Computers and Electronics in Agriculture 194 (2022) 106691

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

LFPNet: Lightweight network on real point sets for fruit classification and segmentation

Qirui Yu [a,1], Huijun Yang [a,c,*], Yangbo Gao [a,1], Xinrui Ma [a], Guochao Chen [a], Xin Wang [b]

[a] College of Information Engineering, Northwest A&F University, Yangling 712100, Shaanxi, China
[b] College of Language and Culture, Northwest A&F University, Yangling 712100, Shaanxi, China
[c] Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling 712100, Shaanxi, China

ARTICLE INFO

Keywords:
Point Cloud
Deep Learning
3D Spatial Transformer Network
Fruit Classification
Fruit Segmentation

ABSTRACT

3D point cloud reconstruction, as the key technology to obtain high-throughput fruit phenotypic data, has solved the problems caused by complex environments, high fruit similarity, and the lack of public datasets suitable for fruit characterization. However, in the process of identifying and segmenting fruit data from point cloud, the existing network architectures lead to problems such as classification error, incomplete segmentation and low efficiency. In this paper, we introduce LFPNet, a novel and efficient lightweight neural network that directly consumes fruit point clouds in the real scene. Our network mainly has the following three advantages: 1) The introduction of voxel-filter based down-sampling preprocessing can help to avoid classification error caused by invalid noise interference. 2) A 3D STN is designed to solve the lack of spatial invariance in convolutional neural network (CNN) when calculating and analyzing fruit point clouds. 3) By introducing spatial pyramid pooling and combining local and global features, a fruit segmentation network is built to improve the integrity of segmentation in fruit scenes. Experimental results show that our LFPNet performs as well as or better than most of its peers in terms of classification accuracy and segmentation integrity.

1. Introduction

With the development of precision agriculture, how to improve the level of digitization and acquire fruit phenotypic structure accurately and timely has become the core issue of current standardized fruit production and quality improvement. At the same time, as 3D point cloud acquisition technology advances, traditional 3D scanners can no longer meet current needs and have problems such as high price, poor portability, low availability and difficulty in popularization. Kinect-based depth camera has the advantages of low cost, good portability, low entry threshold and rapid popularization. All of these advantages make it become the mainstream device for obtaining fruit point cloud in the current point cloud collection process. However, due to the influence of the environment, the obtained fruit point clouds contain many scattered points with complex backgrounds, such as leaves, branches and lighting. In three-dimensional fruit modeling, the background debris information not only reduces the accuracy of fruit measurement, but also increases the amount of data processing. The premise of accurate measurement is thus how to identify and segment debris from low-precision and unordered fruit point clouds with complex background.

Traditional point cloud segmentation algorithms are divided into 6 categories, edge-based (Sappa and Devy, 2001), region-based (Vo et al., 2015), attribute-based (Zhan et al., 2010), model-based (Schnabel et al., 2007); supervoxel-based (Papon et al., 2013) and graph-based (Golovinskiy and Funkhouser, 2009). Their feature detection and object segmentation are primarily based on geometric relationships, focusing on objects with sharp edges and corners, resulting in feature information loss in the area with fuzzy boundary. Meanwhile, prior knowledge of target object to be recognized is required.

In recent years, with the widespread application of deep learning in AI, CNNs have achieved excellent results in almost all 2D visual recognition tasks, and will fundamentally change the pattern of computer vision (CV) (Turaga et al., 2008). Simultaneously, thanks to CNNs' success in translation invariance, the same set of convolution filters can be applied to different positions in the image, reducing the number of parameters and improving the generalization ability. However, existing 3D point clouds are typically unordered, and traditional CNNs are incapable of adapting to such input. One alternative method is to process

* Corresponding author.
E-mail address: yhj740225@nwafu.edu.cn (H. Yang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compag.2022.106691
Received 4 June 2021; Received in revised form 25 November 2021; Accepted 5 January 2022
Available online 29 January 2022

the 3D space as a volume grid, but the volume grid is typically very sparse, making it difficult for CNN to calculate on this grid with high resolution. Inspired by the successful experience of end-to-end deep learning on point cloud, we propose LFPNet to solve the existing point cloud segmentation methods, which are hampered by the defects of fruit point cloud, complex background and sparse data, resulting in difficulty of feature extraction and low recognition rate. Indeed, our method has significant practical implications for the quantitative evaluation of fruit phenotypes, the application of AI in digital orchards, and the development of 3D computer vision, as demonstrated by experiments on apple, lemon, pear. The work of this paper is a critical technical guarantee for phenotypic character extraction, field automatic picking, yield estimation, post-harvesting sorting, and quality grading. Our contributions are as follows:

(1) We propose LFPNet, a lightweight fruit phenotypic processing network, to solve the problems of current 3D deep learning network in fruit phenotypic feature recognition, such as complex calculation and low segmentation accuracy. The extracted phenotypic parameters can be used to achieve accurate classification in practical postharvest scenarios such as fruit grading, fruit sorting, and automatic fruit weighing in unmanned supermarkets by identifying the surface traits of fruits.
(2) Based on 3D spatial transformation, we design a novel fruit segmentation network, which combines spatial pyramid pooling and global feature feedback to construct combined global information. The proposed segmentation network has a significant advantage in segmentation integrity, which is important in high-throughput phenotypic extraction, fruit sortation, automatic picking and other applications.
(3) Unlike existing deep learning methods for ideal labeled datasets, our LFPNet framework provides an effective solution for directly processing the real preharvest and postharvest environment of fruit point cloud with noise acquired by the KinectV2. In the experiment, we show how to use the voxel filtering based preprocessing method to avoid misclassification caused by invalid point interference.

2. Related work

2.1. Segmentation of CV in agricultural scenes

2.1.1. CV based digitization of crops

Guo et al. (Guo et al., 2013) used a series of assumptions to label different pixels in the image and extracted features for training to better segment the plant image. Khan et al. (Khan et al., 2018) proposed a new method for estimating vegetation index from RGB color images. The preceding studies primarily focus on 2D images, which frequently lose important information about real crops. With the advancement of sensors and computer technology, some methods for sensor-based 3D reconstruction for a wide range of applications have been developed (Vázquez-Arellano et al., 2016). A number of stereo vision-based methods for 3D modeling of plants have been developed, but evaluation of 3D sensing technologies and comparison of the most 2D and 3D neural networks show that the accuracy of stereo vision-based methods is unreliable because it is sensitive to crop texture (Song et al., 2014; Louedec et al., 2020). Structured light has been successfully used in high-accurate plant growth monitoring (Nguyena et al., 2016), but it is complicated and expensive, limiting its use in real fruit digitization. Consumer-grade depth sensors (Kinect known as the famous one) have gotten a lot of attention in recent years because of their low price, high frame rate and simultaneous acquisition of color and spatial information. However, there is significant noise in the Kinect-measured point cloud of real fruit (Corti et al., 2016).

2.1.2. Point cloud based crop segmentation

At the moment, there are few methods for crop point cloud segmentation at the farmland environment, while 3D segmentation of crop point cloud is the premise of phenotypic extraction, automatic picking, production estimation and quality classification. A Sweet Pepper Harvesting System (Lehnert et al., 2018) used an RGB-D camera to capture the entire 3D scene and then used a convolutional neural network and a 3D filtering algorithm to locate the target sweet pepper. Paproki et al. developed a cotton segmentation model, but the speed of multi-view stereo (MVS) limits it (Paproki et al., 2012). Lin et al. proposed a column spatial clustering segmentation method for field crops, but it is only suitable for plants with interval sowing (Lin et al., xxxx). Paulus et al. adopted a different method to classify grape leaves and stems, but it is more suitable for plants with considerable phenotypic differences (Paulus et al., 2013). Guo et al. proposed a Kd-trees-ICP algorithm for high-precision registration of 3D point cloud data of apple canopy (Guo et al., 2020). Chen et al. improved RandLA-Net to create a deeper segmentation network for large-scale unstructured agricultural scenes (Chen et al., 2021). It can be seen that robust 3D deep learning segmentation methods for preharvest and postharvest real fruit scenes are still lacking.

2.2. Segmentation of point cloud deep learning

2.2.1. Projection and voxel based methods

Inspired by the success of 2D CNNs, many works project 3D point clouds onto 2D images. Su et al. render 3D point clouds into multi-view 2D images and classify them using a well-designed MVCNN (Su et al., 2015). Alternatively, point clouds can be voxelized into 3D grids, represented by VoxNet, the unordered point clouds are voxelized into regular structures, and then 3D CNN is applied for object detection and semantic segmentation (Maturana and Scherer, 2015). This method solves the problem of unstructured point clouds, but it is constrained by the 3D voxel grid resolution and the computational cost. Graham et al. create a sparse CNN and apply it to 3D segmentation tasks (Graham, 2014). Li et al. try to sample the sparse 3D data and then feed the sampling results into the network for processing to reduce the amount of calculation (Li et al., 2016). Klokov et al. and Riegler et al. propose spatial division methods of Kd-tree (Klokov and Lempitsky, 2017) and Octree (Riegler et al., 2017) respectively, to solve the spatial resolution of the voxel grid. However, these methods only rely on the voxel boundary without considering its local geometric structure. SEGCloud combines the advantages of standard voxel-based 3D-FCNN, trilinear interpolation (TI), and fully connected Conditional Random Fields (FC-CRF) to achieve effective semantic segmentation (Tchapmi et al., 2017). Le et al. propose PointGrid, which uses 3D CNNs to learn grid cells containing fixed points to obtain local geometric details (Le and Duan, 2018). In addition, Hua et al. propose a 3D convolution operator based on a unified grid kernel for point cloud semantic segmentation and target recognition (Hua et al., 2018). Louedec et al. directly apply 2D convolutions and traditional CNN to a point cloud grid and successfully recover surface and manifold from grid points using convolutions and pooling functions (Louedec et al., 2020). These methods, however, are still incapable of resolving the quantization artifacts caused by voxelization.

2.2.2. Point based methods

Influenced by the end-to-end thinking, many recent works use deep learning methods to directly process point clouds. PointNet (Qi et al., 2017) is a pioneering work that uses a shared MLP to extract the characteristics of each point by discussing the disorder, permutation invariance and symmetry of the point set. However, since the feature is learned through the entire point clouds, its local features cannot be captured. As a result, Qi et al. proposed PointNet++ (Qi et al., 2017); a hierarchical network, to integrate the local area division network into the original network, effectively solving the problem of local feature

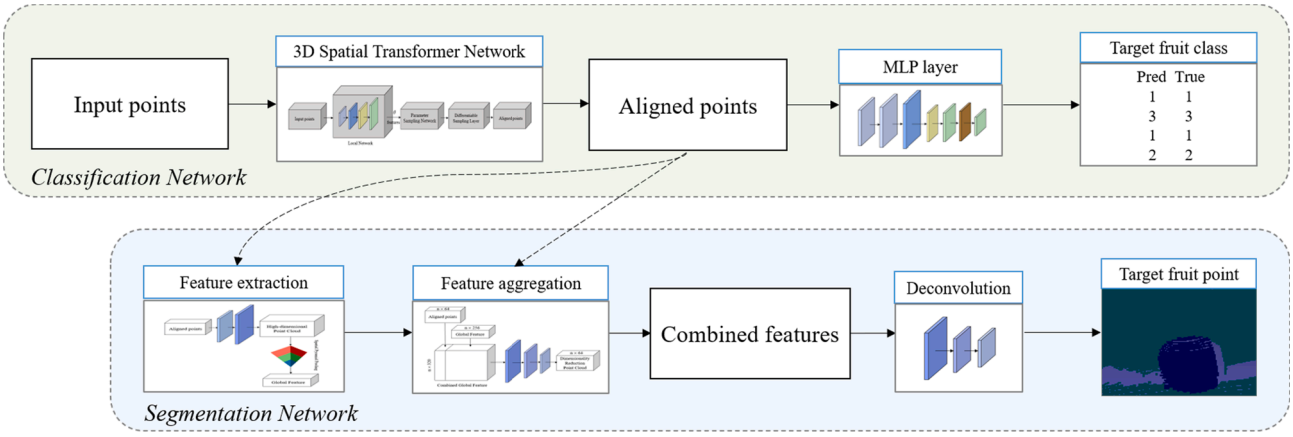


Fig. 1. LFPNet framework. The classification network takes n points as input, applies 3D spatial transformation, and then aggregates point features by max pooling (in MLP layers). The output is classification scores for the target class. The segmentation network is an extension to the classification net. It concatenates global and local features and outputs per point scores. "MLP" stands for multi-layer perceptron, where Dropout layer is used to improve the accuracy in classification net.

extraction and improving its segmentation effect. Furthermore, Hermosilla et al. and Wu et al. propose the 3D convolution Monte Carlo approximation method (Hermosilla et al., 2019) and the PointConv (Wu et al., 2019) method respectively, which both further realize the extraction of local features by taking density into account. PointConv, as opposed to the former, estimates the learning density function by kernel density. Aiming at the unorder of point clouds, Li et al. propose PointCNN (Li et al., 2018), which re-arranges the relevant features of each point through a χ-conv to solve the information loss caused by direct convolution. Jiang et al. propose a PointSIFT (Jiang et al., 2018) module that can be embedded in a basic network. It encodes the information of 8 main directions using a direction encoding unit to obtain multi-scale features.

According to the above, Projection-based methods require projection and 2D or 3D convolution to achieve shape classification, which will inevitably increase the computational complexity, and may result in the loss of some effective information. Point-based methods, on the other hand, work directly on raw point clouds, do not introduce explicit information loss, and are becoming increasingly popular. However, due to irregularity, occlusion, and susceptibility of the preharvest and postharvest environments, the real fruit point cloud with outliers requires a large amount of ideal labeled data to train, resulting in a low efficiency problem of classification and segmentation. To address the high computational complexity and low segmentation accuracy of current 3D deep learning methods in recognizing fruit phenotypic features, a lightweight 3D point cloud network for tasks with low precision, unordered and complex background has become an urgent need. Our LFPNet framework provides an effective solution for directly processing the real fruit point cloud with noise acquired by the KinectV2.

3. LFPNet framework

Affected by the accuracy of the depth camera and the complexity of the environment, the current 3D deep learning methods have issues such as a large amount of calculation, slow convergence speed, poor segmentation integrity, and inaccuracy in understanding and recognition. Inspired by the successful application of PointNet (Qi et al., 2017) in end-to-end 3D point cloud deep learning, we propose LFPNet, which has significantly advanced 3D point cloud deep learning in accurate crop recognition. Our full network architecture is visualized in Fig. 1, where the classification network and the segmentation network share a great portion of structures. Our network mainly includes five key modules: 3D spatial transformer network (3D STN), feature extraction, feature aggregation, classification and segmentation network, which will be discussed in separate section below.

(1) **3D Spatial Transformer Network:** The first step of our framework is to introduce the self-designed 3D spatial transformer network (3D STN) to ensure that the spatial feature map can be adaptively transformed, so as to achieve the spatial invariance of point clouds in convolutional neural network (CNN).
(2) **Feature Extraction:** In order to extract more complete global features, a point cloud dimensionality reduction method based on 3D spatial pyramid pooling (3D SPP) is proposed.
(3) **Feature Aggregation:** By combining aligned points and global features, we can create a combined global information that can be used to achieve accurate segmentation.
(4) **Classification Network:** A lightweight fruit point cloud classification framework, consisting of a 3D STN and a MLP layer, is

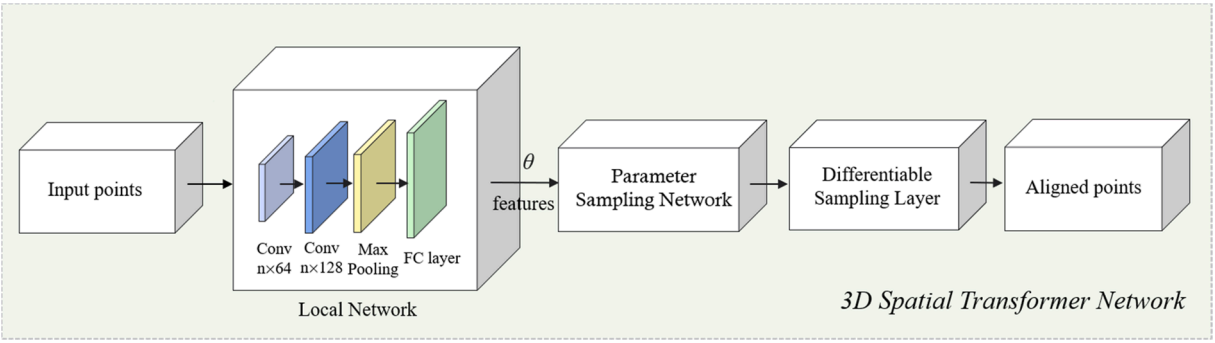


Fig. 2. 3D spatial transformer network. We generate the transformation parameter θ through the local network, then the parameter θ and some features of the original point cloud are used as input of the parameter sampling network. Finally, the aligned points can be obtained by the differentiable sampling layer.

designed on the basis of invalid point removal and down-sampling simplification.

(5) **Segmentation Network:** An efficient framework for fruit point cloud segmentation is designed by combining the 3D STN, feature extraction, feature aggregation and deconvolution module.

3.1. 3D spatial transformer network for unordered point cloud

In light of lacking spatial invariance of convolutional neural networks, before calculating and analyzing, we propose a spatial invariance network: 3D STN on the original point clouds in Eq. (1).

$$\{C_i \in R^n | i = i, 2, \cdots n\} \tag{1}$$

The currently available spatial invariance strategies mainly include: 1) Adding a transfer function for spatial transformation, but the transfer function typically lacks adaptability and requires manual adjustment of the corresponding transformation parameters; 2) Performing spatial transformation through the pooling layer, the received data is fixed and partial, resulting in features being lost during the transformation and cannot be applied to the local network. Based on above, we design a 3D spatial transformer network for unordered point cloud by expanding the transformer network from 2D to 3D, which integrates local network, parameter sampling network, differentiable sampling layer, and obtains the aligned fruit point cloud, as shown in Fig. 2.

3.1.1. Local network

Taking the original point cloud in Eq. (2) as input, we perform affine transformation through the convolutional layers, the max pooling layer and the fully connected layer in the local network to extract the features of the original point cloud and finally generate the transformation parameters θ in Eq. (3).

$$\left\{C_i \in \mathrm{R}^{\mathrm{B}\times\mathrm{H}\times\mathrm{W}\times\mathrm{Ch}} | \mathrm{i} = 1, 2, \cdots \mathrm{n}\right\} \tag{2}$$

$$\theta = floc(C_i)(C_i \in \mathrm{R}^{\mathrm{n}}, \mathrm{n} = \mathrm{B} \times \mathrm{H} \times \mathrm{W} \times \mathrm{Ch}) \tag{3}$$

Where, θ is the transformation parameter, C_i is the input original point cloud, B is the batch size of the original point cloud, H is the height, W is the width, and Ch is the number of channels. In this section, convolutional and pooling layers are employed to implement the function $floc(*)$ (Jaderberg et al., 2015) to ensure that the dimension of θ generated by the 3D point cloud is $[B, Ch, Ch]$ and can perform the spatial transformation on Eq. (2).

3.1.2. Parameter sampling network

To prevent the influence of dirty data from being trapped into the local optimal, the coordinates in Eq. (2) are first converted to the standard coordinates, as in Eq. (4).

$$(X_{si}, Y_{si}, Z_{si}) = k(X_{ti}, Y_{ti}, Z_{ti}) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{bmatrix} (X_{ti}, Y_{ti}, Z_{ti})$$

Where, (X_{ti}, Y_{ti}, Z_{ti}) is the original point cloud coordinates, (X_{si}, Y_{si}, Z_{si}) is the standardized coordinates that defines the sample point, and θ is the transformation parameter in Eq. (3).

Then, by fusing the transformation parameter θ and the original point cloud information, we correct the original point cloud coordinates in Eq. (1), and further transform the position information to achieve spatial alignment.

3.1.3. Differentiable sampling layer

The differentiable sampling layer is used to convert the point set in Eq. (2) into aligned point cloud.in order to ensure the consistency and efficiency of the fruit point cloud spatial transformation, In this section, bilinear interpolation is used to improve the consistency of the point cloud transformation space, and the spatial coordinates of the sampling kernel are used to obtain the coordinates of the aligned points. The weight, on the other hand, may appear in decimal while the subscript appears in integer, which is processed by the rounding method to prevent the gradient from invariance and obtain the converted coordinates. Simultaneously, to maintain the spatial consistency between channels Ch, each channel of the input point is converted in the same manner, as shown in Eq. (5).

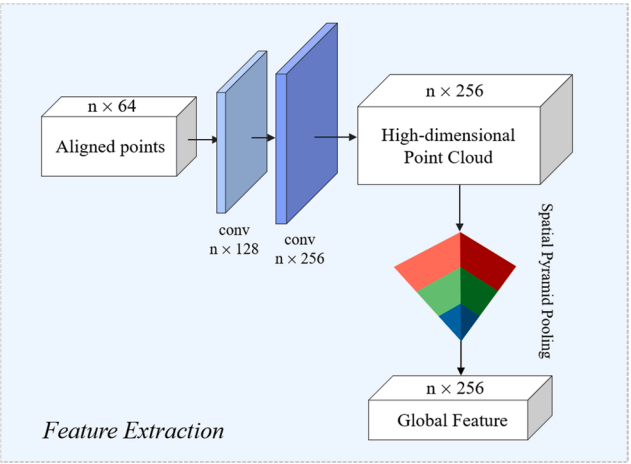


Fig. 3. Global feature extraction. First, 3D convolution is employed to raise the dimension of the aligned point cloud for feature redundancy, and then 3D spatial pyramid pooling is used to accurately extract global features.

$$V_i^C = \sum_n^H \sum_m^W U_{nm}^C \delta\left(\lfloor x_i^s + 0.5 \rfloor - m\right) \delta\left(\lfloor y_i^s + 0.5 \rfloor - n\right) \tag{5}$$

$$\delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \tag{6}$$

$$\{V_i \in \mathrm{R}^{\mathrm{n}} | \mathrm{i} = 1, 2, \cdots, \mathrm{n}\} \tag{7}$$

Where, δ in Eq. (6) is the Kronecker delta function, which converts the coordinates of the nearest point to the output position of (X_{si}, Y_{si}, Z_{si}). Finally, we obtain the aligned points in Eq. (7).

3.2. Spatial pyramid pooling based feature extraction

As shown in Fig. 3, we propose a 3D spatial pyramid pooling (3D SPP) method based on point cloud dimensionality raising to extract the global features of fruit point cloud as accurately and completely as possible, reduce over-fitting, and improve generalization ability.

3.2.1. Preprocessing

To keep integrality of fruit information and avoid over-fitting, a dimensionality raising operation is introduced to extract the global information of the aligned point cloud. In light of the chaotic point cloud scene, a 3D convolution method is introduced to increase the channels Ch in Eq. (2) for feature redundancy, allowing us to obtain complete global features even if some information is lost in the subsequent pyramid pooling process.

3.2.2. Spatial pyramid pooling

The methods available for extracting point cloud information primarily include: max pooling, average pooling, and other pooling strategies. Although the effect of max pooling layer is superior to that of the average pooling layer, the information extracted by both is insufficient to obtain a better fruit point cloud segmentation effect. In order to accurately extract high-dimensional fruit features from preprocessed aligned point cloud scenes while also ensuring the model's generalization ability, we propose a three-level 3D pyramid pooling model. To

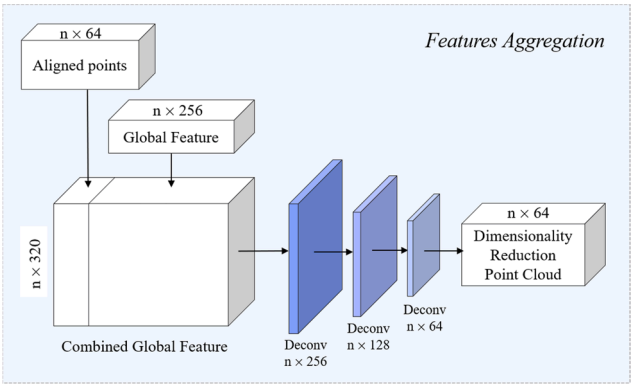


Fig. 4. Fruit feature aggregation. Each fruit point is classified by combining global features and aligned points, and the deconvolution neural network is then used to reduce the dimension of the combined global features for final fruit segmentation.

begin, by aggregating local features from coarse to fine, the entire point cloud can be more precisely divided into blocks, denoted as different subsets with similar features. Based on this, the pyramid pooling layer with three-level pooling performs local feature extraction on each fruit block by designing the adaptive pooling window. Finally, we obtain complete and accurate global fruit features by aggregating the aforementioned features extracted from each subset.

3.3. Feature aggregation

In this section, a novel feature aggregation method for better fruit feature matching is presented, which combines extracted global feature from Section 3.2 and aligned points (local feature) from Section 3.1 to construct complete information for later fruit segmentation.

The combined global information is obtained by feeding back the extracted global feature to the feature of each aligned fruit point. Then, based on the combined global information, each point is recognized. Because of the inaccurate segmentation result caused by high-dimensional label matching, a convolutional neural network based on dimensionality reduction is finally used to restore the fruit point cloud to the correct dimension. As shown in Fig. 4, sophisticated information of the point cloud will support the accurate fruit segmentation.

3.4. Classification of unordered fruit point cloud

A lightweight classification network for post-harvest fruit scenes is designed to address the problem of low accuracy in current 3D depth

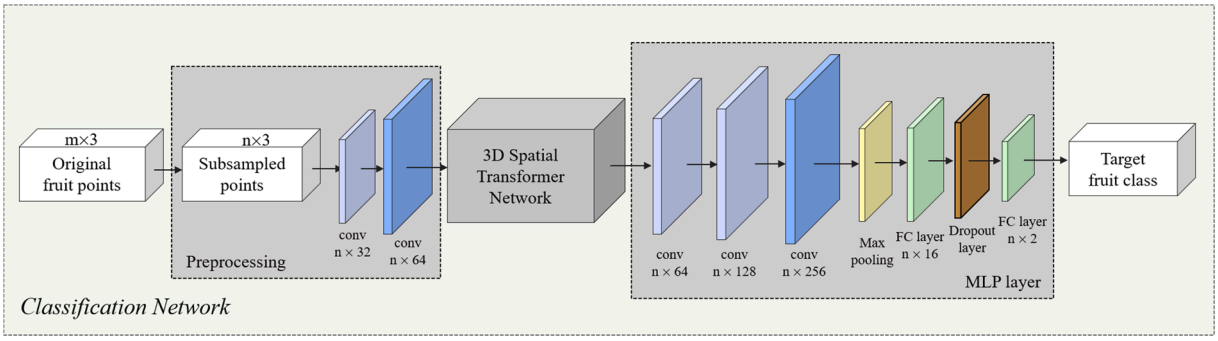


Fig. 5. Fruit classification network. The preprocessed fruit point cloud is spatially transformed to obtain an aligned point cloud. Convolution is used on aligned points for information redundancy and ensure the integrality of feature information. The global features are then extracted using the max pooling layer, and the fully connected layers are finally used for classification.

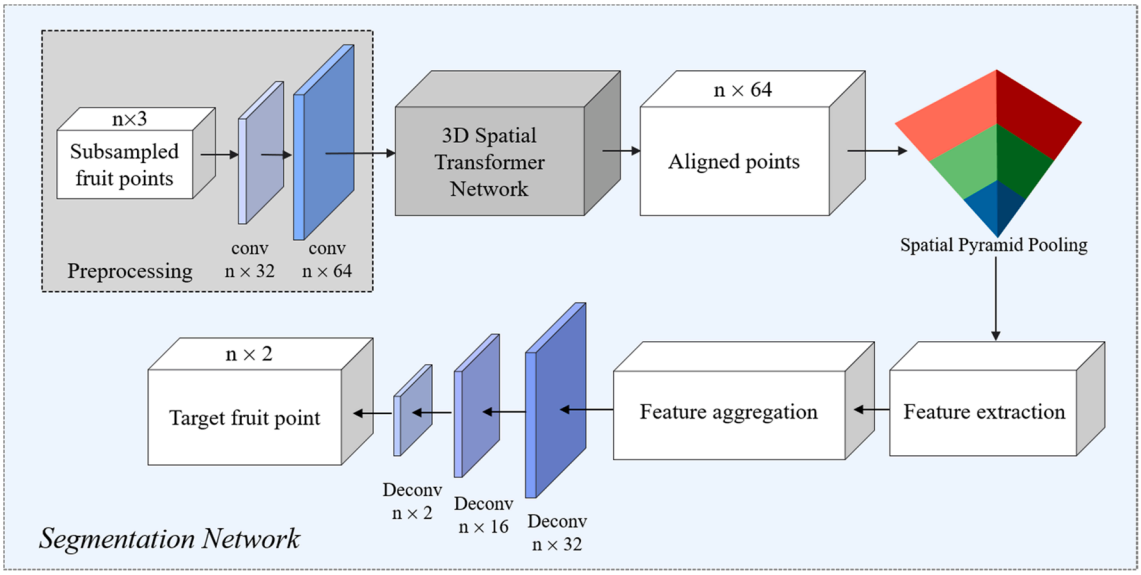


Fig. 6. Fruit segmentation network. The preprocessed fruit point cloud is first transformed using the proposed 3D STN, and then the global fruit features are extracted from high-dimensional space using the dimensionality raising based 3D SPP. Finally, feature aggregation is used to feedback global fruit features to aligned fruit points in order to perform target fruit segmentation.

understanding and perception methods in fruit classification. The network architecture is shown in Fig. 5.

3.4.1. Preprocessing

Restricted to the depth camera, there are many invalid and unordered points which will result in fruit misclassification, we design a preprocessing stage before inputting the classification network. First of all, the interactive method is used to remove invalid points to avoid efficiency loss during the training process. Secondly, in order to avoid excessive memory usage consumed by a large number of points, voxel filtering is introduced to down-sample the fruit point clouds as shown below. ① In 3D space, a voxel lattice is created, and all fruit points within the lattice are roughly replaced by the barycenter to form the filtered point clouds. ② The points are sampled down to a fixed number under the premise of retaining the original fruit geometric information by using voxels size as the threshold for controlling down-sampling. ③ We avoid misclassification and feature loss as a result of removing too many fruit features during uneven sampling in this manner.

3.4.2. Classification

The following is how we designed our lightweight fruit classification network. ① Based on the above preprocessing, the labeled fruit point cloud is aligned by the 3D STN in Section 3.1 to address the problem of the CNNs' lack of spatial invariance when processing the original fruit point cloud. ② A MLP based dimensionality raising layer is designed to produce the target scores of fruit point cloud. The redundancy local features of the fruit point clouds are extracted first by using CNN to perform dimensionality raising. The global features of the fruit points are then extracted by a max pooling layer. We finally map the global features of the fruit point cloud to the aligned points with redundancy features using fully connected layers to obtain target scores of fruit features. A Dropout layer is also introduced between the fully connected layers to address the problem of over-fitting caused by the complexity of the CNN and noise interference in the fruit point cloud.

The extracted phenotypic parameters can be used for later segmentation, variety identification, germplasm resource investigation, and so on, by identifying the surface traits of fruits and then using the acquired phenotypic parameters to achieve accurate classification, As a result, we can avoid the influence of complex backgrounds and improve the recognition accuracy of our classification net, which can obtain an accurate and efficient fruit point cloud classification result.

3.5. Segmentation of unordered fruit point cloud

Although current 3D deep learning methods have made significant progress in regular point cloud segmentation, they are still incapable of addressing the problem of fruit segmentation inaccuracy from unordered point clouds. As a result, we propose a lightweight fruit segmentation network, as shown in Fig. 6.

The method described in Section 3.4 is used to remove the invalid points from the fruit point clouds. In contrast to classification, we design a double voxel filtering to reduce the number of sampling points and thus improve the efficiency. Following that, the segmentation net is designed as follows on the basis of fruit point cloud in Eq. (1).

(1) In Section 3.1, a 3D STN is used to obtain an aligned point cloud in Eq. (7) to solve the CNNs' lack of spatial invariance when processing the original fruit point cloud.
(2) The 3D SPP proposed in Section 3.2 is used to extract global fruit features from aligned points, which can obtain a more complete fruit features by dimensionality raising for the later fruit segmentation from complex background.
(3) The feature aggregation in Section 3.3 is used to feed the global fruit features back to the local features of aligned points. Thus, while in high-dimensional space, we can obtain a well-segmented fruit point cloud.

Table 1
Capture parameters.

Attribute		Parameter
Color	Resolution	1920*1080
	fps	30
Depth	Resolution	512*424
	fps	30
Angle	Horizontal	70°
	Vertical	60°
Range of detection		0.5 ~ 4.5 *m*

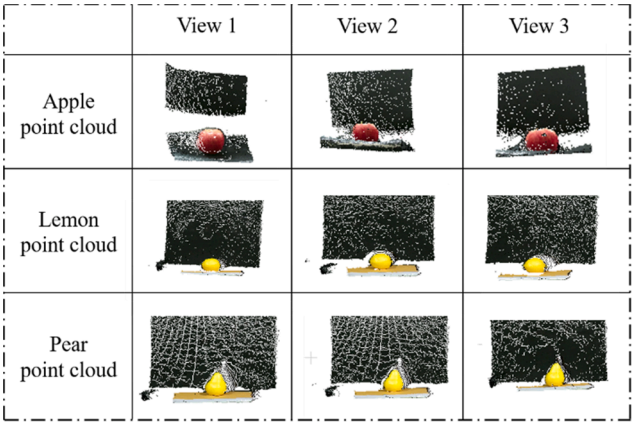


Fig. 7. Part of the original point cloud of postharvest fruit from multi-view.

(4) A dimensionality reduction operation is performed on the high-dimensional fruit point clouds using a deconvolution module to obtain the fruit segmentation results on original dimension.

4. Experiment

In this section, we evaluate the proposed LFPNet on fruit point clouds in four ways. First, a description of the collection and preparation of fruit datasets is provided. Then, we provide detailed training, testing and evaluation to validate our classification and segmentation network. The model is then evaluated using three metrics: accuracy of each category (Acc for classification and segmentation), Intersection over Union (IoU for segmentation) and overall accuracy (OA for segmentation). Finally, an ablation study is performed to demonstrate the comparable or even superior performance of the proposed 3D STN and 3D SPP, and visualization results are provided.

4.1. Dataset collection and preparation

Although several 3D scene datasets have been made available for deep learning, there is no deep learning fruit datasets that are directly applied to agriculture (Chang et al., 2015; Armeni et al., 2017, 2016; Hua et al., 2016; Georgakis et al., 2016). In order to make the proposed algorithm more useful, we created a fruit dataset using the consumptive depth camera KinectV2 and multi-view stereo (MVS) method. The KinectV2 camera is chosen to acquire fruit point clouds of 1000 scenes from 5 to 6 perspectives, with more than 20,000 points per fruit scene. Table 1 shows the collection parameters, and Fig. 7 and Fig. 8 show a portion of the obtained fruit point cloud. We preprocess the acquired point cloud by the voxel subsampling to ensure the accuracy of feature extraction and to avoid misclassification due to many invalid points in the fruit point clouds, where the points number in the original point cloud reached more than 20,000.

In the classification, the label of apple is set to 1, pear is set to 2, and lemon is set to 3. In the segmentation, the semantic segmentation editor is used to mark the preprocessed fruit point cloud, which is classified

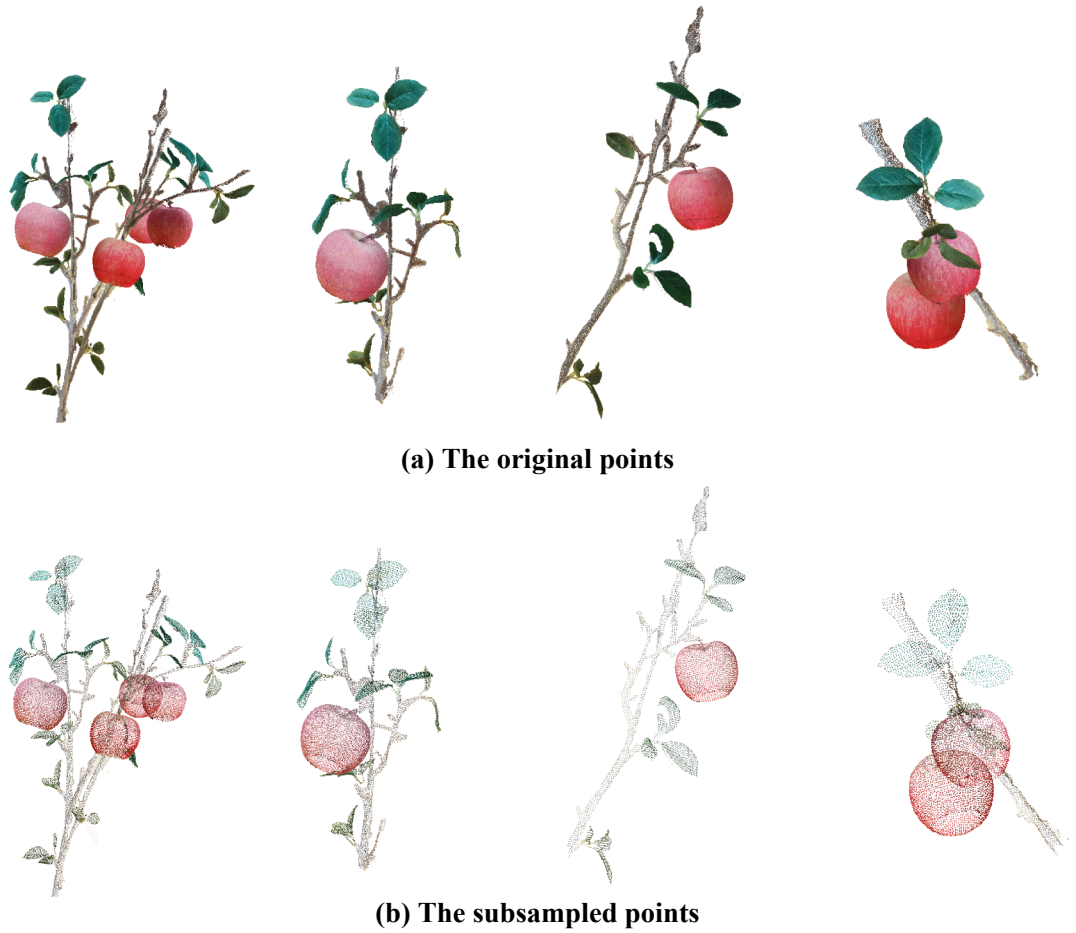


Fig. 8. Part of the original point cloud of preharvest fruit (collected from multi-view). To make it easier to visualize, we divide the entire point cloud into different parts and provide unobstructed views.

Table 2
Label mapping of fruit point cloud.

	Apple	Lemon	Pear	Background
Classification label	1	2	3	–
Segmentation label	1	2	3	0

Table 3
The contrast of fruit point number in preprocessing.

Category	Original points	Invalid point removal	Voxel grid filter
Apple 1	18,788	7912	4096
Apple 2	19,043	8064	4096
Apple 3	19,257	8912	4096
Lemon 1	19,836	7924	4096
Lemon 2	20,584	7962	4096
Lemon 3	21,396	7984	4096
Pear 1	22,456	8082	4096
Pear 2	23,680	7806	4096
Pear 3	24,085	9025	4096
Branch 1	733,344	12,263	4096
Branch 2	362,417	8182	4096
Branch 3	842,882	9981	4096
…	…	…	…

into fruit part and background part. Apple, lemon and pear are marked as the same label with classification, while the background is labeled as 0. Table 2 displays the results. The marked data is saved separately as point cloud files and label files.

As shown in Table 3, we contrast the point number in raw point clouds, outlier removed point clouds, and voxel filtered point clouds.

Based on the aforementioned files, point clouds and the label are encapsulated in the h5 file, as shown in Table 4, and are used for fruit classification and segmentation. Whereas h5 files, as a common format for point cloud deep learning, provide a conversion interface from Python to HDF5 lib, simplifying data processing and storing thousands of datasets in a single file for effective sorting and labeling. Size denotes the scale of the dataset, PoNum is the number of points in the point cloud,

Table 4
Fruit point cloud datasets in h5. (Part).

Type	Name	Size (Mb)	PiNum	PoNum
Segmentation	Fruit_train_0	16.9	120	4096
	Fruit_train_1	16.9	120	4096
	Fruit_train_2	16.9	120	4096
	Fruit_test	20.9	150	4096
Classification	Fruit_train_cla_0	50.6	350	4096
	Fruit_train_cla_1	50.6	350	4096
	Fruit_test_cla	50.6	350	4096

Table 5
Hardware and software configuration.

Name	Parameter
CPU	Intel core i7
GPU	Nvidia GeForce RTX 2080 Ti
RAM System	64 Gb Linux (Ubuntu 16.04)
CUDA cuDNN Pytorch Tensorflow Language	9.0 7.6.1 1.3 1.13.1-gpu Python 3.5

Table 6
Classification results on our dataset.

Model	Input	Apple	Pear	Lemon	mAcc
PointNet (Qi et al., 2017)	point	85.4	73.9	82.2	80.5
PointNet++ (Qi et al., 2017)	point	82.6	**80.9**	84.3	82.6
Ours LFPNet	point	**89.8**	75.0	**88.4**	**84.4**

and PiNum is the number of point cloud pieces in each dataset.

4.2. Network training

In this section, we illustrate how our LFPNet framework can be trained to perform the fruit point cloud classification and segmentation on the same preharvest and postharvest scenes, and our configuration environment is shown in Table 5.

4.2.1. Fruit point cloud classification

Our classification model can learn global features of point clouds that can be used for fruit classification, which is evaluated by comparing the accuracy of each category (Acc). There are 1,000 h5 models from three object categories: apple, pear and lemon, split into 700 for training and 300 for testing. It can be seen from the quantitative results in Table 6, we get a 3.9% improvement in mAcc when compared to PointNet (Qi et al., 2017), and a 1.8% improvement than PointNet++ (Qi et al., 2017). The results show that, when compared to other networks, we can better reduce misjudgments and improve the fruit point cloud classification accuracy.

4.2.2. Fruit point cloud segmentation

In the fruit point cloud segmentation process, our proposed segmentation framework is compared in terms of accuracy of each category (Acc), overall accuracy (OA), and Intersection over Union (IoU). We tested our model on the fruit segmentation dataset created in Section 4.1. There are 700 h5 models divided into 510 for training and 190 for testing from four object categories: apple, pear, lemon, and background.

According to the quantitative results in Table 7, our proposed fruit segmentation network outperforms most typical networks in real-world scenarios in all metrics. Among them, the IoU improves by 1% over PointNet (Qi et al., 2017), 3.8% over Pointwise (Hua et al., 2018), and 4.2% over DGCNN (Wang et al., 2019). We see a 3.2% improvement over PointNet (Qi et al., 2017) and a 1.1% improvement over PointNet++ (Qi et al., 2017) in OA. While the proposed framework improves fruit prediction accuracy by 3.7% over PointNet (Qi et al., 2017) and 3.4% over Pointwise (Hua et al., 2018). The results show that the preprocessing method and 3D SPP module introduced in our segmentation network have benefits in terms of segmentation integrality and efficiency. Table 8 compares the performance of our LFPNet framework to other networks in terms of parameters, conv-layer, pooling and Dropout, demonstrating the lightweight (represented by FLOPs) of our best design.

4.3. Ablation research

We take the apple dataset as an example and compare each design to several baselines to better understand the impact of our various designs on prediction accuracy. Table 9 and Table 10 show the comparison results of various models and our proposed model.

4.3.1. Ablation study on classification

Given the importance of 3D STN in feature mapping, we take the spatial transformation network and preprocessing as the baselines in the classification network to compare no 3D spatial transformer networks with MLP and preprocessing (*NoSTN*), and 3D spatial transformer networks + MLP without preprocessing (*STN + MLP*). Because the 3D STN and MLP modules are all included in the PointNet (Qi et al., 2017) network, they can be compared without preprocessing. As shown in Table 9, without STN, the classification accuracy of the network has dropped by 6.6% compared to our LFPNet, and the overall accuracy (OA) has dropped by 3% compared to ours. This confirms the superiority of STN in solving accurate classification of small-scale low-precision fruit point clouds.

It can also be seen that in models using STN + MLP without preprocessing, the classification accuracy has decreased by 10.2%, while the overall accuracy has decreased by 8%. When compared to PointNet

Table 8
Performance analysis. (The M stands for million.)

Model	Parameter	FLOPs	Conv-layer	Pooling	Dropout	Acc
PointNet (Qi et al., 2017)	4.01 M	284 M	6	Max-Pooling	0.7	76.5
PointNet++ (Qi et al., 2017)	6.21 M	884 M	6	Max-Pooling	0.7	78.6
Ours_1	1.05 M	180 M	4	Max-Pooling	0.7	76.4
Ours_2	1.17 M		6	Mean-Pooling	0.7	79.1
Ours_3	1.25 M		6	Max-Pooling	0.5	80.2
Ours_4	1.75 M		8	Max-Pooling	0.7	80.7
Ours_best	**1.17 M**		**6**	**Max-Pooling**	**0.7**	**80.6**

Table 9
Ablation study on classification.

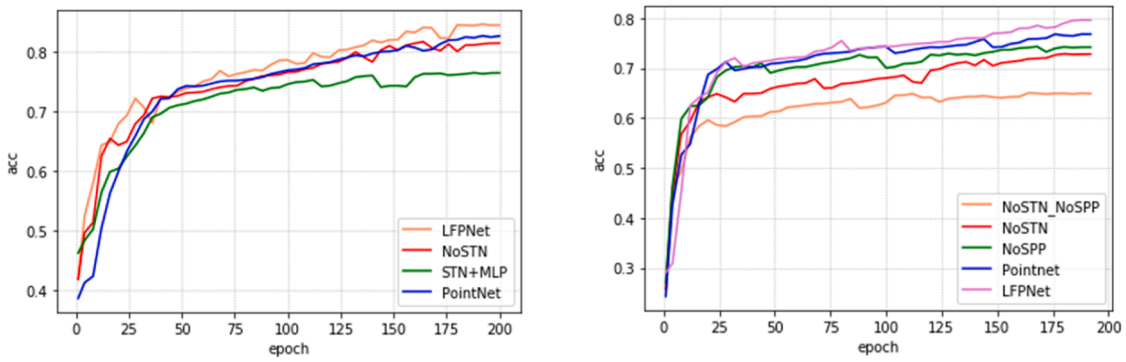
Model	*Preprocessing*	*Acc*	*OA*
NoSTN	√	83.2	81.4
STN + MLP	–	79.6	76.4
PointNet (Qi et al., 2017)	–	85.4	82.6
LFPNet (Ours)	√	**89.8**	**84.4**

Table 10
Ablation research on fruit segmentation.

Model	*OA*	*IoU*
NoSTN, NoSPP	64.9	60.9
NoSTN	72.8	70.5
NoSPP	74.2	72.5
PointNet (Qi et al., 2017)	76.8	75.4
LFPNet (Ours)	**79.6**	**78.2**

Table 7
Results on semantic segmentation in scenes.

Model	Input	Apple	Pear	Lemon	mAcc	OA	mIoU
PointNet (Qi et al., 2017)	point	82.3	71.5	72.4	76.5	76.1	75.4
PointNet++ (Qi et al., 2017)	point	81.2	70.6	**78.6**	78.6	78.2	**76.8**
DGCNN (Wang et al., 2019)	point	79.5	72.3	64.8	75.1	74.3	72.2
Pointwise (Hua et al., 2018)	point	80.1	76.2	61.5	76.8	75.9	72.6
Ours LFPNet	point	**86.2**	**78.4**	64.6	**80.2**	**79.3**	76.4



(a) Classification accuracy (b) Segmentation accuracy

Fig. 9. Comparison of the accuracy of fruit classification and segmentation in ablation study. The horizontal direction represents the number of training layers, and the vertical coordinate represents classification or segmentation accuracy.

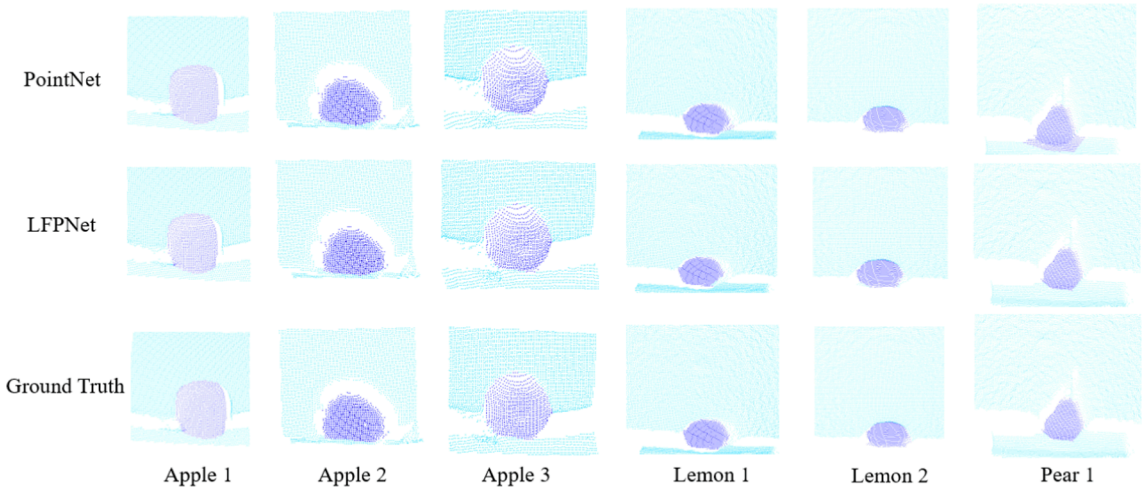


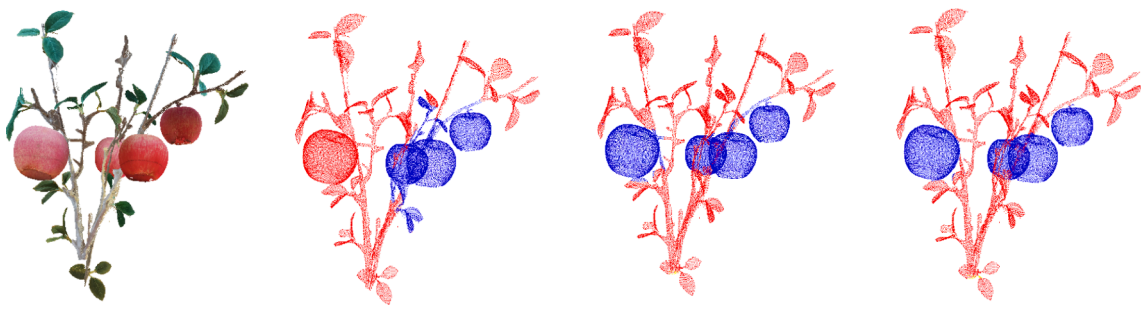
Fig. 10. Visualization of postharvest environment. Comparison of the fruit segmentation between PointNet (Qi et al., 2017); LFPNet (Ours) and the Ground Truth. We color-code the segmented information of apple, lemon and pear from left to right, and demonstrate the superiority of our segmentation network.

(Qi et al., 2017) without preprocessing, our network's classification accuracy has dropped by 4%, and overall accuracy has dropped by 1.8%. These fully validated the significance of our preprocessing for the fruit point clouds acquired by the depth camera.

Simultaneously, a line graph is plotted in Fig. 9(a) to illustrate the overall classification accuracy of the ablation study in the four cases, demonstrating the superiority of our classification model.

4.3.2. *Ablation study on segmentation*

We compare the effects of 3D STN, 3D SPP and various combinations on network performance using 3D STN and 3D SPP as baselines respectively. As shown in Table 10, we removed 3D STN and only considered SPP on fruit segmentation (*NoSTN*) to exploring the benefits of spatial transformation networks. The performance of the network without STN is reduced by 7.3% when compared to our segmentation network, demonstrating the superiority of our STN for segmenting



(a) (b) (c) (d)

Fig. 11. Visualization of preharvest environment. (a): Original point cloud, (b): PointNet segmentation result, (c): LFPNet segmentation result (Ours), (d): The Ground Truth.

small-scale low-precision fruit point clouds. By removing pyramid pooling and only considering the impact of STN on fruit segmentation in no 3D SPP with 3D STN (*NoSPP*), the performance is reduced by 5.7% compared to our segmentation net. By removing both STN and SPP from *NoSTN*, *NoSPP*, the network performance has dropped by 15% when compared to our segmentation net. The results confirmed the benefits of combining STN and SPP in our segmentation network.

In Fig. 9(b), to demonstrate the superiority of our segmentation framework, a line graph is plotted to compare the segmentation accuracy of the ablation study in five cases.

4.4. Visualization of our LFPNet

In this paper, we visualize the segmentation effect in Figs. 10 and 11. As shown in the visualization, our LFPNet can accurately segment fruit from the background and obtain on par or better segmentation results when compared to PointNet (Qi et al., 2017) and the Ground Truth (GT).

5. Conclusion

In this paper, we proposed a novel lightweight 3D point cloud deep learning network, LFPNet, which provides an effective solution for directly processing the real preharvest and postharvest environment of fruit point cloud with noise acquired by the KinectV2. Our results indicates that, by combining preprocessing, 3D STN, 3D SPP and global feature aggregation feedback into deep learning network, the proposed framework can avoid fruit misclassification and ensure the integrity of fruit segmentation caused by invalid noise interference. The proposed fruit classification net is of great value in practical scenarios such as fruit grading, non-destructive traceability and automatic fruit weighing in unmanned supermarkets, while our fruit segmentation net is extremely useful in high-throughput phenotypic extraction, fruit sortation and automatic picking. The experimental results on real fruit scenes demonstrate that, despite the fact that the proposed LFPNet is performed on realistic scenes, our model outperforms state-of-the-art deep learning models in most cases.

Despite the fact that our LFPNet is useful for phenotypic character extraction, field automatic picking, yield estimation, post-harvesting sorting, and quality grading. However, because the collection equipment operates in a facility environment, the data learning effect is limited to natural light environment and requires offline preprocessing, which is difficult to complete in single step. At the same time, our fruit point cloud deep learning framework cannot be deployed to mobile devices and cannot be used for on-site detection directly in this stage. We will continue to perform experimental verification in natural lighting scenes In the future to improve our model's landing in chaotic field scenes.

CRediT authorship contribution statement

Qirui Yu: Methodology, Project administration, Software, Investigation, Writing – original draft, Data curation. **Huijun Yang:** Supervision, Validation, Project administration, Resources, Funding acquisition, Investigation, Writing – review & editing. **Yangbo Gao:** Methodology, Project administration, Software. **Xinrui Ma:** Writing – original draft, Software, Writing – review & editing. **Guochao Chen:** Writing – original draft, Data curation, Investigation, Software. **Xin Wang:** Funding acquisition, Investigation, Resources, Data curation.

Declaration of Competing Interest

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

Acknowledgement

Research supported by Foundation of Key Research and Development Program of Shaanxi province (2021NY-179, 2019ZDLNY07-02-01, 2020NY-205), Undergraduate Training Program for Innovation and entrepreneurship plan (S202010712063, X202110712259, X202110712259, S202110712613).

References

Armeni, I., Sax, S., Zamir, A.R., Savarese, S., 2017. Joint 2D3D-Semantic Data for Indoor Scene Understanding. ArXiv e-prints, Feb. 2017.
Armeni, I., Sener, O., Zamir, A., Jiang, H., Savarese, S., 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces. CVPR 1534–1543.
Chang, A.X., Funkhouser, T., Guibas, L., et al., 2015. ShapeNet: An Information-Rich 3D Model Repository. Computer Sci.
Chen, Y.i., Xiong, Y., Zhang, B., Zhou, J., Zhang, Q., 2021. 3D point cloud semantic segmentation toward large-scale unstructured agricultural scene classification. Comput. Electron. Agric. 190, 106445. https://doi.org/10.1016/j.compag.2021.106445.
Corti, A., Giancola, S., Mainetti, G., Sala, R., 2016. A metrological characterization of the Kinect V2 time-of-flight camera. Robot. Auton. Syst. 75, 584–594 [CrossRef].
Georgakis, G., Reza, M.A., Mousavian, A., et al., 2016. Multiview RGB-D Dataset for Object Instance Detection. IEEE.
Golovinskiy, A., Funkhouser, T., 2009. Min-cut based segmentation of point clouds. 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, IEEE, 39–46.
Graham, B., 2014. Spatially-sparse convolutional neural networks. arXiv preprint arXiv: 1409.6070.
Guo, W., Rage, U.K., Ninomiya, S., 2013. Illumination invariant segmentation of vegetation for time series wheat images based on decision tree model. Comput. Electron. Agric. 96 (6), 58–66.
Guo, N., Zhang, B., Zhou, J., Zhan, K., Lai, S., 2020. Pose estimation and adaptable grasp configuration with point cloud registration and geometry understanding for fruit grasp planning. Comput. Electron. Agric. 179, 105818. https://doi.org/10.1016/j.compag.2020.105818.
Hermosilla, P., Ritschel, T., Vázquez, P.-P., Vinacua, À., Ropinski, T., 2019. Monte Carlo convolution for learning on non-uniformly sampled point clouds. ACM Trans. Graphics (TOG) 37 (6), 1–12.
Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.-K., Yu, L.F., Yeung, S.K., 2016. SceneNN: A scene meshes dataset with annotations. In International Conference on 3D Vision (3DV).
Hua, B.S., Tran, M.K., Yeung, S.K., 2018. Pointwise convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 984–993.
Jaderberg, M., Simonyan, K., Zisserman, A., 2015. Spatial transformer networks. Ad. Neural Informat. Process. Syst. 28, 2017–2025.
Jiang, M., Wu, Y., Zhao, T., et al., 2018. Pointsift: A sift-like network module for 3d point cloud semantic segmentation. arXiv preprint arXiv:1807.00652.
Khan, Z., Rahimi-Eichi, V., Haefele, S., Garnett, T., Miklavcic, S.J., 2018. Estimation of vegetation indices for high-throughput phenotyping of wheat using aerial imaging. Plant Methods 14 (1). https://doi.org/10.1186/s13007-018-0287-6.
Klokov, R., Lempitsky, V., 2017. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. Proc. IEEE Int. Conf. Comput. Vision. 863–872.
Le, T., Duan, Y., 2018. Pointgrid: A deep network for 3d shape understanding. Proceedings of the IEEE conference on computer vision and pattern recognition, 9204–9214.
Lehnert, C., Mccool, C., Sa, I., et al., 2018. A Sweet Pepper Harvesting Robot for Protected Cropping Environments.
Li, Y., Pirk, S., Su, H., et al., 2016. Fpnn: Field probing neural networks for 3d data. Adv. Neur. Inform. Process. Syst. 307–315.
Li, Y., Bu, R., Sun, M., et al., 2018. Pointcnn: Convolution on x-transformed points. Adv. Neural Inform. Process. Syst. 820–830.
Lin, Chengda, Han, Jing, Xie, Liangyi, et al., 2021. Cylinder space segmentation method for field crop population using 3D point cloud. Trans. Chin. Soc. Agric. Eng. 37 (7), 8.
Louedec, J., Li, B., Cielniak, G., 2020. Evaluation of 3D Vision Systems for Detection of Small Objects in Agricultural Environments. 15th International Conference on Computer Vision Theory and Applications.
Louedec, J.L., Montes, H.A,. Duckett, T., et al., 2020. Segmentation and detection from organised 3D point clouds: a case study in broccoli head detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE.
Maturana, D., Scherer, S., 2015. Voxnet: A 3d convolutional neural network for real-time object recognition. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 922–928.
Nguyena, T.T., Slaughtera, D.C., Maloofb, J.N., Sinhab, N., 2016. Plant phenotyping using multi-view stereo vision with structured lights. In: Proceedings of the SPIE Commercial + Scientific Sensing and Imaging, Anaheim, CA, USA, 17 May 2016, p. 986608.
Papon, J., Abramov, A., Schoeler, M., et al., 2013. Voxel cloud connectivity segmentation-supervoxels for point clouds. Proceedings of the IEEE conference on computer vision and pattern recognition, 2027–2034.

Paproki, A., Sirault, X., Berry, S., Furbank, R., Fripp, J., 2012. A novel mesh processing based technique for 3D plant analysis. BMC Plant Biol. 12 (1), 63. https://doi.org/10.1186/1471-2229-12-63.

Paulus, S., Dupuis, J., Mahlein, A.-K., Kuhlmann, H., 2013. Surface feature based classification of plant organs from 3D laser scanned point clouds for plant phenotyping. BMC Bioinformat. 14, 238.

Qi, C.R., Su, H., Mo, K., et al., 2017. PointNet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 652–660.

Qi, C.R., Yi, L., Su, H., et al., 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inform. Process. Syst. 5099–5108.

Riegler, G., Osman Ulusoy, A., Geiger, A., 2017. Octnet: Learning deep 3d representations at high resolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3577-3586.

Sappa, A.D., Devy, M., 2001. Fast range image segmentation by an edge detection strategy. Proceedings Third International Conference on 3D Digital Imaging and Modeling. IEEE, pp. 292–299.

Schnabel, R., Wahl, R., Klein, R., 2007. Efficient RANSAC for point-cloud shape detection. In: Computer graphics forum. Blackwell Publishing Ltd, Oxford, UK, pp. 214–226.

Song, Y., Glasbey, C.A., Polder, G., van der Heijden, G.W.A.M., 2014. Non-destructive automatic leaf area measurements by combining stereo and time-of-flight images. IEEE Multimed. 8, 391–403.

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

Tchapmi, L., Choy, C., Armeni, I., et al., 2017. SEGCloud: Semantic segmentation of 3d point clouds. 2017 international conference on 3D vision (3DV), IEEE, 537–547.

Turaga, P, Veeraraghavan, A., Chellappa, R., 2008. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 1–8.

Vázquez-Arellano, M., Griepentrog, H., Reiser, D., Paraforos, D., 2016. 3-D Imaging Systems for Agricultural Applications—A Review. Sensors 16 (5), 618. https://doi.org/10.3390/s16050618.

Vo, A.-V., Truong-Hong, L., Laefer, D.F., Bertolotto, M., 2015. Octree-based region growing for point cloud segmentation. ISPRS J. Photogramm. Remote Sens. 104, 88–100.

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 Trans. Graphics 38 (5), 1–12.

Wu, W., Qi, Z., Fuxin, L., 2019. Pointconv: Deep convolutional networks on 3d point clouds. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9621–9630.

Zhan, Q., Yu, L., Liang, Y., 2010. A point cloud segmentation method based on vector estimation and color clustering. The 2nd International Conference on Information Science and Engineering, IEEE, 3463–3466.