1)} \sum_{t_A=1}^{|I|} \sum_{\substack{t_B=1 \\ t_A \neq t_B}}^{|I|} [2\delta_d - \|\mu_{tA} - \mu_{tB}\|]_+^2, \tag{11}$$

$$L_{reg} = \frac{1}{|I|} \sum_{t=1}^{|I|} \|\mu_t\|, \tag{12}$$

where $|I|$ is the number of total instances in the point cloud. N_t is the number of points in the t-th instances; μ_t is the cluster center of the t-th instances in the feature embedding. $\|\cdot\|$ stands for a distance measure such as $L2$. e_i is the feature vector of a point i in the embedding. δ_s is the range that allows the points of the same semantic category to be aggregated; $2\delta_d$ is the closest distance between the two centers of different instances. In equations (10) and (11), $[x]_+ = MAX(0, x)$. Each cluster center μ_t is the center of gravity of all feature vectors belonging to the same instance from the current training batch. In this study, we set $\delta_s = 0.5$, and $\delta_d = 1.5$.

In addition, in the feature fusion part of PlantNet, we design fusion features F_{fusion} to combine semantic and instance information. Since F_{fusion} have both semantic and instance features, the distribution of points could be classified into three cases (Eq. (15)). Case 1: the points belonging to different semantic categories should be far away from each other, and the distance should be as large as possible. Case 2: points belonging to the same semantic category but different instances are separated from each other but are generally not far away. Case 3: points belonging to the same instance should be as close as possible in the embedding. Inspired by Wang et al. (2018b), we use Double-hinge Loss (DHL) L_d to constraint F_{fusion}, as shown below,

$$L_d = \sum_{i}^{N} \sum_{j}^{N} d(i,j), \tag{13}$$

where $d(i,j)$ is the distance loss of any two points p_i and p_j, which is defined as below,

$$d(i,j) = \begin{cases} \boldsymbol{S}_{ij} \text{ Case 1} \\ \varepsilon \cdot MAX(0, D_1 - \boldsymbol{S}_{ij}) \text{ Case 2} \\ MAX(0, D_2 - \boldsymbol{S}_{ij}) \text{ Case 3.} \end{cases} \tag{14}$$

where ε, D_1, D_2 are three constants that satisfy $\varepsilon > 1$, $D_2 > D_1$. In this study we set $\varepsilon = 10$, $D_1 = 10$ and $D_2 = 80$ by using the grid searching method and prior knowledge referring to Wang et al. (2018b). The similarity matrix $\boldsymbol{S}$ has a size of $N \times N$. $\boldsymbol{S}_{ij}$ stands for the characteristic distance between points p_i and p_j.

$$\boldsymbol{S}_{ij} = \left\| F_{fusion}(i) - F_{fusion}(j) \right\|_2, \tag{15}$$

where $F_{fusion}(i)$ and $F_{fusion}(j)$ are the feature vectors of the i-th and j-th points in the fusion feature, respectively. $\| \ \|_2$ represents the $L2$ distance. The smaller the distance is, the greater the probability that the two points belong to the same instance.

The final loss function L of PlantNet is the weighted sum of the semantic cross-entropy loss, the comprehensive discrimination loss, and the DHL.

$$L = \varphi L_c + \gamma L_f + L_d. \tag{16}$$

where φ and γ are both given parameters and were both set as 10 after tuning by grid selection in the range of 4.0–30.0 with a step of 2.0 in the study.

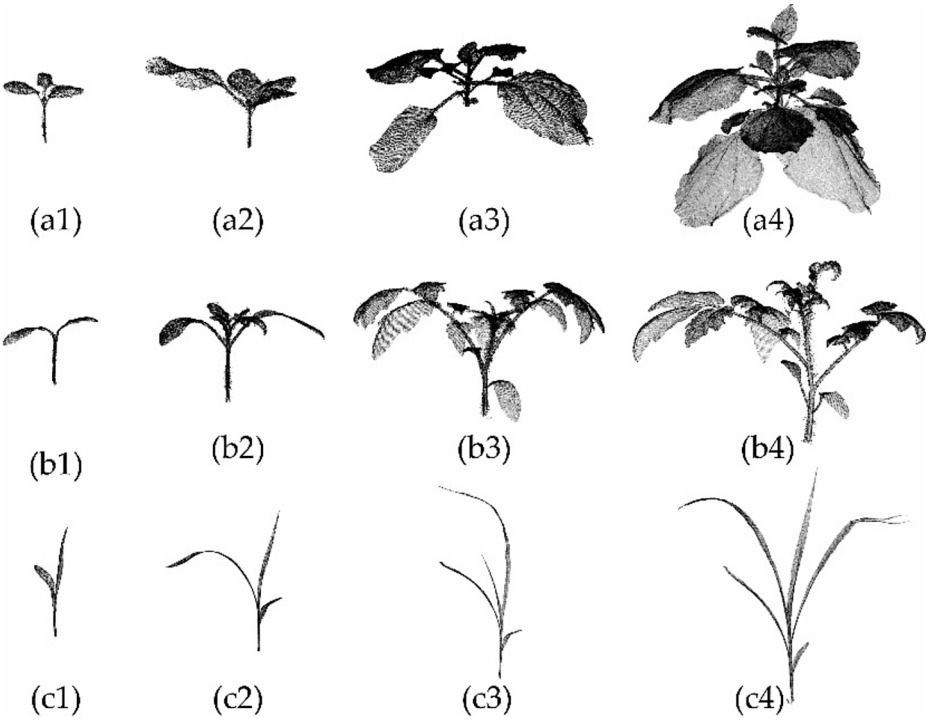


Fig. 6. A demonstration of three different crop species from the original dataset published by Conn et al. (2017a) and Conn et al. (2017b). (a1)-(a4), (b1)-(b4), (c1)-(c4) show the tobacco, tomato, and sorghum point clouds at four different growth stages, respectively.

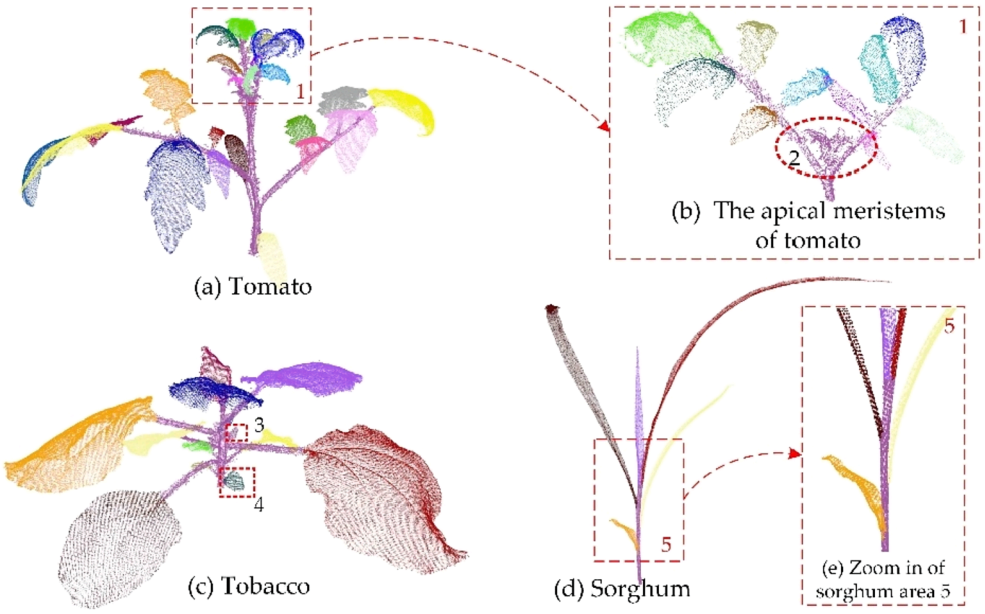


Fig. 7. Examples of the manual labeling of the three plants in the dataset. (a) shows a labeled point cloud of a tomato plant, in which each leaflet is rendered with a unique color; (b) is an enlarged view of the area 1 inside (a). In the elliptical region (area 2) in (b), the top bud of the plant was labeled as a part of the stem; (c) is a labeled tobacco crop, in which area 3 was labeled as stem points and area 4 was label as a leaf; (d) is a labeled sorghum crop; and (e) is an enlarged demonstration of area 5 in Fig. 7(d).

3. Experiment

3.1. Plant dataset preparation

A well-labeled dataset is critical for segmenting plant point clouds using deep learning. The dataset should possess the following features: complete structure in plants, high precision, coverage of multiple species, and coverage of several growth periods. This study uses the 3D plant dataset released by Conn et al. (2017a) and Conn et al. (2017b) that is scanned by a non-destructive sensor (i.e., Edge Scan Arm HD) with very high scanning accuracy (less than1mm). It includes 558 single-plant point clouds of three types of crops (tobacco, tomato, and sorghum) under 3 to 5 different growth environments (ambient light, shade, high heat, high light, drought) during a 20-day growth process. The three species in the dataset differ a lot in structure and leaf characteristics. Tobacco is a dicotyledonous plant with broad leaves and a rosette-like structure in early growth stages. Tomato is also a dicotyledonous plant, but it has compound leaves and the small leaflets usually have jagged boundaries. The sorghum is a common monocotyledonous plant with a high stem height and the leaves are thin and long. Fig. 6 shows point clouds of the three species in the dataset at several growth stages. The tobacco plants have broad leaves that may easily cause

occlusion. The tomato plants usually have a canopy structure with a main single layer, whose canopy structure is simpler than tobacco. However, tomato leaves are compound leaves; the tiny side leaflets in each compound structure and the young buds on top of the main stem can easily cause occlusion and segmentation error. The sorghum plants have longer leaves and stems than the first two, which brings more challenges.

Based on such a representative original dataset, we manually added point semantic labels belonging to stems and leaves, and instance labels belonging to all individual leaves by using the Semantic Segmentation Editor (SSE) tool (https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor/). Points of plant canopies were first labeled to the stem category and the leaf category, and then the points of different leaves were separately labeled as different instances. In this study, buds were not labeled as a separate category because not all samples have buds, and buds of some types cover only a small portion of points that are insufficient for deep learning.

The labeling process of three kinds of plants is shown in Fig. 7. Different colors on the plants represent different labels, and each leaf instance is rendered with a unique color. The labeling of a tomato plant is shown in Fig. 7(a)-(b). For the meristem at the top of the main stem, only the leaves with evident leaf-style morphological characteristics were labeled, while the tiny leaflets and buds that cannot be effectively distinguished were labeled as a part of the stem category. For the tomato plant, each leaflet in the compound leaf was defined as a leaf instance. Fig. 7(c) shows the labeled result of a tobacco plant, where Areas 3 and 4 are two stages of budding that are commonly seen on tobacco plants. In Area 3, the axillary bud is still small and has a different shape from mature leaves, so we labeled this bud as a part of stems. In Area 4, the bud has developed into a leaf shape and covered enough points, so it is labeled as a leaf instance. Compared with dicotyledonous plants (tomato, tobacco), the monocotyledonous sorghum was difficult to label because the connection between the leaf and the main stem was hard to determine even for an experienced worker due to the existence of sheath. The labeling results of a sorghum plant in an early growth stage are shown in Fig. 7(d)-(e), in which we mark the cylindrical leaf sheath as stem and the non-cylindrical leaf sheath as a part of the leaf.

For the original dataset, we removed 12 point clouds in the early growth period of sorghum because the period only has one bud and it is almost impossible to distinguish leaves from the stem. The number of final labeled individual plants is 546. For each species, about two-thirds of the point clouds in each growth period were used as the training set, and the rest were taken as the testing set. Finally, the training dataset consists of 364 (tomato: 208, tobacco: 70, and sorghum: 86) labeled individual plants and the testing dataset consists of 182 (tomato: 104, tobacco: 35, and sorghum: 43) individual plants. The minimum and maximum point number of an individual plant in the labeled dataset is around 5000 and 100000, respectively.

3.2. Data augmentation

To enhance the training of the segmentation network and to avoid overfitting, the dataset is further augmented by the proposed 3DEPS down-sampling method. The 3DEPS lifts the proportion of the edge points in the result. The only parameter in 3DEPS is the "*ratio*", which is defined as the proportion of the edge points to the total sampled points. The edge points N_{edge} and non-edge points N_{centre} in a sampled point cloud by 3DEPS can then be represented as:

$$N_{edge} = \lceil N \cdot ratio \rceil, \tag{17}$$

$$N_{centre} = N - \lceil N \cdot ratio \rceil, \tag{18}$$

in which the operator means round-up to an integer. The input number of points of PlantNet N is fixed to 4096 throughout the study, which requires the result of 3DEPS to be a point cloud of 4096 points for each single plant sample. 3DEPS can easily realize data augmentation due to the randomness in down-sampling. Therefore, we can always obtain a different sampling result by changing the start point. The data augmentation can be divided into three steps: (i) first, SBF algorithm is used to separate the edge points from non-edge parts in all the original crop point clouds; (ii) then, 3DEPS was carried out to sample 820 points from the edge part and sample 3276 points from the non-edge part according to the sampling ratio (e.g., 0.2), to form a down-sampled crop point cloud containing 4096 points; (iii) finally, we independently repeat the above process for 10 times for each point cloud to make a 10-times data augmentation. To improve the stability of the network training, we normalized the scale for all plant point clouds to a 1-meter cube space, and the plant center is moved to the origin of the point cloud coordinate system. Finally, we obtained a new augmented dataset that contains 3640 training point clouds, and 1820 testing point clouds. The details of the dataset are given in Table 1.

Table 1
The number of point clouds in the manually labeled dataset and their usage in this study.

Number of point clouds	Tobacco	Tomato	Sorghum	Total	
				Training	Testing
Original	105	312	129	364	182
After Augmentation & Normalization	1050	3120	1290	3640	1820

3.3. Network training and testing

All experiments in this study are carried out on a server that includes a CPU of 16 cores and 32 threads, a memory of 128 GB, 4 GPUs of the NVIDIA GeForce RTX 2080Ti. The server runs under the Ubuntu operating system, and the training framework is TensorFlow. At the training stage, all input point clouds to PlantNet only contain the 3D XYZ coordinates without color, the number of input points is fixed at 4096. The training batch size is 10, the initial learning rate is 0.002, and the learning rate decreases by 30% for every 10 Epoch. The Adam Solver is used to optimize the network. The momentum is set to 0.9. The network was trained until its loss stayed stable after 100 epochs, and the network parameters were determined according to the lowest loss.

During testing, the batch size is fixed at 1. Each plant point cloud in the testing set is also augmented 10 times by using 3DEPS, and the quantitative results are averaged on the augmented testing samples to suppress the errors caused by random sampling. For the instance segmentation task, the instance pathway of the network outputs a $N \times 5$ feature embedding, which was used to cluster the final instance segmentation results (i.e., individual leaves) by using the MeanShift algorithm (with a bandwidth of 0.6). If the number of points in an instance cluster is less than a certain threshold, the instance is abandoned to avoid over-segmentations. The instance point threshold is set to be 1% of the average number of points of the largest semantic category in a point cloud.

3.4. Evaluation criteria

In this study, we perform semantic segmentation and instance segmentation tasks at the same time for three types of crops. In terms of semantic segmentation, we calculate four quantitative measures—*Precision*, *Recall*, *F1-score*, and Intersection over Union (*IoU*) on each semantic class. For all four measures, the higher value means better segmentation. For each semantic class, IoU is a standard intersection over union representation. *Precision* reflects the proportion of points correctly classified by the network to the total predicted points of a semantic class. *Recall* is the proportion of correctly predicted points in a semantic class by the network to the total number of ground truth points in the semantic class. *F1-score* is a harmonic average of *Precision* and

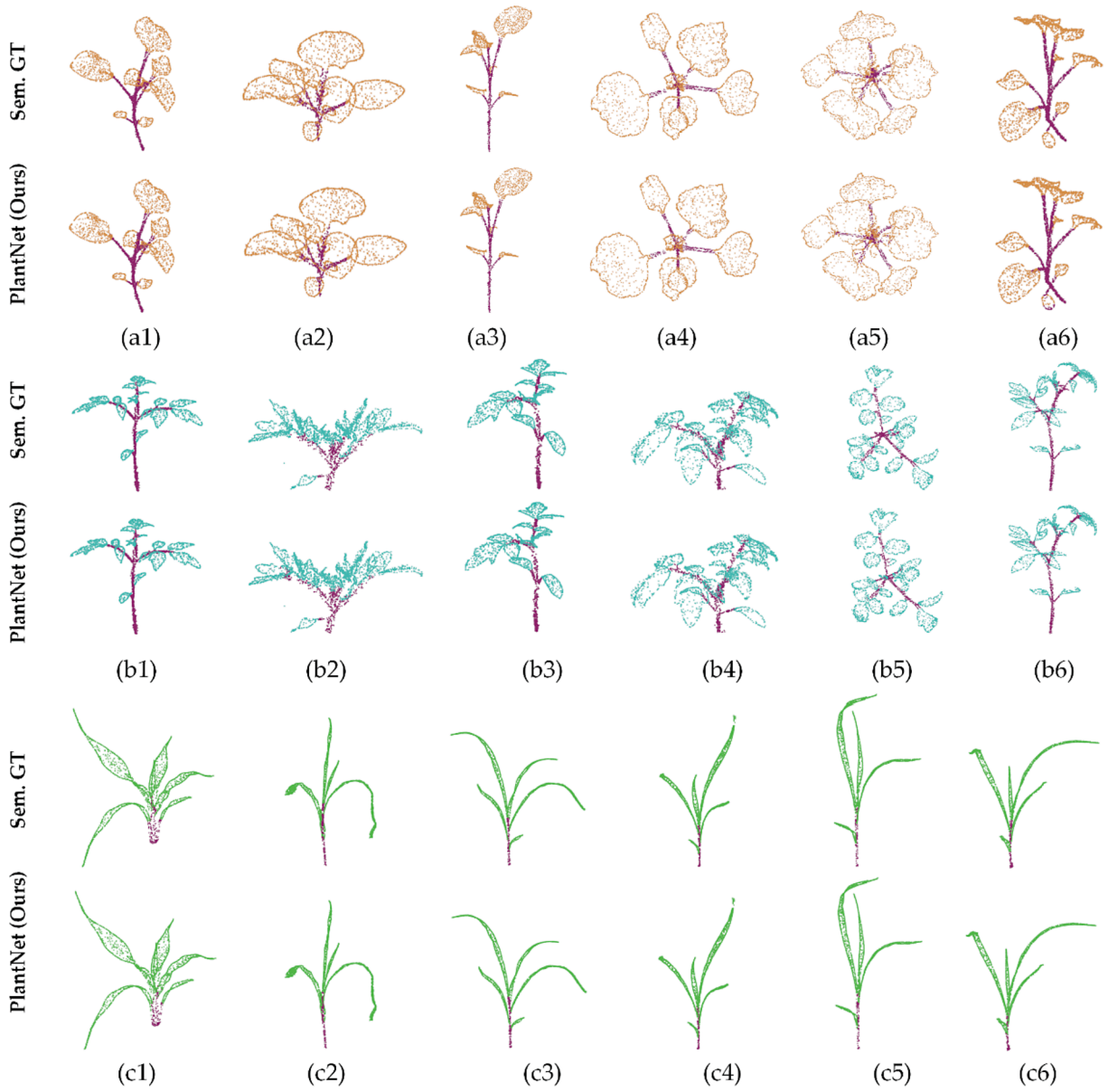


Fig. 8. Demonstrations of some qualitative results of the semantic segmentation of the three plants by PlantNet. The displayed samples are intentionally selected for covering individuals that are with different growth environments and also different structures from the dataset. The semantic ground truths and the semantic segmentation results on several tobacco samples are shown on the 1st row and the 2nd row, respectively. The semantic ground truths and the semantic segmentation results on several tomato samples are shown on the 3rd row and the 4th row, respectively. The semantic ground truths and the semantic segmentation results on several sorghum samples are shown on the 5th row and the 6th row, respectively.

Recall by considering both measures, and its value ranges between 0 and 1. The four measures are defined as follows:

$$IoU = \frac{TP}{TP + FP - FN}, \tag{19}$$

$$Precision = \frac{TP}{TP + FP}, \tag{20}$$

$$Recall = \frac{TP}{TP + FN}, \tag{21}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{22}$$

where *TP*, *FP*, and *FN* are the number of true positive, false positive, and false negative points of a semantic class, respectively.

For instance segmentation, we employ the mean coverage (*mCov*) and the mean weighted coverage (*mWCov*) (Liu et al., 2017; Ren and Zemel, 2017; Zhuo et al., 2017) as the point-level evaluation measures. *mCov* is defined as the average point-level *IoU* of instance prediction matched with ground truth, and *mWCov* is a weighted version of *mCov* defined below,

$$mCov(I, P) = \frac{1}{|I|} \sum_{m=1}^{|I|} \max_{n} IoU(I_m, P_n), \tag{23}$$

$$mWCov(I, P) = \sum_{m=1}^{|I|} \omega_m \max_{n} IoU(I_m, P_n), \tag{24}$$

$$\omega_m = I_m / \sum_{k=1}^{|I|} I_k, \tag{25}$$

Table 2
Quantitative results of semantic segmentation of PlantNet on plants.

	Precision (%)		*Recall* (%)		*F1-score* (%)		*IoU* (%)	
	Stem	Leaf	Stem	Leaf	Stem	Leaf	Stem	Leaf
Tobacco	91.47	95.04	85.29	94.28	88.27	94.55	79.01	89.66
Tomato	93.83	96.44	93.03	97.14	93.26	96.59	87.37	93.40
Sorghum	80.15	98.03	84.67	97.82	82.24	97.89	69.84	95.87
Mean	88.48	96.50	87.66	96.41	87.92	96.34	78.74	92.98

where I_m represents the m-th instance region in the ground truth instance collection, which can also be perceived as the number of points in the instance regions, P_n represents the n-th predicted instance region. $|I|$ is the number of all instances of a semantic category in the ground truth. Besides the point-level measures, two instance-level measures—the mean precision (*mPrec*) and the mean recall (*mRec*) are calculated for all instances with an IoU higher than 0.5 in each semantic category (Conn et al. (2017b)).

$$mPrec = \frac{1}{C}\sum_{i=1}^{C}\frac{|TP_i^{ins}|}{|P_i^{ins}|}, \tag{26}$$

$$mRec = \frac{1}{C}\sum_{i=1}^{C}\frac{|TP_i^{ins}|}{|G_i^{ins}|}, \tag{27}$$

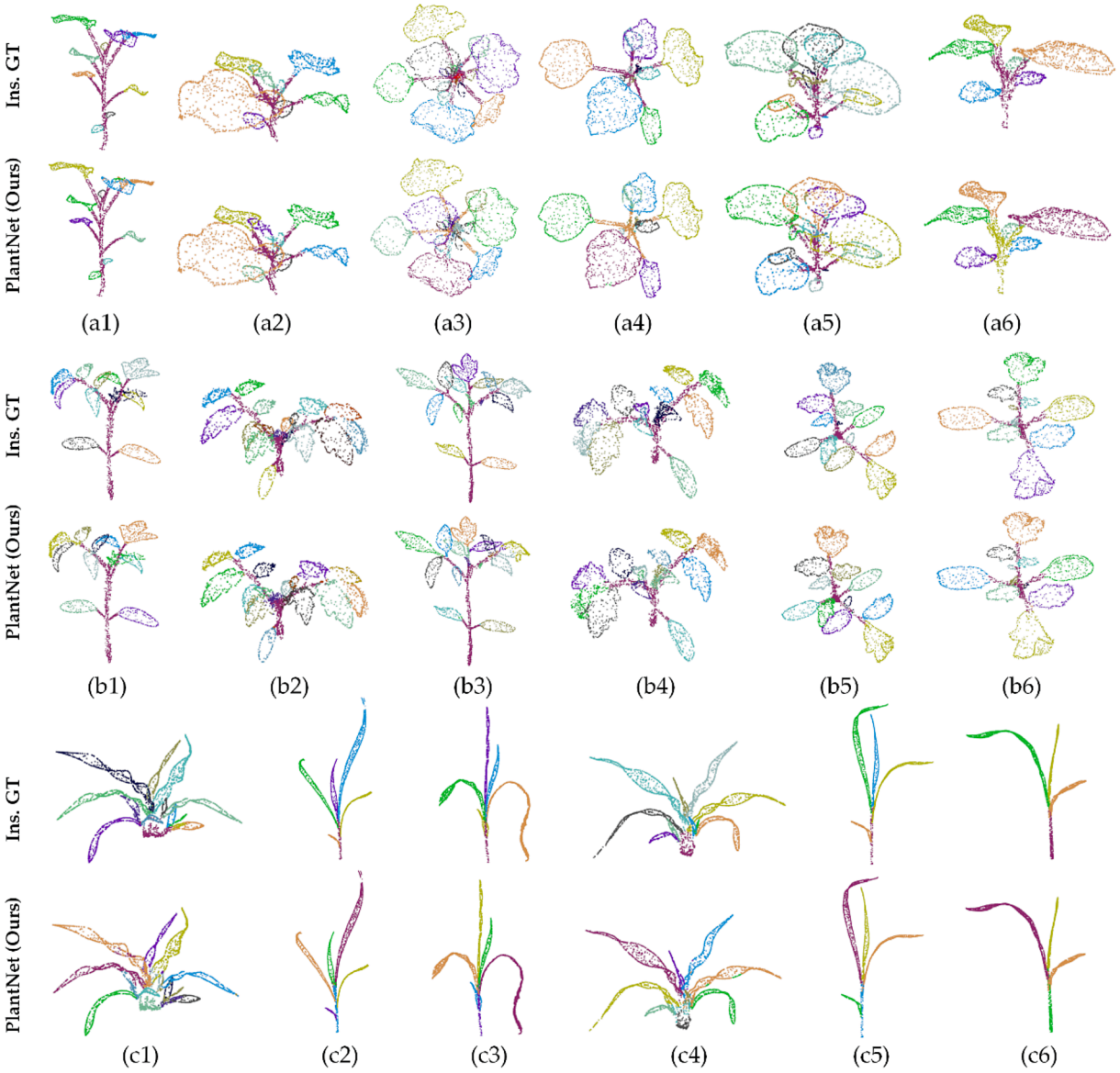


Fig. 9. Demonstration of the qualitative segmentation results for the three types of plants by PlantNet. The leaf instance ground truths and the corresponding outputs of PlantNet of several samples of tobacco are shown on the 1st and the 2nd rows, respectively. The leaf instance ground truths and segmentation results of several tomato samples are shown on the 3rd and the 4th rows, respectively. The last two rows show the leaf instance ground truths and PlantNet results of several sorghum samples, respectively. Each leaf instance is rendered by a different color. Those different leaf colors are only used for visualization, and there is no specific label meaning for each color.

where $\left|TP_i^{ins}\right|$ is the number of instances that both belong to a semantic category i and have an IoU with ground truth larger than 0.5; $\left|P_i^{ins}\right|$ is the total number of predicted instances of the category i, and $\left|G_i^{ins}\right|$ is the number of the instances of category i in the ground truth. As we only evaluate the leaf instances of three species of plants, thereby C was set as 3 in (26) and (27).

4. Results

4.1. Semantic segmentation results

The semantic segmentation results of the three plants at several different growth stages were assessed qualitatively and quantitatively. The displayed samples in Fig. 8 are chosen with diversified spatial structures and under different growth environments to exhibit the good generalization ability and accuracy of PlantNet. The results show that PlantNet has a high sensitivity to detecting small leaves. For example, the network successfully detected the small leaf around the bifurcation point on the main stem of the tobacco samples in Fig. 8(a6).

The tomato plant has a more complex canopy structure and more leaves compared with the other two kinds of crops in the dataset. The tomato crop samples have large variances not only on the spatial structure but also on the degree of the leaf curvature, and even the shapes of leaflets are not unified in the same compound leaf (see Fig. 8 (b1)-(b6)). Our PlantNet still shows good segmentation results on tomato samples. For sorghums, it is difficult to distinguish the exact junction (sheath) between the leaf and the stem. However, the PlantNet still shows a high discernibility on the sheath areas.

Meanwhile, we also noticed that under the same number of input points and a fixed training batch, PlantNet has better segmentation accuracy in simpler plant structures. For example, the sample of Fig. 8(a5) is older (more complex structure) than the sample of Fig. 8(a1), leading to a sparser point coverage on each leaf. Therefore, some petiole points at the end of several leaves are falsely classified as leaf points in Fig. 8 (a5). An easy solution to this problem is to increase the input number (N) of points for the network. The influence of N on the performance of the network will be further discussed in section 5.2.

The quantitative results of the semantic segmentation on the testing set are shown in Table 2. The quantitative results of the tomato are generally better than the other two species, and almost all quantitative measures of the tomato are higher than 93% except for the stem *IoU*. The most likely explanation for the good quantitative results on tomato plants is that the training data of the tomato is richer than that of the tobacco and the sorghum. For most neural networks, the amount of training data directly affects the performance of the prediction. As shown in the last row of Table 2, the mean value of the stem is lower than the value of the leaf for all *Precision*, *Recall*, *F1-score*, and *IoU* measures. This phenomenon may have two reasons: (i) the stem systems have diverse and complex spatial structures, making it harder to be segmented than leaves by the network. (ii) the number of stem points is much smaller than that of the leaf class in the training set, resulting in a training imbalance for semantic segmentation.

4.2. Instance segmentation results

The instance segmentation results of the three plants at several different growth stages were also assessed qualitatively and quantitatively. The qualitative results of leaf instance segmentation for some representative samples are shown in Fig. 9. The instance segmentation results of tobacco samples at different growing stages show that all leaves in the hierarchical canopies of the six tobacco plants are well separated, and even the small leaves near the bottom of the main stems are perfectly segmented Fig. 9(a1), (a5), and (a6)). With the help of the proposed 3DEPS strategy, PlantNet also shows good instance segmentation performance for mature plants with more overlapping leaves in the dataset (e.g., Fig. 9(a5)). Meanwhile, the instance segmentation results of tomato samples show that the PlantNet successfully segments leaflets in compound leaves on the right branch of Fig. 9(b2) and the left branch of Fig. 9(b4), although the biggest challenge in leaf instance segmentation is the overlapping of leaves. Moreover, how to effectively define the exact junction (sheath) between the main stem and the leaf is the key to improving the segmentation accuracy on the monocotyledonous plant. The instance segmentation results of some sorghum samples show that the PlantNet accurately locates all leaf sheath positions (Fig. 9(c1)-(c6)). In particular, all single leaves in the two complex samples in Fig. 9(c1) and (c4) are well segmented with PlantNet.

Table 3
Quantitative results of leaf instance segmentation of PlantNet on plants.

	Tobacco Leaf	Tomato Leaf	Sorghum Leaf
mPrec (%)	84.13	84.66	81.11
mRec (%)	67.48	73.58	81.19
mCov (%)	71.11	81.35	83.41
mWCov (%)	81.06	85.45	86.62

The quantitative results of leaf instance segmentation for the three different species calculated from the testing set are listed in Table 3. The *mPrec* values of the three crops shown in Table 3 are all higher than 80%, indicating that the contour description of the segmented leaves is relatively accurate. The *mRec* values of tobacco and tomato are smaller than their corresponding *mPrec* values, this is mainly because the samples of these two dicotyledonous plants in the testing set have crowded canopies, which are prone to serious overlapping. The *mCov* measure can be understood as the average *IoU* of all leaf instances. The *mCov* measures of the tobacco, tomato, and sorghum plants are 71.11%, 81.35%, and 83.41%, respectively. Comparing to *mCov*, the *mWCov* is a weighted average *IoU* of all leaf instances, which can better reflect the segmentation performance on mature leaves. The *mWCov* values are all higher than 80% (Table 3), showing the satisfactory instance segmentation performance of PlantNet.

4.3. Comparison with other methods

Several popular point cloud segmentation networks are selected for comparison with PlantNet, and we used recommendations from their original papers for parameter selection in training and testing. Among them, PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b) can only conduct semantic segmentation, so we trained and tested them with only semantic labels. SGPN (Wang et al., 2018b) and ASIS (Wang et al., 2019a) can realize semantic segmentation and instance segmentation at the same time, so we trained and tested the two networks with the same data as used by the PlantNet. In the accuracy and efficiency comparison experiment, the experimental environment and the parameter configuration of PlantNet are the same as those specified in Section 3.3. Table 4 shows the quantitative comparison for semantic segmentation across the five networks.

It can be observed that PlantNet has the best semantic segmentation performance in most cases across all methods compared in Table 4. The averages of *Precision*, *Recall*, *F1–score*, and *IoU* of PlantNet are all around 1% higher than that of the second-best model (PointNet++), respectively. The highest difference between PlantNet and the second-best (PointNet++) of the four quantitative measures falls on the average Precision (with a 1.44% gap). PlantNet has balanced advantages on both dicotyledonous and monocotyledonous plants, showing good adaptability on different species. The dual-function networks—SGPN and ASIS perform relatively poorly on the semantic task (Table 4) due to the interactions on their networks to balance the semantic segmentation task and the instance segmentation task in training. By contrast, PlantNet balances the semantic segmentation and the instance segmentation tasks while keeping the best semantic segmentation results (Table 4), because PlantNet also shows the highest mean quantitative measures in

Table 4
The quantitative comparison for semantic segmentation across the five methods. The best results are in boldface, and the 2nd best results are underlined.

		Tobacco		Tomato		Sorghum		Mean
		Stem	leaf	Stem	leaf	Stem	leaf	
Precision (%)	PointNet	80.80	89.50	90.49	95.15	71.54	97.06	87.42
	PointNet++	88.35	94.23	93.02	95.79	76.04	**98.86**	91.05
	SGPN	75.59	89.49	85.07	96.04	40.44	98.48	80.85
	ASIS	82.28	93.37	93.36	96.30	67.91	97.13	88.39
	PlantNet	**91.47**	**95.04**	**93.83**	**96.44**	**80.15**	98.03	**92.49**
Recall (%)	PointNet	77.18	91.38	88.51	96.04	62.04	**98.07**	85.54
	PointNet++	**89.94**	93.42	92.42	96.23	79.05	98.05	91.52
	SGPN	77.97	88.17	90.97	93.20	82.13	90.54	87.16
	ASIS	84.95	93.87	88.86	96.74	75.32	97.94	89.61
	PlantNet	85.29	**94.28**	**93.03**	**97.14**	**84.67**	97.82	**92.04**
F1-score (%)	PointNet	78.95	90.43	89.49	95.60	66.45	97.56	86.41
	PointNet++	**89.13**	93.83	92.92	96.01	77.52	**98.20**	91.27
	SGPN	76.76	88.82	87.92	94.60	54.19	94.34	82.77
	ASIS	83.59	93.62	91.06	96.52	71.42	97.53	88.96
	PlantNet	88.27	**94.55**	**93.26**	**96.59**	**82.24**	97.89	**92.13**
IoU (%)	PointNet	65.22	82.53	80.98	91.56	49.76	95.24	77.55
	PointNet++	**80.40**	89.16	86.77	93.19	63.29	**96.47**	84.88
	SGPN	62.29	79.89	78.45	89.75	37.17	89.29	72.81
	ASIS	71.81	88.00	83.58	93.27	55.55	95.18	81.23
	PlantNet	79.01	**89.66**	**87.37**	**93.40**	**69.84**	95.87	**85.86**

Table 5
The quantitative comparison for instance segmentation across three methods. The best results are in boldface; the second-best results are underlined.

		Tobacco Leaf	Tomato Leaf	Sorghum Leaf	Mean
mPrec (%)	SGPN	36.37	50.77	34.21	40.45
	ASIS	81.72	84.15	79.54	81.80
	PlantNet	**84.13**	**84.66**	**81.11**	**83.30**
mRec (%)	SGPN	40.32	50.36	39.36	43.35
	ASIS	67.20	70.52	78.51	72.08
	PlantNet	**67.48**	**73.58**	**81.19**	**74.08**
mCov (%)	SGPN	46.69	62.95	47.06	52.24
	ASIS	70.87	79.77	79.62	76.75
	PlantNet	**71.11**	**81.35**	**83.41**	**78.62**
mWCov (%)	SGPN	55.64	69.13	53.76	59.51
	ASIS	80.94	83.82	82.92	82.56
	PlantNet	**81.06**	**85.45**	**86.62**	**84.38**

Table 6
Comparison of the five methods on both training speed and testing speed.

	PointNet	PointNet++	ASIS	SGPN	PlantNet
Training time (s)	4224.1	7168.5	17192.0	65738.9	74724.9
Testing time per point cloud (ms)	36.0	68.3	124.7	167.8	247.6

instance segmentation (Table 5).

The quantitative comparison for instance segmentation among PlantNet, ASIS, and SGPN on our dataset shows that the PlantNet obtains the best results on all four measures—*mPrec*, *mRec*, *mCov*, and *mWCov*, not only for all single species but also for the total mean of the testing set (Table 5). PlantNet also has more obvious accuracy improvements, compared to ASIS, on sorghum leaf than the other two species, which reveals that PlantNet is more good at the instance segmentation task of the complex monocotyledonous leaves.

The comparison of the training and testing time of the five methods (Table 6) shows that all methods have to be trained for several hours (within one day), of which PlantNet has the longest training time, similar to SGPN, followed by ASIS, PointNet++, and PointNet that have simpler architectures or only make semantic segmentation. As we know, the hour-level difference of training time is very common and should not be a big deal for training a deep learning network, which is usually a long process and can be trained offline. For the application of a deep learning model, the testing time is more important. Although PlantNet was slower than other methods, it takes less than 1 s for segmenting a single point cloud, satisfying most real-time requirements.

To sum up, PlantNet outperformed the state-of-the-art deep learning networks including PointNet, PointNet++, SGPN, and ASIS, which achieved an average improvement of 5.56%, 3.58%, 4.78%, and 6.74% in *Precision, Recall, F1-score, IoU* on semantic segmentation, and an average improvement of 22.18%, 16.37%, 14.13%, and 13.35% in *mPrec*, *mRec*, *mCov*, and *mWCov* on instance segmentation. Besides, it demonstrates good generalization ability on plant varieties, and also reflects the effectiveness of the proposed 3DEPS strategy for data preparation. More importantly, the training time of PlantNet is comparable to the compared methods, and testing time should be feasible for most real-time applications. To the best of our knowledge, PlantNet is the first end-to-end point cloud deep learning network that can perform dual-function segmentations for three different plant species simultaneously.

5. Discussion

5.1. The sampling ratio

This subsection discusses the optimal value of the *ratio* parameter for PlantNet. The average ratio distribution of edge points to all points of a plant point cloud in the original dataset before using 3DEPS is 0.074 (see methods in Appendix A.1). To analyze the influence of the sampling ratio, a parameter tuning experiment is constructed to find the optimal *ratio* of 3DEPS on PlantNet segmentation, with 0.05 as the *ratio* step size changing from 0.0 to 1.0. In the parameter tuning experiment, different datasets were formed separately by using different sampling ratios, which were used to train different versions of PlantNets.

The parameter tuning results for the *ratio* of 3DEPS on semantic and instance segmentation are shown in Fig. 10. As the proportion of edge points gradually increases, the performance of semantic segmentation increases at first, then falls with fluctuations. The highest semantic segmentation accuracy is achieved when *ratio* is between 0.1 and 0.2, and the performance drops after 0.2 (Fig. 10(a)). Similarly, the instance segmentation accuracy of PlantNet rises rapidly when *ratio* increases from 0.0 to 0.2, and then decreases gradually after 0.2 (Fig. 10(b)). All metrics reach the peak at the same ratio of 0.2. This high degree of consistency not only shows that the trend is reliable but also implies statistical significance.

Because both segmentation performances of PlantNet peaked around

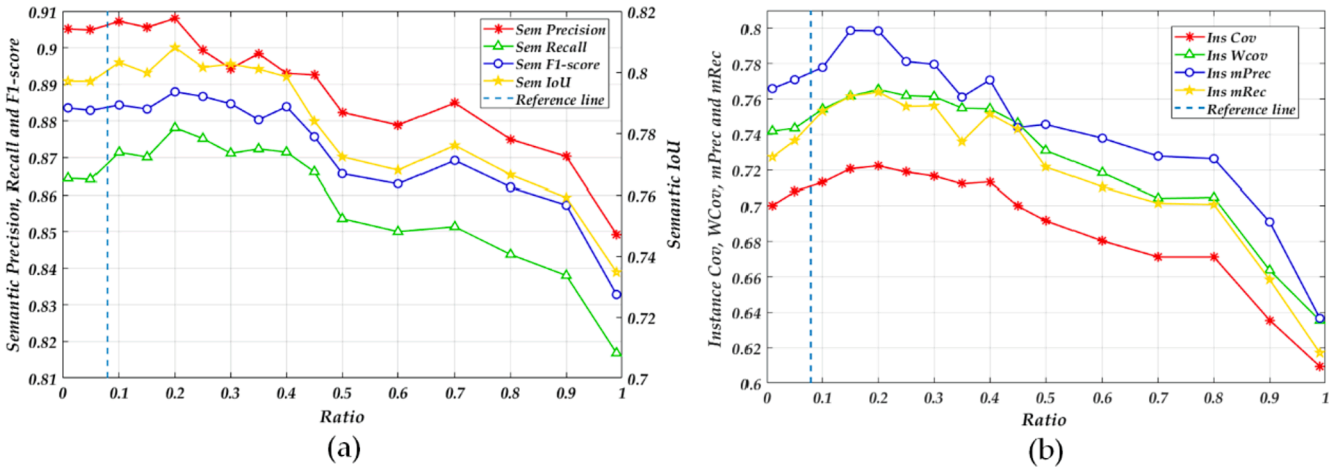


Fig. 10. The parameter tuning of the *ratio* parameter of the 3DEPS strategy. (a) exhibits the influence of the changes of *ratio* on the four semantic segmentation measures of PlantNet; (b) displays the influence of the changes of *ratio* on the four instance segmentation measures of PlantNet. The blue dotted line perpendicular to the x-axis in the figure is the case of *ratio* = 0.074, which is equivalent to using IRFPS on the original point cloud for down-sampling. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

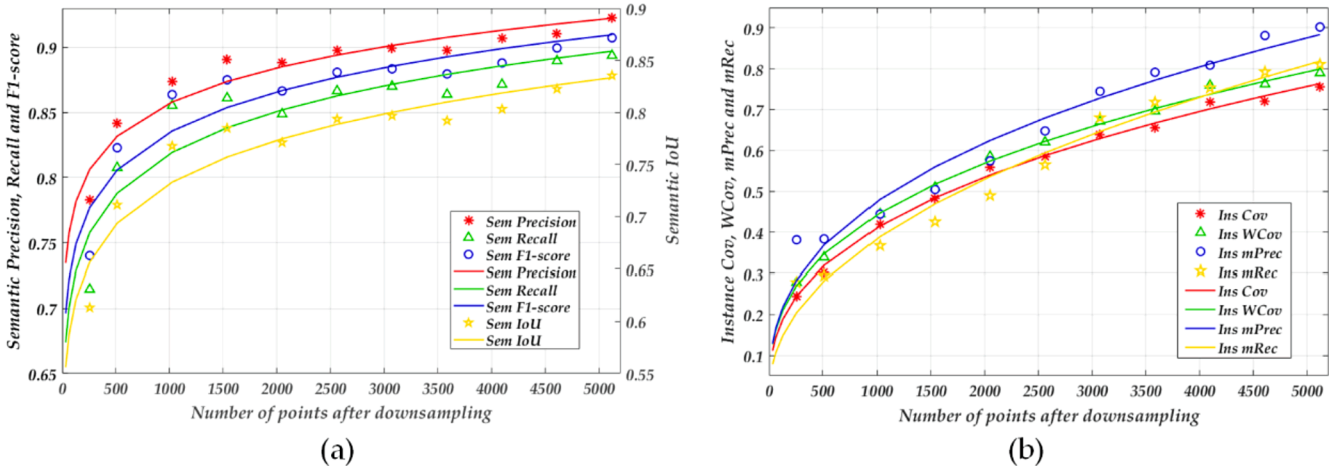


Fig. 11. The influence of the number of down-sampled points on PlantNet. (a) shows the changes of the four quantitative measures for semantic segmentation with the increase of *N*, and we perform curve fitting for all tested values to exhibit the trend. The measures are computed with *N* set at 256, 512, 1024, 1536, 2048, 2560, 3072, 3584, 4096, 4608, and 5120, respectively. (b) shows the changes of the four quantitative measures for instance segmentation with the increase of N, and the curve fitting is also performed.

ratio = *0.2*, the *ratio* for 3DEPS in all experiments is fixed to 0.2 in this study. All eight quantitative measures of the 3DEPS case (*ratio* = 0.2) outperform the case of using the *ratio* of original dataset or sampled by IRFPS strategy, i.e., *ratio* = *0.074* , as the blue dotted line shown in Fig. 10. This result proves that introducing adequately more edge points in down-sampling can enhance the understanding of point clouds on PlantNet. In other words, 3DEPS can be interpreted as a generalized IRFPS strategy. 3DEPS could degenerate to a standard IRFPS when the "*ratio*" is set the same as the proportion of edge points to total points before down-sampling. Moreover, because the "*ratio*" has a noticeable influence on segmentation accuracies, 3DEPS should also outperform other similar simple sampling methods (e.g., random sampling) that do not consider the difference between edge and non-edge points. Additionally, we visually found that the uniformity of the internal points of small leaves after 3DEPS could be improved (Fig. 3(b2)), perhaps a regularization method can be introduced for further improvement of the sampled inner sections of plants in the future.

5.2. The number of down-sampled points

Under a fixed number of input points for each sample, a more complicated plant (e.g., more branches and leaves) tends to have unsatisfactory segmentation results by PlantNet. One reason for this phenomenon is that the limited number of sampling points (*N*) is insufficient to represent the structural details, especially on small leaves. To solve this problem, despite designing effective down-sampling strategies such as 3DEPS, increasing *N* for each single point cloud while satisfying the limit of the computation capacity determined by the hardware system, is also important. In this section, the influence of *N* after 3DEPS with the *ratio* fixed at 0.2 on PlantNet segmentation performance is explored.

As shown in Fig. 11, the performances of semantic segmentation and instance segmentation both increase with the number of input points. When *N* is less than 512, the measures of both segmentation tasks are low because insufficient input points may cause difficulty on semantic association in the local connectivity in the point cloud. When *N* grows from 256 to 1024, the semantic segmentation accuracy of the network increases significantly. Within the range of 1024 to 5120 points, the semantic segmentation accuracy gradually increases. At *N* = 1024, the

Table 7
The ablation analysis for semantic segmentation of PlantNet. The best results are in boldface.

		Tobacco		Tomato		Sorghum		Mean
		Stem	leaf	Stem	Leaf	Stem	leaf	
Precision (%)	PlantNet\FFM	90.05	85.87	92.75	95.26	79.70	96.29	89.99
	PlantNet\DHL	89.71	93.87	92.05	95.03	78.31	97.41	91.06
	PlantNet\LFEO	89.56	93.01	92.21	95.58	78.12	96.54	90.84
	PlantNet	**91.47**	**95.04**	**93.83**	**96.44**	**80.15**	**98.03**	**92.49**
Recall (%)	PlantNet\FFM	84.25	91.16	90.59	94.58	75.61	96.99	88.86
	PlantNet\DHL	84.25	90.60	92.69	96.29	80.15	97.21	90.20
	PlantNet\LFEO	84.36	91.45	90.68	95.23	78.54	97.01	89.55
	PlantNet	**85.29**	**94.28**	**93.03**	**97.14**	**84.67**	**97.82**	**92.04**
F1-score (%)	PlantNet\FFM	87.11	89.82	91.10	94.92	79.89	97.13	90.00
	PlantNet\DHL	87.95	93.16	92.87	96.15	77.65	97.31	90.85
	PlantNet\LFEO	87.56	92.22	91.58	96.35	78.56	97.22	90.58
	PlantNet	**88.27**	**94.55**	**93.26**	**96.59**	**82.24**	**97.89**	**92.13**
IoU (%)	PlantNet\FFM	78.75	81.52	86.09	90.33	66.52	94.43	82.94
	PlantNet\DHL	78.49	87.19	86.69	92.58	63.47	94.76	83.86
	PlantNet\LFEO	78.86	85.92	86.58	91.34	65.86	94.01	83.76
	PlantNet	**79.01**	**89.66**	**87.37**	**93.40**	**69.84**	**95.87**	**85.86**

Table 8
The ablation analysis for instance segmentation of PlantNet. The best results are in boldface.

		Tobacco Leaf	Tomato Leaf	Sorghum Leaf	Mean
mPrec (%)	PlantNet\FFM	78.32	82.79	76.27	79.13
	PlantNet\DHL	80.28	83.86	77.98	80.71
	PlantNet\LFEO	89.32	83.51	76.95	83.26
	PlantNet	**84.13**	**84.66**	**81.11**	**83.30**
mRec (%)	PlantNet\FFM	65.57	70.46	77.18	71.07
	PlantNet\DHL	66.62	71.44	75.54	71.20
	PlantNet\LFEO	66.01	70.86	76.95	71.27
	PlantNet	**67.48**	**73.58**	**81.19**	**74.08**
mCov (%)	PlantNet\FFM	70.24	79.98	79.45	76.56
	PlantNet\DHL	**71.15**	80.85	78.87	76.96
	PlantNet\LFEO	70.56	80.75	80.32	77.21
	PlantNet	71.11	**81.35**	**83.41**	**78.62**
mWCov (%)	PlantNet\FFM	80.44	83.09	82.52	82.02
	PlantNet\DHL	80.29	84.97	82.26	82.51
	PlantNet\LFEO	80.51	84.22	82.19	82.31
	PlantNet	**81.06**	**85.45**	**86.62**	**84.38**

four semantic measures are all above 82%. Therefore, 1024 downsampled points can already generate satisfactory semantic segmentation results for PlantNet, which is friendly to the system with a limited GPU memory. Similar to semantic segmentation, the four quantitative measures of instance segmentation increase rapidly with the increase of *N*. The instance segmentation task seems to be more sensitive to *N* than the semantic segmentation task of PlantNet. The reason may be the network needs more points to cover each instance for effective deep learning, considering the number of instance classes is more than the semantic class, and each instance target is smaller than any semantic target. When *N* becomes larger in Fig. 11(b), the points of each instance also increase, which brings in better instance segmentation results. In practice, we also have to consider the trade-off between accuracy and efficiency influenced by the point number (Appendix A.2).

5.3. Ablation analysis

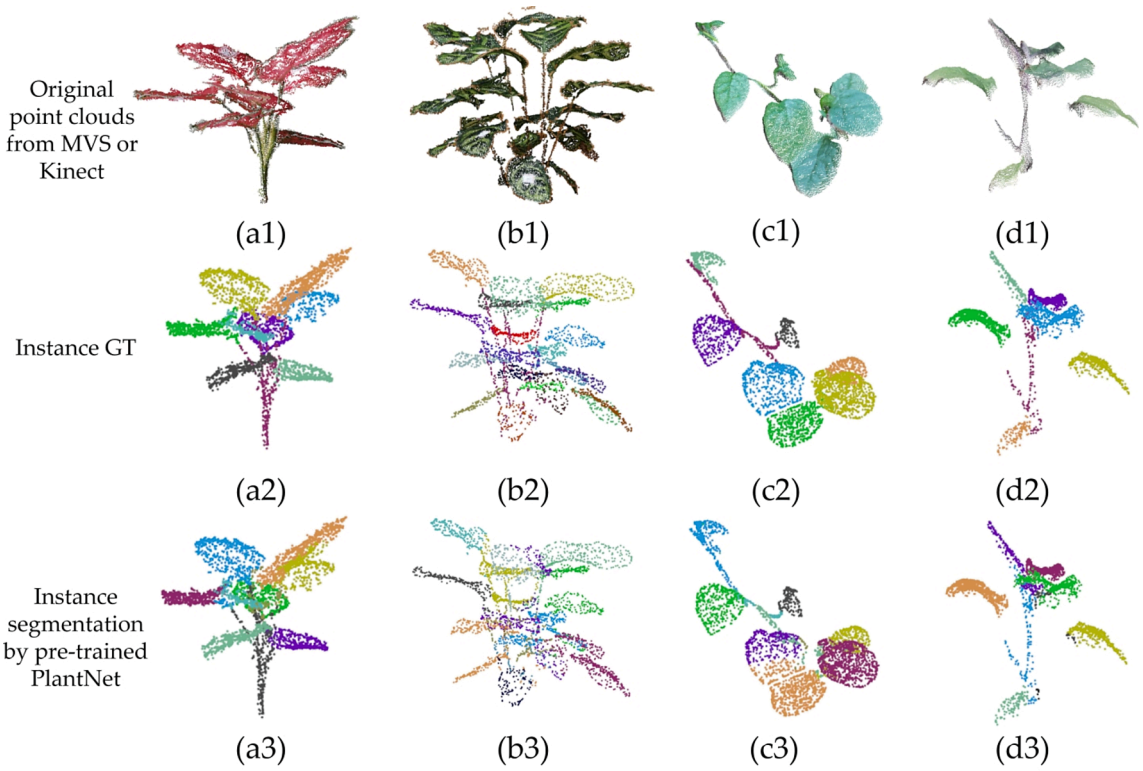
The quantitative ablation analysis (see methods in Appendix A.3) for semantic segmentation and instance segmentation are shown in Table 7 and Table 8, respectively. All measures show that the removal of DHL ("PlantNet\DHL"), LFEO sub-module ("PlantNet\LFEO"), or the FFM ("PlantNet\FFM") will all deteriorate the performance of PlantNet. For semantic segmentation, the mean decrease of *Precision*, *Recall*, *F1-score*, and *IoU* can be up to 2.5%, 3.18%, 2.13%, and 2.92%, respectively. For instance segmentation, the mean decrease of *mPrec, mRec, mCov,* and *mWCov* can be up to 4.17%, 3.01%, 2.06%, and 2.36%, respectively. We also compare the time costs for models that appear in the ablation analysis and find all versions of PlantNet have similar time costs in the training and testing process. Additionally, feature visualization analyses on different function modules of PlantNet are given to provide a more vivid understanding of their effectiveness in Appendix A.4.

5.4. PlantNet on different types of 3D data

All original point clouds in our dataset were collected by an expensive scanner and the data acquisition process was also time-consuming. Inexpensive and reliable 3D imaging techniques provide potential new data sources for the widespread use of point clouds in the field of plant phenotypes in the future. Thus, it is necessary to check out whether the pre-trained PlantNet can work directly on other types of point clouds captured from several different imaging techniques.

This subsection mainly verifies two assumptions via experiments: (i) pre-trained PlantNet on the current dataset can be applied to point clouds of new crop species with similar structures; and (ii) pre-trained PlantNet can be applied to the leaf instance segmentation task of crop point clouds obtained by other imaging techniques; e.g., the Multi-view Stereo (MVS) and Kinect V2 imaging. The details of the two imaging systems are given in Appendix A.5. When validating the pre-trained PlantNet on a point cloud of a new species, only the leaf instance segmentation task is tested because it is more challenging and can reflect the performance of semantic segmentation.

In this test, two ornamental plants—*Dieffenbachia picta* and *Maranta arundinacea* are reconstructed by MVS (Fig. A4 (a), (b)). A total of 47 and 52 images with a resolution of 2736 × 3648 are captured for a *Dieffenbachia picta* and a *Maranta arundinacea* using the camera on a mobile phone, respectively. Then, MVS is used to generate two point clouds shown in Fig. 12(a1) and (b1). Fig. 12(a3) and (b3) are the instance segmentation results on (a1) and (b1) by the pre-trained PlantNet, respectively. In terms of the qualitative aspect, the pre-trained PlantNet can already produce satisfactory leaf instance segmentation results for MVS-imaged crop species that have not been trained before. Fig. 12(c1)



	Dieffenbachia picta	*Maranta arundinacea*	*Vigna unguiculata*	*Solanum melongena*
mPrec	80.0%	68.8%	70.0%	75.0%
mRec	88.9%	52.8%	87.5%	85.7%
mCov	82.9%	43.8%	74.4%	89.5%
mWCov	84.1%	47.1%	75.8%	94.4%

Fig. 12. Using pre-trained PlantNet to conduct leaf instance segmentation on plant point clouds generated by MVS and Kinect V2. (a1) and (b1) are a *Dieffenbachia picta* and a *Maranta arundinacea* point clouds from the MVS system, respectively. (c1) and (d1) are a *Vigna unguiculata* and a *Solanum melongena* point clouds from the Kinect V2 system, respectively. All of the original clouds have colors. (a2)-(d2) are the instance ground truths corresponding to (a1)-(d1), respectively. (a3)-(d3) shows the predicted leaf instance segmentation results by pre-trained PlantNet on the four point clouds, respectively. Different colors in the 2nd row and the 3rd row of this figure are only used to help visualizing different leaves, and they are not associated with specific leaf indices. The quantitative measures of the four plants are given at the bottom of the figure.

and (d1) show point clouds of a Cowpea plant (*Vigna unguiculata*) and an eggplant (*Solanum melongena*) collected by the Kinect V2 platform. Fig. 12(c3) and (d3) are the instance segmentation results of Fig. 12(c1) and (d1) using the pre-trained PlantNet, respectively. The results of Fig. 12(c3) and (d3) show the pre-trained network can directly segment leaves on new plant species scanned by the Kinect V2 platform, which again proves the good generalization ability of PlantNet. Moreover, the quantitative comparison results show that half of the measures are larger than 80%. The accuracies of *Maranta arundinacea* (Fig. 12(b3)) are not as good as the others because its canopy structure contains multiple main stems, different from the three single-main-stem species in our training dataset. However, the average quantitative measures for the four plants still surpass 70%. The average *mPrec*, *mRec*, *mCov*, and *mWCov* are 73.4%, 78.7%, 72.7%, and 75.4%, respectively.

6. Conclusion

Although AI applications in agriculture have made rapid progress in recent years, three challenges still exist for the 3D organ segmentation of plants. First, we are in shortage of well-labeled 3D plant datasets. Second, it is difficult to simultaneously achieve point-level organ semantic segmentation and instance segmentation using one model. Third, domain adaptation is a prominent problem in deep learning, making the existing deep learning networks mainly focus on single species and cannot be generalized to other species. To address these challenges, we first constructed a labeled 3D dataset containing 5460 individual-plant point clouds of three different crop species under different growth environments covering a 20-day growth process after data augmentation. Secondly, we designed a novel point cloud down-sampling strategy (i.e., 3DEPS), which preserves a better global 3D structure by intentionally increasing the proportion of edge points, and hence enhances the accuracy of organ segmentation without changing the total number of sampled points. Finally, we proposed a dual-function deep learning network (PlantNet) that can simultaneously conduct semantic segmentation of leaves and stems and leaf instance segmentation for serval plant species with different architecture, including the dicotyledonous (i.e.,

tobacco and tomato) and monocotyledonous (i.e., sorghum) plants. The semantic segmentation results of tobacco, tomato, and sorghum in average *Precision*, *Recall*, *F1-score*, and *IoU* reached 92.49%, 92.04%, 92.13%, and 85.86%, respectively; and the instance segmentation results in the mean precision (*mPrec*), the mean recall (*mRec*), the mean coverage (*mCov*), *and* the mean weighted coverage (*mWCov*) reached 83.30%, 74.08%, 78.62%, and 84.38%, respectively. Compared with several mainstream point cloud segmentation networks, PlantNet has better segmentation results on both qualitative and quantitative aspects due to the self-designed 3DEPS, network submodules (i.e., LEFO, FFM), and the novel loss function.

In the future, we will use several different 3D imaging and reconstruction tools to enrich the plant datasets and introduce new plant species to the dataset to form a more accurate, robust, and universal plant organ segmentation network. We believe the efforts of this study will contribute to the development of high-throughput plant phenotyping and intelligent agriculture.

Author contributions

D. Li, G. Shi, and S. Jin wrote the paper and crafted all figures and tables. D. Li and G. Shi designed the whole architecture of PlantNet with comments from S. Jin. G. Shi and J. Li carried out experiments. J. Li, Y. Chen, and S. Xiang prepared the dataset. S. Zhang organized the references. All authors read and approved the final manuscript.

Funding

This work was supported in part by the Shanghai Rising-Star Program (No.21QA1400100), Jiangsu Agricultural Science and Technology Independent Innovation Fund Project (No. CX(21)3107), Shanghai Natural Science Foundation (No. 20ZR1400800), Shanghai Sailing Program (No. 20YF1401600), Fundamental Research Funds for the Central Universities of China (No. 2232020D-47), High Level Personnel Project of Jiangsu Province (JSSCBS20210271), and in part by the Plant Phenomics Research Program of Science and Technology Department of Jiangsu Province (No. BM2018001).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

A.1 Method for deriving the ratio of the whole dataset before down-sampling

The linear regression method was used to derive the average proportion of the edge points to the non-edge points based on the magnitude of the slope. The result shows that the average proportion of the edge points to the non-edge points is about 0.0798 for the whole dataset (Fig. A1). This means that the edge points account for about 7.4% of all points before down-sampling.

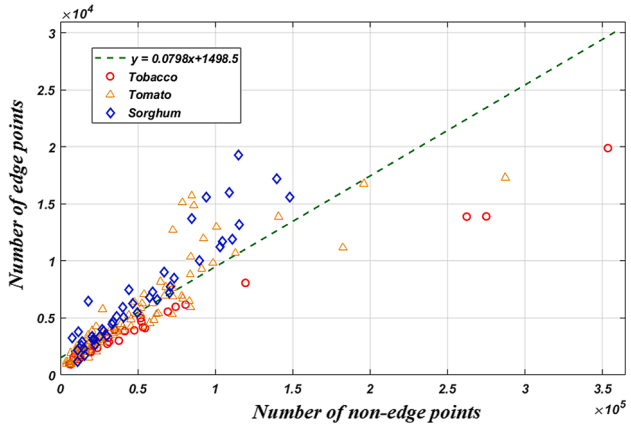


Fig. A1. The distribution about the ratio of the edge points to the non-edge points of all point cloud samples before down-sampling. Each point in the figure represents an original plant point cloud. The red circles are tobacco plants; the yellow triangles are the tomato plants, and the blue rhombi are sorghums. Note that there is a 10 times relation between the scales of the x-axis and the y-axis, and the dotted green line is the regression line fitted to all points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A.2 The influence of point number on the efficiency and accuracy

This study analyzes the influence of the point number of each sample on the model performance and training time. The different point numbers include 1024, 2048, and 4096, of which the maximum number (4096) was determined according to the limitation of GPU memory in this study. The comparison results (Table A1) showed that both the training and testing times of PlantNet grow (almost linearly) with the number of points. Meanwhile, we can see a dramatic increase in all accuracy measures. Although there is a small sacrifice in time with the increase of point number, the testing time is almost at the millisecond level that does not affect most real-time applications. We acknowledge that point number is an import parameter and should be considered according to the trade-off between efficiency and performance.

Table A1
The speed and accuracy comparisons of PlantNet under different numbers of down-sampled points.

Number of points	1024	2048	4096
Training time (s)	14446.22	25392.48	74724.92
Testing time per point cloud (ms)	48.8	74.0	247.6
Precision	87.3%	89.1%	92.0%
Recall	85.5%	85.0%	87.2%
F1-score	86.0%	86.6%	88.9%
IoU	82.4%	82.9%	85.1%
mPrec	44.8%	58.0%	80.7%
mRec	37.5%	49.6%	74.8%
mCov	41.6%	56.2%	71.5%
mWCov	44.9%	59.1%	77.3%

A.3 Methods for ablation analysis

To verify the effectiveness of the designed DHL, the LFEO sub-module, and the FFM for semantic and instance segmentation. The network that removes DHL from the total loss function is named as "PlantNet\DHL", which corresponds to remove L_d from (16). The ablation of LFEO modules ("PlantNet\LFEO") is realized by replacing all the three EdgeConvs in each LFEO in the shared encoder with common MLP operations, and the network's overall depth is kept unchanged after ablation to ensure the fairness of comparison. The ablation of FFM ("PlantNet\FFM") is realized by directly removing the feature layer F_{fusion} from the standard PlantNet architecture while keeping all other parts unchanged. The above three ablation networks are compared with the standard PlantNet under the same parameter configuration and the training environment.

A.4 Feature visualization

To visually verify the effectiveness of the feature extraction and aggregation of some modules in the network, we focus on two visualization tasks: (i) are there some traces of the semantic and instance feature abstraction on the mid-level feature layers of PlantNet (e.g., FFM)? (ii) whether the EdgeConv-based LFEOs can gradually concentrate and extract nascent semantic information from low-level features?

A.4.1 Visualizing the feature fusion space

The third part of PlantNet is the FFM for the combination of semantic information and instance information. To search for the traces of semantic and instance feature abstraction on mid-level layers and explain the effectiveness of FFM, the feature fusion space F_{fusion} is visualized. More specifically, three phenomena are expected to be seen in this fusion space: (i) points from different semantic categories should be as far as possible; (ii) points from different instances but the same semantic category should congregate to several nearby clusters; (iii) points belonging to the same instance should be as close as possible.

The mapping from the semantic/instance ground truths to the fusion feature embedding after dimensionality reduction by T-SNE (Van der Maaten and Hinton, 2008) is shown in Fig. A2. In semantic visualizations (the 2nd column of Fig. A2), the points of the stem class and the leaf class in each sample are already well separated in the fusion embedding after dimension reduction, which proves the effectiveness of semantic abstraction of FFM. In instance visualizations (the 4th column of Fig. A2), different leaves form different clusters in the embedding, and all leaf instances fall within the corresponding leaf semantic region on the second column, indicating that the instance features have been effectively learned. The visualizations of F_{fusion} indirectly prove that the FFM and the DHL function of PlantNet are effective and reasonable.

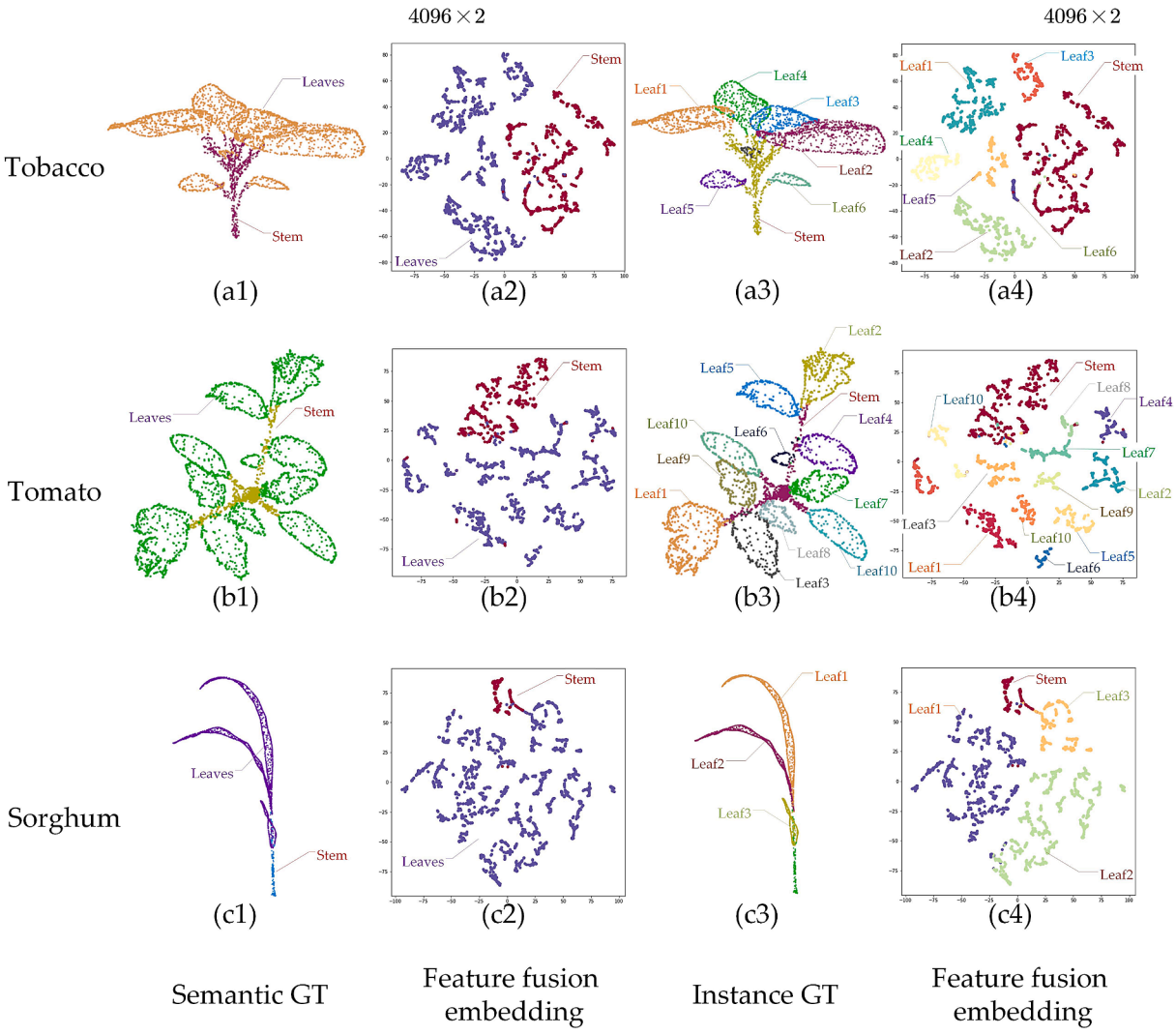


Fig. A2. The correspondence between the feature fusion embedding and the semantic/instance ground truths for three plant samples. Since the dimension of the embedding of F_{fusion} is 128, T-SNE is used to reduce the 128-D space to a 2D space for visualization. (a1), (b1), and (c1) are the semantic ground truths of a tobacco, a tomato, and a sorghum sample, respectively. (a2), (b2), and (c2) are the color-rendered feature fusion spaces by semantic categories in (a1), (b1), and (c1), respectively; and the stem points are rendered in dark red while the leaf points are rendered in purple. (a3), (b3), and (c3) are the instance ground truths of the same three plant samples, respectively. (a4), (b4), and (c4) are the color-rendered feature fusion spaces by leaf instances in (a3), (b3), and (c3), respectively; and different colors represent different leaf instances. Please note that for the same leaf instance, the color of the leaf instance in the ground truth (on the 3rd column) and the color of the leaf instance feature points are not necessarily the same. This is due to the restriction of our visualization tool. Therefore, we label the exact leaf indices on the 3rd and the 4th columns to better present the correspondences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A.4.2 Visualizing the concentration of high-level features in LFEO

The LFEO is composed of three consecutive EdgeConvs. Each EdgeConv operation is a dynamic process of finding local adjacent points in the feature space on the previous EdgeConv output. Theoretically, the gradual concentration of high-level semantic information should be observed in the feature space after applying EdgeConvs.

The normalized *L2* distance in feature space between all other points and the target point at three different stages of the 1st LFEO in the encoder of PlantNet are visualized in Fig. A3. In Fig. A3(a1-c1), the feature distance highly coincides with the 3D distance in the XYZ space at first, which means the feature space is still a low-level one at that time. In Fig. A3(a3) and (b3), it can be seen that the target point at the tip of the leaf (the big black point) becomes very different from the points on the stem in dotted circles. The red points in the dotted areas of Fig. A3(a1) and (b1) turn to bluish points in Fig. A3(a3) and (b3), respectively. At the last stage of the 1st LFEO, the network has also noticed that the target point at the bottom of the main stem is different from the points at the end of the leaf; therefore, some leaf points (areas in the dotted circles) in the middle of the sorghum change from red in Fig. A3(c1) to blue in Fig. A3(c3).

In summary, when the information goes deeper on the encoder of PlantNet, the proximity in the feature space changes in two ways: (i) the feature receptive field gradually converges to local areas, which reflects the local information abstraction functionality of LFEO; (ii) the feature space gradually transforms from a low-level 3D spatial structure to a high-level semantic relation, indicating the concentration of high-level features in LFEO by convolutions.

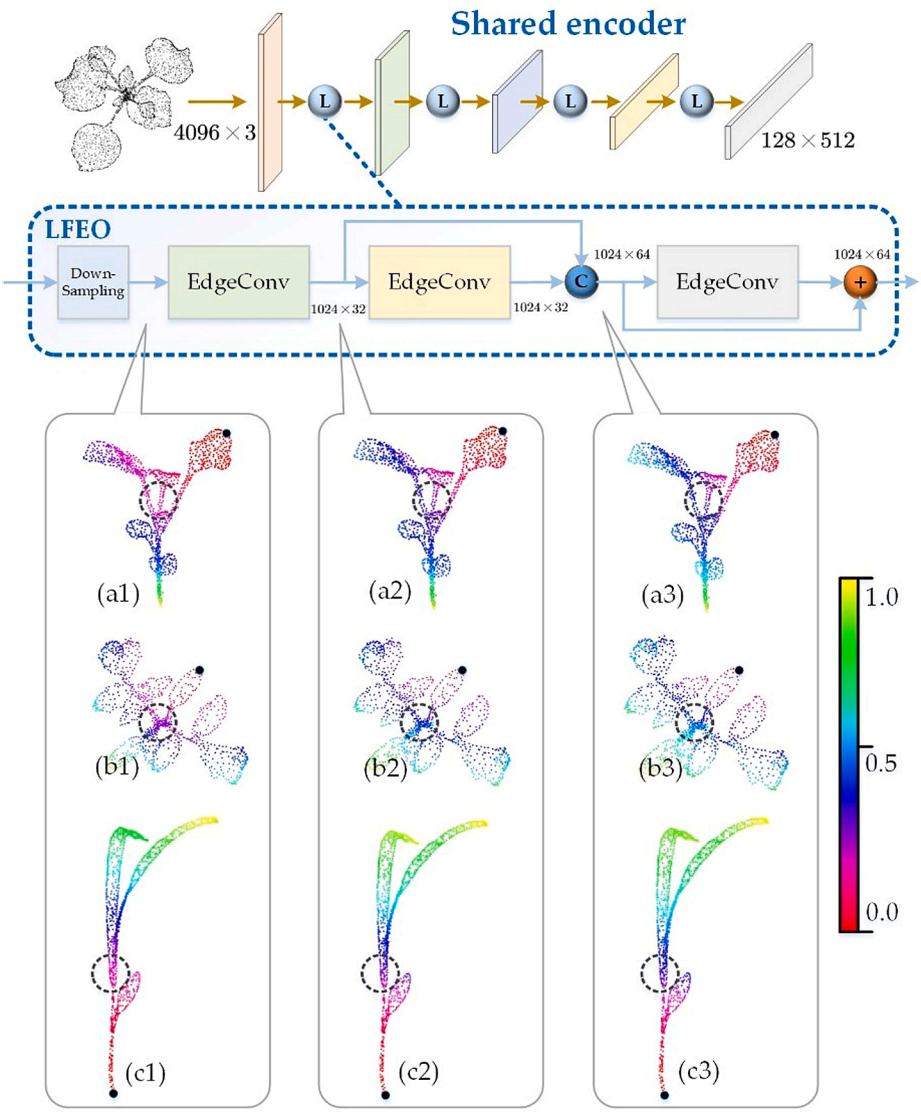


Fig. A3. A visualization of the normalized L2 distance between a target point with all other points in feature space at different stages of the first LFEO of the encoder for a plant sample. We rendered all points with a heat map, in which red indicates a small distance to the target point in the L2-measured feature space and yellow indicates a large distance. (a1), (a2), and (a3) are the visualizations of the same tobacco plant in different stages. (b1), (b2), and (b3) are the visualizations for a tomato plant. (c1), (c2), and (c3) are the visualizations of a sorghum plant. The black points in all sub-figures are the target points for the three plants. During the LFEO feature extraction, we can see the distances change gradually in areas labeled in dotted circles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A.5 Two imaging systems for the generalization analysis of PlantNet

The MVS technique can generate dense point clouds with low cost and high precision and has become a popular method for 3D reconstruction of plant models in recent years (Li et al., 2020). By collecting many multi-view images of a plant for feature point matching, an accurate and complete 3D colorful point cloud can be reconstructed even by an ordinary cell phone. In this study, image samples and the positions of the camera in MVS imaging for a plant are shown in Fig. A4(a) and (b), respectively. Besides, Kinect V2 is a commercial depth sensor based on structured-light technology. The hardware platform of a Kinect V2 system used for our experiment is shown in Fig. A4(c).

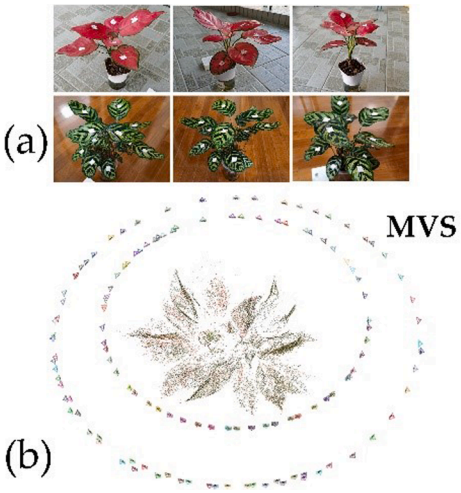


Fig. A4. Two other types of 3D imaging systems for the generation of plant point clouds that are used to test the generalization ability of PlantNet. The two imaging sensors scanned several new plant species to generate new point clouds that are not involved in the training set, and these point clouds were tested on pre-trained PlantNet to verify the generalization ability. (a) shows different input images (taken by a cell phone) from different positions of the Multi-view Stereo (MVS) imaging technique for two species; (b) shows the reconstructed sparse plant point cloud of a *Dieffenbachia picta* plant by using the MVS tool; (c) shows the hardware platform of a Kinect V2 imaging sensor.

References

Ben-Shabat, Y., Lindenbaum, M., Fischer, A., 2017. 3D point cloud classification and segmentation using 3d modified fisher vector representation for convolutional neural networks. arXiv preprint arXiv:1711.08241, 1-13.

Boulch, A., Guerry, J., Le Saux, B., Audebert, N., 2018. SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. Comput. Graph. 71, 189–198.

Brown, T.B., Cheng, R., Sirault, X.RR., Rungrat, T., Murray, K.D., Trtilek, M., Furbank, R. T., Badger, M., Pogson, B.J., Borevitz, J.O., 2014. TraitCapture: genomic and environment modelling of plant phenomic data. Curr. Opin. Plant Biol. 18, 73–79.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24 (5), 603–619.

Conn, A., Pedmale, U.V., Chory, J., Navlakha, S., 2017a. High-resolution laser scanning reveals plant architectures that reflect universal network design principles. Cell Syst. 5 (1), 53–62.e3.

Conn, A., Pedmale, U.V., Chory, J., Stevens, C.F., Navlakha, S., 2017b. A statistical description of plant shoot architecture. Curr. Biol. 27 (14), 2078–2088.e3.

Duan, T., Chapman, S.C., Holland, E., Rebetzke, G.J., Guo, Y., Zheng, B., 2016. Dynamic quantification of canopy structure to characterize early plant vigour in wheat genotypes. J. Exp. Bot. 67 (15), 4523–4534.

Engelmann, F., Kontogianni, T., Hermans, A., Leibe, B., 2017. Exploring spatial context for 3D semantic segmentation of point clouds. In: Proceedings of the IEEE international conference on computer vision workshops, pp. 716–724.

Guerry, J., Boulch, A., Le Saux, B., Moras, J., Plyer, A., Filliat, D., 2017. Snapnet-r: Consistent 3D multi-view semantic labeling for robotics. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 669–678.

Guo, Q., Jin, S., Li, M., Yang, Q., Xu, K., Ju, Y., Zhang, J., Xuan, J., Liu, J., Su, Y., Xu, Q., Liu, Y.u., 2020. Application of deep learning in ecological resource research: Theories, methods, and challenges. Sci. China Earth Sci. 63 (10), 1457–1474.

Hua, B.-S., Tran, M.-K., Yeung, S.-K., 2018. Pointwise convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 984–993.

Huang, J., You, S., 2016. Point cloud labeling using 3D convolutional neural network. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 2670–2675.

Huang, Q., Wang, W., Neumann, U., 2018. Recurrent slice networks for 3D segmentation of point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2626–2635.

Jin, S., Su, Y., Gao, S., Wu, F., Hu, T., Liu, J., Li, W., Wang, D., Chen, S., Jiang, Y., Pang, S., Guo, Q., 2018a. Deep learning: individual maize segmentation from terrestrial lidar data using faster R-CNN and regional growth algorithms. Front. Plant Sci. 9, 866–875.

Jin, S., Su, Y., Wu, F., Pang, S., Gao, S., Hu, T., Liu, J., Guo, Q., 2018b. Stem-leaf segmentation and phenotypic trait extraction of individual maize using terrestrial LiDAR data. IEEE Trans. Geosci. Remote Sens. 57 (3), 1336–1346.

Jin, S., Su, Y., Gao, S., Wu, F., Ma, Q., Xu, K., Ma, Q., Hu, T., Liu, J., Pang, S., Guan, H., Zhang, J., Guo, Q., 2019. Separating the structural components of maize for field phenotyping using terrestrial lidar data and deep convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 58 (4), 2644–2658.

Jin, S., Su, Y., Zhao, X., Hu, T., Guo, Q., 2020. A point-based fully convolutional neural network for airborne LiDAR ground point filtering in forested environments. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3958–3974.

Jin, S., Sun, X., Wu, F., Su, Y., Li, Y., Song, S., Xu, K., Ma, Q., Baret, F., Jiang, D., Ding, Y., Guo, Q., 2021. Lidar sheds new light on plant phenomics for plant breeding and management: Recent advances and future prospects. ISPRS J. Photogramm. Remote Sens. 171, 202–223.

Kalogerakis, E., Averkiou, M., Maji, S., Chaudhuri, S., 2017. 3D shape segmentation with projective convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3779–3788.

Koma, Z.s., Rutzinger, M., Bremer, M., 2018. Automated segmentation of leaves from deciduous trees in terrestrial laser scanning point clouds. IEEE Geosci Remote S 15 (9), 1456–1460.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4558-4567.

Li, D., Cao, Y., Shi, G., Cai, X., Chen, Y., Wang, S., Yan, S., 2019. An overlapping-free leaf segmentation method for plant point clouds. IEEE Access 7, 129054–129070.

Li, D., Shi, G., Kong, W., Wang, S., Chen, Y., 2020. A leaf segmentation and phenotypic feature extraction framework for multiview stereo plant point clouds. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 2321–2336.

Li, D., Shi, G., Wu, Y., Yang, Y., Zhao, M., 2021. Multi-scale neighborhood feature extraction and aggregation for point cloud segmentation. IEEE Trans. Circuits Syst. Video Technol. 31 (6), 2175–2191.

Li, S., Dai, L., Wang, H., Wang, Y., He, Z., Lin, S., 2017. Estimating leaf area density of individual trees using the point cloud segmentation of terrestrial LiDAR data and a voxel-based model. Remote Sensing 9, 1202–1217.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B., 2018. Pointcnn: Convolution on x-transformed points. Adv. Neural Information Processing Systems 31, 820–830.

Li, Y., Pirk, S., Su, H., Qi, C.R., Guibas, L.J., 2016. Fpnn: Field probing neural networks for 3D data. Adv. Neural Inform. Process. Syst. 29, 307–315.

Liu, S., Jia, J., Fidler, S., Urtasun, R., 2017. Sgn: Sequential grouping networks for instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3496–3504.

Livny, Y., Yan, F., Olson, M., Chen, B., Zhang, H., El-Sana, J., 2010. Automatic reconstruction of tree skeletal structures from point clouds, ACM SIGGRAPH Asia 2010 papers. ACM Trans. Graph. 29 (6), 1–8.

Marulanda, F.G., Libin, P., Verstraeten, T., Nowé, A., 2018. IPC-Net: 3D Point-Cloud Segmentation Using Deep Inter-Point Convolutional Layers. In: 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 293–301.

Maturana, D., Scherer, S., 2015. Voxnet: A 3D convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 922–928.

McCormac, J., Handa, A., Davison, A., Leutenegger, S., 2017. Semanticfusion: Dense 3D semantic mapping with convolutional neural networks. In: 2017 IEEE International Conference on Robotics and automation (ICRA). IEEE, pp. 4628–4635.

Nguyen, T., Slaughter, D., Max, N., Maloof, J., Sinha, N., 2015. Structured light-based 3D reconstruction system for plants. Sensors 15 (8), 18587–18612.

Pham, Q.-H., Hua, B.-S., Nguyen, T., Yeung, S.-K., 2019a. Real-time progressive 3D semantic segmentation for indoor scenes. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1089–1098.

Pham, Q.-H., Nguyen, T., Hua, B.-S., Roig, G., Yeung, S.-K., 2019b. Jsis3D: Joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8827–8836.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. 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.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413, 1-14.

Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R., 2017c. 3D graph neural networks for rgbd semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5199–5208.

Ren, M., Zemel, R.S., 2017. End-to-end instance segmentation with recurrent attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6656–6664.
Rusu, R.B., Cousins, S., 2011. 3D is here: Point cloud library (pcl), 2011 IEEE international conference on robotics and automation. IEEE 1–4.
Schulze-Brüninghoff, D., Hensgen, F., Wachendorf, M., Astor, T., 2019. Methods for LiDAR-based estimation of extensive grassland biomass. Comput. Electron. Agric. 156, 693–699.
Shen, Y., Feng, C., Yang, Y., Tian, D., 2018. Mining point cloud local structures by kernel correlation and graph pooling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4548–4557.
Shi, W., van de Zedde, R., Jiang, H., Kootstra, G., 2019. Plant-part segmentation using deep learning and multi-view vision. Biosyst. Eng. 187, 81–95.
Simonovsky, M., Komodakis, N., 2017. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3693–3702.
Su, W., Zhang, M., Liu, J., Sun, Z., 2018. Automated extraction of corn leaf points from unorganized terrestrial LiDAR point clouds. Int. J. Agric. Biol. Eng. 11 (3), 166–170.
Su, Y., Wu, F., Ao, Z., Jin, S., Qin, F., Liu, B., Pang, S., Liu, L., Guo, Q., 2019. Evaluating maize phenotype dynamics under drought stress using terrestrial lidar. Plant methods 15, 11–26.
Sun, D., Xu, Y., Cen, H., 2021. Optical sensors: deciphering plant phenomics in breeding factories. Trends Plant Sci. https://doi.org/10.1016/j.tplants.2021.06.012.
Sun, S., Li, C., Paterson, A.H., 2017. In-field high-throughput phenotyping of cotton plant height using LiDAR. Remote Sensing 9, 377–397.
Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.-Y., 2018. Tangent convolutions for dense prediction in 3d. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3887–3896.
Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.
Vázquez-Arellano, M., Reiser, D., Paraforos, D.S., Garrido-Izard, M., Burce, M.E.C., Griepentrog, H.W., 2018. 3-D reconstruction of maize plants using a time-of-flight camera. Comput. Electron. Agric. 145, 235–247.
Wang, D.Z., Posner, I., 2015. Voting for voting in online point cloud object detection, Robotics: Science and Systems. Rome, Italy, pp. 10-15.
Wang, S., Suo, S., Ma, W.-C., Pokrovsky, A., Urtasun, R., 2018a. Deep parametric continuous convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2589–2597.
Wang, W., Yu, R., Huang, Q., Neumann, U., 2018b. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2569–2578.
Wang, X., Liu, S., Shen, X., Shen, C., Jia, J., 2019a. Associatively segmenting instances and semantics in point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4096–4105.
Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019b. Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (tog) 38 (5), 1–12.
Wolf, D., Prankl, J., Vincze, M., 2015. Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters. 2015 IEEE International conference on robotics and automation (ICRA) IEEE 4867–4873.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.
Wu, W., Qi, Z., Fuxin, L., 2019. Pointconv: Deep convolutional networks on 3D point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9621–9630.
Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3D shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1912–1920.
Xu, Y., Fan, T., Xu, M., Zeng, L., Qiao, Y., 2018. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 87–102.
Yang, L., Zhang, L., Dong, H., Alelaiwi, A., Saddik, A.E., 2015. Evaluating and improving the depth accuracy of Kinect for Windows v2. IEEE Sens. J. 15 (8), 4275–4285.
Yang, S., Huang, Y., Scherer, S., 2017. Semantic 3D occupancy mapping through efficient high order CRFs. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 590–597.
Ye, X., Li, J., Huang, H., Du, L., Zhang, X., 2018. 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 403–417.
Yuan, H., Bennett, R.S., Wang, N., Chamberlin, K.D., 2019. Development of a peanut canopy measurement system using a ground-based LiDAR sensor. Front. Plant Sci. 10, 203–214.
Zermas, D., Morellas, V., Mulla, D., Papanikolopoulos, N., 2017. Estimating the leaf area index of crops through the evaluation of 3D models. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 6155–6162.
Zhao, L., Tao, W., 2020. Jsnet: Joint instance and semantic segmentation of 3D point clouds, Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12951-12958.
Zhuo, W., Salzmann, M., He, X., Liu, M., 2017. Indoor scene parsing with instance segmentation, semantic labeling and support relationship inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5429–5437.