ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-1/W2-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

Graph Self-Attention Network with Semantic Embedding for Stem-Leaf Separation from 3D Point Clouds

Anhao Yang[1], Haiyang Wu[1], Juntao Yang[1,*], Zhenhai Li[1], Bo Bai[2], Guowei Li[2]
[1]College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China
[2]Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan 250100, China

Keywords: 3D plant point clouds, Stem–leaf segmentation, Graph Self-Attention Network, Semantic-Guided Learning

Abstract:

In the context of agricultural modernization, precise 3D organ segmentation has become indispensable for automated extraction of phenotypic traits. In particular, the precise delineation of stem and leaf structures from 3D point clouds is critical for monitoring plant growth and supporting high-throughput breeding programs. However, the intricate structure of crops and the blurred boundaries between stems and leaves present significant challenges, leading to the poor segmentation performance. To tackle these problems, we propose a Semantic Embedding-Guided Graph Self-Attention Network for stem-leaf separation in 3D point clouds, to tackle weak feature representation and low inter-class separability in complex plant structures. During the encoding stage, a multi-scale feature extraction module captures fine-grained local geometries, while a feature fusion module integrating graph convolution and self-attention facilitates deep fusion of local and global semantic information. In the decoding stage, hierarchical upsampling combined with multi-level feature fusion reconstructs high-resolution representations to achieve fine-grained segmentation. Furthermore, we introduce a joint loss function that integrates inter-class discriminative loss with cross-entropy, aiming to optimize intra-class uniformity and reinforce class boundary delineation. Validation experiments on the Plant-3D dataset demonstrate that our methodology attains superior performance, with mean precision, recall, and IoU achieving 96.47%, 96.39%, and 93.50%, respectively. The proposed approach demonstrates high robustness and generalizability across diverse plant species and growth stages, providing an effective solution for high-throughput plant phenotyping.

1. Introduction

Plant phenotyping aims to measure and analyze structural traits, growth conditions, and physiological–biochemical properties of plants to uncover the interaction mechanisms among genotype, environment, and phenotype, thereby providing a scientific basis for crop growth regulation, quality improvement, and yield prediction (Watt et al., 2020; Tester and Langridge, 2010). In recent years, driven by advances in agricultural digitalization and precision farming, high-throughput and automated phenotyping has become a key research focus at the convergence of agriculture, biology, and computer vision (Singh et al., 2018). Accurate and efficient acquisition of the three-dimensional morphology and spatial distribution of plant organs (e.g., stems and leaves) is crucial for advancing precise crop monitoring, breeding optimization, and intelligent agronomic management.

Conventional manual measurements suffer from high destructiveness, low efficiency, and restricted accuracy (Furbank and Tester, 2011). Two-dimensional imaging approaches are vulnerable to lighting variations, occlusions, and background noise, hindering accurate reconstruction of plants' intricate spatial architecture. In contrast, three-dimensional imaging techniques—including LiDAR, structured light, and time-of-flight (ToF) cameras—offer richer depth information and spatial detail, garnering considerable attention in plant phenotyping (Rajapaksha et al., 2024). Nevertheless, the diverse morphologies of plant organs and severe stem–leaf occlusions present formidable obstacles to accurate stem–leaf segmentation, thereby constraining the broader adoption of precision agriculture (Weyler et al., 2024). Consequently, enhancing the accuracy, robustness, and generalizability of stem–leaf segmentation in complex scenarios has become a critical challenge.

In recent years, deep learning-based point cloud segmentation methods have been widely employed for plant organ segmentation owing to their strong feature learning capabilities and high computational efficiency. Existing methods are typically classified into three categories: voxel-based, projection-based, and point-based approaches. Voxel-based approaches encode point clouds into structured 3D voxel grids and utilize 3D convolutional neural networks to extract features(Du et al., 2023); whereas projection-based methods segment projected 2D images using 2D convolutional neural networks (2D CNNs). Although both approaches benefit from well-established network architectures, they often suffer from significant loss of spatial and geometric information during transformation, thereby limiting their ability to accurately represent complex plant structures (Bai et al., 2023). In contrast, point-based methods extract features directly from raw point clouds, thus avoiding the spatial information loss associated with voxelization or projection and providing enhanced adaptability and representational capacity for complex plant architectures (Qi et al., 2017). DGCNN (Wang et al., 2019) enhances point-wise feature interactions through the construction of dynamic adjacency graphs, thereby enabling effective capture of local geometric relationships in complex regions, such as stem-leaf junctions; KPConv employs deformable convolutional kernels to enhance feature extraction from point clouds exhibiting non-uniform densities (Thomas et al., 2019); PointTransformer integrates self-attention to dynamically model semantic relationships among distant points, thereby facilitating improved global representation in occluded or structurally complex regions (Zhao et al., 2021); RandLA-Net (Hu et al., 2020) adopts efficient random sampling and feature aggregation strategies to preserve high segmentation accuracy while reducing computational costs, making it suitable for large-scale point cloud processing. In addition, methods such as Graph Convolutional Networks (GCNs), Recurrent Neural Networks (RNNs), and Conditional Random Fields (CRFs) have also been widely adopted for fine-grained semantic segmentation of point

* Correspondence author. E-mail: jtyang@sdust.edu.cn

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W2-2025-239-2025 | © Author(s) 2025. CC BY 4.0 License.

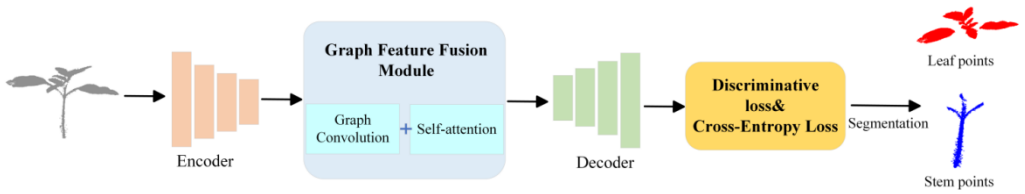


Figure 1. Illustrates the overall workflow of the proposed approach.

clouds, as they model neighborhood structures and spatial dependencies to further enhance representational capability and segmentation performance.

Although existing methods have achieved notable progress in plant stem-leaf separation tasks (Yang et al., 2024b) and demonstrated satisfactory segmentation quality, they still face numerous challenges in real-world applications. The intrinsic characteristics of point cloud data—such as sparsity, noise, and structural complexity—pose significant challenges for existing methods in accurately capturing both local and global features and distinguishing between structurally similar plant organs (Yang et al., 2024a). Moreover, current models exhibit limited generalization across different crop varieties, growth stages, and morphological types. This limitation becomes particularly evident when handling multi-layered stem-leaf structures or heavily occluded scenes, where segmentation accuracy degrades significantly. How to further improve segmentation accuracy while enhancing model robustness and generalizability in complex scenarios remains a key challenge in plant structural point cloud segmentation. Therefore, there is a pressing need to develop point cloud segmentation models that are more robust, efficient, and generalizable, capable of handling intricate plant architectures and supporting high-throughput phenotypic analysis.

To address these challenges, we propose a semantic embedding–guided graph self-attention network for the separation of stems and leaves in 3D plant point clouds. Our primary research contributions include:
(1) A graph attention-based feature enhancement module is proposed to improve the expressiveness and discriminative capacity of point cloud features. This module utilizes a dynamic graph convolutional network to capture local geometric structures, while the self-attention derives attention weights to enhance the modeling of long-range contextual relationships.
(2) A multi-task collaborative optimization strategy is devised, integrating cross-entropy loss with an inter-class discriminative loss to optimize the distribution of high-dimensional features. Classification objectives focus on performance improvement, whereas discriminative objectives promote intra-category clustering and inter-category separation.
(3) Comprehensive experimental validation is performed using the Plant-3D dataset, demonstrating our approach outperforms current leading methods in segmentation precision and robustness (Conn et al., 2017).

2. Methodology

To address the critical challenges associated with stem–leaf segmentation in 3D plant point clouds, we present a semantic embedding-guided encoder–decoder framework, as depicted in Figure 1. The proposed network comprises four principal modules: (1) an encoder that performs multi-scale geometric feature extraction by progressively enlarging the receptive field to capture detailed local structural information; (2) a graph-based feature enhancement module that combines graph convolution with self-attention mechanisms to effectively integrate local geometric relationships and global contextual cues; (3) a decoder that employs hierarchical upsampling and skip connections to incrementally reconstruct high-resolution features while preserving critical semantic cues from the encoding phase; (4) a multi-task optimization module that jointly applies inter-class discriminative and cross-entropy losses to refine the feature space, promoting intra-class compactness and inter-class separability, thereby enhancing point-wise classification accuracy and enabling precise stem–leaf segmentation.

2.1 Multi-scale coding and feature extraction

In 3D point cloud analysis, the non-uniform point distribution, varied local geometries, and intricate global structures pose considerable challenges for effective multi-scale feature extraction. To overcome these difficulties, we propose a multi-scale feature extraction approach that combines downsampling with self-attention mechanisms, aiming to maximally preserve local spatial structures while capturing rich geometric information.

During the feature extraction stage, a farthest point sampling strategy is employed to construct point cloud subsets, thereby reducing data redundancy and mitigating sampling bias. Subsequently, for each sampled point p_i, a local neighborhood $\mathcal{N}(p_i)$ is then established using K-nearest neighbor (K-NN) search, and both spatial coordinates and point-wise features are extracted to encode detailed local geometric information. Specifically, the feature representation of each sampled center point p_i is defined as:

$$\boldsymbol{X_i} = \mathrm{Concat}[p_i, p_i^k, ||p_i - p_i^k||, (p_i - p_i^k)] \tag{1}$$

Where p_i and p_i^k denote the 3D coordinates of the center point and its k neighboring, and $||\cdot||$ denotes the Euclidean distance. A multi-layer perceptron (MLP) followed by the Softmax function is used to compute attention weights α_{ij}, facilitating dynamic aggregation of neighborhood features:

$$\begin{aligned} \hat{\boldsymbol{X}}_i &= \sum_{j \in N(i)} \alpha_{ij} \cdot \boldsymbol{X}_{ij} \\ \alpha_{ij} &= \mathrm{Softmax}\left(\mathrm{MLP}(\boldsymbol{X}_{ij})\right) \end{aligned} \tag{2}$$

where $\boldsymbol{X}_{ij}$ represents the feature difference between point p_i and its neighbor p_i^k. This operation enhances the representation of critical local structures while mitigating the influence of weak or noisy points.

To alleviate the degradation and feature attenuation caused by increasing network depth, residual connections are introduced into each encoding layer, as defined as follows.

$$\boldsymbol{X}_t = f(\boldsymbol{X}_t) + g(\boldsymbol{X}_{t-1}) \tag{3}$$

Where $f(\cdot)$ denotes the nonlinear transformation at the current encoder layer, and $g(\cdot)$ is a 1×1 convolution used to preserve feature dimensionality and enhance feature stability.

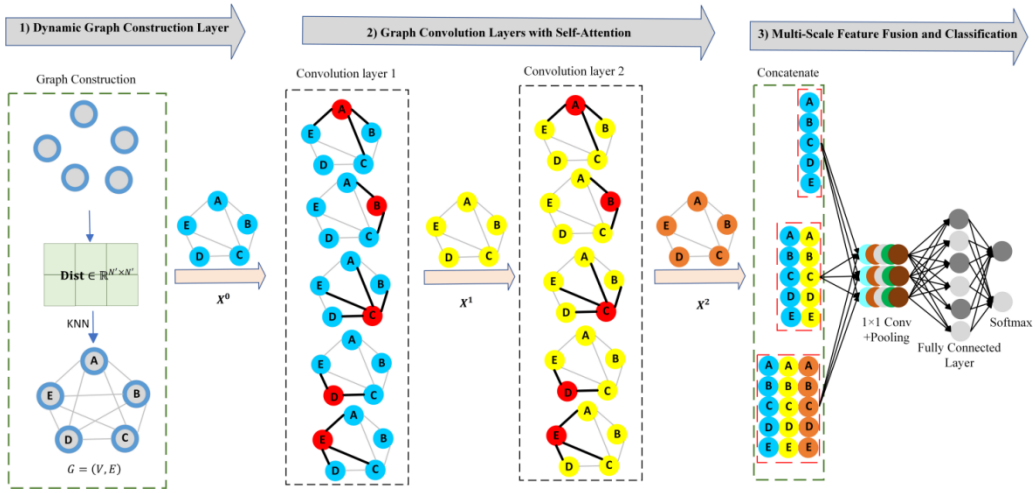


Figure 2. Graph Attention-based Feature Aggregation Module

To simultaneously capture fine-grained local details and global semantic context, we incorporate multi-scale feature fusion following downsampling, using cross-layer residual connections. Specifically, after obtaining the output of the *t*-th layer, features sampled from the preceding layer are concatenated with the encoded features of the current layer, as expressed below:

$$\boldsymbol{X}_t = [\mathrm{DS}(\boldsymbol{X}_{t-1}), \mathrm{Encode}(\boldsymbol{X}_t)] \tag{4}$$

Where $\mathrm{DS}(\cdot)$ denotes the downsampling operation, while $\mathrm{Encode}(\boldsymbol{X}_t)$ represents the feature mapping function at the current layer. This design enables deep networks to retain low-level local geometric information while fully leveraging high-level semantic features, thereby enhancing the overall precision and robustness of feature representation.

2.2 Graph Attention Based Feature Aggregation and Optimization

Owing to the irregular nature and complex spatial structure of point cloud data, conventional global feature representations—such as the encoded feature $\boldsymbol{X}$—are often inadequate for capturing fine-grained category boundary information. This limitation undermines both the overall shape representation and the accurate delineation of complex category boundaries. To address this, and inspired by prior work (Tian and Li, 2022), A graph attention-based feature aggregation module is incorporated prior to the decoding stage, aiming to comprehensively capture correlations between local geometric details and global contextual information. The detailed processing pipeline of this module is illustrated in Figure 2.

2.2.1 Graph Structure: This module transforms the visual features output by the encoder into a graph representation $G = (\boldsymbol{V}, \boldsymbol{E})$, where the node set $\boldsymbol{V} = \{p_1, p_2, \cdots, p_{N'}\}$ represents sampled points, with each point p_i corresponding to a graph node. To accurately capture point-wise spatial topology, we first compute the Euclidean distance matrix $\boldsymbol{Dist} \in \mathbb{R}^{N' \times N'}$, and apply the KNN algorithm to determine the local neighborhood $\mathcal{N}(p_i)$ of each node p_i. The edge set $\boldsymbol{E}$ is then constructed as:

$$\boldsymbol{E} = \{(p_i, p_j) | p_j \epsilon \mathcal{N}(p_i)\} \tag{5}$$

To effectively encode local geometric structures and feature discrepancies in the point cloud, each edge feature $\boldsymbol{e}_{ij} = (p_i, p_j)$ is defined by concatenating the spatial coordinate difference and feature difference between adjacent nodes:

$$\boldsymbol{e}_{ij} = \mathrm{concat}(p_j - p_i, \boldsymbol{x}_j - \boldsymbol{x}_i) \tag{6}$$

where, p_i and p_j denote 3D spatial coordinates, while $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent the corresponding downsampled point features. This design significantly enhances the descriptive power of edge features in representing local geometry and provides informative cues for subsequent attention-based aggregation.

2.2.2 Self-Attention: After constructing the graph structure, a self-attention mechanism is introduced to iteratively update node features through adaptive aggregation of contextual information from local neighborhoods. Specifically, this mechanism learns attention weights between each node and its neighbors, enabling dynamic acquisition of informative context to facilitate effective local-global feature interaction. The node feature update is formally defined as:

$$\boldsymbol{x}_i^{(k+1)} = \max_{(i,j)\in\varepsilon} h_\theta(\mathrm{concat}[\boldsymbol{x}_i^{(k)}, \boldsymbol{x}_{ij}^{(k)}]) \tag{7}$$

Where, $\boldsymbol{x}_i^{(k)}$ represents the feature associated with node i upon completion of the k-th iteration, $\boldsymbol{x}_{ij}^{(k)}$ is the corresponding edge feature, h_θ is a mapping function comprising a linear transformation, instance normalization, and a LeakyReLU non-linear activation, while $\max(\cdot)$ indicates a max-pooling aggregation operation within the local neighborhood. By stacking multiple layers of this architecture, the model progressively refines point representations to capture fine-grained local structures while simultaneously incorporating global spatial context.

2.2.3 Cross-scale feature fusion: To improve the expressiveness of node features, this work introduces a cross-scale feature fusion mechanism. Specifically, we concatenate multiple intermediate features (e.g., $\boldsymbol{x}_i^{(1)}$, $\boldsymbol{x}_i^{(2)}$) obtained from iterative graph attention aggregation with the initial features (e.g., $\boldsymbol{x}_i^{(0)}$). The concatenated features are then projected to a specified dimension through a learnable linear mapping (e.g., an MLP or linear layer), producing the final graph attention-enhanced feature:

$$\boldsymbol{x}_i^{\mathrm{GNN}} = h_\theta(\mathrm{concat}[\boldsymbol{x}_i^{(0)}, \boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}]) \tag{8}$$

This operation not only effectively integrates fine-grained local structures and overarching semantic information across different

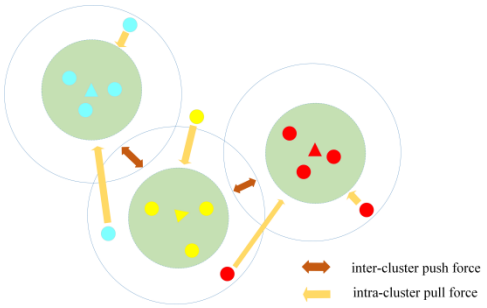


Figure 3. Illustration of the discriminative loss with intra-cluster pull and inter-cluster push forces.

feature levels but also substantially improves the capacity of node features to distinguish complex geometric structures. Overall, the proposed graph attention-based feature aggregation module more efficiently captures local geometric details and global contextual relationships in 3D point cloud data, thereby exhibiting improved adaptability and robustness in tasks such as plant stem–leaf segmentation.

2.3 Multi-scale decoding and feature reconstruction

In the decoding phase, to efficiently recovers fine-grained geometric details as well as global spatial semantics lost during downsampling, we propose a hierarchical feature reconstruction and fusion decoding module that incrementally restores the point cloud's spatial resolution while ensuring consistency across multiple scales. Specifically, the module initially employs a K-nearest neighbor interpolation strategy to project low-resolution features onto their corresponding high-resolution coordinates, thereby mitigating feature deviation during interpolation and preserving spatial continuity and consistency. Let the decoding layer input features be $\boldsymbol{X}^{\mathrm{GNN}}$. The upsampling process is formulated as:

$$\boldsymbol{X}_{\mathrm{up}}(p_i) = \boldsymbol{X}^{\mathrm{GNN}}(p_{j^*}), \quad j^* = \underset{j}{\operatorname{argmin}}\, d(p_i, p_j) \tag{9}$$

Where p_i and p_j denote high- and low-resolution points, respectively, and $d(\cdot)$ indicates the Euclidean distance measure. In view of potential attenuation of local geometric features during interpolation, we further introduce a multi-scale feature fusion strategy combined with an attention mechanism to weight cross-scale information. Specifically, at the decoding stage, we fuse the current layer's features $\boldsymbol{X}_{\mathrm{de}}^{t}$ with the previous layer's features $\boldsymbol{X}_{\mathrm{de}}^{t-1}$. An attention-based pooling strategy is adopted to dynamically allocate fusion weights across multi-scale features, enabling effective cross-scale interaction and feature optimization. Furthermore, to strengthen the spatial detail awareness and semantic expressiveness of the features, the decoding module employs a skip-connection architecture to directly transfer high-resolution features from the matching encoder layer to the corresponding decoder layer. A multi-layer perceptron (MLP) subsequently learns the fusion weights, fully exploiting the complementarity and structural consistency of features across different stages. With these designs in place, the proposed decoding module effectively restores the spatial resolution and fine-grained structure of the point cloud, enhancing detail integrity and semantic consistency during feature reconstruction. Consequently, it offers more stable and accurate feature support for subsequent point-level classification.

2.4 Loss Function

The formulation of a loss function directly impacts the efficiency of network optimization and the overall performance. To address this, we propose an optimization strategy that integrates cross-entropy loss with a Semantic-aware discriminative loss, thereby jointly enhancing classification accuracy and the discriminative capacity of the feature space.

2.4.1 Cross-entropy: The cross-entropy loss measures the difference between the network's predicted outputs and the true labels. It is formulated as:

$$\mathcal{L}_{\mathrm{ce}} = -\sum_{n=1}^{C} t_n \log(\hat{p}_n) + (1 - t_n)\log(1 - \hat{p}_n) \tag{10}$$

where, C denotes the total number of categories, t_n represents the ground-truth label, and $\hat{p}_n$ is the predicted probability from the model. Minimizing this difference via backpropagation enhances the classification performance.

2.4.2 Semantic-aware discriminative loss: Even though semantic and instance segmentation tasks in 3D point clouds often treat category recognition and instance separation as separate objectives, they are intrinsically related: instance-level features can guide category identification, whereas intra-category features may vary significantly across different instances. Accordingly, we draw on inter-class separability to propose an inter-class feature discrimination loss $\mathcal{L}_{\mathrm{DL}}$, which unifies semantic and instance segmentation within a single embedding space to improve discriminative power. As illustrated in Figure 3, the proposed loss enforces feature compactness within each class and repulsion between different classes, thereby improving the separability of semantic categories.

This embedding loss comprises three components—an intra-class pull term, an inter-class push term, and a feature-space regularization term—formally expressed as:

$$\mathcal{L}_{\mathrm{DL}} = \alpha \cdot \mathcal{L}_{\mathrm{pull}} + \beta \cdot \mathcal{L}_{\mathrm{push}} + \gamma \cdot \mathcal{L}_{\mathrm{reg}} \tag{11}$$

where the intra-class pull term $\mathcal{L}_{\mathrm{pull}}$ clusters features belonging to the same instance:

$$\mathcal{L}_{\mathrm{pull}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{j=1}^{N_k}[||\mu_k - e_j|| - \delta_v]_+^2 \tag{12}$$

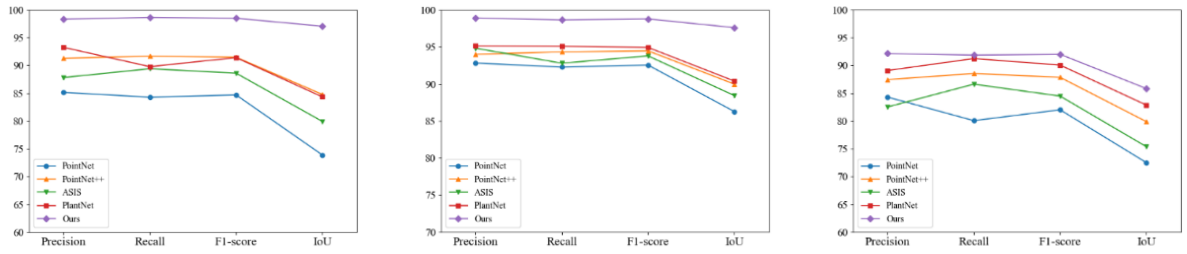
The inter-class push term $\mathcal{L}_{\mathrm{push}}$ boosts inter-category separability:

$$\mathcal{L}_{\mathrm{push}} = \frac{1}{K(K-1)}\sum_{k=1}^{K}\sum_{m=1,m\neq k}^{K}[2\delta_d - ||\mu_k - \mu_m||_2]_+^2 \tag{13}$$

The feature regularization term $\mathcal{L}_{\mathrm{reg}}$ stabilizes the overall feature distribution:

$$\mathcal{L}_{\mathrm{reg}} = \frac{1}{K}\sum_{k=1}^{K}||\mu_k||_2 \tag{14}$$

Where, $[x]_+ = \max(0, x)$. The parameters δ_v and δ_d denote the thresholds for intra-class compactness and inter-class separation, respectively, μ_k denotes the centroid of the k-th instance's features, and e_j corresponds to the embedding associated with that instance. We set $\alpha = \beta = 1$ and $\gamma = 0.001$.



(a)Tobacco (b)Tomato (c)Sorghum

Figure 4. Performance comparison of different methods across four evaluation metrics in stem-leaf separation.

By jointly leveraging these terms, the model significantly boosts feature discriminability across categories, improving performance on point cloud segmentation tasks.
Finally, we define the overall loss function as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{DL}} \tag{15}$$

Where λ represents a weighting factor that controls the balance between the cross-entropy loss and the inter-class feature discrimination loss.

3. Experimentation

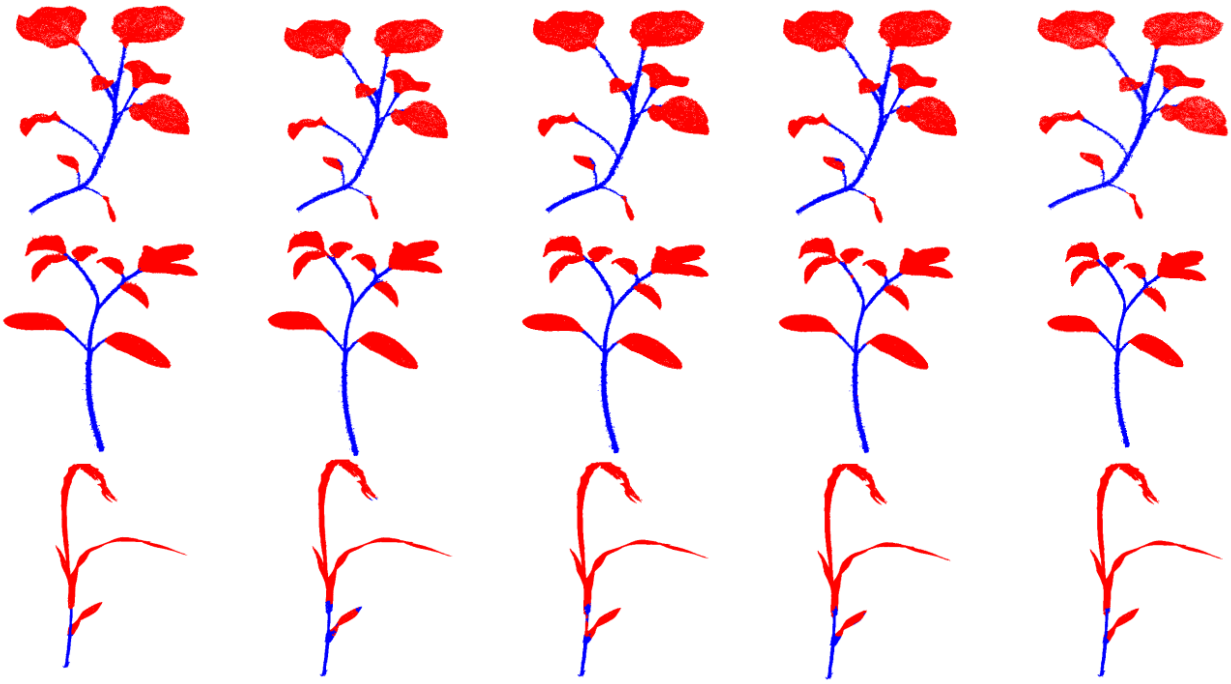
High-quality, meticulously annotated, and structurally preserved point cloud datasets form the foundation for advancing deep learning–driven 3D plant segmentation. Ideally, these datasets should exhibit comprehensive plant structures, high sampling resolution, broad species diversity, and multiple growth stages. In the present work, we employed the Plant-3D dataset released by Conn et al. (2017), which comprises 558 individual plant samples of tobacco, tomato, and sorghum, captured using a non-contact 3D laser scanner. The dataset spans approximately 20 growth stages and encompasses a wide range of environmental conditions, including natural lighting, shadows, high temperature, strong light, and drought. While ensuring structural integrity, it also reflects rich morphological and environmental diversity, thus providing a reliable basis for developing and evaluating plant point cloud segmentation methods.

3.1 Evaluation indicators

To comprehensively evaluate model performance, four widely used metrics are employed: precision, recall, F1-score, and IoU (Yang et al., 2025). Precision measures the fraction of points correctly predicted among all points assigned to a particular class. Recall measures the fraction of ground-truth points that are correctly identified for a given class. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced assessment of overall model performance. IoU measures the spatial overlap between predicted regions and the corresponding ground-truth areas. In addition, to assess model performance across all semantic categories, the mean intersection over union (mIoU) is calculated to provide a comprehensive evaluation of semantic segmentation accuracy.

3.2 Quantitative analysis

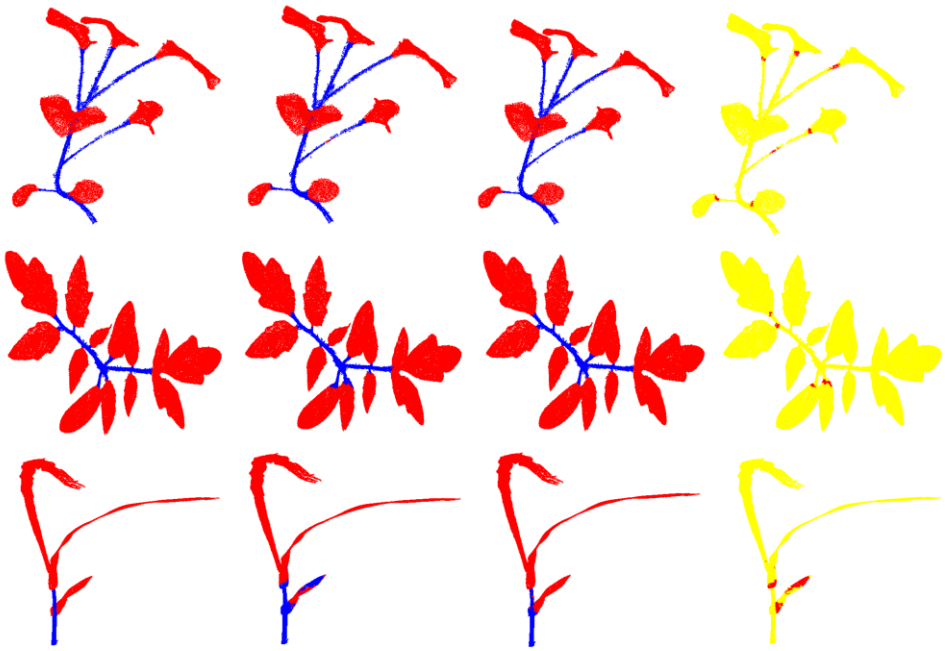
The results of the model's semantic segmentation for three plant species, tobacco, tomato, and sorghum, are presented in Table 1. Overall, tomato achieved the highest segmentation performance,



(a) Ground Truth (b) Initial (c) Embed (d) GCN (e) E+G

Figure 5. Qualitative comparison of stem-leaf separation results for three plant species (red represents leaves, blue represents stems). Embed denotes the semantic-aware discriminative loss module, GCN denotes the GCN self-attention module, and E+G denotes the collaboration of both modules.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W2-2025-239-2025 | © Author(s) 2025. CC BY 4.0 License.



(a) Ground Truth (b) Baseline (c) Ours (d) Improved Area

Figure 6. Comparison of segmentation results and improvement effects on the Plant-3D dataset. The red areas in the subfigure (d) highlight the regions where our proposed method outperforms the baseline.

particularly for the leaf class, where all evaluation metrics exceeded 98%, while the IoU for stems remained comparatively lower. This superior performance is primarily attributable to the larger sample size and broader coverage of growth stages, which facilitated the learning of more comprehensive and generalizable feature representations. In contrast, sorghum exhibited relatively lower segmentation accuracy, especially in stem regions where complex structures and tightly connected leaf sheaths posed significant challenges. Notably, across all plant species, the model consistently underperformed in segmenting stems relative to leaves, as reflected by similar trends across all evaluation metrics. This phenomenon is likely caused by class imbalance in the dataset, where the overwhelming number of leaf points compared to stem points biases the model towards learning leaf-specific features during training. To improve stem structure recognition, future work could explore optimized data sampling strategies or incorporate class-level augmentation techniques to mitigate performance gaps arising from class imbalance.

To further validate the effectiveness of the proposed method for stem–leaf segmentation in plant point clouds, we compared it with mainstream methods such as PointNet, PointNet++, ASIS and PlantNet, and systematically evaluates it on three crop datasets, namely tobacco, tomato and sorghum. Figure 4 demonstrates the quantitative performance of each method under four metrics: Precision, Recall, F1-score and IoU. The results indicate that the proposed method surpasses the comparison approaches across all evaluation metrics, with particularly notable improvements observed on the structurally complex sorghum dataset. Meanwhile, the proposed method also demonstrates good modeling ability for fine-grained structures and boundary regions in tobacco and tomato data, especially in Precision and IoU metrics, which fully verifies its robustness and generalization ability under diverse plant structures.

Metrics		Tobacco	Tomato	Sorghum	Mean
Precision (%)	Stem	97.13	98.50	85.96	93.86
	Leaf	99.55	99.34	98.33	99.07
Recall (%)	Stem	97.94	97.72	85.33	93.66
	Leaf	99.36	99.57	98.41	99.11
F1-score (%)	Stem	97.54	98.11	85.64	93.76
	Leaf	99.46	99.46	98.37	99.10
IoU (%)	Stem	95.19	96.29	74.89	88.79
	Leaf	98.92	98.92	96.78	98.21

Table 1. Quantitative results of the network for plant semantic segmentation

3.3 Qualitative analysis

To systematically assess the model's segmentation capability across different crop types and developmental stages, a systematic qualitative visualization was performed on tobacco, tomato, and sorghum. As shown in Figure 5, the semantic segmentation outcomes for these three crops are illustrated under complex structural patterns and diverse environmental conditions. Each row of the figure represents a plant sample, with columns showing the ground truth, the initial output (Initial), the result after introducing the discriminative loss module (Embed), the output with the graph attention module (GCN), and the final output combining both modules (E+G). The visualizations indicate that the baseline network exhibits clear limitations when processing structurally complex plant regions. For example, in tobacco, bifurcation regions between stems and leaves are poorly delineated; tomato samples reveal blurred boundaries and uncertain class transitions within leaf structures; and sorghum often shows misclassification in the tightly connected stem–sheath areas. With the progressive introduction of the Embed and GCN modules, the model demonstrates substantially improved spatial awareness and semantic discrimination in 3D point clouds. Organ boundaries become sharper, and the separation between stems and leaves is considerably enhanced, with the final predictions closely matching the ground truth annotations. The findings demonstrate that the proposed framework is both robust and effective in segmenting complex plant architectures from 3D point clouds, laying a strong foundation for subsequent analyses of detailed phenotypic traits.

Metrics	Embed	GCN	Tobacco			Tomato			Sorghum		
			Stem	Leaf	Mean	Stem	Leaf	Mean	Stem	Leaf	Mean
Precision (%)	×	×	96.31	99.32	97.82	95.69	99.46	97.58	76.73	97.71	87.22
	√	×	95.84	99.45	97.65	97.18	99.36	98.27	**87.17**	96.21	91.69
	×	√	96.47	99.44	97.96	**98.73**	98.61	98.67	73.25	**99.08**	86.17
	√	√	**97.13**	**99.55**	**98.34**	98.50	**99.34**	**98.92**	85.96	98.33	**92.15**
Recall (%)	×	×	96.76	99.22	97.99	**98.17**	98.69	98.43	82.06	96.85	89.46
	√	×	97.39	99.11	98.25	97.85	99.16	98.51	70.78	**98.98**	84.88
	×	√	97.37	99.25	98.31	95.23	99.54	97.30	**91.96**	96.26	**94.11**
	√	√	**97.94**	**99.36**	**98.65**	97.72	**99.57**	**98.65**	85.33	98.41	91.87
F1-score (%)	×	×	96.54	99.27	97.91	96.91	99.07	97.99	79.31	97.28	88.30
	√	×	96.61	99.28	97.95	97.52	99.26	98.39	78.68	97.89	88.29
	×	√	96.92	99.35	98.14	96.95	99.13	98.04	81.54	97.65	89.60
	√	√	**97.54**	**99.46**	**98.50**	**98.11**	**99.46**	**98.79**	**85.64**	**98.37**	**92.01**
IoU (%)	×	×	93.31	98.55	95.93	94.01	98.16	96.09	65.71	94.71	80.21
	√	×	93.43	98.56	96.00	95.15	98.54	96.85	64.86	95.86	80.36
	×	√	94.02	98.7	96.36	94.08	98.27	96.18	68.84	95.4	82.12
	√	√	**95.19**	**98.92**	**97.06**	**96.29**	**98.92**	**97.61**	**74.89**	**96.78**	**85.84**

Table 2. Ablation study of different modules on the Plant-3D dataset. The best performance for each metric is highlighted in BOLD fonts.

To further assess the effectiveness of the proposed method, we performed a comparative evaluation against the baseline model on tobacco, tomato, and sorghum, as illustrated in Figure 6. Subfigure (a) displays the ground truth annotations; (b) and (c) depict the segmentation outputs of the baseline and proposed models, respectively; (d) highlights the regions where our model outperforms the baseline. The results indicate that the baseline model suffers from evident mis-segmentation at blade edges and in structurally complex regions. In contrast, the proposed approach exhibits stronger capability in capturing geometric structures under challenging scenarios, such as overlapping or bending blades and stem–leaf junctions, thereby correcting segmentation errors of the baseline model and markedly improving local segmentation accuracy.

4. Ablation Study

To evaluate the individual contributions of each component to overall performance, we performed ablation experiments on the Plant-3D dataset (Table 2), focusing on the discriminative module (Embed) and the graph-based self-attention module (GCN). The results indicate that the integration of these two modules yields performance gains of 2.26%, 1.1%, 1.7%, and 2.76% in precision, recall, F1-score, and IoU, respectively. Specifically, the Embed module improves class separability and intra-class compactness within the feature space; its removal causes a marked decline in recall and F1-score. For crops with complex structural morphology, including tomato and sorghum, the average recall declined to 97.30% and F1-score fell to 89.60%, highlighting inadequate feature aggregation and increased confusion between stems and leaves. Meanwhile, the GCN module proves essential for both local structural modeling and global semantic perception. Removing this module leads to a 1.06% drop in tobacco's IoU, a 3.72% reduction in sorghum's F1-score, and an almost 10% decrease in stem IoU, suggesting that the model's ability to aggregate neighborhood features and maintain segmentation consistency is severely weakened. In summary, both modules are indispensable for strengthening the model's capability to capture intricate plant architectures in 3D point clouds and ensuring high-quality stem–leaf separation.

5. Summary and outlook

With the accelerated advancement of modern agriculture and smart farming, efficient and precise segmentation of plant organs has become a fundamental prerequisite for high-throughput phenotyping and comprehensive crop monitoring. This study introduces a semantic Embedding-Guided Graph Self-Attention Network for stem–leaf separation in 3D plant point clouds. By integrating graph convolution with self-attention, the framework effectively models intricate geometric structures while capturing long-range contextual dependencies within point clouds. During feature space optimization, cross-entropy and inter-class discriminative loss functions are jointly optimized to make features within a class more similar and features between classes more distinct, thereby enhancing classification performance and strengthening the separability of the feature space. Experimental validation on the Plant-3D dataset demonstrates that the proposed approach achieves substantial performance gains in stem–leaf separation, yielding improvements of 3.97%, 4.35%, and 7.64% in precision, recall, and IoU, respectively. The proposed framework offers an effective solution for fine-grained plant structure recognition and intelligent phenotypic analysis. Future studies will focus on extending the proposed approach to a wider range of plant species and systematically assessing its generalization ability and practical effectiveness on large-scale datasets with complex structures.

Acknowledgement

This work was jointly funded by the Shandong Provincial Key Research and Development Program (grant numbers 2022LZGC021 and 2021LZGC026), the Higher Education Institutions Youth Innovation and Science & Technology Support Program of Shandong Province under Grant 2024KJH062.

References

Bai, Y., Durand, J.-B., Vincent, G., & Forbes, F. (2023). Semantic segmentation of sparse irregular point clouds for leaf/wood discrimination. *Advances in Neural Information Processing Systems, 36*, 48293-48313.

Conn, A., Pedmale, U. V., Chory, J., & Navlakha, S. (2017). High-resolution laser scanning reveals plant architectures

that reflect universal network design principles. *Cell systems, 5*(1), 53-62. e53. http://dx.doi.org/10.1016/j.cels.2017.06.017

Du, R., Ma, Z., Xie, P., He, Y., & Cen, H. (2023). PST: Plant segmentation transformer for 3D point clouds of rapeseed plants at the podding stage. *ISPRS Journal of Photogrammetry Remote Sensing, 195*, 380-392. http://dx.doi.org/10.1016/j.isprsjprs.2022.11.022

Furbank, R. T., & Tester, M. (2011). Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in plant science, 16*(12), 635-644. http://dx.doi.org/10.1016/j.tplants.2011.09.005

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., et al. (2020). *Randla-net: Efficient semantic segmentation of large-scale point clouds.* Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. http://dx.doi.org/10.1109/CVPR42600.2020.01112

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems, 30*. http://dx.doi.org/10.48550/arXiv.1706.02413

Rajapaksha, U., Sohel, F., Laga, H., Diepeveen, D., & Bennamoun, M. (2024). Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. *ACM Computing Surveys, 56*(12), 1-51. http://dx.doi.org/10.1145/3677327

Singh, A. K., Ganapathysubramanian, B., Sarkar, S., & Singh, A. (2018). Deep learning for plant stress phenotyping: trends and future perspectives. *Trends in plant science, 23*(10), 883-898. http://dx.doi.org/10.1016/j.tplants.2018.07.004

Tester, M., & Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. *Science, 327*(5967), 818-822. http://dx.doi.org/10.1126/science.1183700

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., & Guibas, L. J. (2019). *Kpconv: Flexible and deformable convolution for point clouds.* Paper presented at the Proceedings of the IEEE/CVF international conference on computer vision. http://dx.doi.org/10.1109/ICCV.2019.00651

Tian, Z., & Li, S. (2022). Graph-based leaf–wood separation method for individual trees using terrestrial LiDAR point clouds. *IEEE Transactions on Geoscience Remote Sensing, 60*, 1-11. http://dx.doi.org/10.1109/TGRS.2022.3218603

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics, 38*(5), 1-12. http://dx.doi.org/10.1145/3326362

Watt, M., Fiorani, F., Usadel, B., Rascher, U., Muller, O., & Schurr, U. (2020). Phenotyping: new windows into the plant for breeders. *Annual Review of Plant Biology, 71*(1), 689-712. http://dx.doi.org/10.1146/annurev-arplant-042916-041124

Weyler, J., Magistri, F., Marks, E., Chong, Y. L., Sodano, M., Roggiolani, G., et al. (2024). PhenoBench: A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *IEEE transactions on pattern analysis machine intelligence, 46*, 9583 - 9594. http://dx.doi.org/10.1109/TPAMI.2024.3419548

Yang, A., Yang, J., Wu, H., Li, Z., Bai, B., & Li, G. (2025). Semantic embedding-guided graph self-attention network for plant stem–leaf separation from 3D point clouds. *Computers and Electronics in Agriculture, 238*, 110857. http://dx.doi.org/https://doi.org/10.1016/j.compag.2025.110857

Yang, B., Dong, Z., Liang, F., & Mi, X. (2024a). *Ubiquitous Point Cloud: Theory, Model, and Applications*: CRC Press.

Yang, X., Miao, T., Tian, X., Wang, D., Zhao, J., Lin, L., et al. (2024b). Maize stem–leaf segmentation framework based on deformable point clouds. *ISPRS Journal of Photogrammetry Remote Sensing, 211*, 49-66. http://dx.doi.org/10.1016/j.isprsjprs.2024.03.025

Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V. (2021). *Point transformer.* Paper presented at the Proceedings of the IEEE/CVF international conference on computer vision. http://dx.doi.org/10.1109/ICCV48922.2021.00755