Scientia Horticulturae 331 (2024) 113158

Contents lists available at ScienceDirect

Scientia Horticulturae

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

Research Paper

PSTNet: Transformer for aggregating neighborhood features in 3D point cloud semantic segmentation of eggplant plants

Linqian Ma [a], Lingyuan Kong [a], Xingshuo Peng [a], Keyuan Wang [a], Nan Geng [a,b,c,*]

[a] College of Information Engineering, Northwest Agriculture and Forestry University, Shaanxi Yangling 71210, PR China
[b] Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Northwest A&F University, Yangling 712100, PR China
[c] Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service, Northwest A&F University, Yangling 712100, PR China

ARTICLE INFO

Keywords:
Sementic segmentation
Plant phenotyping
Point cloud
3D scanning

ABSTRACT

Improving the quality of plant point cloud data and achieving precise segmentation is essential for effective plant phenotyping analysis and plant breeding. To ensure data quality and accuracy of segmentation, we employed a self-developed three-dimensional scanning device based on binocular vision to acquire eggplant plant point clouds with high spatial resolution. And we improved down-sampling algorithm Octree Farthest Point Sampling (OFPS) to process the original point cloud data. Furthermore, we proposed a neural network called Plant Segmentation Transformer Network (PSTNet) to achieve semantic segmentation of high spatial resolution eggplant plant datasets. PSTNet comprises the following components: (i) a Neighborhood Feature Aggregator (NPA) for storing and aggregating local neighborhood features of input points, and (ii) a cascaded Point Self-Attention module (PSA) for capturing contextual information. Experimental results demonstrate the remarkable performance of PSTNet in semantic segmentation tasks, with IoU, Precision, Recall, F1-score, and Accuracy values of 92.20 %, 95.30 %, 95.57 %, 95.43 %, and 95.15 %, respectively. Additionally, compared to the Point Cloud Transformer (PCT), which exhibited the second-best performance, PSTNet achieved improvements of 4.37, 1.47, 2.87, 2.17, and 4.22 percentage points in the abovementioned metrics. This method achieves high-precision segmentation of plant point clouds, expands the semantic segmentation method of plants, and lays a solid foundation for plant phenotypic analysis.

1. Introduction

Plant phenotypes refer to observable characteristics or traits that are influenced by both the genetic makeup of the plant and the surrounding environment. Analyzing plant phenotypic data allows for the assessment of the impact of biotic factors, including weeds, microorganisms, and related species, as well as abiotic factors, such as moisture, salinity, and other environmental conditions, on plant traits (Kolhar and Jagtap, 2023). Consequently, plant phenotypic data serves as a crucial source of information for applications in plant breeding and precision agriculture (Tran et al., 2017). Eggplant is one of the most important vegetable crops in tropical and subtropical regions. The main challenge in cultivating eggplants with high yield and stress tolerance lies in the accuracy of plant phenotypic data. Hence, the precise acquisition of high-quality phenotypic data for eggplants is of utmost importance for successful eggplant breeding endeavours (Gosa et al., 2019; Martínez-Ispizua et al., 2021; Zhang and Zhang, 2018).

Nevertheless, the manual measurement of plant phenotypic information has drawbacks such as low efficiency, high labor intensity, and limited accuracy. These limitations harm the outcomes of plant breeding efforts and hinder the comprehensive understanding of plant growth processes (Jin et al., 2022). In recent years, imaging technology and computer processing techniques have emerged as practical tools for acquiring plant phenotypic data. Various technologies such as depth cameras, hyperspectral imaging, lidar, and thermal imaging have been employed for non-destructive, high-precision, and high-throughput acquisition of plant phenotypic information. Leveraging computer technology, researchers can efficiently and quickly measure plant phenotypic data through these methods, significantly enhancing the capabilities of data collection and phenotypic analysis (Cen et al., 2020; Zhang et al., 2023).

The data acquired from measurement equipment can be classified into two-dimensional (image) and three-dimensional (point cloud) data. The prevailing approach entails segmenting plant organs based on two-

* Corresponding author.
E-mail addresses: 2022056083@nwafu.edu.cn (L. Ma), nangeng@nwafu.edu.cn (N. Geng).

https://doi.org/10.1016/j.scienta.2024.113158
Received 24 November 2023; Received in revised form 7 March 2024; Accepted 26 March 2024
Available online 8 April 2024

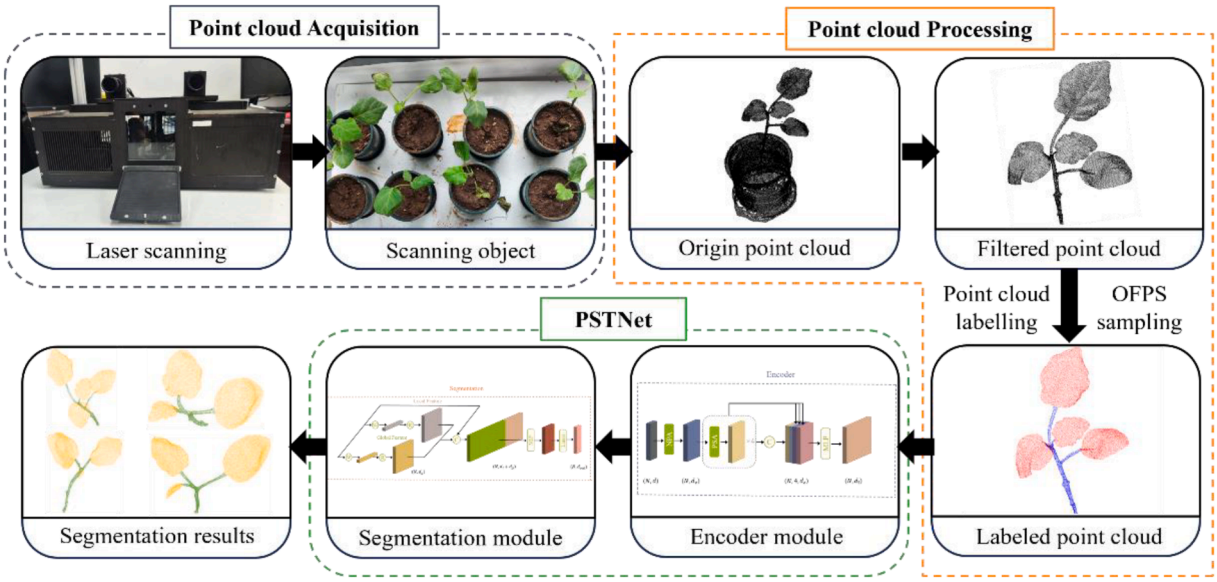

Fig. 1. The framework for semantic segmentation of plants.

dimensional images. However, challenges such as leaf overlap, similarities between plants and the background, illumination variations, and canopy shadows impose limitations on this method. On the other hand, three-dimensional sensing technologies, such as three-dimensional scanners (Rist et al., 2018), lidar (Ao et al., 2022), and structured light (Zhang, 2018), enable the collection of more precise three-dimensional plant data. These technologies alleviate issues related to leaf occlusion and enhance the accuracy of extracting plant phenotypic parameters.

Moreover, accurately segmenting plant organs from high-precision point cloud datasets poses a significant challenge in plant phenotype research. Traditional plant organ segmentation methods typically rely on plants' geometric characteristics. These approaches utilize normal vectors and curvature (Vo et al., 2015), incorporate shape prior knowledge (Wang et al., 2015), or employ octrees to represent point cloud data (Bassier et al., 2017). However, these methods often depend on expert knowledge, exhibit limited efficiency and accuracy, and are not well-suited for large-scale point cloud segmentation tasks (Jin et al., 2019). They tend to perform poorly when dealing with different plant species or leaves of varying shapes, lacking generalization capabilities. Consequently, the segmentation of individual plant organs, such as stems and leaves, faces significant challenges and continues to be an active area of research.

In recent years, with the success of neural networks in two-dimensional image processing, more and more people are exploring how to apply deep learning to three-dimensional point cloud processing and analysis. Initially, the research mainly focused on voxel segmentation, where the point cloud is partitioned into numerous voxels, and convolution operations are applied to learn features from these voxels (Huang and You, 2016). While voxelization transforms point clouds into structured data, it is essential to note that this process introduces specific issues, including the loss of information. PointNet (Qi et al., 2017a) revolutionized the field by pioneering feature learning directly from raw point clouds. Building upon this pioneering work, researchers have focused on incorporating local features into the model in subsequent work. PointNet++ (Qi et al., 2017b) addresses the limitation of PointNet by capturing the neighborhood point characteristics through ball queries, although it overlooks the inter-point relationships. DGCNN (Wang et al., 2019) introduces a dynamically evolving graph convolutional network by constructing local graphs based on neighborhood proximity and assigning features to neighboring points using the local graph structure. The advent of attention mechanisms has sparked interest among researchers in incorporating them into point cloud segmentation models. Cui et al. explored the fusion of geometric attention with DGCNN, enabling learning of more comprehensive intrinsic features (Cui et al., 2021). Similarly, Guo et al. introduced the Point Transformer model, which leverages the Transformer architecture for point cloud segmentation, showcasing the efficacy of attention-based mechanisms in this domain (Zhao et al., 2021).

The advancements made in point cloud segmentation have also significantly impacted the segmentation of plant point clouds. Ao et al. (2022) employed PointNet to segment corn organs and developed the DeepSeg3DMaize software for corn organ segmentation and labelling. Masuda (2021) applied PointNet++ for tomato plant point cloud segmentation and estimation of leaf area. Jin et al. (2020) developed the VCNN model for corn semantic segmentation, utilizing voxel-based convolution to learn point cloud features. Shi et al. (2019) constructed a three-dimensional point cloud from multi-view two-dimensional images. They mapped the plant images into the three-dimensional space to achieve organ segmentation of the point cloud. These studies highlight the rapid development and diverse plant point cloud segmentation research approaches.

One of the challenges in plant point cloud segmentation research is the scarcity of open-source datasets available for evaluating the performance of segmentation models. Acquiring plant point cloud data is a challenging task, and the complex structure of plants necessitates extensive manual annotation efforts. Thus, the lack of accessible and annotated datasets hinders the development and evaluation of plant segmentation models. To address this challenge, Dutagaci et al. (2020) obtained 11 annotated real rose point clouds through X-ray scanning, providing a small but valuable dataset for model evaluation. However, this dataset still has limitations in terms of its size and diversity. Conn et al. (2017) collected point clouds from three different plant species, including tomato, tobacco, and sorghum, grown in five distinct environments, resulting in 546 plant point clouds. Another notable dataset is the Pheno4D dataset created by Schunck et al. (2021), which consists of 4D point clouds of corn and tomato plants specifically designed for the evaluation of plant phenotyping analysis. The availability of high-quality point cloud data is crucial for accurate model training, making the creation of such datasets an essential task in plant point cloud segmentation research (Zhang et al., 2019).

In order to address the challenges related to obtaining high-quality point clouds and achieving accurate segmentation of plant organs in the field of plant point cloud segmentation, we propose a novel plant

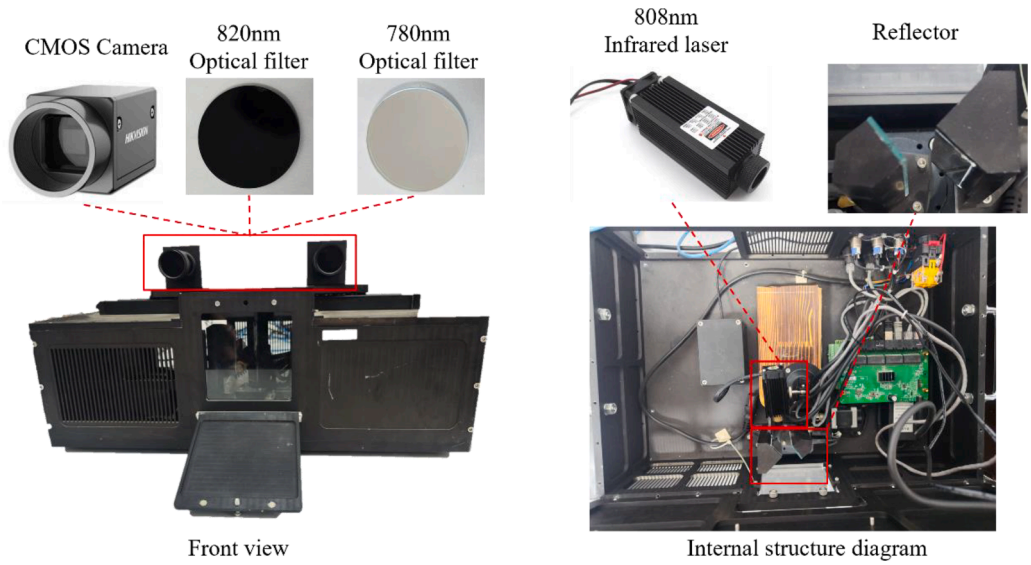

Fig. 2. Structure diagram of 3D scanning equipment.

point cloud semantic segmentation model called PSTNet. The specific contributions of our work are as follows:

(i). Dataset Creation: We present a fully annotated dataset comprising 50 samples of eggplant plants. To acquire high-precision point cloud data of eggplant plants, we developed a three-dimensional scanning device based on binocular vision. This device enables us to capture detailed and accurate point cloud representations of the plants.

(ii). Improved Point Cloud Downsampling: We employ an improved point cloud downsampling method called OctreeFPS. This method effectively reduces the number of points in the original point cloud data while preserving the overall shape and structure of the plant. By downsampling the point cloud to a desired number of points, denoted as N, we create a high-quality dataset tailored explicitly for plant organ segmentation tasks.

(iii). Plant Segmentation Transformer Network: We propose the Plant Segmentation Transformer Network (PSTNet), an end-to-end model for semantic segmentation of eggplant plant point clouds. Specifically, we propose the Neighbor Points Aggregator (NPA) module, designed to capture and aggregate local neighborhood features of input points. Furthermore, we introduce the cascaded Point Self-Attention (PSA) module to capture contextual information in our plant point cloud organ segmentation model.

2. Materials and methods

2.1. Data acquisition

The plant cultivation for our study was conducted in the experimental field located at the coordinates (108.06 N, 34.26E) of the North Campus of Northwest A&F University in Xianyang City, Shaanxi Province, China, during the period from July to August 2023. Hangqie No. 1 seedling stage plants were utilized as the experimental materials. The overall process of plant stem and leaf segmentation is shown in Fig. 1. In order to ensure the acquisition of high-precision three-dimensional point cloud data for eggplant, a dedicated three-dimensional point cloud scanning platform was developed, integrating line laser technology, binocular vision, and a rotating platform.

Fig. 2 illustrates the components of the 3D scanning equipment utilized in this study. The structure of the 3D scanning equipment used in this research is shown in Fig. 2, which mainly consists of two CMOS cameras, an 808 nm infrared laser emitter, two 820 nm filters, two 780 nm filters and two reflectors. To ensure the acquisition of high-quality

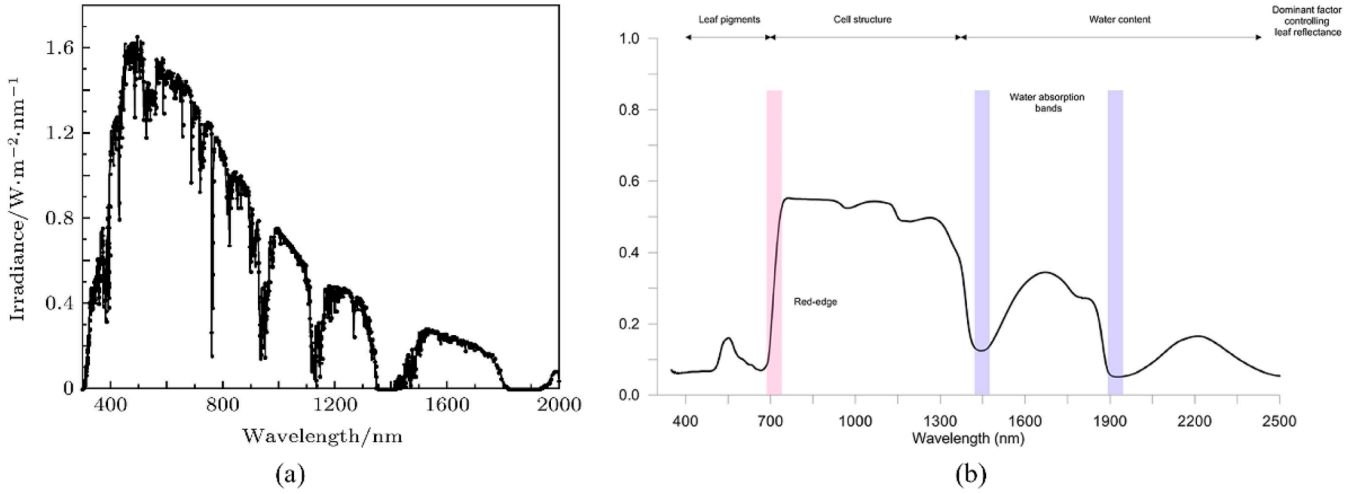

Fig. 3. (a) Intensity of sunlight at various wavelengths; (b) Reflectance of plants at various wavelengths .

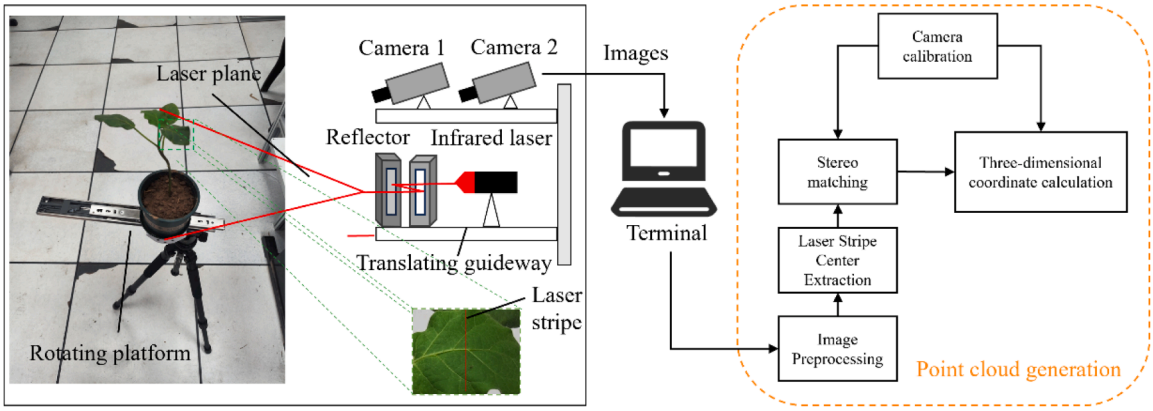

Fig.4. Schematic of point cloud generation.

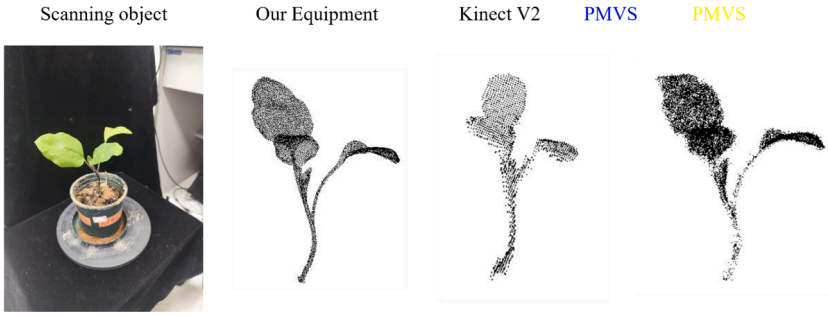

Fig.5. Comparison of different point cloud acquisition methods: (a) our device; (b) Kinect V2; (c) PMVS.

3D point clouds in a natural environment, minimizing the interference caused by sunlight on the laser stripes and camera imaging was essential. The intensity of sunlight across different wavelengths is depicted in Fig. 3(a), indicating that the range of 760nm-890 nm corresponds to low light intensity (Xue-Tong et al., 2020). Furthermore, considering the characteristics of plants, light waves within the band exhibiting the highest reflectance on plant surfaces were selected. Fig. 3(b) displays the reflectivity of the plant surface at various wavelengths, with the highest reflectivity observed within the range of 720nm-900 nm (Galieni et al., 2021). Considering the characteristics above and practical considerations such as cost and other factors, we opted for an 808 nm infrared laser as the primary laser source in the device. We also added 820 nm low-pass filter and 780 nm high-pass filter to the CMOS camera, which filter out light waves with wavelengths above 820 nm and below 780 nm to reduce their influence on the 808 nm laser streak imaging.

The generation process of the 3D point cloud is shown in Fig. 4. The device is controlled through a computer terminal. The operating procedure consists of several steps. Initially, an infrared laser generator

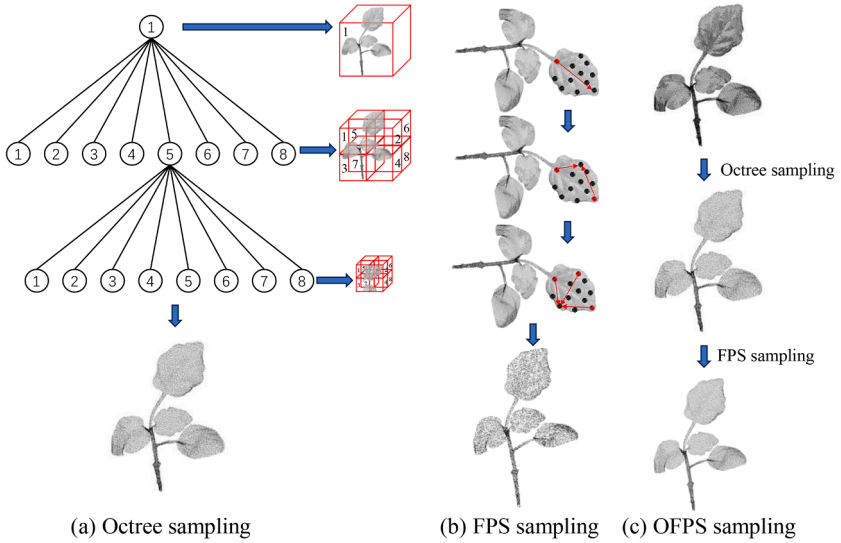

Fig. 6. Illustration of Octree Farthest Point Sampling (OFPS). (a) The point cloud data is represented using an octree structure until the number of leaves exceeds a predefined threshold; (b) FPS selects the farthest point at each iteration until the desired number of points is reached, as specified by the threshold; (c) the OFPS process.

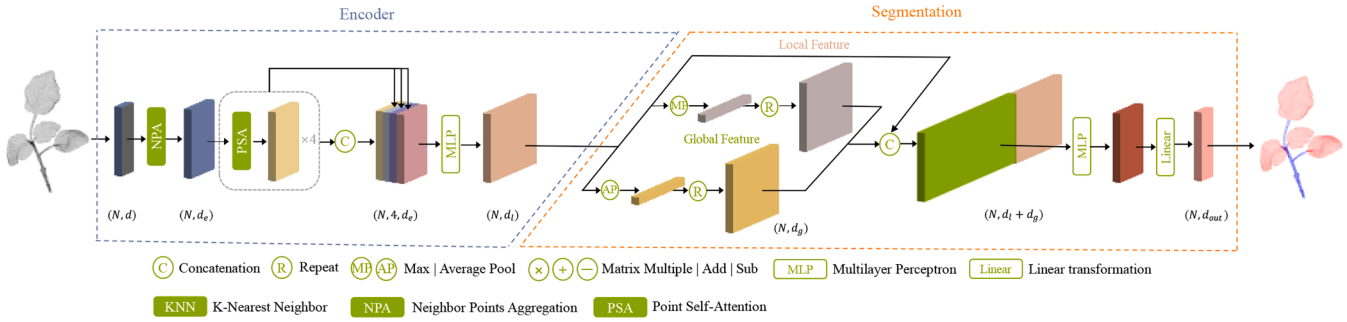

Fig. 7. PSTNet architecture. The network is composed of two parts. The encoder mainly comprises a Neighbor Points Aggregation module and four stacked Point Self-Attention modules. The segmentation module can be regarded as a decoder comprising multiple Linear layers. The number beneath each feature represents the output dimension of the corresponding module.

emits a linear laser light through two reflectors onto the object's surface to be measured. This is subsequently captured by the image sensor of a CMOS camera, which produces two images containing the laser streaks, and the centre-of-gravity method is used to extract the pixel coordinates of the laser streaks. Subsequently, the correspondence between the laser stripes in the two images is established using image stereo matching, resulting in a set of matching point pairs. By employing the concept of similar triangles, the coordinates of an actual 3D point can be computed using any pair of matching point pairs. By rotating the mirror, the laser can transversally scan the entire surface. As the laser light bar moves, multiple laser stripes can be extracted to calculate multiple actual 3D points. This process yields a point cloud representing the eggplant plants from a single viewpoint. As shown in Fig. 5, which demonstrates the plant point cloud data acquired by our device, Kinect V2(He et al., 2018) and PMVS(Furukawa and Ponce, 2010), respectively, it can be seen that for plant objects, our device is capable of acquiring high-density, high-precision and highly robust point cloud data on the plant surface. Meanwhile, binocular vision technology eliminates the need for laser plane calibration during each acquisition, simplifying device operation. To obtain complete eggplant point cloud data, we scanned the eggplant from multiple angles using a slide and rotating platform to capture different viewpoints.

To obtain complete eggplant point cloud data, we scanned the eggplant from multiple angles using a slide and rotating platform to capture different viewpoints. This process generates a set of point cloud data that represents the eggplant from various viewpoints. To achieve a unified representation by integrating the multiple views of the point cloud data, we employ the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992).

2.2. Point cloud processing

To ensure the quality and reliability of our point cloud data, we utilize two filtering techniques: straight-through filtering and CSF ground filtering (Zhang et al., 2016). These filtering methods are employed to eliminate outliers and ground point clouds, thereby improving the accuracy of the subsequent segmentation process. The annotation work for point cloud segmentation is performed using (Girardeau-Montaut, 2016). The amount of point cloud data plays a crucial role in the accuracy of segmentation results. By utilizing a fixed number of points, the model can better adapt to different point cloud densities and achieve enhanced generalization performance (Boogaard et al., 2022). Hence, this study introduces the Octree furthest sampling strategy, which combines Octree sampling with Farthest Point Sampling (FPS) (Qi et al., 2017b) to optimize the point cloud data. The process and results of Octree sampling and FPS sampling are shown in Fig. 6.

By employing Octree sampling, our method effectively reduces the number of points in the point cloud while preserving its overall structure and geometry. As depicted in Fig. 6, it is evident that FPS sampling has resulted in a loss of some shape characteristics of the plant. In contrast, OFPS sampling has successfully retained the distinctive characteristics of the eggplant plant. Furthermore, it ensures consistent sample point counts, contributing to the reliability of the segmentation results.

Accurate data labelling plays a crucial role in model training. In this study, we employed the point cloud processing plug-in provided by CloudCompare(Girardeau-Montaut, 2016) to perform point cloud labelling. Each 3D point was assigned a semantic label through manual visual inspection, categorizing them into different categories. Given that the primary objective of our semantic segmentation was to facilitate the subsequent extraction of phenotypic parameters of leaves (such as leaf area, leaf length, and leaf width), we classified the petiole and shoot apex as plant stems, assigning them a semantic label of 0. Leaves were assigned a semantic label of 1. This categorization allowed us to segment stems and leaves accurately, enabling the extraction of relevant phenotypic parameters.

Consequently, we generated the eggplant point cloud dataset, complete with the corresponding labelling information. The format of the eggplant point cloud dataset resembles that of other publicly available point cloud datasets like ShapeNet(Chang et al., 2015). The point cloud file contains each point's XYZ coordinates and their respective labelling information.

2.3. Network architecture

We designed a general point cloud semantic segmentation algorithm, PSTNet, to segment plant stems and leaves precisely. The overall architecture of PSTNet is depicted in Fig. 7. The architecture primarily comprises two main components: the encoder and the Segmentation module. The encoder is responsible for converting the input points, represented as three-dimensional coordinates, into a high-dimensional feature space. The input to the encoder module is a point cloud set $P \in \mathbb{R}^{N\times 3}$, where N represents the number of points in the set. Each point in the set is represented by 3-dimensional features (XYZ coordinates). In the first step, the point cloud set P is fed into the Neighbor Points Aggregator (NPA), which effectively captures and integrates information from neighboring points. This aggregation process generates a high-dimensional feature representation $F_e \in \mathbb{R}^{N\times d_e}$. Subsequently, the feature representation F_e is passed through four cascaded Point Self-Attention (PSA) modules. The outputs of the PSA modules are then concatenated and fed into a linear transformation layer. The linear transformation layer processes the concatenated outputs and produces the local features of the point cloud, denoted as $F_l \in \mathbb{R}^{N\times d_e}$. We adopt a method combining maximum pooling and average pooling to extract more effective global features from the point cloud. By connecting the outputs of these pooling operations, denoted as $F_M \in \mathbb{R}^{d_g}$ and $F_A \in \mathbb{R}^{d_g}$, respectively, we obtain a comprehensive global feature representation$F_g \in \mathbb{R}^{N\times d_g}$.

The Segmentation module of our approach aims to partition the

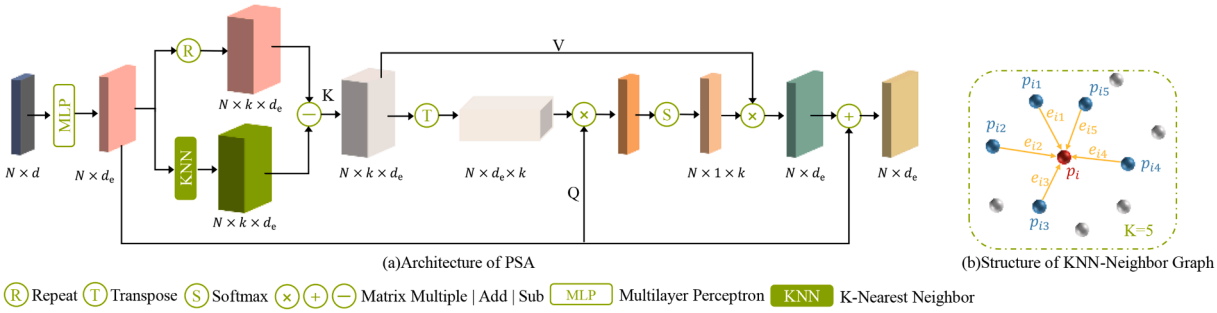

Fig. 8. (a) The Architecture of Neighbor Points Aggregation. (b) An example of a neighbor graph constructed using the k-nearest neighbors (KNN) algorithm.

point cloud into n parts, such as stems and leaves, and assign a global label to each point. We combine the global features F_g, obtained after connecting the local features F_l obtained from the aggregation step, and provide them as input to a Multilayer Perceptron (MLP) network. The MLP network consists of two cascaded Feed-forward Neural Networks (FNN). Each FNN consists of linear layers, batch normalization, rectified linear unit (ReLU) activation, and dropout. Finally, the probability score for each point is computed through a linear transformation, resulting in a matrix $S \in \mathbb{R}^{N\times n}$. The index of the matrix S determines the global label of a point.

2.3.1. Neighbor points aggregator

Point Self-Attention (PSA) is a module specially used to extract global features. However, for the specific task of plant point cloud segmentation, capturing the relationships and information among local neighboring points is equally crucial. We draw on the idea of attention mechanism (Vaswani et al., 2017), the dynamic directed graph structure of DGCNN (Wang et al., 2019), and the graph attention of GAT (Veličković et al., 2018) to design a Neighbor Points Aggregator (NPA) is a method for aggregating point features in a neighborhood directed graph. In Fig. 8(a), for the point cloud set $P \in \mathbb{R}^{N\times\ 3}$, the input feature $F \in \mathbb{R}^{N\times d}$ undergoes initial processing through a Multilayer Perceptron (MLP) to obtain a high-dimensional feature representation $F_e \in \mathbb{R}^{N\times d_e}$. The MLP architecture consists of two cascaded Feed-forward Neural Networks (FNN), each comprising linear layers, batch normalization, and rectified linear unit (ReLU) activation functions.

$$F_{p_i\in P}(p_i) = ReLU(BN(Linear(p_i))) \tag{1}$$

We construct a neighborhood graph $G_i \in \mathbb{R}^{K\times d_e}$ for each point $p_i \in P$ using the K-Nearest Neighbor (KNN) algorithm. The structure of the graph is depicted in Fig. 8(b) . For a given centre point p_i and its neighboring points $P_i = \{p_i, p_{i1}, p_{i2}, \ldots, p_{ij}\}, P_i \subseteq P$, we establish a neighborhood graph $G_i = (V, E)$. Here, V represents the set of K points closest to p_i obtained through KNN, and $E \in V \times V$ denotes the directed edge set $\{(p_i, p_{j1}), (p_i, p_{j2}), \ldots, (p_i, p_{jk})\}$. To capture the relationship and importance between the centre point p_i and its neighbor points, we define the edge attention coefficient $e_{ij} = h^l_{ij}(p_i, p_{ij})$. This coefficient reflects the significance or importance of the neighbor point p_{ij} with respect to p_i. By leveraging graph attention, we compute the local features for each point in the neighborhood graph,

$$F_{p_j\in KNN(p_i,K)}(p_{ij}) = concat\left(F\left(p_j - p_i\right)\right) \tag{2}$$

$$(Q, K, V) = F(p_{ij})\cdot(W_q, W_k, W_v) \tag{3}$$

$$h^l(p_i, p_{ij}) = Q^T \cdot K \tag{4}$$

Where function $F(\cdot)$ transforms the input point set into higher-level features. In the neighborhood graph, p_i represents the centre point, and p_j represents its K nearest neighbor points. We adopt the centre point feature to serve as the query input, and the feature difference $F(p_j - p_i)$ between the neighbor point and the centre point is used as both the key input and value input. Notably, our approach capitalizes on the use of feature differences as inputs, which has translation invariance properties, can satisfy the translation invariance properties of point clouds.

To make the attention coefficients e_{ij} easier to compare, we apply a normalization operation:

$$A^l_i = softmax\left(\frac{h^l(p_i, p_{ij})}{\sqrt{d_e}}\right) = \frac{\exp\left(h^l(p_i, p_{ij})\right)}{\sum_k \exp\left(h^l(p_i, p_{ij})\right)} \tag{5}$$

The neighborhood aggregation feature F_{na} of point p_i is defined as:

$$F_{na}(p_i) = A^l_i \cdot V \tag{6}$$

Notably, the independence of the Q, K, V, and softmax operations allows the attention coefficients to remain unaffected by the point order. This property makes the attention mechanism particularly well-suited for handling the inherent disorder in point clouds. Finally, the neighborhood aggregation feature F_{na} for each point p_i is connected and added to the input feature F_e, resulting in the generation of the local feature F_l.

$$F_l = NPA(F_e) = F_{na} + F_e \tag{7}$$

2.3.2. Point self-attention

After the Neighbor Points Aggregator (NPA) module, each point already incorporates the characteristics of its surrounding neighbor points. However, it may still lack crucial contextual information to

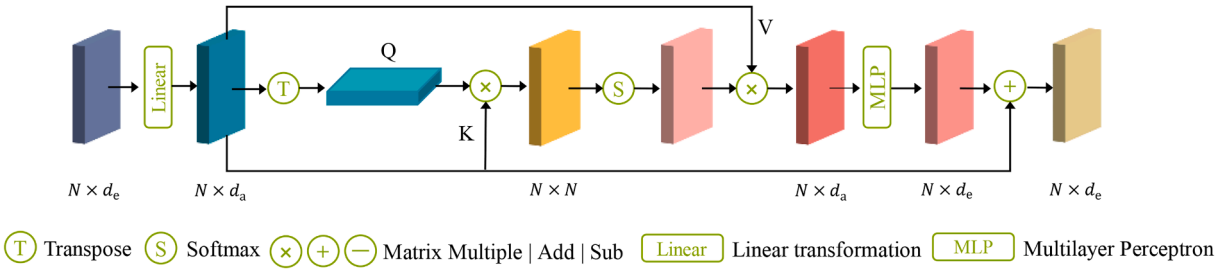

Fig. 9. Architecture of Point Self-Attention.

Table 1
Calculation Methods for Five Quantitative Indicators.

Metrics	Equations
Accuracy	$Acc = \frac{1}{N}\sum_{i=1}^{N} f_i,\ f_i\{\begin{smallmatrix}1 & y_i = x_i\\ 0 & y_i \neq x_i\end{smallmatrix}$
Recall	$Rec = \frac{TP}{TP+FN}$
Precision	$Prec = \frac{TP}{TP+FP}$
*F*1 Score	$F1 = \frac{Pre \cdot Rec}{Pre + Rec}$
IoU	$IoU = \frac{TP}{TP+FP+FN}$

capture global features. Therefore, we draw inspiration from the self-attention mechanism in Transformer(Vaswani et al., 2017). We adapt self-attention to the three-dimensional point cloud, allowing us to calculate the attention between points to capture contextual information. In Point Self-Attention(PSA), we draw an analogy between the points in the point cloud and the words in a sentence. In order to perform a de-dimensionalization operation on the local features F_e of point $p_i \subseteq P$, we utilize a Linear layer to transform it into a feature vector $F_a \in \mathbb{R}^{N\times d_a}$ (Fig. 9).

$$F_a = Linear(F_e) \tag{8}$$

The Query(Q), Key(K) and Value(V) are then generated by matrix multiplication. These matrixs are used to calculated the correlation $h_{ij}^g(\cdot)$ between each pair of points, which determines the weight between the two points. The specific definitions of $Q,\ K,\ V$, and $h_{ij}^g(\cdot)$ are as follows:

$$(Q, K, V) = F_a \cdot (W_q, W_k, W_v) \tag{9}$$

$$h_{ij}^g = Q^T \cdot K \tag{10}$$

Where $F_a \in \mathbb{R}^{N\times d_e}$ is the input feature. $Q, K \in \mathbb{R}^{N\times d_a}$ and $V \in \mathbb{R}^{N\times d_e}$ are query matrix, key matrix and value matrix, respectively. W_q, W_k, and W_v are shared learnable weight matrices. To improve efficiency, we reduce the dimensionality of the feature vectors by setting $d_a = d_e/4$. We perform a dot product between the Query and Key matrix to calculate the correlation weight.

$$h_{ij}^g = Q^T \cdot K \tag{11}$$

To make the attention-effective comparison, we apply a normalization step:

$$A_i^g = softmax\left(\frac{h_{ij}^g}{\sqrt{d_a}}\right) = \frac{\exp\left(h_{ij}^g\right)}{\sum_N \exp\left(h_{ij}^g\right)} \tag{12}$$

The self-attention feature F_{pa} of point p_i can be calculated from the normalized correlation weight A_i^g and the value matrix K:

$$F_{pa} = A_i^g \cdot V \tag{13}$$

Finally, we pass the self-attention feature F_{pa} through a Multilayer Perceptron (MLP) for further processing.

$$F_g = PSA(F_a) = MLP\left(F_{pa}\right) + F_a \tag{14}$$

The MLP consists of linear layers, batch normalization and rectified linear unit (ReLU).

2.4. Loss function

The loss function is a critical component in model training and result prediction. In our approach, we utilize the weighted cross-entropy function (L) as the loss function for PSTNet because it balances the imbalance in the number of points for each category of stems and leaves in the dataset. The class with the least number of classes is assigned a weight w_i, which is determined inversely by the class frequency. The weights allow the model to prioritize the least number of classes. The weighted cross-entropy function L is then calculated as follows:

$$L = -\sum_{i=1}^{N} w_i y_i \log(x_i) + (1 - y_i)\log(1 - x_i) \tag{15}$$

2.5. Evaluation metrics

To assess the performance of the PSTNet model in point cloud segmentation, we conducted a comparison between the predicted labels generated by the model and the ground truth labels. We employed five quantitative indicators, namely *Accuracy*, *Recall*, *Precision*, *F*1 *Score*, and *IoU*, to evaluate the quality of the semantic segmentation. Table 1 presents the calculation methods for each of the indicators used to evaluate the performance of the model. *Accuracy* is calculated as the proportion of correctly predicted points out of the total number of points. *Precision* is calculated as the proportion of correctly predicted points to the total number of points predicted. *Recall* is calculated as the proportion of correctly predicted points to the total number of ground truth points. *F*1 *Score* is a metric that combines Precision and Recall by calculating their harmonic mean. *IoU* quantifies the degree of overlap between the predicted point and the ground truth.

3. Experiments and results

3.1. Data preparation

The eggplant point cloud dataset, described in Sec. 2.1, was utilized in this study. The dataset comprises 50 individual eggplant point clouds, with sizes ranging from the largest point cloud containing 200,000 points to the smallest one containing 50,000 points. To ensure consistency and optimize the dataset for analysis, each eggplant point cloud was downsampled to a fixed size of $N = 15,000$ points using the Octree Farthest Point Sampling (OFPS) method.

Before proceeding with network training and testing, it is essential to partition the dataset appropriately. In the case of the eggplant dataset, we divided it into three subsets: the training set, the validation set and the test set. More specifically, samples 1–40 were assigned to the training set, samples 41–55 were designated as the validation set, and samples 46–50 were exclusively reserved for the test set. This division ensures a balanced distribution of data across the subsets and facilitates accurate evaluation of the model's performance.

3.2. Network train and test

All experiments and computations presented in this paper were conducted on an Ubuntu 18.04 operating system. The hardware configuration used for the experiments consisted of a CPU equipped with 16 cores and 32 threads, 128GB of RAM, and an NVIDIA Tesla V100 GPU. The code implementation was carried out using the PyTorch 1.8.1 framework.

During model training, the number of training iterations was set to 250, the batch size was 8, the initial learning rate was 0.001, and the learning rate was adjusted by decreasing the learning rate every 10 iterations using the cosine annealing algorithm with a minimum learning rate of 0.0001. The network parameters were updated using the Adam optimizer, with a momentum of 0.9, and the weight decay was set to 0.0001. In the encoder component of the PSTNet model, we set $K = 32$, embedded feature dimension $d_e = 128$, and local feature dimension $d_l = 128$. In the Segmentation component of the PSTNet model, the global feature dimension indicated as d_g, was set to 1024.

We compare with three mainstream methods, PCT (Guo et al., 2021), DGCNN (Wang et al., 2019), and PointNet++ (Qi et al., 2017b), which are evaluated using the five quantitative metrics in Sec. 2.7.

Table 2
The comparison of Semantic Segmentation Performance for Four Networks. The Best Results are Highlighted in Bold.

Method	IoU(%)	Prec(%)	Rec(%)	F1(%)	Acc(%)
PointNet++	83.94	94.05	85.25	88.64	87.18
DGCNN	85.73	92.09	87.80	89.69	88.81
PCT	87.23	93.83	92.70	93.26	90.93
PSTNet(Ours)	**92.20**	**95.30**	**95.57**	**95.43**	**95.15**

3.3. Semantic segmentation results

Table 2 presents a comprehensive comparison of the segmentation results obtained by the PSTNet model and other algorithms. Notably, the PSTNet model exhibits superior segmentation performance, surpassing all the compared models in terms of *IoU*, Precision, Recall, F1-score, and Accuracy. Specifically, the PSTNet model achieves an impressive IoU of 92.20 %, Precision of 95.30 %, Recall of 95.57 %, F1-score of 95.43 %, and Accuracy of 95.15 %. These results outperform the second-ranked PCT (Guo et al., 2021) by 4.37, 1.47, 2.87, 2.17, and 4.22 percentage points, respectively. The trends of IoU and Accuracy for the four compared models are illustrated in Fig. 10.

Fig. 11 showcases the segmentation results of the four models on the test set. Despite the challenges posed by the unevenly sized leaves of eggplant plants at the seedling stage, the PSTNet model demonstrates its effectiveness in accurately distinguishing between them. PointNet++ (Qi et al., 2017b), though an improved version of PointNet, struggles to accurately identify stems and leaves, displaying the weakest performance among all the networks. PCT (Guo et al., 2021), based on the Transformer model (Vaswani et al., 2017) and utilizing self-attention to learn global features, exhibits better recognition than the other two models at the stem-leaf connection. DGCNN (Wang et al., 2019) introduces local features through directed graphs and achieves intermediate performance, falling between PointNet++ and PCT. It can roughly distinguish between stems and leaves.

3.4. Ablation experiment of PSTNet

The results clearly demonstrate that utilizing both the NPA and PSA modules ensures that PSTNet achieves the best performance, particularly in accurately segmenting data after OFPS processing. Conversely, using FPS-processed data negatively impacts the Intersection over Union (IoU) metric on the validation set, resulting in decreased model performance. This can be attributed to the loss of original plant traits caused by FPS, leading to the model's inability to learn correct shape features (Table 3).

4. Discussion

4.1. Multi-view point cloud collection and generation

To address the point cloud quality problem, we developed our 3D scanning equipment based on binocular vision and designed a dedicated 3D scanning platform specifically for capturing multi-view point cloud data. By acquiring plant point clouds in a controlled laboratory

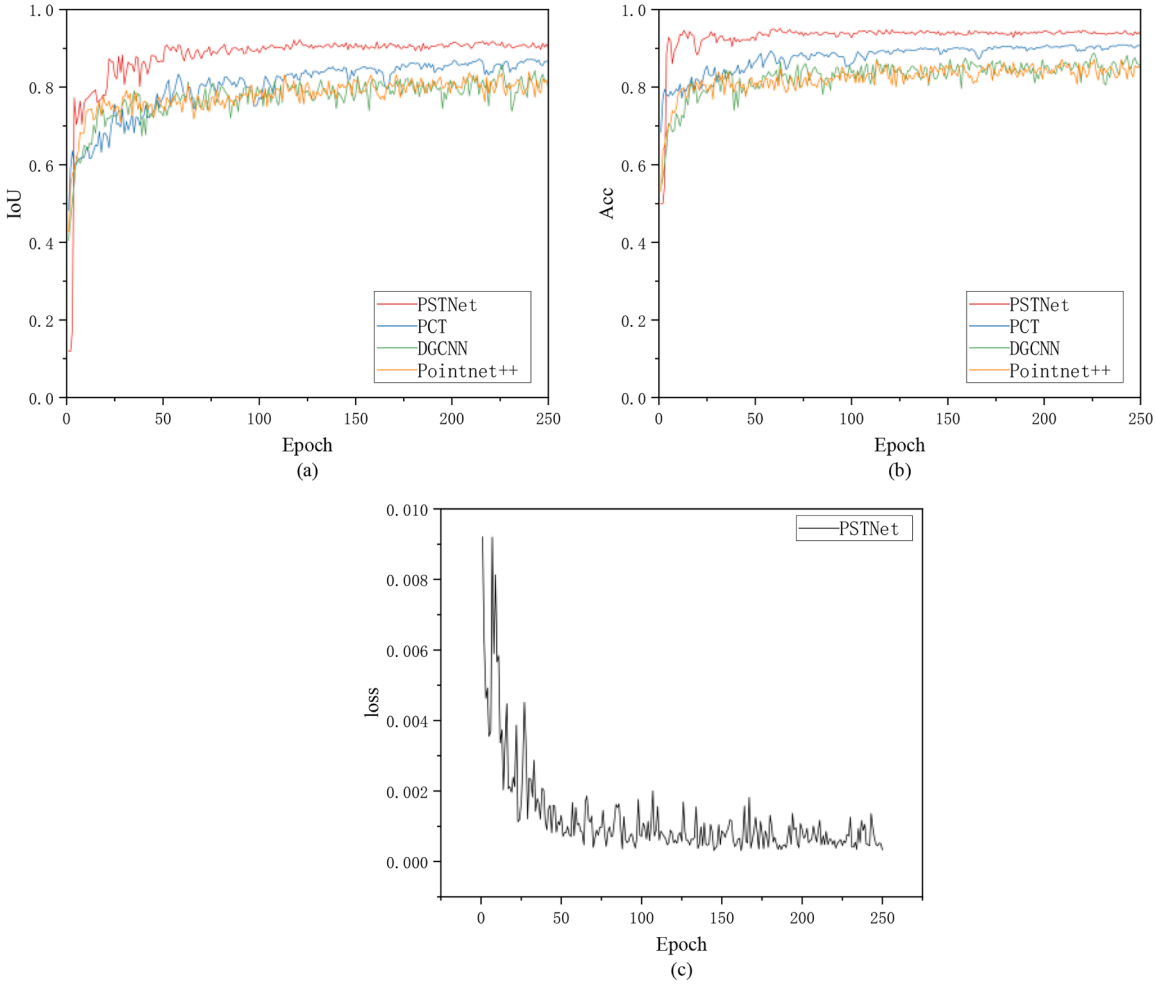

Fig. 10. Semantic segmentation results of four models. (a) IoU, (b) Acc, (c) PSTNet loss.

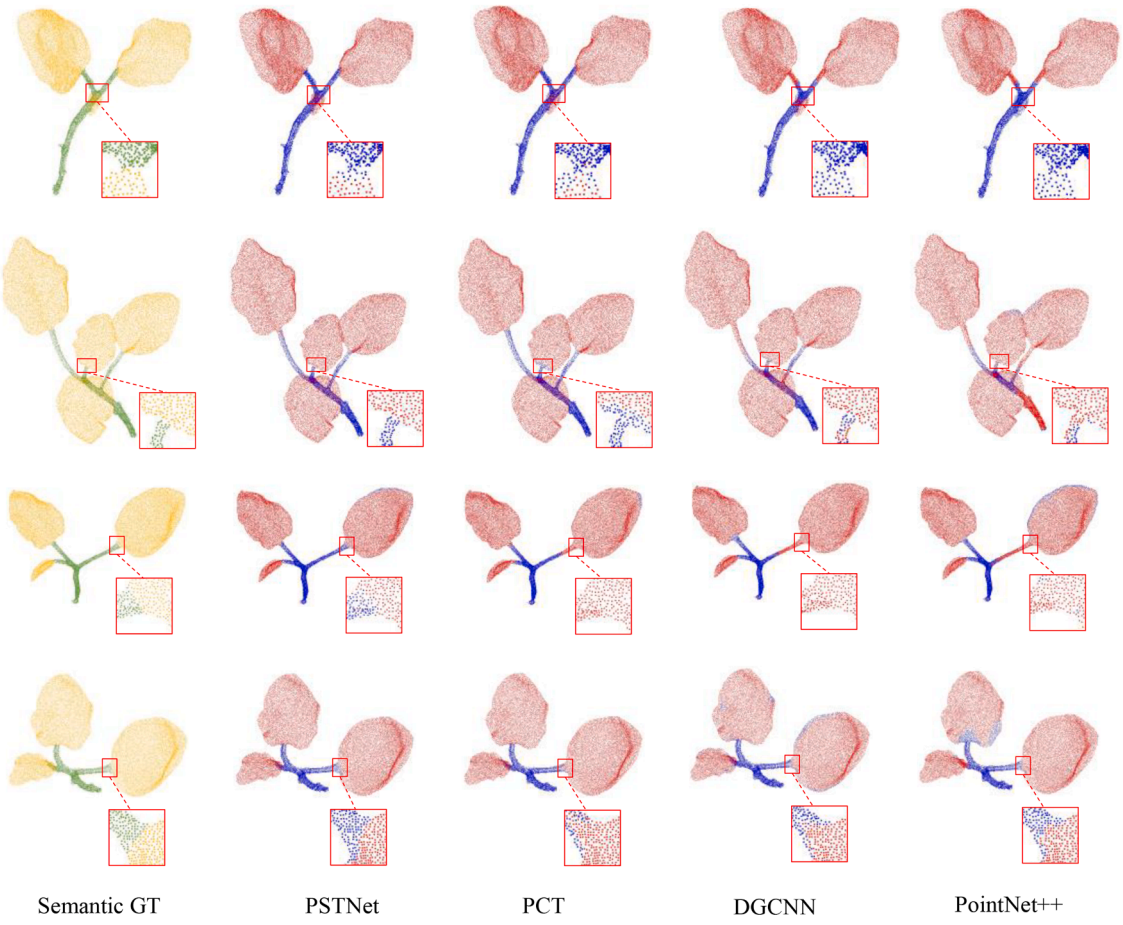

Fig. 11. Illustration of the qualitative results on the testing set by PSTNet, PCT, DGCNN, and PointNet++.

Table 3
Ablation experiments of PSTNet with different functions on the validation set.

OFPS	FPS	NPA	PSA	IoU(%)
	✓	✓	✓	90.34
✓		✓		89.78
✓			✓	88.57
✓		✓	✓	**92.20**

environment, we were able to mitigate the impact of natural conditions such as lighting and wind. This approach significantly improved the quality of the point cloud data by reducing background noise and enhancing data accuracy.

However, our approach also has some limitations. The 3D scanning platform used in this study relies on a rotating turntable to capture multi-view point clouds while preserving the shape of the plant. As a result, the scanning process requires a slow rotation speed, which can be time-consuming. Additionally, capturing point clouds for large-scale plantings in natural outdoor environments presents specific challenges that need to be considered, such as variations in lighting conditions, occlusions, and the potential effects of wind on the plants.

4.2. Analysis of experiment results

During the seedling stage, eggplant plants display a wide range of leaf sizes, including small, medium, and medium-to-large leaves. Moreover, these leaves can exhibit significant variations in shape, posing a challenge for accurately segmenting different leaf types. Nonetheless, PSTNet effectively addresses these challenges and achieves precise segmentation.

PointNet++ (Qi et al., 2017b) employs a recursive sampling grouping technique to learn feature representations from point clouds. However, this approach presents limitations when it comes to handling stem-leaf junctions in plant segmentation tasks. The recursive nature of the sampling and grouping process can lead to incorrect segmentation expansion, resulting in the misclassification of stem-leaf junctions. Consequently, this may cause inaccuracies in segmenting excessively long stems. DGCNN (Wang et al., 2019) addresses the limitations of PointNet (Qi et al., 2017a) by incorporating local features into point cloud processing. However, DGCNN still struggles to accurately distinguish stem-leaf connections. On the other hand, PCT (Guo et al., 2021) introduces the Transformer model to point cloud processing, marking its first application in this context. PCT demonstrates better recognition of stems and leaves compared to PointNet++ and DGCNN by considering both local and global features. However, even PCT falls short when it comes to achieving accurate segmentation at the joints of tiny leaves and stems.

These observations highlight the effectiveness of PSTNet in addressing the segmentation challenges posed by diverse leaf sizes and shapes during the eggplant plant's seedling stage. While other models, such as PointNet++, DGCNN, and PCT, have made notable contributions, they still exhibit limitations in accurately segmenting stem-leaf junctions and tiny leaf-stem connections.

4.3. Future work

While we have developed a 3D scanning platform, the scanning equipment can only achieve its best results in an indoor environment. Additionally, the scanning process itself may be slow due to the need to

maintain plant shape during rotation. These factors should be taken into account when designing the scanning equipment. Furthermore, the proposed PSTNet model for plant organ segmentation has certain limitations. It requires input point clouds with equal points, which restricts its generalization ability to handle point clouds of varying densities. Future research will focus on algorithmic enhancements and the development of new equipment to overcome this limitation and improve the accuracy and efficiency of point cloud acquisition.

5. Conclusion

In conclusion, we developed a specialized 3D point cloud scanning device specifically designed for capturing precise and reliable eggplant plant point clouds. To segment the eggplant plant point cloud, we incorporated neighborhood feature aggregation and point self-attention in the design of our PSTNet model. Our results show that PSTNet outperforms the state-of-the-art counterparts in the semantic and instance segmentation of eggplant plant point clouds. Specifically, the evaluation metrics for PSTNet are as follows: IoU achieves 92.20 %, Precision reaches 95.30 %, Recall is 95.57 %, F1-score achieves 95.43 %, and Accuracy attains 95.15 %.

PSTNet represents a novel approach to plant point cloud segmentation and holds great promise for future applications in the field of plant phenotyping analysis. Particularly, PSTNet has proven its effectiveness in accurately segmenting the organs of eggplant plants. Future work should be performed on capturing finer feature differences for more precise segmentation between plant stem tips, main stems, and petioles. Improving the feature characterization module will help extract more plant phenotypic parameter features.

CRediT authorship contribution statement

Linqian Ma: Writing – original draft, Supervision, Software, Methodology. **Lingyuan Kong:** Validation, Software, Resources. **Xingshuo Peng:** Investigation, Formal analysis, Data curation. **Keyuan Wang:** Visualization, Validation. **Nan Geng:** Writing – review & editing, Funding acquisition, Formal analysis, Data curation.

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.

Data availability

The data that has been used is confidential.

Acknowledgment

This work was supported by the Key Research and Development Program of Shaanxi (Grant No. 2019ZDLNY07–06–01). Thanks for all the help of the teachers and students.

References

Ao, Z., Wu, F., Hu, S., Sun, Y., Su, Y., Guo, Q., Xin, Q., 2022. Automatic segmentation of stem and leaf components and individual maize plants in field terrestrial LiDAR data using convolutional neural networks. Crop J., Crop Phenotyp. Stud. Appl. Crop Monitor. 10, 1239–1250. https://doi.org/10.1016/j.cj.2021.10.010.

Bassier, M., Bonduel, M., Van Genechten, B., Vergauwen, M., 2017. Segmentation of large unstructured point clouds using octree-based region growing and conditional random fields. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Presented at the 5th International Workshop LowCost 3D. Copernicus Publications, Hamburg, pp. 25–30. https://doi.org/10.5194/isprs-archives-XLII-2-W8-25-2017. Date: 2017/11/28 - 2017/11/29, Location.

Besl, P.J., McKay, N.D., 1992. Method for registration of 3-D shapes. Sensor fusion IV: control paradigms and data structures. presented at the sensor fusion IV: control paradigms and data structures, pp. 586–606. https://doi.org/10.1117/12.57955. SPIE.

Boogaard, F.P., van Henten, E.J., Kootstra, G., 2022. Improved point-cloud segmentation for plant phenotyping through class-dependent sampling of training data to battle class imbalance. Front. Plant Sci. 13.

Cen, H., Nunez-Sanchez, S., Sarua, A., Bickerton, I., Fox, N.A., Cryan, M.J., 2020. Solar thermal characterization of micropatterned high temperature selective surfaces. J. Photonics Energy 10, 024503. https://doi.org/10.1117/1.JPE.10.024503.

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F., 2015. ShapeNet: an Information-rich 3D model repository. https://doi.org/10.48550/arXiv.1512.03012.

Conn, A., Pedmale, U.V., Chory, J., Stevens, C.F., Navlakha, S., 2017. A statistical description of plant shoot architecture. Curr. Biol. 27, e3. https://doi.org/10.1016/j.cub.2017.06.009, 2078-2088.

Cui, Y., Liu, X., Liu, H., Zhang, J., Zare, A., Fan, B., 2021. Geometric attentional dynamic graph convolutional neural networks for point cloud analysis. Neurocomputing. 432, 300–310. https://doi.org/10.1016/j.neucom.2020.12.067.

Dutagaci, H., Rasti, P., Galopin, G., Rousseau, D., 2020. ROSE-X: an annotated data set for evaluation of 3D plant organ segmentation methods. Plant Methods 16 (28). https://doi.org/10.1186/s13007-020-00573-w.

Furukawa, Y., Ponce, J., 2010. Accurate, dense, and robust multiview stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1362–1376. https://doi.org/10.1109/TPAMI.2009.161.

Galieni, A., D'Ascenzo, N., Stagnari, F., Pagnani, G., Xie, Q., Pisante, M., 2021. Past and future of plant stress detection: an overview from remote sensing to positron emission tomography. Front. Plant Sci. 11.

Girardeau-Montaut, D., 2016. CloudCompare. Fr. EDF RD Telecom Paris Tech 11.

Gosa, S.C., Lupo, Y., Moshelion, M., 2019. Quantitative and comparative analysis of whole-plant performance for functional physiological traits phenotyping: new tools to support pre-breeding and plant stress physiology studies. Plant Sci., The 4th Internat. Plant Phenotyping Symposium 282, 49–59. https://doi.org/10.1016/j.plantsci.2018.05.008.

Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M., 2021. PCT: point cloud transformer. Comput. Vis. Media 7, 187–199. https://doi.org/10.1007/s41095-021-0229-5.

He, H., Wang, H., Sun, L., 2018. Research on 3D point-cloud registration technology based on Kinect V2 sensor, 2018. In: Chinese Control And Decision Conference (CCDC). Presented at the 2018 Chinese Control And Decision Conference (CCDC), pp. 1264–1268. https://doi.org/10.1109/CCDC.2018.8407323.

Huang, J., You, S., 2016. Point cloud labeling using 3d convolutional neural network, 2016. In: 23rd International Conference on Pattern Recognition (ICPR). Presented at the 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 2670–2675. https://doi.org/10.1109/ICPR.2016.7900038.

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., 2020. Separating the structural components of maize for field phenotyping using terrestrial LiDAR data and deep convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 58, 2644–2658. https://doi.org/10.1109/TGRS.2019.2953092.

Jin, S., Su, Y., Wu, F., Pang, S., Gao, S., Hu, T., Liu, J., Guo, Q., 2019. Stem–Leaf Segmentation and Phenotypic Trait Extraction of Individual Maize Using Terrestrial LiDAR Data. IEEE Trans. Geosci. Remote Sens 57, 1336–1346. https://doi.org/10.1109/TGRS.2018.2866056.

Jin, X., Yang, W., Doonan, J.H., Atzberger, C., 2022. Crop phenotyping studies with application to crop monitoring. Crop J., Crop Phenotyp. Stud. Appl. Crop Monitor. 10, 1221–1223. https://doi.org/10.1016/j.cj.2022.09.001.

Kolhar, S., Jagtap, J., 2023. Plant trait estimation and classification studies in plant phenotyping using machine vision – A review. Inf. Process. Agric. 10, 114–135. https://doi.org/10.1016/j.inpa.2021.02.006.

Martínez-Ispizua, E., Calatayud, Á., Marsal, J.I., Mateos-Fernández, R., Díez, M.J., Soler, S., Valcárcel, J.V., Martínez-Cuenca, M.R., 2021. Phenotyping local eggplant varieties. Commit. Biodiversit. Nutrit. Qual. Preservat.. Front. Plant Sci. 12.

Masuda, T., 2021. Leaf area estimation by semantic segmentation of point cloud of tomato plants. In: Presented at the proceedings of the ieee/cvf international conference on computer vision, pp. 1381–1389.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. PointNet: deep learning on point sets for 3d classification and segmentation. https://doi.org/10.48550/arXiv.1612.00593.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space. https://doi.org/10.48550/arXiv.1706.02413.

Rist, F., Herzog, K., Mack, J., Richter, R., Steinhage, V., Töpfer, R., 2018. High-precision phenotyping of grape bunch architecture using fast 3D sensor and automation. Sensors 18 (763). https://doi.org/10.3390/s18030763.

Schunck, D., Magistri, F., Rosu, R.A., Cornelißen, A., Chebrolu, N., Paulus, S., Léon, J., Behnke, S., Stachniss, C., Kuhlmann, H., Klingbeil, L., 2021. Pheno4D: a spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. PLoS. One 16, e0256340. https://doi.org/10.1371/journal.pone.0256340.

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. https://doi.org/10.1016/j.biosystemseng.2019.08.014.

Tran, D.T., Hertog, M.L.A.T.M., Tran, T.L.H., Quyen, N.T., Van de Poel, B., Mata, C.I., Nicolaï, B.M., 2017. Population modeling approach to optimize crop harvest strategy. The case of field tomato. Front. Plant Sci. 8.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems. Curran Associates, Inc.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y., 2018. Graph attention networks. https://doi.org/10.48550/arXiv.1710.10903.

Vo, A.V., Truong-Hong, L., Laefer, D.F., Bertolotto, M., 2015. Octree-based region growing for point cloud segmentation. ISPRS J. Photogramm. Remote Sens. 104, 88–100. https://doi.org/10.1016/j.isprsjprs.2015.01.011.

Wang, T., Li, J., An, X., 2015. An efficient scene semantic labeling approach for 3D point cloud. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. Presented at the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp. 2115–2120. https://doi.org/10.1109/ITSC.2015.342.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph CNN for learning on point clouds. https://doi.org/10.48550/arXiv.1801.07829.

Xue-Tong, L., Su-Hui, Y., Xin, W., Zhuo, L., Jin-Ying, Z., 2020. Theoretical study of eye-safe 2μm laser directly pumped by sunlight. 物理学报 69, 094202–094209. https://doi.org/10.7498/aps.69.20191985.

Zhang, H., Wang, L., Jin, X., Bian, L., Ge, Y., 2023. High-throughput phenotyping of plant leaf morphological, physiological, and biochemical traits on multiple scales using optical sensing. Crop J. https://doi.org/10.1016/j.cj.2023.04.014.

Zhang, J., Zhao, X., Chen, Z., Lu, Z., 2019. A review of deep learning-based semantic segmentation for point cloud. IEEe Access. 7, 179118–179133. https://doi.org/10.1109/ACCESS.2019.2958671.

Zhang, S., 2018. High-speed 3D shape measurement with structured light methods: a review. Opt. Lasers Eng. 106, 119–131. https://doi.org/10.1016/j.optlaseng.2018.02.017.

Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X., Yan, G., 2016. An easy-to-use airborne LiDAR data filtering method based on cloth simulation. Remote Sens. 8, 501. https://doi.org/10.3390/rs8060501.

Zhang, Y., Zhang, N., 2018. Imaging technologies for plant high-throughput phenotyping: a review. Front. Agric. Sci. Eng. 5, 406–419. https://doi.org/10.15302/J-FASE-2018242.

Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V., 2021. Point transformer. https://doi.org/10.48550/arXiv.2012.09164.