

Article

Semantic Segmentation Method for High-Resolution Tomato Seedling Point Clouds Based on Sparse Convolution

Shizhao Li ¹, Zhichao Yan ¹ , Boxiang Ma ¹, Shaoru Guo ^{1,*}  and Hongxia Song ²

¹ School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China; 202101204122@email.sxu.edu.cn (S.L.); 202312407023@email.sxu.edu.cn (Z.Y.); 202412407017@email.sxu.edu.cn (B.M.)

² College of Horticulture, Shanxi Agricultural University, Jinzhong 030801, China; 13834836584@sxau.edu.cn

* Correspondence: gsr@sxu.edu.cn; Tel.: +86-15735171023

Abstract: Semantic segmentation of three-dimensional (3D) plant point clouds at the stem-leaf level is foundational and indispensable for high-throughput tomato phenotyping systems. However, existing semantic segmentation methods often suffer from issues such as low precision and slow inference speed. To address these challenges, we propose an innovative encoding-decoding structure, incorporating voxel sparse convolution (SpConv) and attention-based feature fusion (VSCAFF) to enhance semantic segmentation of the point clouds of high-resolution tomato seedling images. Tomato seedling point clouds from the Pheno4D dataset labeled into semantic classes of ‘leaf’, ‘stem’, and ‘soil’ are applied for the semantic segmentation. In order to reduce the number of parameters so as to further improve the inference speed, the SpConv module is designed to function through the residual concatenation of the skeleton convolution kernel and the regular convolution kernel. The feature fusion module based on the attention mechanism is designed by giving the corresponding attention weights to the voxel diffusion features and the point features in order to avoid the ambiguity of points with different semantics having the same characteristics caused by the diffusion module, in addition to suppressing noise. Finally, to solve model training class bias caused by the uneven distribution of point cloud classes, the composite loss function of Lovász-Softmax and weighted cross-entropy is introduced to supervise the model training and improve its performance. The results show that mIoU of VSCAFF is 86.96%, which outperformed the performance of PointNet, PointNet++, and DGCNN, respectively. IoU of VSCAFF achieves 99.63% in the soil class, 64.47% in the stem class, and 96.72% in the leaf class. The time delay of 35ms in inference speed is better than PointNet++ and DGCNN. The results demonstrate that VSCAFF has high performance and inference speed for semantic segmentation of high-resolution tomato point clouds, and can provide technical support for the high-throughput automatic phenotypic analysis of tomato plants.

Keywords: 3D point clouds; semantic segmentation; tomato; sparse convolution



Academic Editor: Ruirui Zhang

Received: 3 December 2024

Revised: 25 December 2024

Accepted: 27 December 2024

Published: 31 December 2024

Citation: Li, S.; Yan, Z.; Ma, B.; Guo, S.; Song, H. Semantic Segmentation Method for High-Resolution Tomato Seedling Point Clouds Based on Sparse Convolution. *Agriculture* **2025**, *15*, 74. <https://doi.org/10.3390/agriculture15010074>

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As one of the most important vegetable varieties, tomato ranks first in the world in terms of export volume and value, and is important in agricultural production and trade around the world, accounting for about 30 percent of the world’s total planting area [1]. To meet the demand for increased tomato production, breeding experts have adopted efficient genetic breeding programs to cultivate high-yield and high-quality tomato varieties [2]. High-throughput phenotypic analysis of tomato plants can effectively help researchers

analyze and track the growth of tomato plants. In the past few decades, computer vision-based techniques for high-throughput analysis of plant phenotypes have received ever-increasing research interest from various communities, such as the computer science and modern intelligent agricultural breeding communities [3,4]. In order to automatically obtain plant organ parameters, it is necessary to accurately segment the stems and leaves, which are the primary components of plant for photosynthesis, evaporation and growth characterization. Recently, with the development of genetic breeding for tomato, the stem-leaf segmentation problem for the spatial tomato structures has gained particular research attention [2].

Recently, two-dimensional (2D) imaging techniques have been applied to obtain structural parameters of plants [5,6], particularly in vegetation studies. It is possible to monitor plant growth, assess health status, and detect diseases through the analysis of obtained images using 2D image-based processing methods [7,8]. Because a 2D image is essentially a projected image in a direction and plants have complicated three-dimensional (3D) structures, techniques based on 2D imaging have significant limitations in stem-leaf segmentation when dealing with complex backgrounds, occlusions and overlaps, structural complexity of plants, image resolution and so on [9,10]. To overcome the problem of insufficient spatial information of 2D imaging, 3D point cloud technology has been utilized to obtain spatial information about plants [11–13]. For example, lettuce leaf segmentation and fresh weight estimation have been achieved in [14] by using 3D point cloud technology. The stem-leaf segmentation of individual corn has been studied based on the 3D point cloud data of lidar (light detection and ranging) [15]. A tomato stem-leaf segmentation method is proposed according to point cloud data obtained through 3D laser scanning [16]. Although the aforementioned traditional plant stem-leaf segmentation algorithm have been achieved success for specific growth stages or plant types with few leaves, it is difficult to implement adaptive segmentation and improve accuracy.

With the development of machine learning methods, deep learning techniques are applied to process plant 3D data and significant results have been achieved [17,18]. The SegNet model was adopted to segment leaves from individual poplar seedlings using 3D point cloud data under the condition of strong interference [19]. Stem-leaf segmentations of maize are obtained by using voxel-based convolutional neural network and 3D point cloud data technology [20,21]. The semantic segmentation method based on PointNet++ is proposed in [22] to realize automatic fruit segmentation through multi-sensor 3D point cloud data fusion. Researchers primarily focus on utilizing deep learning and point cloud techniques for the automatic segmentation of plant leaf and organ, however the point cloud resolution (approximately tens of thousands of points per plant) is relatively low. In high-resolution plant 3D point clouds applications, there are still several shortcomings such as insufficient processing capacity, poor environmental adaptability, and high computational complexity. Although the valuable methodologies is provided for plant phenotypic analysis using existing research, further improvements will help to meet the practical requirements of high-resolution plant 3D point cloud analysis and to enhance the performance of models in agricultural applications. As such, we are motivated to further study the stem-leaf problem for tomato with high-resolution 3D point clouds.

In this work, we propose the encoding-decoding structure with the voxel sparse convolution (SpConv) and attention-based feature fusion (VSCAFF) for semantic segmentation of 3D point clouds generated from high-throughput imaging of tomato seedlings. In the proposed network structure, the point clouds for the tomato seedling are rasterized and voxelified to facilitate deep learning analysis. At the same time, an feature fusion module based on attention for the semantic ambiguity of points in voxels is designed to effectively fuse voxel features with point features. The proposed model can be efficiently performed

and provides technical support for semantic segmentation of point clouds without data preprocessing such as downsampling. Semantic segmentation of tomato seedlings provides a powerful tool for plant phenotypic analysis, whose application can guide tomato breeding in precision agriculture. The main contributions of this paper can be highlighted as follows:

- (1) The SpConv based on 3D Skeleton convolution kernel is designed to strengthen the weights of Conv kernel skeleton and reduce the number of parameters.
- (2) The feature fusion module based on attention mechanism is achieved to effectively fuse the different branch features and suppress noise.
- (3) The composite loss function of Lovász–Softmax and weighted cross-entropy is introduced to effectively alleviate unevenness and improve model performance.
- (4) The the VSCAFF based on encoding-decoding network structure for semantic segmentation is constructed.

2. Materials and Methods

2.1. Point Cloud Data Acquisition

For tasks such as point cloud annotation and semantic segmentation, the spatiotemporal datasets of the tomato seedling point clouds employs Pheno4D [23] obtained using the high-precision 3D laser scanning system, which includes Perceptron Scan Works V5 (Perceptron Inc., Plymouth, MI, USA) as laser triangulation scanner connected to the ROMER Infinite 2.0 measurement arm (Hexagon Metrology Services Ltd., London, UK). The system can completely scan the surface of the tomato plant at different positions and angles, with a spatial accuracy of less than one-tenth of a millimeter. This dataset is scanned by Perceptron Scan Works V5 laser triangulation scanner, whose measuring distance is 100 mm. Thus, in order to scan the tomato plants with a large size in practice, the scanner is mounted on a measuring arm, a ROMER Infinite 2.0 designed by Hexagon Metrology Services Ltd. (Esplinge, Sweden). This mechanical arm consists of seven joints and has a spherical measurement volume with a radius of 140 cm. The measurements are performed in the greenhouse, where the plants are grown. Each plant is scanned separately, using a mounting system, which ensured, that the pot has the same position and orientation relative to the measuring arm every day. In this way, the plant's position and orientation are consistent during the whole series of scans. The measuring volume is sufficient to scan the plant from different positions and angles. Due to the non-invasive nature of the scanning system, the measurements can be carried out with the motion of the plant. A total of 140 high-resolution point clouds of tomato seedlings were acquired over 20 days from 7 tomato plants collected, starting when tomato seedlings issued their first buds.

2.2. Data Annotation

Using CloudCompare (v2.12.2), the datasets of tomato seedling point clouds have been annotated with semantic labels, which are categorized into “Leaf”, “Stem”, and “Soil”, as shown in Figure 1. 3D raw point clouds of the tomato are shown in Figure 1a, and the semantic annotation point clouds is illustrated in Figure 1b, where the green part, the purple part and the yellow part are respectively indicated “leaf”, “stem” and “soil”.

The point clouds of single high-resolution image from tomato seedlings generally contain millions of points, which can finely represent the details of tomato organs. The average point cloud number of seven tomato seedlings is collected in 20 days. The point cloud number varies with the size of tomato seedling at different growth stages. The statistics of 20-day point cloud quantity are shown in Table 1. As illustrated in Table 1, the average number is 2.43 millions, among which the plants scanned in 19th day has the maximum number of points of 3.68 millions and the 17th day has the minimum number

1.88 millions. The standard deviation of point number is 0.47 millions which denotes the imbalance distribution of point along the days.

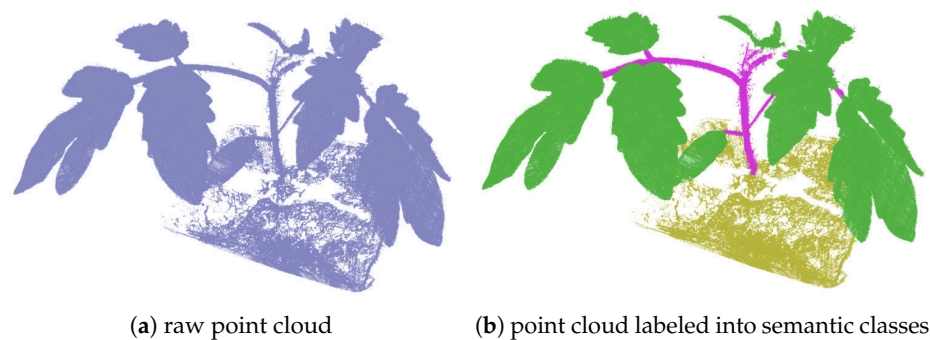


Figure 1. The raw tomato seedling point cloud and point cloud labeled into semantic classes of ‘leaf’, and ‘stem’, and ‘soil’.

Table 1. 20-day tomato point cloud quantity statistics.

Name	Measures	Max	Min	Avg	SD
Num		3.68	1.88	2.43	0.47
Day		19	17	1–20	1–20

Avg represents average. SD represents Standard Deviation.

In this study, we used the 20-day point cloud datasets numbered 1 to 5 as the training set, comprising a total of 100 point clouds. The point cloud datasets numbered 6 and 7, also covering 20 days, are designated as the test set, consisting of 40 point clouds. The datasets distribution is shown in Table 2. As evident in Table 2, the point set for leaf, stem, and soil demonstrate extremely uneven distribution, because tomato seedlings exhibit significant changes in size and morphology during different growth stages, resulting in variations in the number of point clouds. In the semantic class distribution of the training and test set, soil and leaf points each account for more than 45% of the total plant point clouds, while stem points make up less than 5%. In point-based methods that utilize raw point clouds as input, the highly uneven distribution of points among different semantic classes poses a significant challenge for semantic segmentation tasks.

Table 2. The distribution of the datasets in the training and test set.

Class	Number of Plants (Plant)	Leaf (%)	Stem (%)	Soil (%)
Training set	100	49.35	4.56	46.09
test set	40	50.73	4.32	44.95

2.3. Network Structure of Semantic Segmentation

Convolutional Neural Network (CNN) can excel at feature extraction from 2D images and has been extensively utilized in various tasks, including semantic segmentation and object detection in 2D images, demonstrating significant effectiveness [24,25]. During the computation of CNN, convolutional kernels interact with input data using local receptive fields, whereby the principle of weight sharing substantially decreases the number of parameters, thereby reducing the overall complexity of the model. The convolution operation relies on a fixed grid structure to extract features, such as the pixel matrix of image, where the relationships between adjacent pixels are well-defined. Due to the disordered nature and non-structured characteristics difficultly formed fixed connectivity relationships of 3D point cloud data, hinder the direct application of traditional 3D convolution methods to

point clouds. This limitation has spurred researchers to explore novel network structures and algorithms, such as PointNet and its variations, to more effectively process and analyze point cloud data. An effective method that has been widely applied and has shown better performance in 3D point cloud recognition tasks, involves point clouds voxelized into the voxel representation, followed by the application of standard 3D convolution for feature extraction [26]. During voxelization, point cloud data is greatly influenced by resolution, so it is difficult to balance model performance and efficiency. When higher voxel resolution is used, the model demonstrates excellent performance but requires more time and exhibits poor efficiency. Conversely, when lower voxel resolution is employed, the model executes quickly with high efficiency but suffers a significant drop in performance. One of the key reasons contributing to the issue is that when higher voxel resolution is used, an increase in empty voxels leads to excessive sparsity in the 3D point clouds, resulting in a significant number of empty operations when performing regular 3D convolution [27].

The point cloud structures of tomato plants are more complex and irregular in shape, leading to greater sparsity after voxelization. The application of conventional 3D convolution can result in significant detailed deformation and incurs a substantial computational burden. In order to solve the challenge of directly applying convolutions to tomato plant point clouds, an encoding-decoding network structure based on SpConv is designed, as illustrated in Figure 2. The core components of the network includes three parts, such as encoding-decoding structure based on SpConv, skeleton convolution kernel, attention-based feature fusion method. The network structure first performs voxelization on the point clouds of tomato plants, followed by feature extraction using SpConv. SpConv represents the sparse convolution module, while DeConv denotes the deconvolution module, which is used to restore the voxel shape prior to sparse convolution. The skeleton convolution kernel significantly reduces the number of parameters while enhancing the model's performance. The voxel-based encoding-decoding structure is employed to ultimately generate the same class labels for all points within the same voxel. The expand module representing feature diffusion is used to spread voxel features to point features. The diffusion strategy used in this paper shares the voxel features within the same voxel. It is necessary to fuse point branch features for all points to ensure the uniqueness of points within the same voxel using the attention-based feature fusion, thereby eliminating ambiguity.

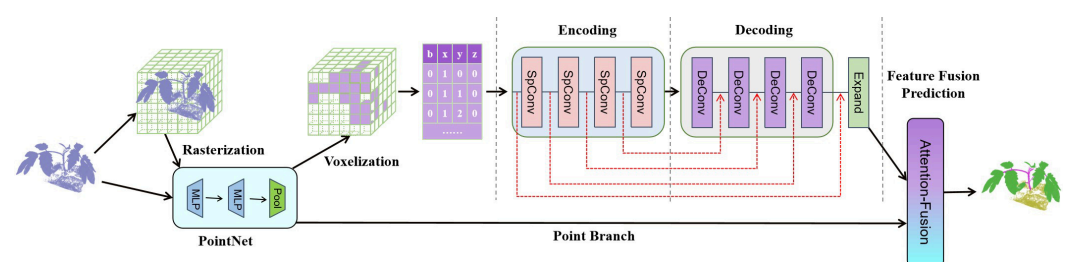


Figure 2. The structure of network.

2.3.1. Encoding-Decoding Structure Based on SpConv

During the process of rasterizing tomato plant point clouds, their sparsity tends to increase with the increase in voxel resolution. A higher voxel resolution for high-resolution point cloud data is often required to accurately extract semantic information and the geometric details of plant organs such as stems, which will increase the time required for convolution operations. So SpConv is employed as a local feature extractor to perform convolution operations on the point clouds, constructing a hierarchical feature extraction module.

In this paper, a SpConv-based feature encoder is constructed using TorchSparse [28], which is an efficient 3D SpConv library that accelerates the process of SpConv through

adaptive matrix multiplication and vectorization. The encoder utilizes the feature sparse matrix of the point clouds as input, and its generation process is as follows. Firstly, the tomato point clouds is rasterized with 0.2 as the voxel unit, and the points within each voxel are input into Pointnet subnet to learn detailed features from the points. Then, the points within the same voxel are pooled to generate voxel features and construct a feature sparse matrix of the point clouds. Finally, the sparse matrix is used as input to the backbone network, which performs feature encoding.

In the decoding section, the feature code is deconvoluted, and the voxel shape is restored layer by layer using the encoding convolution table [29]. The features with the same voxel shape from the encoding and decoding stages are concatenated as the input for the subsequent deconvolutional layers to fully utilize the semantic information of high-level features and the detail information of low-level features. The voxel features are diffused into point features through the diffusion module, meaning that all points within the same voxel share the voxel features. The encoding-decoding structure is shown in Figure 3, where N_i indicates the number of voxels output by the current module. Due to the different sparsity levels of point clouds from various tomato plants, the number of voxels output by different data is also different. The voxel features are diffused to all points using the diffusion module, simultaneously restoring the raw point cloud shapes.

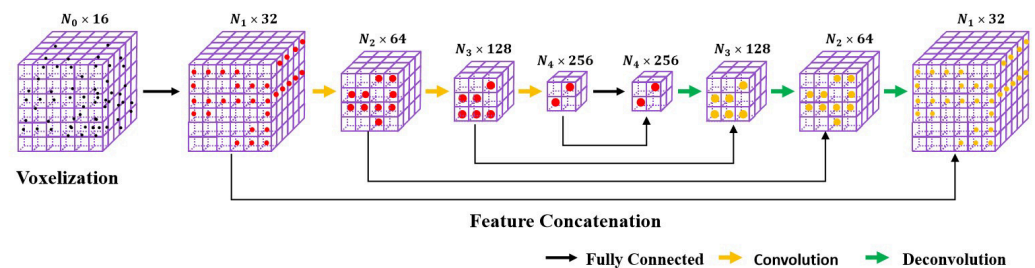


Figure 3. Encoding-decoding architecture based on SpConv.

2.3.2. Skeleton Convolution

In order to further improve the inference speed of the network, we design 3D skeleton convolution (Skeletal Conv), which strengthens the weight of the convolution kernel skeleton, while reducing the model parameters, as illustrated in Figure 4a. Compared to traditional 3D convolutional kernel [30], the parameters of the single skeleton convolution kernel are reduced by nearly 74%, and the parameters of the SpConv module designed in the paper (shown in Figure 4b) are reduced by 37% compared with the equivalent convolution module (shown in Figure 4c).

3D Skeleton Conv extracts key structures from conventional convolutional kernels to enhance core weights, effectively matching the organ structures of tomato plant point clouds. During the actual convolution process, the overlap between the convolution kernel and the point clouds of tomato plants is concentrated on the horizontal and vertical skeletal parts. Through the residual concatenate between the skeleton and conventional convolution kernel, the feature extraction of the key parts from tomato plant point clouds can be satisfied while preserving edge information.

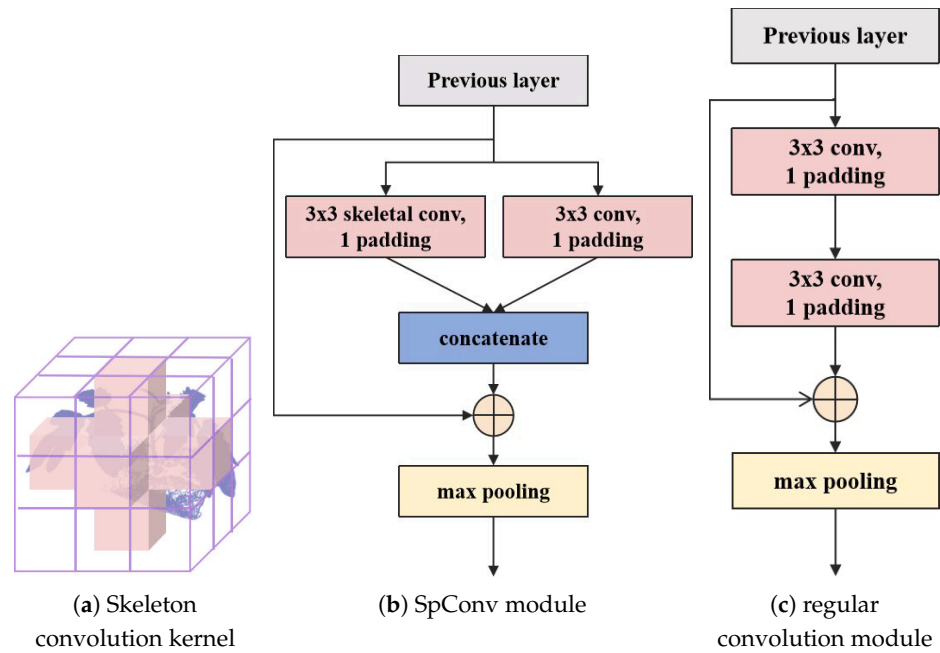


Figure 4. Three kinds of convolution kernel structures.

2.3.3. Attention-Mechanism Feature Fusion Method

At the later period of the decoding stage, point features are generated for voxel points by the expand module. However, the module generates identical features for all points within the same voxel, which can result in ambiguity as a result of different semantic points sharing the same features. To avoid the issue, the point features generated by the expand module are fused with the point features generated at the voxelization stage using PointNet. Traditional feature fusion methods typically rely on accumulation or concatenation of different features, but these approaches often fail to effectively filter out noise in the features.

Therefore, the feature fusion method designed in the paper is based on attention mechanism, integrating channel attention and self-attention to calculate channel attention. Using the fusion method, useful information can be effectively filter from a large amount of irrelevant data. The attention mechanism adopted by the model automatically learns the allocation of attention weights. The allocation of attention weights can be dynamically adjusted according to different inputs, demonstrating the certain flexibility and adaptability [31]. Compared to traditional attention mechanisms, the incorporation of multi-layer perceptron (MLP) can learn more complex attention patterns, enabling better capture of intricate relationships and features within the input data.

The attention scores for feature type is calculated according to voxel feature and point feature of each point, namely the feature weight vector, as shown in Formula (1):

$$s_i = \text{softmax} \left(\sum_{b \in v, p} \text{MLP}_b(f_i^b) \right) \quad (1)$$

where b is the feature type, namely voxel feature (v) or point feature (p). Then the product of weight vector and corresponding point feature vector is calculated, which is the final point feature, as shown in Formula (2):

$$\tilde{f} = [f_i^v, f_i^p] \times s_i \quad (2)$$

where f_i^v is the voxel feature of points, f_i^p is the feature of points, and s_i is the weight vector. The calculation process is shown in Figure 5.

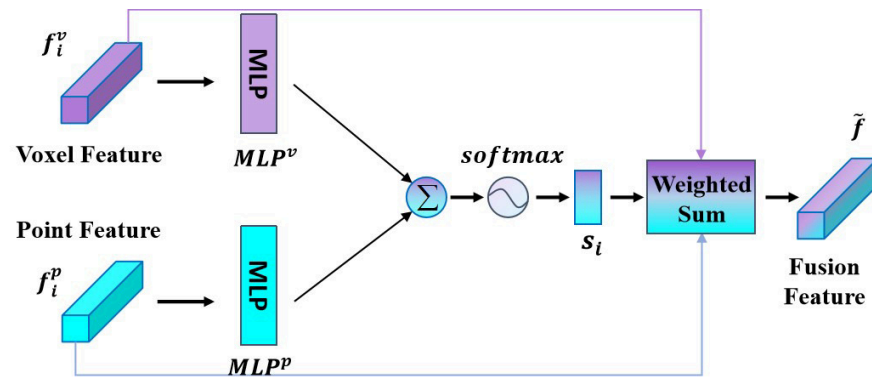


Figure 5. Attention-based feature fusion method.

2.3.4. Loss Function

With stems of tomato plant point clouds accounting for only 4.56% of the total, the class distribution is extremely uneven. The imbalance can lead to a bias towards the classes with higher proportions during the model training, thereby negatively impacting its performance. To effectively alleviate the unevenness, weighted cross-entropy loss function which assigns different weights to different classes is adopted. In addition, Lovász-Softmax loss directly optimizes the intersection over union (IoU) and dynamically adjusts the contribution of each pixel, improving the learning of minority classes and enhancing segmentation accuracy, especially in challenging scenarios like small object detection [32]. Through integrating these two loss functions, The model benefits from classification accuracy while optimizing overall segmentation quality. The composite loss function focuses on both pixel-level classification capabilities and simultaneously optimizing overall segmentation performance, as shown in Formula (3):

$$Loss = Loss_v + Loss_p \quad (3)$$

where $Loss_p$ is responsible for the training of point features. $Loss_v$ is responsible for optimizing the training of voxel branches, consisting of the weighted cross-entropy loss and the Lovász-Softmax loss, as shown in Formula (4):

$$Loss_v = - \sum_i \alpha_i p(y_i) \log(p(\tilde{y})_i) + Lovasz(y, \tilde{y}) \quad (4)$$

where $\alpha_i = \frac{1}{\sqrt{f_i}}$, represents the class weight, calculated from the class proportion f_i . For the training of the point features, only the weighted cross-entropy loss function is adopted.

2.4. Experimental Platform

The computer in the experiment is configured with the CPU of Intel Core i9-12900, the GPU of NVIDIA RTX 3090, the memory of 64 GB, and 2 T solid-state disk. The operating system of Ubuntu20.04 and the deep learning framework of Python 1.18 is used. The model implementation is based on Python.

2.5. Model Evaluating Indicator

The model performance is evaluated by IoU (per-class Intersection over union) and mIoU (mean IoU) [33], as shown in Formula (5):

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i} \quad (5)$$

where i is the class (soil, stem or leaf), TP_i (True Positives) represents the number of samples where both the prediction and the ground truth are of the same class i . FP_i (False Positives) denotes the number of samples where the prediction is of class i , but the ground truth is of a different class. FN_i (False Negatives) indicates the number of samples where the prediction is of a different class, but the ground truth is of the same category i . mIoU is calculated by averaging the IoU values across all classes, as shown in Formula (6):

$$mIoU = \frac{1}{N} \sum_{i=1}^N IoU_i \quad (6)$$

where, N represents the total number of classes.

To verify the model performance using IoU and mIoU, the proposed model is compared with three classical models based on different principles, such as point-based PointNet [34], neighbor retrieval-based PointNet++ [35] and graph-based DGCNN [36].

2.6. Ablation Experiments

To further verify the effectiveness of VScaff, the ablation experiments are conducted to analyze the performance of each individual module. First, the foundational network consisting solely of an encoder and decoder with conventional sparse convolution is constructed, wherein the semantic segmentation module based entirely on fully connected layers. Then, upon the infrastructure, skeleton convolution kernel integrated into the Sp-Conv, simple fusion of point feature extraction branch and voxel branch, and feature fusion module based on attention are respectively introduced. Following the incorporation of the point feature extraction branch, various feature fusion techniques including concatenation, accumulation, and the attention-based method delineated in the paper, are respectively explored. Ultimately, the weighted cross-entropy loss function (WCE) is compared with the combination of WCE and Lovász-softmax loss function in the experiment.

3. Results

3.1. Performance Comparison of Different Semantic Segmentation Models

The results of comparison with other models are shown in Table 3.

Table 3. Performance comparison of different models.

Model	mIoU (%)	IoU (%)			Time Delay (ms)
		Leaf	Stem	Soil	
PointNet	73.74	93.25	28.35	99.61	28
PointNet++	81.03	94.91	48.78	99.41	268
DGCNN	83.96	95.49	56.73	99.67	178
VScaff	86.94	96.72	64.47	99.63	35

As can be seen from Table 3, the VScaff model achieves mIoU of 86.96%, surpassing the other three models. Specifically, it outperforms DGCNN by 2.98%, PointNet++ by 5.91%, and PointNet by 13.2%. In comparison to other models, PointNet demonstrates the least effective performance, because the MLP-based networks lack the learning of neighbor

features, especially for the identification of small classes (e.g., Stem). PointNet-based PointNet++ introduces neighbor feature learning and hierarchical structures, significantly enhancing its ability to learn details, such as improving the identification of stems by nearly 20%. DGCNN benefits from the powerful representation capabilities of graphs. By dynamically updating the relationship graph at each layer, it learns semantic features of points, exhibiting superior abilities in learning local neighbor features. VScaff adopts voxel features extracted by SpConv and point features extracted by PointNet branch, which can take into account the geometric relationship of points and the unique features of the points themselves. Additionally, the attention-based feature fusion method can effectively fuse the features extracted from two different views, and more accurately identify the details. Although the stem identification using VScaff is unsatisfactory due to an insufficient number of stem points, it still has an advantage over other models, especially 7.74% higher than DGCNN. It means that in certain application, VScaff may still be the better choice, given its excellent performance in soil and leaf segmentation. However, it could lead to inaccurate stem recognition in practical applications, potentially affecting tasks that require precise stem analysis, such as evaluating plant health and detecting damage caused by insect pests.

In the inference speed, high-resolution point clouds typically incur substantial computational time due to the excessive number of points. In contrast, PointNet, with straightforward architecture that only relies on MLP for feature extraction, exhibits notably shorter processing times, as shown in Table 3. However, PointNet++ introduces farthest point sampling and K-nearest neighbors as neighbor retrieval algorithms, leading to a time complexity of $O(n^2)$. Moreover, the extensive use of hierarchical feature extraction layers significantly increases latency when handling high-resolution point clouds. DGCNN, which adopts the same neighbor retrieval strategy as PointNet++ and leverages a dynamic graph structure during hierarchical feature construction, substantially accelerates processing and consequently reduces latency compared to PointNet++. In order to optimize the model performance, the proposed VScaff incorporates a PointNet-like structure for point feature extraction, while integrating SpConv and skeleton convolution kernels for voxel feature extraction. Notably, these operations run in parallel, resulting in an average processing latency of point clouds per plant that is merely 7 ms (millisecond) greater than PointNet.

The semantic segmentation results of tomato plant using different model across various growth stages is shown in Figure 6, and the red box represents misrecognition. It becomes evident that PointNet struggles to detect connection points between stems and leaves, in which the semantic information consequently fails to be precisely extracted. The shortcoming arises from PointNet's inability to model inter-point relationships, thus hindering its capacity to learn the local geometric details of tomato plant point clouds. In contrast, PointNet++, by incorporating hierarchical feature extractors, effectively models neighbor relationships, leading to a significant improvement in the stem-leaf segmentation quality. However, constrained by the inherent limitations of MLP in feature representation, a large number of stem organs cannot be identified by PointNet++. DGCNN, leveraging dynamic graphs, adeptly extracts local relationship features and excels in representing detailed characteristics. As evidenced in the column d of Figure 6, DGCNN surpasses PointNet++ in accurately segmenting stems and leaves. Nevertheless, it occasionally misclassifies leaves. In contrast to these methods, the VScaff model proposed in the paper employs multiple branches to extract features at various scales. This method not only captures the detailed features of individual points, but also effectively achieves local relationship features. Consequently, VScaff delivers more precise segmentation of stem and leaf details, demonstrating superior performance across different growth stages of tomato plants.

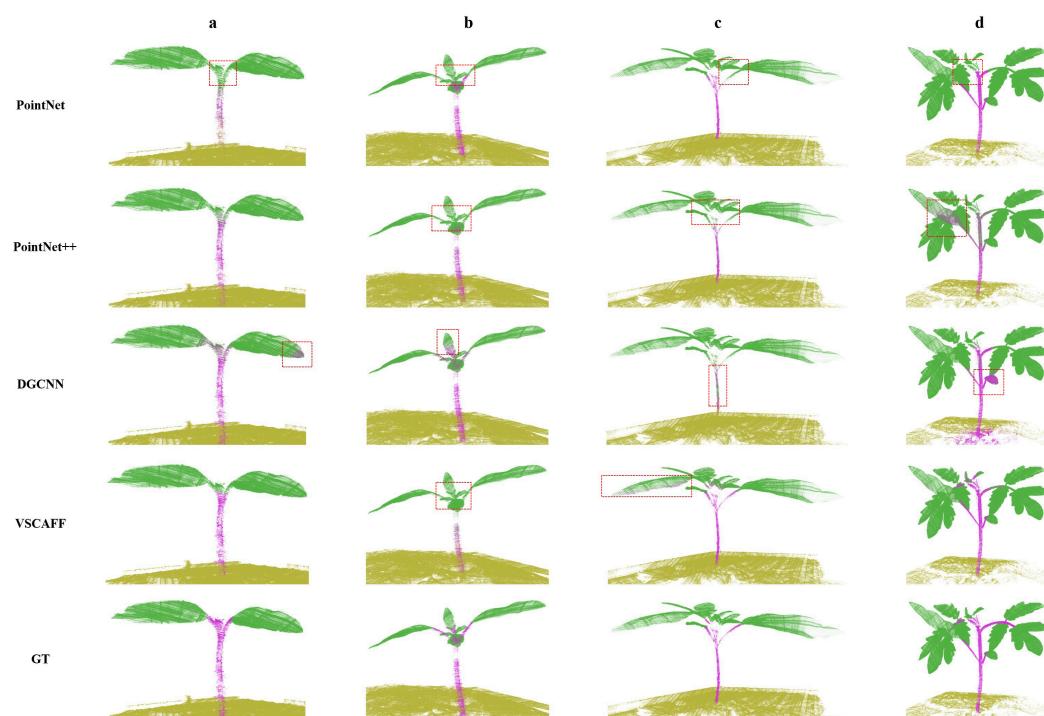


Figure 6. Semantic segmentation of the tomato plant point clouds. Note1: GT represents ground truth. Note2: Four seedling point clouds scanned in four discrete days represented.

3.2. Ablation Results

Each individual module are added or removed in the foundational network. The ablation results are detailed in Table 4.

Table 4. Performance comparison of different model.

Basic Network	Skeleton Convolution Kernel	Point Feature Branch			Loss Function		mIoU (%)
		Concatenation	Accumulation	Attention	WCE	Compound	
✓					✓		74.3
✓	✓				✓		76.25
✓	✓	✓			✓		78.62
✓	✓		✓		✓		80.83
✓	✓			✓	✓		84.72
✓	✓			✓		✓	86.94

As can be seen from Table 4, the introduction of skeleton convolution kernel effectively strengthens the weights of the original convolution kernel along horizontal and vertical axes, significantly enhancing the recognition of the tomato plant structure for stems and leaves. Notably, This modification improvement boosts the mIoU by 0.95% over the foundational network while reducing parameters. Furthermore, the incorporation of the point feature extraction branch differentiates the model's point recognition by infusing detailed point features alongside voxel features, eliminating ambiguity of different semantic points within the same voxel, thereby enhancing model's precision. During feature fusion, the use of concatenation improves mIoU by 5.37%, and accumulation by 3.58%. In contrast, the proposed attention-based fusion method achieves substantial 7.47% increase in mIoU, outperforming both, which is due to its ability to focus on relevant information, minimizing noise and irrelevant data's impact. The attention-based method effectively can extract valuable information from two distinct views through training and reduce the impact of useless information, thereby more efficiently enhancing model's performance. On the contrary, concatenation and accumulation, which can still be interfered by noise, fail to capture es-

sential features. Finally, the modification of the loss function from a weighted cross-entropy to a composite loss function combining Lovász-softmax and weighted cross-entropy leads to a further 1.46% improvement in model performance. This result underscores the effectiveness of Lovász-softmax in addressing the challenges posed by imbalanced datasets, validating its integration into the model.

4. Discussion

4.1. Efficiency of Methods

To achieve efficient semantic segmentation of 3D point clouds of high-resolution tomato plant, the SpConv-based network model of semantic segmentation integrating PointNet is constructed. By employing skeleton convolution kernels, the parameter number of the model is reduced, which in turn enhances the inference speed. Additionally, an attention mechanism is utilized to fuse features extracted from two distinct views, thereby successfully accomplishing semantic segmentation for high-resolution point clouds of tomato plant. Finally, the composite loss function of Lovász-softmax and weighted cross-entropy efficiently solves the problems caused by the uneven datasets for tomato plant point clouds to enhance the model performance.

The comparative analysis with several different models, including the point-based PointNet, the neighbor feature-based PointNet++, the graph-based DGCNN, shows that the point-based model, due to the lack of neighbor information, deliver the worst segmentation performance, albeit with the fastest inference speed. While incorporating neighbor retrieval significantly improves segmentation accuracy, it also introduces substantial memory and computational overhead. In order to overcome this issue, the proposed model successfully achieves efficient and accurate semantic segmentation of high-resolution tomato plant point clouds by fusing features extracted from two distinct views.

Due to the non-invasive and physically harmless feature of semantic segmentation for tomato seedling point clouds, measurements can be repeated without disrupting plant growth, achieving dynamic monitoring. The precise segmentation facilitates the extraction of not only plant morphological features, such as leaf area, plant height, and stem diameter, but also 3D structural features like branch angles and leaf distribution [37]. It is significant for understanding plant growth mechanisms and genetic variations in precision agriculture. Phenotype-based decisions can optimize planting density, irrigation strategies, and harvest timings, thereby enhancing crop yield and quality. Analysis of Phenotype helps identify plant varieties with desirable traits, expedite the breeding process, and cultivate varieties that are better adapted to specific environments and market demands [38]. Continuous monitoring of phenotypic changes allows for the early detection of disease or environmental stress, facilitating timely interventions to minimize losses. Furthermore, personalized management for each plant can achieve refined management practices. In summary, segmentation of tomato seedling point clouds offers a robust tool for agricultural plant phenotypic analysis, and phenotypic application in precision agriculture is pivotal for fostering sustainable agricultural practices and boosting agricultural productivity.

Improving the study of tomato seedling phenotype can guide tomato breeding. Through high-precision phenotypic analysis techniques, the growth characteristics of tomato seedlings can be more accurately identified and quantified, helping breeders select plants with desirable traits [39]. By integrating phenotypic data with genotypic data, the relationships between specific genes or gene combinations and their associated phenotypic traits can be revealed, which help breeders optimize breeding strategies by selecting genes that positively contribute to desired traits [38]. Leveraging high-throughput phenotypic analysis techniques, plants exhibiting favorable traits can be rapidly identified

and rapidly propagated, thereby significantly reducing the breeding cycle and improving breeding efficiency.

4.2. Limitations of Methods

On the other hand, there are still some limitations in the research, such as fewer stem points, lower IoU for the stem segmentation. The IoU of stems using these compared methods falls into a relatively large range (28–64%), which causes the differences of stem segmentation between these methods. Similar to the stem, petiole and pod are both slender and tiny, which needs more effective and generative methods to be developed. When the stem points is too low, it may lead to the following issues: (1) Data sparsity: insufficient points in space or time make it difficult to accurately capture the growth patterns and deformation of the stem. (2) Low statistical analysis accuracy: when performing statistical analysis such as stem shape analysis or growth rate calculations, the reliability and precision of the results are limited due to the scarcity of data points. (3) Difficulty in Model fitting: a low number of points makes it challenging for the model to learn accurate shape features, affecting the accuracy of predictions and analysis [40]. To address the issue of too few points on the stem, we can enhance the frequency of data collection to increase the density of points in both time and space, thereby obtaining more stem points. The interpolation algorithms (e.g., linear interpolation, spline interpolation) are also used to generate new points between existing ones, increasing data density and continuity. To improve the model's generalization ability and data diversity, virtual augmented reality data can be create combining it with real data for model training.

4.3. Future Work

In the future, suboptimal stem segmentation can affect applications that require precise analysis of stem. To improve the accuracy of stem segmentation, it is feasible to add more feature extraction layers to the existing model, use stronger encoder-decoder structures, and introduce multi-scale feature fusion mechanisms [41], which can enhance the model's ability to capture stem details and handle the complex relationships between the stem and its surroundings. Using the improvements, the performance of the VScaff model for stem segmentation can be significantly enhanced to make it more practical for agriculture applications. The semantic segmentation of high-resolution 3D point clouds for tomato plant can explore more advanced deep learning architectures and integrate multi-modal data. At the same time, the lightweight models are focused to adapt to practical applications, improving segmentation efficiency and accuracy. These provide strong support for precision agriculture, promoting the development of agricultural intelligence.

5. Conclusions

Because semantic segmentation methods often suffer from low precision and slow inference speed, the encoding–decoding structure is proposed, incorporating voxel-sparse convolution (SpConv) and attention-based feature fusion (VScaff). The number of parameters is reduced using the SpConv module to further enhance the inference speed. The ambiguity arising from points with different semantics is avoided using feature fusion module based on the attention mechanism. The model training class bias is solved using the composite loss function to improve its performance.

The experimental results indicate: (1) the VScaff model exhibits the best segmentation performance, with mIoU 86.96%, and IoU in the soil, stem, leaf, reaching 99.63%, 64.47%, 96.72%, respectively, significantly outperforming other conventional models. (2) The proposed model achieves the most efficiency (35 ms) with the least parameters and memory consumption. (3) The VScaff model fully utilizes computational resources,

achieving the optimal balance between segmentation performance and speed. The results demonstrate that VSCAFF has significant advantages in semantic segmentation of high-resolution tomato point clouds, providing an efficient and practical solution for the high-throughput automatic phenotypic analysis.

Author Contributions: Conceptualization, S.L.; methodology, S.L., Z.Y. and S.G.; software, S.L., Z.Y. and B.M.; validation, Z.Y. and S.G.; formal analysis, S.L. and Z.Y.; investigation, S.L. and Z.Y.; resources, S.G. and H.S.; data curation, S.L., Z.Y., H.S. and B.M.; writing—original draft preparation, S.L.; writing—review and editing, S.L., S.G. and B.M.; visualization, S.L.; supervision, S.G. and H.S.; project administration, H.S.; funding acquisition, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the key research and development project of Shanxi Province, grant numbers 202102020101008; 202302010101003.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Chitwood-Brown, J.; Vallad, G.E.; Lee, T.G.; Hutton, S.F. Breeding for resistance to Fusarium wilt of tomato: A review. *Genes* **2021**, *12*, 1673. [\[CrossRef\]](#) [\[PubMed\]](#)
- Roohanitaziani, R.; de Maagd, R.A.; Lammers, M.; Molthoff, J.; Meijer-Dekens, F.; van Kaauwen, M.P.; Finkers, R.; Tikunov, Y.; Visser, R.G.; Bovy, A.G. Exploration of a resequenced tomato core collection for phenotypic and genotypic variation in plant growth and fruit quality traits. *Genes* **2020**, *11*, 1278. [\[CrossRef\]](#) [\[PubMed\]](#)
- He, P.F. Comprehensive Evaluation and Primary Metabolite Analysis of Tomato Quality Characters. Master's Thesis, Fujian Agriculture and Forestry University, Fuzhou, China, 2023. [\[CrossRef\]](#)
- Zhao, C.J. Big data of plant phenomics and its research progress. *J. Agric. Big Data* **2019**, *2*, 5–18.
- Li, K.; Feng, Q.; Zhang, J. Co-segmentation algorithm for complex background image of cotton seedling leaves. *J. Comput. Aided Des. Comput. Graph.* **2017**, *10*, 1871–1880.
- Singh, V.; Misra, A.K. Detection of plant leaf diseases using image segmentation and soft computing techniques. *Inf. Process. Agric.* **2017**, *1*, 41–49. [\[CrossRef\]](#)
- Hasan, U.; Sawut, M.; Chen, S. Estimating the leaf area index of winter wheat based on unmanned aerial vehicle RGB-image parameters. *Sustainability* **2019**, *23*, 6829. [\[CrossRef\]](#)
- Rasti, S.; Bleakley, C.J.; Holden, N.M.; Whetton, R.; Langton, D.; O'Hare, G. A survey of high resolution image processing techniques for cereal crop growth monitoring. *Inf. Process. Agric.* **2022**, *2*, 300–315. [\[CrossRef\]](#)
- Shi, W.; van de Zedde, R.; Jiang, H.; Kootstra, G. Plant-part segmentation using deep learning and multi-view vision. *Biosyst. Eng.* **2019**, *187*, 81–95. [\[CrossRef\]](#)
- Ma, X.; Jiang, Q.; Guan, H.; Wang, L.; Wu, X. Calculation Method of Phenotypic Traits for Tomato Canopy in Greenhouse Based on the Extraction of Branch Skeleton. *Agronomy* **2024**, *12*, 2837. [\[CrossRef\]](#)
- Itakura, K.; Hosoi, F. Automatic Leaf Segmentation for Estimating Leaf Area and Leaf Inclination Angle in 3D Plant Images. *Sensors* **2018**, *10*, 3576. [\[CrossRef\]](#)
- Yuan, P.S.; Li, R.L.; Ren, S.G.; Gu, X.J.; Xu, H.L. State-of-the-Art review for representation learning and its application in plant phenotypes. *Trans. Chin. Soc. Agric. Mach.* **2020**, *6*, 1–14.
- Zhu, C.; Miao, T.; Xu, T.Y.; Li, N.; Deng, H.; Zhou, Y. Ear segmentation and phenotypic trait extraction of maize based on three-dimensional point cloud skeleton. *Trans. CSAE* **2021**, *6*, 295–301.
- Mortensen, A.K.; Bender, A.; Whelan, B.; Barbour, M.M.; Sukkarieh, S.; Karstoft, H.; Gislum, R. Segmentation of lettuce in coloured 3D point clouds for fresh weight estimation. *Comput. Electron. Agric.* **2018**, *154*, 373–381. [\[CrossRef\]](#)
- Jin, S.; Su, Y.; Wu, F.; Pang, S.; Gao, S.; Hu, T.; Liu, J.; Guo, Q. Stem-leaf segmentation and phenotypic trait extraction of individual maize using terrestrial LiDAR data. *IEEE Trans. Geosci. Remote Sens.* **2018**, *3*, 1336–1346.

[CrossRef]

16. Peng, C.; Li, S.; Miao, Y.L.; Zhang, Z.; Zhang, M.; Li, H. Stem-leaf segmentation and phenotypic trait extraction of tomatoes using three-dimensional point cloud. *Trans. Chin. Soc. Agric. Eng.* **2022**, *9*, 187–194.
17. Weng, Y.; Zeng, R.; Wu, C.M.; Wang, M.; Wang, X.J.; Liu, Y.J. A survey on deep-learning-based plant phenotype research in agriculture. *Sci. Sin. Vitae* **2019**, *6*, 698–716. [CrossRef]
18. Guo, X.; Sun, Y.; Yang, H. FF-Net: Feature-Fusion-Based Network for Semantic Segmentation of 3D Plant Point Cloud. *Plants* **2023**, *12*, 1867. [CrossRef]
19. Hu, C.H.; Liu, X.; Ji, M.J.; Li, Y.J.; Li, P.P. Single Poplar Leaf Segmentation Method Based on SegNet and 3D Point Cloud Clustering in Field. *Trans. Chin. Soc. Agric. Mach.* **2022**, *6*, 259–264.
20. Jin, S.; Su, Y.; Gao, S.; Wu, F.; Ma, Q.; Xu, K.; Hu, T.; Liu, J.; Pang, S.; Guan, H.; et al. Separating the structural components of maize for field phenotyping using terrestrial LiDAR data and deep convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *4*, 2644–2658. [CrossRef]
21. Li, Y.; Wen, W.; Miao, T.; Wu, S.; Yu, Z.; Wang, X.; Guo, X.; Zhao, C. Automatic organ-level point cloud segmentation of maize shoots by integrating high-throughput data acquisition and deep learning. *Comput. Electron. Agric.* **2022**, *193*, 106702. [CrossRef]
22. Kang, H.; Wang, X. Semantic segmentation of fruits on multi-sensor fused data in natural orchards. *Comput. Electron. Agric.* **2023**, *204*, 107569. [CrossRef]
23. David, S.; Federico, M.; Rosu, A.R.; Cornelißen, A.; Nived, C.; Stefan, P.; Jens, L.; Sven, B.; Cyrill, S.; Heiner, K.; et al. Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis. *PLoS ONE* **2021**, *8*, e0256340.
24. Zhang, Q.; Pang, Y.S.; Li, B. Visual positioning and picking pose estimation of tomato clusters based on instance segmentation. *Trans. Chin. Soc. Agric. Mach.* **2023**, *10*, 205–215.
25. Li, T.; Sun, M.; He, Q.; Zhang, G.; Shi, G.; Ding, X.; Lin, S. Tomato recognition and location algorithm based on improved YOLOv5. *Comput. Electron. Agric.* **2023**, *208*, 107759. [CrossRef]
26. Wang, S.P.; Liu, Y.; Guo, Y.X.; Sun, C.Y.; Tong, X. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Trans. Graph. TOG* **2017**, *4*, 1–11. [CrossRef]
27. Riegler, G.; Osman Ulusoy, A.; Geiger, A. Octnet: Learning deep 3d representations at high resolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3577–3586.
28. Tang, H.; Liu, Z.; Li, X.; Lin, Y.; Han, S. Torchspase: Efficient point cloud inference engine. *Proc. Mach. Learn. Syst.* **2022**, *4*, 302–315.
29. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
30. Zhu, T.; Ma, X.; Guan, H.; Wu, X.; Wang, F.; Yang, C.; Jiang, Q. A method for detecting tomato canopies' phenotypic traits based on improved skeleton extraction algorithm. *Comput. Electron. Agric.* **2023**, *214*, 108285. [CrossRef]
31. Bahdanau, D. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
32. Berman, M.; Triki, A.R.; Blaschko, M.B. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4413–4421.
33. Sun, Y.; Guo, X.; Yang, H. Win-Former: Window-Based Transformer for Maize Plant Point Cloud Semantic Segmentation. *Agronomy* **2023**, *11*, 2723. [CrossRef]
34. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
35. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114. [CrossRef]
36. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.* **2019**, *5*, 1–12. [CrossRef]
37. Sun, Y.Z.; Zhang, Z.X.; Sun, K.; Li, S.; Yu, J.L.; Miao, L.X.; Zhang, Z.G.; Li, Y.; Zhao, H.J.; Hu, Z.B.; et al. Soybean-MVS: Annotated three-dimensional model dataset of whole growth period soybeans for 3D plant organ segmentation. *Agriculture* **2023**, *7*, 1321. [CrossRef]
38. Darwish, M.A.; Elkot, A.F.; Elfanah, A.M.; Selim, A.I.; Yassin, M.M.; Abomarzoka, E.A.; El-Maghraby, M.A.; Rebouh, N.Y.; Ali, A.M. Evaluation of wheat genotypes under water regimes using hyperspectral reflectance and agro-physiological parameters via genotype by yield*trait approaches in sakha station, delta, egypt. *Agriculture* **2023**, *7*, 1338. [CrossRef]
39. Wang, X.W.; Wang, Z.F.; Qin, X.Z.; Chen, J.; Dai, J.P.; Zhang, H.T.; Xie, Y.Q.; Feng, L.; Gurinur, A.; Guo, W.C.; et al. The Growth Phenotype of Tomato: Response to Physical and Chemical Properties of Cotton Straw Mixed Seedling Matrix. *Chin. Agric. Sci. Bull.* **2020**, *20*, 36–43.

40. Xu, M.; Yoon, S.; Fuentes, A.; Park, D.S. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognit.* **2023**, *137*, 109347. [[CrossRef](#)]
41. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.