Automatic segmentation of plant organs from point clouds

Qiwei Shen
student #5687500
Delft University of Technology | Architecture and the Built Environment

1st supervisor: Liangliang Nan
2nd supervisor: Jantien Stoter

January 15, 2024

Contents

1 Introduction

Plant phenotyping, the quantitative acquiring, modeling, and analyzing of plant traits that are formed by the dynamic interaction of genotype and environment, could bridge the gap between genotype and phenotype and reveal the contribution of genotype to phenotypic variation through quantitative trait locus (QTL) mapping and genome-wide association studies (GWASs) (Yang et al., 2020; Ninomiya et al., 2019; Junker et al., 2015; Xiao et al., 2017). Moreover, plant phenotyping is also essential in plant genetic gain and plant variety improvement by characterizing desired phenotypes in the breeding programme (Watt et al., 2020). However, unlike the advanced high-throughput DNA sequencing technology in plant genomics, the plant phenotyping approaches mainly remain in the infancy stage (Hu et al., 2021). Confronted with large genotype data, the lack of efficient high-throughput plant phenotyping technology to acquire corresponding reliable phenotype data has emerged as the bottleneck of agronomy and plant sciences (Mir et al., 2019). Traditional plant phenotyping methods, characterized by their labor-intensive, time-consuming, and often invasive nature, coupled with a reliance on subjective and manual measurements, are increasingly inadequate for the demands of modern plant phenomics (Tardieu et al., 2017). Therefore, to break the ”phenotypic bottleneck”, the development of automatic, non-invasive, high-throughput plant phenotyping technology has attracted much attention worldwide.

In the last several decades, the integration of computer vision in plant phenotyping has advanced significantly, which is evident in the efficient extraction of plant traits and reduction of manual labor (Das Choudhury et al., 2019). Among those computer vision-based plant phenotyping pipelines, the importance of high-precision segmentation of plant organs is self-evidence. Early works mainly focus on developing two-dimensional (2D) image-based methods. For instance, using threshold-based approaches, Hartmann et al. (2011) and De Vylder et al. (2012) established organ segmentation from barley and arabidopsis 2D images, respectively; Zhang et al. (2017) and Das Choudhury et al. (2018) successfully segmented individual components of maize from its 2D image sequences by using graph-based methods; based on edge detection, Yin et al. (2017) segmented all leaves of arabidopsis. What is more, the rapid development of the application of convolutional neural networks (CNNs) in image classification and segmentation has led to breakthroughs in plant phenotyping, as demonstrated by Aich and Stavness (2017) and Hasan et al. (2018) who achieved segmentation of arabidopsis and tobacco leaf, and detection of wheat spikes, respectively. These 2D phenotyping pipelines, being either automatic or semiautomatic, have significantly reduced the manual labor required in conventional phenotyping practice. However, they are not without limitations. Due to the lack of depth information, self-occlusions and leaf crossover problems are difficult to address; it is hard to describe plants with complex morphological structures accurately (Sun et al., 2020). Consequently, the 2D phenotyping methods are mainly restricted to the aforementioned simple monocotyledonous or rosette plants (Li et al., 2022b).

Those limitations spurred the studies of three-dimensional (3D) phenotyping methods, which involve segmenting plant organs from the 3D model reconstructed by using technologies like time of flight (ToF) cameras (Xiang et al., 2019), structure from motion (SfM) (Jay et al., 2015), light detection and ranging (LiDAR) (Jin et al., 2021), and neural radiance fields (NeRFs) (Jignasu et al., 2023). The advantages of 3D phenotyping are notable, addressing occlusion and overlapping issues, allowing for accurate trait extraction from plants with more complicated structures, such as rapeseed (Du et al., 2023), rice (Gong et al., 2021), tomato (Li et al., 2022b), which shows great potential in plant phenotyping.

Focusing on 3D point cloud data, the common traditional organ-segmentation methods include model-based algorithms (Gélard et al., 2017), clustering-based algorithms (Xu et al., 2018; Thapa et al., 2018), normal vector-based method (Lin et al., 2016; Li et al., 2017), and skeleton-based algorithms (Xiang et al., 2019; Gaillard et al., 2020; Miao et al., 2021; Ma et al., 2023). Although those methods could be efficient on multiple plant species through parameter tuning, the segmentation performance on the boundary area between two parts can be unstable (Miao et al., 2021; Peng et al., 2022). Moreover, the parameter tuning processes are time-consuming and highly rely on researchers' prior knowledge of plant morphology structure (Li et al., 2022b). Therefore, those common traditional organ-segmentation methods lack generalizability and cannot meet modern plant phenomics' evolving demands. Concerning that, developing a general method that could segment plant organs across multiple plant species is one of the critical research directions in plant phenotyping.

In recent studies, deep learning-based methods have shown their high generalizability and accuracy in multiple 3D point cloud tasks, including shape classification, object tracking, and semantic/instance segmentation (Guo et al., 2020). Regarding that, utilizing deep learning methods in plant organ segmentation has the potential to tackle the limitations of traditional 3D segmentation methods. However, unlike regular 2D image pixels, point cloud data with its unordered and uneven natures can not be directly input into most deep learning models widely used in 2D tasks (Yang et al., 2023). To tackle that problem, previous works mainly transform 3D point cloud data into multi-view (Su et al., 2015; Shi et al., 2019) or volumetric (Maturana and Scherer, 2015; Jin et al., 2019; Zhang et al., 2023) representations, before feeding them into a deep learning model (Qi et al., 2017a). However, the projection angles in multi-view-based methods are hard to determine, and voxelization parameters in voxel-based methods are also tough to balance between performance and computation complexity, and geometry information will be lost during such transformations. Point-based methods, extract point features in an end-to-end fashion, which could minimize the effect of those issues by directly taking point cloud data as input (Qi et al., 2017a,b). For example, Li et al. (2022c), and Patel et al. (2023) adopted PointNet (Qi et al., 2017a)/PointNet++ (Qi et al., 2017b) to segment organs on maize and sorghum plants, respectively, and both of them obtained outperformed performance; Li et al. (2022b) proposed PlantNet, a point-wise network, established tobacco, tomato, and sorghum organ segmentation successfully. However, the computation complexity is sensitive to the number of the input point cloud, therefore, the down-sampling operation is normally necessary to balance the training speed and segment performance in point-based methods (Li et al., 2022a).

Besides those approaches, Transformer (Vaswani et al., 2017), originally designed for natural language processing, with its well-designed self-attention module, has achieved impressive results in 3D point cloud segmentation (Zhao et al., 2021; Guo et al., 2021). Therefore, the improvement or modification of the self-attention module is one of the main research directions of 3D point cloud Transformers and is still in its infancy (Lu et al., 2022; Li et al., 2022a). Adopting the shifted window attention in Swin-Transformer (Liu et al., 2021) and Single-stride Sparse Transformer (Fan et al., 2022), Du et al. (2023) built the dual window sets attention to capture neighborhood information, which performed well in rapeseed plant organ segmentation; inspired by convolutional block attention module (Woo et al., 2018), Li et al. (2022a) designed spatial and channel attention modules in its PSegNet, which improved its training efficiency on multiple plants organ segmentation tasks; similarly, by introducing cross-window self-attention module, Win-Former, proposed by Sun et al. (2023), conducting maize point cloud efficiently.

In conclusion, the field of plant organ segmentation is increasingly adopting deep learning methodologies. However, there is a notable scarcity of Transformer-based approaches tailored for the automatic segmentation of plant organs from point clouds. Existing Transformers are predominantly species-specific, casting doubt on their efficacy for cross-species segmentation. Furthermore, most multi-scale attention modules integrated into these Transformers are borrowed from non-plant segmentation networks. These modules typically treat all regions equally, lacking specific emphasis or attention guidance. To the best of our knowledge, existing self-guided attention modules for point clouds remain cumbersome and inefficient. In light of these limitations, this research draws inspiration from the multi-scale attention modules and the concept of the plant skeleton. We will propose the development of a novel skeleton-guided attention module. This initiative culminates in the introduction of the Skeleton-aware Attention Transformer, which has the potential to handle organ segmentation across multiple plant species.

2 Related work

To confine the range of the literature review within our core research scope, three parts will be examined in this section, the common 3D point cloud skeletonization algorithms will be discussed in section Section 2.1, the pioneer Transformer networks in plant organ segmentation domain will be presented in Section 2.2, and lastly, the novel multi-scale attention modules will be explored in section Section 2.3.

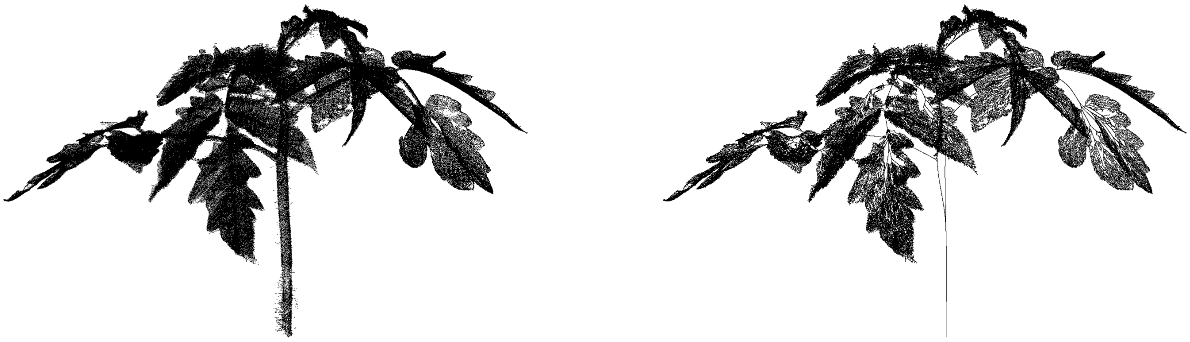
2.1 Skeleton extraction algorithms

The skeleton is one of the simplified representations of a 3D point cloud model, which can intuitively reflect the morphology structure and topology structure of itself. However, with the unordered, uneven, and unconnected natures of the point cloud data, and the normally complexity structure of plant leaves and branches, how to extract the correct skeleton from the plant point cloud model is a question.

Bao et al. (2019) introduced a simple and efficient algorithm to extract the skeleton graph of Maize, which involves slicing the plant point cloud into thin layers along the plant growth direction (the X-axis) first and clustering the points on the same layer into several Euclidean clusters according to a Euclidean distance threshold, and then the point which is closest to the centroid of the cluster will be labeled as the skeleton point. The skeleton graph is generated using the minimum spanning tree (MST). This algorithm is efficient for plants with generally vertical structures and without horizontal parts, like maize or sorghum (Bao et al., 2019; Xiang et al., 2019). For plants with significant horizontal branches (e.g., tomato, rapeseed), the extracted skeleton graph can contain unexpected errors.

The L_1-Medial Skeleton algorithm constructs the skeletal framework of a point cloud by iteratively computing local L_1 median points (Huang et al., 2013). This process is governed by a balance of attractive and repulsive forces within a progressively expanding local neighborhood. Through this iterative method of contraction, the algorithm efficiently identifies and extracts the skeleton points. This approach allows for a robust representation of the point cloud's underlying structure, even in the presence of noise, outliers, or incomplete data. Utilizing L_1-medial skeleton algorithm, Ma et al. (2023) successfully extracted the refined skeleton of rapeseed plants, which have strong scattering structures (Du et al., 2023). L_1-medial skeleton algorithm could feasibly process cylindrical-shaped point clouds, adeptly extracting skeletons from objects like tree branches (Su et al., 2019) or rapeseed siliques (Ma et al., 2021, 2023). However, it is unsuitable for flat or planar structures, such as plant leaves. In these cases, the extracted leaf skeleton may not accurately represent the leaf veins, and can sometimes fall outside the leaf's point cloud (Wu et al., 2019).

In the paper by Du et al. (2019), AdTree, an innovative method that integrates Delaunay triangulation (DT) and graph theory for extracting tree skeletons, was introduced. This approach utilizes DT for connectivity graph generation from the point cloud and further extracts the initial skeleton by establishing a MST using the Dijkstra shortest-path algorithm on that connectivity graph. The refined skeleton is obtained by iterative pruning on the initial skeleton to remove redundant parts. Despite the effectiveness in processing cylinders of this graph-based algorithm, especially tree branches (Du et al., 2019; Wang et al., 2021), similar to L_1-medial skeleton algorithm, this algorithm shows limitations in accurately representing flat and wide plant leaves. Figure 1 depicts a tomato skeleton generated by AdTree, which fails on skeleton extraction from leaves and the stem skeleton also contains errors.



(a) Raw tomato point cloud

(b) Tomato skeleton

Figure 1: Skeleton extraction using AdTree

Laplacian-based contraction algorithm, proposed by Cao et al. (2010), which iteratively contracts a point cloud using the cotangent-weighted Laplacian operator constructed by the one-ring Delaunay neighborhood (Au et al., 2008; Meyer et al., 2003). With the contraction and attraction weights to balance the shrinkage processing, the contracted point cloud could preserve the original geometry's characteristics while minimizing volume. The algorithm generates a connectivity graph from key points identified through clustering in the contracted cloud. Subsequently, a skeleton graph is derived using a MST on this connectivity graph. Notably effective in extracting curve skeletons from both cylinder objects (Li et al., 2020) and surfaces with boundaries (e.g., plant stems and leaves), this method has been widely adopted in plant skeleton extraction. For instance, Wu et al. (2019), Miao et al. (2021), and Peng et al. (2022) successfully applied it to extract refined skeletons of maize and tomato plants, respectively.

2.2 Transformers in plant organ segmentation

The Transformer architecture, known for its exceptional ability to learn global features and execute permutation-equivariant operations, is inherently suitable for point cloud processing and analysis (Lu et al., 2022). This suitability has led to its widespread adoption in a variety of 3D point cloud processing tasks, as evidenced by several studies (Zhao et al., 2021; Guo et al., 2021). Despite its success in these areas, the application of Transformers in the field of 3D phenotyping, particularly in the segmentation of plant organs, remains relatively unexplored. And some pioneer studies in the past two years will be discussed in this section.

MASPC_Transform, standing for Multi-head Attention Separation and Position Code, represents a pioneering application of Transformer architecture in 3D plant organ segmentation (Li and Guo, 2022). This model primarily builds upon the Point Transformer framework (Engel et al., 2021). It introduces a position-coding network (which contains absolute and relative positions) to mitigate the effects of point cloud disorder and irregularity during feature extraction and integrates this network within the local and global feature extraction modules of the Point Transformer. Additionally, MASPC_Transform incorporates a multi-head attention separation loss based on spatial similarity. This loss function effectively segregates attention positions, thereby facilitating the creation of distinct attention feature spaces and preventing overlap among them. Comparative semantic segmentation results on the ROSE-X dataset (Dutagaci et al., 2020) demonstrate that MASPC_Transform outperforms not only the Point Transformer but also other common networks like PointNet, DGCNN (Zhang et al., 2021), PointCNN (Li et al., 2018).

Du et al. (2023) introduced PST (Plant Segmentation Transformer) for extracting silique phenotypes in rapeseed plants, achieving both semantic and instance segmentation. Inspired by dynamic voxelization (Zhou et al., 2020) and voxel feature encoding (Zhou and Tuzel, 2018), PST incorporates a dynamic voxel feature encoder (DVFE), which efficiently converts point-wise inputs into voxel-wise embeddings with a learned feature, reducing information loss. Its self-attention module was adopted from the shifted-window self-attention (Liu et al., 2021; Fan et al., 2022), which could efficiently capture the neighbor context for voxel feature learning. Additionally, PST integrates an instance segmentation head from PointGroup (Jiang et al., 2020), enabling its instance segmentation capabilities. The results show that, for both semantic and instance silique segmentation, PST obtained the best performance among PointNet++, PAConv (Xu et al., 2021), and DGCNN.

In the article by Li et al. (2022a), PSegNet is introduced as an advanced network for point cloud segmentation in plants, featuring a Double-Granularity Feature Fusion Module (DGFFM) and an Attention Modules (AMs) with spatial and channel components. DGFFM adeptly decodes and fuses features of varying granularities, enhancing the network's segmentation capabilities. The unique double-flow structure of PSegNet with AMs, with its upper and lower branches dedicated to instance and semantic segmentation respectively, leverages AMs for focused feature processing. This innovative architecture allows PSegNet to surpass established networks like PointNet, PointNet++, ASIS, and PlantNet in both organ semantic and leaf instance segmentation tasks across tobacco, tomato, and sorghum plants

Sun et al. (2023) innovatively proposed Win-Former, a model utilizing Sphere Projection and Window Transformer for local feature aggregation. This method projects point clouds onto a spherical surface, dividing it into sphere windows for local self-attention computations. It uniquely adopts a Cross-Window self-attention mechanism, akin to the shifted-window approach in Liu et al. (2021) and Fan et al. (2022), by altering sphere window positions along azimuth and elevation angles. This design facilitates hierarchical feature extraction through both Window Transformer and Cross-Window attention. The results demonstrate Win-Former's superior performance over established models like PointNet, PointNet++, and DGCNN on the maize dataset from Pheno4D (Schunck et al., 2021) dataset.

2.3 Multi-scale attention modules

From the above pioneer Transformer architectures, we can find that the ability to extract features in multi-scale is crucial for plant organ segmentation performance. Moreover, as the core component of Transformers, the importance of the self-attention module is self-evidence. Herein, the recent studies on multi-scale attention modules will be discussed in this section. Regarding the infancy stage of Transformer application in 3D tasks, both 2D and 3D scenarios will be included.

Inspired by the notable performance of multi-scale CNNs, Chen et al. (2021) proposed CrossViT (Cross-Attention Multi-Scale Vision Transformer), which could efficiently extract multi-scale feature representations for 2D image classification. And the core component of that operation is cross-attention module. By partitioning the image with different patch sizes, the large branch patches operated by coarse-grained patch size, and the small branch patches clipped by fine-grained patch size can obtained. After adding an additional classification token to two branches, as in the original BERT (Devlin et al., 2018), they will be fed into cross-attention module. As illustrated in Figure 2, the additional classification token of the large branch will be projected to match the feature dimension of the small-scale group to serve as a **Query**, and interact with the **Key** and **Value** derived from the small branch's embedded feature tokens. The small branch follows the same procedure but swaps the additional classification token and embedded feature tokens from another branch. By enabling multi-scale features in Transformer, CrossViT obtained impressive performance. Moreover, Yang et al. (2023) proposed PointCAT (Cross-Attention Transformer for Point Cloud), a dual-branch cross-attention transformer network. By replacing multi-scale image partitions with multi-scale point grouping, which is established by farthest point sampling (FPS), K-nearest neighbor (KNN) search, and max-pooling aggregation, such cross-attention module was able to be utilized in 3D scenarios. And the ablation shows that cross-attention module is also efficient in 3D scenarios, and has demonstrated improvements in segmentation tasks on ModelNet40 (Wu et al., 2015) and S3DIS (Armeni et al., 2016) dataset.

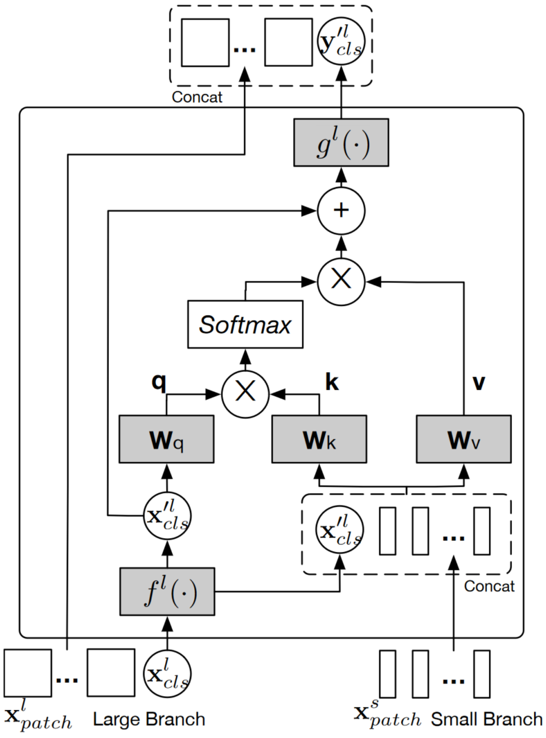


Figure 2: Cross-attention module for large branch (Chen et al., 2021)

The computation cost of self-attention grows quadratically with the token length, which limits the self-attention application for large-scale input. Concerning that, Ren et al. (2023) proposed

SG-Former (Self-guided Transformer). This network introduced a token allocation mechanism guided by a significance map, namely self-guided attention module. According to the significance map, the self-guided attention module will allocate more tokens to regions of high salience and fewer to less significant areas 3. Consequently, focusing on the inherent significance difference of different tokens, the self-guided attention module enables the extraction of more comprehensive and efficient multi-scale features. The SG-Former demonstrated superior performance over many existing vision Transformers, striking a balance between computational efficiency and the extraction of high-scale features.

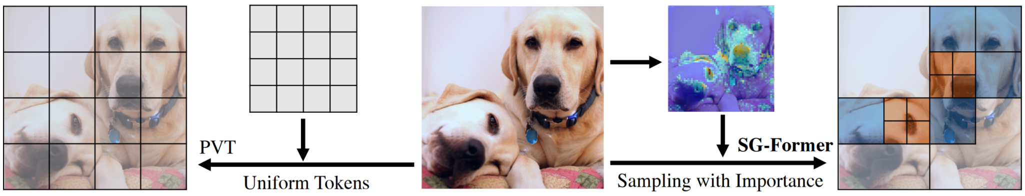


Figure 3: The idea of SG-Former (Ren et al., 2023)

3 Research questions

3.1 Objective

The main research question of this project is:

> *To what extent will the skeleton-guided attention module improve the performance and generalizability of the plant organ segment network?*

This research aims to design a deep-learning network for plant organ semantic and instance segmentation, which could be performed across plant species. To achieve this research goal, the main research question can be divided into the following sub-questions:

1) *How can we extract refined skeletons from various plant species?*

2) *How to use the skeleton information to guide attention computation?*

3) *How to establish multi-scale feature extraction according to skeleton information?*

4) *How accurate is the segmentation result of our network compared to traditional methods and other pioneer networks?*

3.2 Research scope

This research will focus on the organ semantic and instance segmentation for plant shoots, the plant root will be excluded from our research. And the network architecture will be limited to Transformer architecture, other architectures will not be discussed. Furthermore, as there is no official dataset for the 3D plant segmentation task, our network will be evaluated on the hybrid dataset (including public, unpublic, and self-built datasets) described in Section 6.3.

4 Methodology

In this research, the skeleton of the plant will be extracted using Laplacian-based contraction algorithm. The baseline of our deep-learning network will be PointCAT, and the cross-attention layer in PointCAT, self-guided attention in SG-Former will be used as inspiration for the design of the skeleton-aware attention module. The transformer decoder of OneFormer3D (Kolodiazhnyi et al., 2023) will be combined with our network, to establish semantic and instance segmentation.

4.1 Skeleton extraction

Compared with skeleton extraction algorithms mentioned in Section 2.1, Laplacian-based contraction algorithm demonstrates notable proficiency when handling objects of horizontal, cylindrical, and flat geometries. Given this advantage, the Laplacian-based contraction algorithm has been selected as the primary technique for extracting plant skeletons in the present study. The core process involves an iterative contraction of the point cloud, achieved through the resolution of a linear system, as defined in Equation 1:

$$\begin{bmatrix} W_L L \\ W_H \end{bmatrix} P' = \begin{bmatrix} 0 \\ W_H P \end{bmatrix} \tag{1}$$

Here, P represents the initial point cloud, while P' denotes its contracted form. The matrix L, a cotangent-weighted Laplacian matrix, is constructed using one-ring Delaunay neighbors. The diagonal matrices W_L and W_H regulate the intensity of the contraction and the preservation of the original position, respectively, ensuring the movement of the point cloud along the estimated normal direction.

The contraction process unfolds iteratively. Each iteration involves solving Equation 1 to obtain P'. The matrices W_L and W_H are subsequently updated as per Equation 2, where S_i^t and S_i^0 represent the current and original neighborhood extent of the point p_i, respectively. This results in the generation of a new point cloud P^{t+1} from the current point cloud P^t. The updated Laplacian matrix, L^{t+1}, is reconstructed using P^{t+1}. The iteration terminates when $W_L^{t+1}/W_L^t < 0.01$, or after exceeding 15 iterations. Typically, the point cloud can contract to the skeleton's shape within 10 iterations.

$$W_L^{t+1} = S_L W_L^t, \quad W_{H,i}^{t+1} = W_{H,i}^0 \frac{S_i^0}{S_i^t} \tag{2}$$

Upon obtaining the contracted point cloud, the final skeleton graph is subsequently derived by applying a set of graph-based operations.

4.2 Down-sampling

For most Transformer architectures, down-sampling is necessary for the point cloud before being fed into. The computation cost is highly sensitive to the number of input points. It is almost impossible to train a Transformer network with the full-scale input point cloud. With the advanced sensor technology, the average number of points in common plant point cloud datasets is exceedingly high, which can exceed 100,000 (ROSE-X), and even above 1,000,000 (Pheno4D). Contrastingly, the input capacity for many Transformer networks is capped at less than 5,000 points, which means at least 95% of the points need to be deleted, and this poses a risk of losing the original geometric characteristics information of the plants. Consequently, the selection of an effective down-sampling methodology is critical. A well-chosen down-sampling strategy should not only mitigate noise impact but also preserve geometry information as much as possible.

Although the proportion of edge points of the plant point cloud is often less than 10% of the total points, they can efficiently represent the overall structure of the plant, indicating that the edge points contain important global information of the point cloud (Li et al., 2019). The 3D Edge-Preserving Sampling (3DEPS) approach, as proposed in Li et al. (2022b), could preserve the edge/boundary points of the object during down-sampling, and has the potential ability to preserve the object's geometry information. 3DEPS employs Surface Boundary Filter (SBF) (Klasing et al., 2009) to partition the point cloud into two distinct categories: edge points and internal points. The steps of SBF are as follows:

1) Start with a point in the input point cloud, and find its k-nearest neighbors;

2) Calculate the principal components of that point with its k-nearest neighbors, and project them on the PCA plane constructed by the first two principal components ($\mathbf{u}$, $\mathbf{v}$), the projected k-nearest neighbors $X = \{x_k\}_{k \in K} \subset \mathbb{R}^2$, and projected point $x_i \subset \mathbb{R}^2$;

3) For each projected k-nearest neighbors, compute the angle θ by equation 3:

$$\theta_j = \arccos\left((x_k - x_i), \mathbf{u}\right), \quad j \in K \tag{3}$$

4) The sign of θ is assigned by the sign of $(x_k - x_i) \cdot \mathbf{u}$ (i.e., assign the same sign of $(x_k - x_i) \cdot \mathbf{u}$ to θ), therefore, the range of θ becomes $[-\pi, \pi]$, and push the θ with the assigned sign into the angle set Θ;

5) Sort the angle set $\Theta = \{\theta_j\}_{j \in K}$ in ascending order. If the maximum angle difference satisfies $\max(\theta_{j+1} - \theta_j) > \theta_{\text{threshold}}$ (default $\theta_{\text{threshold}} = \frac{\pi}{2}$), then the corresponding point of x_i in the point cloud is labeled as a edge point;

6) Repeat Step 1-5 with the next point in the input point cloud. After all points have performed the above steps, SBF is completed.

After labeling the edge points in the point cloud, 3DEPS applies FPS separately on edge points and non-edge points to form the final sampled point cloud. The proportion of edge points and non-edge points in the final sampled point cloud is user-defined. By purposefully increasing the edge points ratio in the final sampled point cloud, geometry loss during the down-sampling operation can be minimized.

4.3 Baseline network

As discussed in Section 2.3, PointCAT could efficiently capture long-range dependencies and multi-scale information among sampled points through its cross-attention module. This module is adept at fusing and learning features from multiple scales. What is more, our research aims to develop a Skeleton-aware Attention Transformer, which will leverage a skeleton-guided attention module to capture multi-scale features. Given the effectiveness of the cross-attention module in PointCAT in learning features from multi-scale data, it is an ideal baseline for our research, offering a robust foundation for our network design.

In the original PointCAT network, point patch tokens are derived through FPS and KNN grouping. While this method is efficient, it is somewhat arbitrary and lacks a targeted approach. Incorporating skeleton information to guide the generation of point patch tokens can potentially enhance the network's performance. Therefore, in our proposed modification, the multi-scale grouping module in PointCAT will be substituted with a skeleton-guided attention module. This approach aims to provide a more structured and informed method for token generation, potentially leading to improvements in the model's performance.

4.3.1 Skeleton-guided attention module

The skeleton-guided attention module is inspired by the self-guide attention module in SG-Former. The self-guided attention module could allocate the token according to the significance map, which aims to extract detailed features from salient regions by allocating more tokens, and cursory features from inconspicuous areas by allocating fewer tokens. This mechanism established a comprehensive patch token generation compared with the multi-scale grouping module in PointCAT.

In this self-guided attention module, token distribution is determined by a significance graph derived from self-attention computations. To incorporate skeleton information for token allocation, it's necessary to devise an algorithm capable of computing a skeleton-based significance map. Inspired by skeleton-based methods used in plant organ segmentation, as discussed in Section 1; those methods typically consider the skeleton graph as a connectivity graph, utilizing its junction vertices to guide segmentation. Following this logic, we could create a significance map that references these junction vertices, attributing higher significance to areas near the junctions and lower significance to those further away. Such a skeleton-based significance map would enable the skeleton-guided attention module to allocate tokens more effectively.

4.3.2 Semantic and instance segmentation

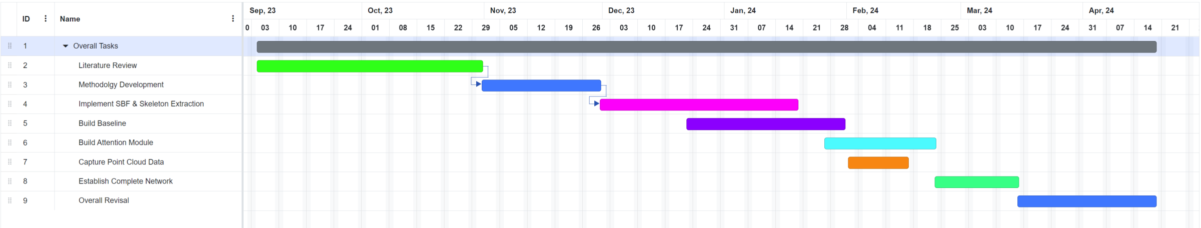
Kolodiazhnyi et al. (2023) proposed OneFormer3D, which demonstrated significant accomplishments in 3D semantic and instance segmentation with a Query Decoder. This is a Transformer decoder that could tackle both 3D semantic and instance segmentation tasks at the same time. Thus, we will directly combine its Query Decoder with our network to establish plant organ semantic/instance segmentation. Moreover, the settings for loss functions will be as same as those in OneFormer3D.

4.4 Evaluation metrics

In terms of semantic segmentation, we will calculate Precision, Recall, F1-score, Intersection over Union (IoU), and overall accuracy on each semantic class to evaluate the network's performance. For instance segmentation, we will employ mean precision (mPrec), mean recall (mRec), mean coverage (mCov), and the mean weighted coverage (mWCov).

5 Time planning

The following Gantt chart shows the initial schedule of this research.



6 Tools and datasets used

6.1 Programming language

C++ is renowned for its efficiency in executing computationally intensive tasks, making it an ideal choice for handling large-scale numerical computations, particularly in the domain of point cloud processing. The language's performance is further enhanced by robust libraries such as Eigen, Open3D, PCL, and OpenMP. These tools collectively facilitate efficient point cloud processing, including crucial operations like down-sampling and skeleton extraction. Accordingly, these aspects of our research will be implemented in C++.

With excellent deep learning libraries, like PyTorch and TensorFlow, and user-friendly syntax, Python has been widely used in the deep learning area. Therefore, the construction and optimization of our Transformer network, as well as the training and testing of the network, will be implemented in Python.

6.2 Platform

The network building, module designing, and initial training/testing will be conducted on a laptop with an i7-12700H CPU and NVIDIA GeForce RTX 3060 GPU (6G) under the Ubuntu operation system. For experiments in the final stage, like performance comparison, and ablation study, the High-Performance Computer, DelftBlue (Delft High Performance Computing Centre , DHPC), will also be utilized.

6.3 Datasets

As there is no official benchmark dataset for 3D plant organ segmentation, a hybrid dataset consisting of public, unpublic, and self-built datasets will be used for network training and testing. The components of this hybrid dataset are discussed as follows.

6.3.1 Pheno4D

Pheno4D, as introduced by Schunck et al. (2021), represents a significant advancement in the acquisition of plant point cloud data through the use of a laser triangulation scanner. The dataset encompasses detailed measurements of plant growth, featuring observations of 7 maize plants over 12 days and 7 tomato plants monitored over 20 days. This comprehensive data collection resulted in 84 point clouds for maize and 140 for tomato plants.

By using a laser triangulation scanner, Schunck et al. (2021) produced a high-quality plant point cloud dataset, namely Pheno4D. The authors measured 7 maize plants on 12 days, and 7 tomatoes measured on 20 days. This gives 84 maize and 140 tomato point clouds. However, not all of them were labeled by the authors, only 49 maize and 77 tomato point clouds were labeled. For all labeled point clouds, soil, stem, and leaves points were added semantic labels, and instance labels were assigned to all individual leaves.

6.3.2 ROSE-X

The ROSE-X (Dutagaci et al., 2020) dataset consists of 11 annotated 3D models of rosebush plants acquired through X-ray tomography. However, those plant models only contain semantic labels. Therefore, in the subsequent processing, the instance label will be added manually to every individual leaf by using the Semantic Segmentation Editor (SSE) tool `https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor/`.

6.3.3 Unpublic dataset

Conn et al. (2017) introduced Plant3D, a data repository dedicated to the storage of three-dimensional scans of plants. Currently, this database comprises 714 point clouds over four plant species (tomato, tobacco, sorghum, and Arabidopsis). However, a notable limitation of this dataset was the absence of labels, which restricted its usage in the deep learning domain.

To address this gap, Li et al. (2022b) made a significant contribution by manually annotating the point cloud in the Plant3D dataset with semantic and instance labels. Their efforts resulted in the labeling of a substantial portion of the Plant3D dataset, encompassing 105 tobacco, 312 tomato, and 129 sorghum specimens. However, the enhanced dataset with Li et al. (2022b) annotations was not publicly released. By communicating with the authors, we were granted access to this labeled dataset under specific terms and conditions. This means we are not permitted to distribute this dataset, either publicly or privately. Concerning that, to make our research result reproducible by any other researchers, this dataset will be limited to use in the preliminary training and testing stage, for parameter pruning, and in the final stage, only the public and self-built datasets will be utilized.

6.3.4 Self-built dataset

The current 3D plant data are still lacking, and those data are mainly limited to species like maize, tomato, and sorghum. To enrich and diversify the current 3D plant data, we decide to develop a self-built dataset. This work will focus on capturing the 3D point cloud data of *Polygonum lapathifolium* L., an annual herbaceous plant, by using SfM, or NeRFs-based method such as NeUDF (Liu et al., 2023).

7 Expected outcomes

When completing this research, we expect to have below outcomes:

1. A dataset that merges several public datasets and our self-built dataset will be public, to boost the development of 3D plant phenotyping.
2. An academic article is expected to be produced and ready to be published.
3. The complete code and trained weight will be public on GitHub.

References

S. Aich and I. Stavness. Leaf counting with deep convolutional and deconvolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.

O. K.-C. Au, C.-L. Tai, H.-K. Chu, D. Cohen-Or, and T.-Y. Lee. Skeleton extraction by mesh contraction. *ACM transactions on graphics (TOG)*, 27(3):1–10, 2008.

Y. Bao, L. Tang, S. Srinivasan, and P. S. Schnable. Field-based architectural traits characterisation of maize plant using time-of-flight 3d imaging. *Biosystems Engineering*, 178:86–101, 2019.

J. Cao, A. Tagliasacchi, M. Olson, H. Zhang, and Z. Su. Point cloud skeletons via laplacian based contraction. In *2010 Shape Modeling International Conference*, pages 187–197. IEEE, 2010.

C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

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

S. Das Choudhury, S. Bashyam, Y. Qiu, A. Samal, and T. Awada. Holistic and component plant phenotyping using temporal image sequence. *Plant methods*, 14(1):1–21, 2018.

S. Das Choudhury, A. Samal, and T. Awada. Leveraging image analysis for high-throughput plant phenotyping. *Frontiers in plant science*, 10:508, 2019.

J. De Vylder, F. Vandenbussche, Y. Hu, W. Philips, and D. Van Der Straeten. Rosette tracker: an open source image analysis tool for automatic quantification of genotype effects. *Plant physiology*, 160(3):1149–1159, 2012.

Delft High Performance Computing Centre (DHPC). DelftBlue Supercomputer (Phase 1). `https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase1`, 2022.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

R. Du, Z. Ma, P. Xie, Y. He, and H. Cen. Pst: Plant segmentation transformer for 3d point clouds of rapeseed plants at the podding stage. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:380–392, 2023.

S. Du, R. Lindenbergh, H. Ledoux, J. Stoter, and L. Nan. Adtree: Accurate, detailed, and automatic modelling of laser-scanned trees. *Remote Sensing*, 11(18):2074, 2019.

H. Dutagaci, P. Rasti, G. Galopin, and D. Rousseau. Rose-x: an annotated data set for evaluation of 3d plant organ segmentation methods. *Plant methods*, 16(1):1–14, 2020.

N. Engel, V. Belagiannis, and K. Dietmayer. Point transformer. *IEEE access*, 9:134826–134840, 2021.

L. Fan, Z. Pang, T. Zhang, Y.-X. Wang, H. Zhao, F. Wang, N. Wang, and Z. Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8458–8468, 2022.

M. Gaillard, C. Miao, J. Schnable, and B. Benes. Sorghum segmentation by skeleton extraction. In *European conference on computer vision*, pages 296–311. Springer, 2020.

W. Gélard, A. Herbulot, M. Devy, P. Debaeke, R. F. McCormick, S. K. Truong, and J. Mullet. Leaves segmentation in 3d point cloud. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18*, pages 664–674. Springer, 2017.

L. Gong, X. Du, K. Zhu, K. Lin, Q. Lou, Z. Yuan, G. Huang, and C. Liu. Panicle-3d: efficient phenotyping tool for precise semantic segmentation of rice panicle point cloud. *Plant Phenomics*, 2021.

M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021.

Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

A. Hartmann, T. Czauderna, R. Hoffmann, N. Stein, and F. Schreiber. Htpheno: an image analysis pipeline for high-throughput plant phenotyping. *BMC bioinformatics*, 12(1):1–9, 2011.

M. M. Hasan, J. P. Chopin, H. Laga, and S. J. Miklavcic. Detection and analysis of wheat spikes using convolutional neural networks. *Plant Methods*, 14:1–13, 2018.

T. Hu, N. Chitnis, D. Monos, and A. Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021.

H. Huang, S. Wu, D. Cohen-Or, M. Gong, H. Zhang, G. Li, and B. Chen. L1-medial skeleton of point cloud. *ACM Trans. Graph.*, 32(4):65–1, 2013.

S. Jay, G. Rabatel, X. Hadoux, D. Moura, and N. Gorretta. In-field crop row phenotyping from 3d modeling performed using structure from motion. *Computers and Electronics in Agriculture*, 110:70–77, 2015.

L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020.

A. Jignasu, E. Herron, T. Z. Jubery, J. Afful, A. Balu, B. Ganapathysubramanian, S. Sarkar, and A. Krishnamurthy. Plant geometry reconstruction from field data using neural radiance fields. In *2nd AAAI Workshop on AI for Agriculture and Food Systems*, 2023.

S. Jin, Y. Su, S. Gao, F. Wu, Q. Ma, K. Xu, T. Hu, J. Liu, S. Pang, H. Guan, et al. Separating the structural components of maize for field phenotyping using terrestrial lidar data and deep convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4): 2644–2658, 2019.

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

A. Junker, M. M. Muraya, K. Weigelt-Fischer, F. Arana-Ceballos, C. Klukas, A. E. Melchinger, R. C. Meyer, D. Riewe, and T. Altmann. Optimizing experimental procedures for quantitative evaluation of crop plant performance in high throughput phenotyping systems. *Frontiers in plant science*, 5:770, 2015.

K. Klasing, D. Althoff, D. Wollherr, and M. Buss. Comparison of surface normal estimation methods for range sensing applications. In *2009 IEEE international conference on robotics and automation*, pages 3206–3211. Ieee, 2009.

M. Kolodiazhnyi, A. Vorontsova, A. Konushin, and D. Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation, 2023.

B. Li and C. Guo. Maspc_transform: A plant point cloud segmentation network based on multi-head attention separation and position code. *Sensors*, 22(23):9225, 2022.

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

D. Li, J. Li, S. Xiang, and A. Pan. Psegnet: Simultaneous semantic and instance segmentation for point clouds of plants. *Plant Phenomics*, 2022a.

D. Li, G. Shi, J. Li, Y. Chen, S. Zhang, S. Xiang, and S. Jin. Plantnet: A dual-function point cloud segmentation network for multiple plant species. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:243–263, 2022b.

S. Li, L. Dai, H. Wang, Y. Wang, Z. He, and S. Lin. Estimating leaf area density of individual trees using the point cloud segmentation of terrestrial lidar data and a voxel-based model. *Remote sensing*, 9(11):1202, 2017.

Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

Y. Li, Y. Su, X. Zhao, M. Yang, T. Hu, J. Zhang, J. Liu, M. Liu, and Q. Guo. Retrieval of tree branch architecture attributes from terrestrial laser scan data using a laplacian algorithm. *Agricultural and Forest Meteorology*, 284:107874, 2020.

Y. Li, W. Wen, T. Miao, S. Wu, Z. Yu, X. Wang, X. Guo, and C. Zhao. Automatic organ-level point cloud segmentation of maize shoots by integrating high-throughput data acquisition and deep learning. *Computers and Electronics in Agriculture*, 193:106702, 2022c.

Y. Lin, Z. Ruifang, S. Pujuan, and W. Pengfei. Segmentation of crop organs through region growing in 3d space. In *2016 Fifth international conference on agro-geoinformatics (Agro-Geoinformatics)*, pages 1–6. IEEE, 2016.

Y.-T. Liu, L. Wang, J. Yang, W. Chen, X. Meng, B. Yang, and L. Gao. Neudf: Leaning neural unsigned distance fields with volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 237–247, 2023.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

D. Lu, Q. Xie, M. Wei, K. Gao, L. Xu, and J. Li. Transformers in 3d point clouds: A survey, 2022.

Z. Ma, D. Sun, and H. Cen. A novel skeletonization algorithm combined with hierarchical segmentation for phenotyping siliques of oilseed rape. In *2021 ASABE Annual International Virtual Meeting*, page 1. American Society of Agricultural and Biological Engineers, 2021.

Z. Ma, R. Du, J. Xie, D. Sun, H. Fang, L. Jiang, and H. Cen. Phenotyping of silique morphology in oilseed rape using skeletonization with hierarchical segmentation. *Plant Phenomics*, 5: 0027, 2023.

D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015.

M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr. Discrete differential-geometry operators for triangulated 2-manifolds. In *Visualization and mathematics III*, pages 35–57. Springer, 2003.

T. Miao, C. Zhu, T. Xu, T. Yang, N. Li, Y. Zhou, and H. Deng. Automatic stem-leaf segmentation of maize shoots using three-dimensional point cloud. *Computers and Electronics in Agriculture*, 187:106310, 2021.

R. R. Mir, M. Reynolds, F. Pinto, M. A. Khan, and M. A. Bhat. High-throughput phenotyping for crop improvement in the genomics era. *Plant Science*, 282:60–72, 2019.

S. Ninomiya, F. Baret, and Z.-M. M. Cheng. Plant phenomics: emerging transdisciplinary science. *Plant Phenomics*, 2019:2765120, 2019.

A. K. Patel, E.-S. Park, H. Lee, G. L. Priya, H. Kim, R. Joshi, M. A. A. Arief, M. S. Kim, I. Baek, and B.-K. Cho. Deep learning-based plant organ segmentation and phenotyping of sorghum plants using lidar point cloud. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

C. Peng, S. Li, Y. Miao, Z. Zhang, M. Zhang, and H. Li. Stem-leaf segmentation and phenotypic trait extraction of tomatoes using three-dimensional point cloud. *Transactions of the Chinese Society of Agricultural Engineering*, 38(9):187–194, 2022.

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

C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

S. Ren, X. Yang, S. Liu, and X. Wang. Sg-former: Self-guided transformer with evolving token reallocation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6003–6014, 2023.

D. Schunck, F. Magistri, R. A. Rosu, A. Cornelißen, N. Chebrolu, S. Paulus, J. Léon, S. Behnke, C. Stachniss, H. Kuhlmann, et al. Pheno4d: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *Plos one*, 16(8):e0256340, 2021.

W. Shi, R. van de Zedde, H. Jiang, and G. Kootstra. Plant-part segmentation using deep learning and multi-view vision. *Biosystems Engineering*, 187:81–95, 2019.

H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015.

Z. Su, S. Li, H. Liu, and Z. He. Tree skeleton extraction from laser scanned points. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 6091–6094. IEEE, 2019.

S. Sun, C. Li, P. W. Chee, A. H. Paterson, Y. Jiang, R. Xu, J. S. Robertson, J. Adhikari, and T. Shehzad. Three-dimensional photogrammetric mapping of cotton bolls in situ based on point cloud segmentation and clustering. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160:195–207, 2020.

Y. Sun, X. Guo, and H. Yang. Win-former: Window-based transformer for maize plant point cloud semantic segmentation. *Agronomy*, 13(11):2723, 2023.

F. Tardieu, L. Cabrera-Bosquet, T. Pridmore, and M. Bennett. Plant phenomics, from sensors to knowledge. *Current Biology*, 27(15):R770–R783, 2017.

S. Thapa, F. Zhu, H. Walia, H. Yu, and Y. Ge. A novel lidar-based instrument for high-throughput, 3d measurement of morphological traits in maize and sorghum. *Sensors*, 18 (4):1187, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

D. Wang, X. Liang, G. I. Mofack, and O. Martin-Ducup. Individual tree extraction from terrestrial laser scanning data via graph pathing. *Forest Ecosystems*, 8:1–11, 2021.

M. Watt, F. Fiorani, B. Usadel, U. Rascher, O. Muller, and U. Schurr. Phenotyping: new windows into the plant for breeders. *Annual review of plant biology*, 71:689–712, 2020.

S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

S. Wu, W. Wen, B. Xiao, X. Guo, J. Du, C. Wang, and Y. Wang. An accurate skeleton extraction approach from 3d point clouds of maize plants. *Frontiers in plant science*, 10:248, 2019.

Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

L. Xiang, Y. Bao, L. Tang, D. Ortiz, and M. G. Salas-Fernandez. Automated morphological traits extraction for sorghum plants via 3d point cloud data analysis. *Computers and Electronics in Agriculture*, 162:951–961, 2019.

Y. Xiao, H. Liu, L. Wu, M. Warburton, and J. Yan. Genome-wide association studies in maize: praise and stargaze. *Molecular plant*, 10(3):359–374, 2017.

M. Xu, R. Ding, H. Zhao, and X. Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021.

Q. Xu, L. Cao, L. Xue, B. Chen, F. An, and T. Yun. Extraction of leaf biophysical attributes based on a computer graphic-based algorithm using terrestrial laser scanning data. *Remote Sensing*, 11(1):15, 2018.

W. Yang, H. Feng, X. Zhang, J. Zhang, J. H. Doonan, W. D. Batchelor, L. Xiong, and J. Yan. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Molecular Plant*, 13(2):187–214, 2020.

X. Yang, M. Jin, W. He, and Q. Chen. Pointcat: Cross-attention transformer for point cloud, 2023.

X. Yin, X. Liu, J. Chen, and D. M. Kramer. Joint multi-leaf segmentation, alignment, and tracking for fluorescence plant videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1411–1423, 2017.

K. Zhang, M. Hao, J. Wang, X. Chen, Y. Leng, C. W. de Silva, and C. Fu. Linked dynamic graph cnn: Learning through point cloud by linking hierarchical features. In *2021 27th international conference on mechatronics and machine vision in practice (M2VIP)*, pages 7–12. IEEE, 2021.

X. Zhang, C. Huang, D. Wu, F. Qiao, W. Li, L. Duan, K. Wang, Y. Xiao, G. Chen, Q. Liu, et al. High-throughput phenotyping and qtl mapping reveals the genetic architecture of maize plant growth. *Plant physiology*, 173(3):1554–1564, 2017.

Y. Zhang, W. Su, W. Tao, Z. Li, X. Huang, Z. Zhang, and C. Xiong. Completing 3d point clouds of thin corn leaves for phenotyping using 3d gridding convolutional neural networks. *Remote Sensing*, 15(22):5289, 2023.

H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.

Y. Zhou, P. Sun, Y. Zhang, D. Anguelov, J. Gao, T. Ouyang, J. Guo, J. Ngiam, and V. Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, pages 923–932. PMLR, 2020.