Computers and Electronics in Agriculture 200 (2022) 107233

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Mature pomegranate fruit detection and location combining improved F-PointNet with 3D point cloud clustering in orchard

Tao Yu [a], Chunhua Hu [a,*], Yuning Xie [a], Jizhan Liu [b], Pingping Li [c]

[a] College of Information Science and Technology, Nanjing Forestry University, Nanjing, Jiangsu Province 210037, China
[b] Key Laboratory of Modern Agricultural Equipment and Technology, Ministry of Education, Jiangsu University, Jiangsu 212013, China
[c] College of Biology and the Environment, Nanjing Forestry University, Nanjing, Jiangsu Province 210037, China

ARTICLE INFO

Keywords:
Frustum PointNet
Fruit detection and location
Feature fusion
Point cloud
Overlapping fruit segmentation

ABSTRACT

Fruit detection and localization is of great significance for horticulture work and robotic harvesting in orchards. Although the existing studies of fruit detection have achieved good results based on 2D image analysis, accurate fruit detection on trees is still challenging because of illumination changes, shielding of leaves and branches, overlapping of fruits and so on. To improve the accuracy of fruit detection and location, this paper proposes a novel ripe pomegranate fruit detection and location method based on improved F-PointNet and 3D clustering method, which is consisting of: (1) RGB-D feature fusion Mask R-CNN was used to realize fruit detection and segmentation; (2) PointNet combined with OPTICS algorithm based on manifold distance and PointFusion was used to segment point clouds in the frustum fruit region, and 3D box was placed in the region of interest; (3) The sphere fitting was performed to obtain the position and the size of a pomegranate. The comparative experiments have been carried out and analyzed, the RGB-D feature fusion Mask R-CNN has the best performance with the F1 score of 0.845 and the AP score of 0.952 respectively, and the improved F-PointNet has better performance than the classical F-PointNet. The measurement radius experiment results of 100 pomegranate samples randomly selected demonstrate that the RMSE is 0.235 cm, the R^2 is 0.826, and the position error is less than 5 mm. These results validate that the proposed detection and location method can effectively detect and locate a single ripe pomegranate under unstructured orchard environment.

1. Introduction

With the increase of population, the demand for fruits is increasing. The original labor-intensive and time-consuming picking is gradually replaced by various automatic picking robots. Many studies about robots working in orchards have been reported (Auat Cheein et al., 2017; Zhao et al., 2016; Arad et al., 2020; Wang et al., 2022). Automatic fruit harvesting robot technology needs the support of target detection technology to detect and locate fruit position, so the target detection technology based on computer vision is very important in the task (Wang et al. (2016); Wang et al., 2022). The crucial step of fruit harvesting robot is to detect the target fruit. The accuracy of target detection has a decisive influence on the success rate of fruit harvesting (Wu et al., 2020; Kapach et al., 2012). However, it is challenging to implement a robust and efficient fruit detection algorithm in orchards due to differences in illumination and shade among fruits, branches and leaves. In order to improve the success rate of fruit harvesting robot, it is significant to develop an algorithm for fruit detection and location with good performance in orchard with complex environment.

Previous researches of the task of in-field fruit detection methods detection based on machine vision are mainly combining features with machine learning to detect fruits. The appearance features including color threshold, shape feature, and texture were used as fruit detection features, and the fruit classifier was established such as Bayesian classifier, K-means clustering, k-nearest neighbours (KNN) clustering, artificial neural network (ANN), random forest (RF) and support vector machineSVM (Tan et al. 2018; Tsouvaltzis et al. 2019; Liu et al. 2018; Feng et al. 2019; Syazwani et al. 2022). Tan et al. (2018) recognized the blueberry fruit of different maturity using histogram oriented gradients (HOG) and colour features in outdoor scenes. The HOG features of the samples were extracted and a linear SVM (Support Vector Machine) classifier was trained to detect fruit-like regions rapidly. Then, a* and b* features in the L*a*b* colour space were utilized to remove non-fruit regions. The KNN (K-nearest Neighbour) and a newly developed

* Corresponding author.
E-mail address: huchunhua@njfu.edu.cn (C. Hu).

https://doi.org/10.1016/j.compag.2022.107233
Received 1 December 2021; Received in revised form 16 June 2022; Accepted 17 July 2022
Available online 29 July 2022

TMWE (Template Matching with Weighted Euclidean Distance) classifiers were employed to identify the fruit of different maturity. Liu et al. (2018) used watershed algorithm to segment apple images into irregular images. Then, according to the color and texture features of the image, support vector machine is used to divide the image into images containing fruit and images without fruit. Syazwani et al. (2022) proposed an automatic method for identifying and recognizing the pineapple's crown images, which used the shape, colour and texture of pineapple as features and ANN as classifier to recognize the pineapple with fruit counting accuracy of 94.4%. Although these classical algorithms have made some achievements, it is difficult to extract and select features, especially in unstructured environment.

The image processing technology based on color, shape and texture needs accurate feature information of target fruits. However, this method cannot maintain high detection accuracy in the case of occlusion of fruits under changing light conditions and similar color of fruits with background. Machine learning is used for target detection and recognition, but machine learning needs to set parameters in advance, and parameters have a great impact on target classification results. In recent years, deep learning has been widely applied in the field of fruit detection in view of its strong ability to extract target features (Fu et al., 2018; Gené-Mola et al., 2019). Tian et al. (2019a) used the improved YOLO-V3 to detect apples at different growing stages, and the YOLO-V3 model could effectively provide detection of apples under overlapping and occlusion conditions. Yu et al. (2019) used Mask R-CNN network to segment ripe strawberry image instances and locate picking points. The detection accuracy and recall rate were 95.78% and 95.41%, respectively. The MIoU (intersection over union) rate was 89.85%, and the average prediction error of picking points was 1.2 mm. Tu et al. (2020) used their proposed MS-FRCNN network to detect RGB images. the algorithm's F1 score was 91.6%.

However, the position relationship between the picking robot and the target is unknown when a RGB image is used to detect the target. Gené-Mola et al. (2020) used RGB images to achieve fruit detection and 3D (three dimensions) positioning by SFM (structure-from-motion) method, but this method requires a lot of computer resources, and the calculation speed is slow. Fruit trees 3D reconstruction method based on the SFM average single fruit tree reconstruction of point cloud take 45 min, so far, still no unified clear indicators to measure the business performance of harvesting robot, but studies have shown that at present most fruit harvesting robot harvest the success rate of about 66% (number) between 40% and 86%, the average cycle time of each fruit is 33 s(Tang et al., 2020). Due to the large amount of computation, SFM cannot process the collected data in a short time, which cannot meet the efficiency requirements of the algorithm when the harvesting robot performs the real-time harvesting task. At present, there are three methods to obtain 3 dimension (3D) point cloud: (1) large, high-precision, high-cost equipment, 3D laser scanner, lidar and ultrasound. Laser scanners and lidar are generally used to reconstruct large scenes such as 3D maps rather than individual fruit trees. But lidar also has some applications in fruit detection. Gene-mola proposed an algorithm for detecting and positioning Fuji Apples by using mobile ground laser scanner (MTLS) to generate 3D point cloud of apple trees for reflectance analysis of tree elements. Through experiments, it was found that reflectance information collected based on lidar plays a positive role in apple detection and 3D positioning. In the experiment, the success rate of fruit positioning is 87.5%, the success rate of recognition is 82.4%, and the F1 score is 85.8%. These data prove that the detection effect of this method is similar to that based on RGB information, but it has the advantage of directly providing 3D fruit positioning information, which is not affected by light. However, due to its high cost, its application in general scene reconstruction is limited. (2) RGB-D cameras with small range, medium accuracy and low cost (Wang et al., 2017; Chiu et al., 2019; Gené-Mola et al., 2019). Depth cameras are generally used for 3D reconstruction of small scenes, such as indoor scene reconstruction and human body 3D reconstruction. Their range is generally no more than 10 m, and they have high accuracy in small-scale scene reconstruction. Kinect V2 camera can collect point cloud data with good quality in the range of 0.5–4.5 m. (3) 3D point cloud reconstruction based on multiple images (Dey et al., 2012; Zhou et al., 2019;Jay et al., 2015). In this method, two-dimensional images of scenes can be obtained directly by ordinary cameras, but the reconstruction accuracy is affected by the algorithm and camera resolution. Therefore, it is an effective way use RGB-D information to realize fast 3D location. Therefore, using RGB-D camera for fruit detection and location is a method with low cost, high speed and high precision (Fu et al. 2020). Song et al. (2016) used Kinect depth camera to fuse 3D information with the coordinates of image feature points. Firstly, the point cloud was preprocessed by distance filtering and color filtering, and then the point cloud was segmented based on Euclidean clustering algorithm. Finally, the apple point cloud was successfully segmented and the 3D coordinates of the fruit were obtained. The apples in apple point cloud were segmented by apple detection and location algorithm based on color and shape features, and the error was less than 10 mm when comparing the obtained point cloud center position with the real position (Nguyen et al., 2016). Tian et al. (2019b) obtains the center of the target fruit by rotating the depth gradient map obtained from the depth information to determine the location of the fruit. The target recognition rate of this fast detection method is 96.61%. Arad et al. (2020) detects objects using shape and color, and calculates the volume of the detected area using depth information. Yang et al. (2019) realized the recognition and classification of obstacles and picked fruits by improving YOLOv3 recognition algorithm, combined depth information to convert the 2D center position of citrus into 3D center position, with a positioning error of 5.9 mm. Realizing the fast identification and precise positioning of target citrus and obstacles. Zhang et al.(2019) used regions of interest in RGB images to segment fruit regions in point clouds. The purity of red apple and green apple was 96.7% and 96.2%, respectively. The fruit point cloud after segmentation is used to locate the fruit, and the positioning error is less than 5 mm. Yang et al. (2019) combined depth information to convert the 2D center position of citrus into 3D center position, with a positioning error of 5.9 mm. Mai et al. (2015) segmented the point cloud obtained according to the difference between color thresholds by Kinect V2. And obtain the three-dimensional location information and radius of the fruit. The average error of fruit positioning by this method was 8.1 mm, and the average error of fruit radius estimation was 4.5 mm. F-PointNet (Qi et al. 2018) is a network that uses RGB image detection to drive object detection in the 3D point cloud. The detection results of the 2D detector can guide better instance segmentation in object detection in three-dimensional space without spatial geometric loss, which is conducive to positioning in 3D space. Because of the complex environment in the orchard, leaves and fruits are easy to block each other, so it is essential to adopt a new positioning method to deal with the complex occlusion situation in the orchard.

Despite numerous studies of fruit detection and location, there are still some problems: (1) The traditional fruit detection based on RGB image is susceptible to the influence of fruit appearance, crop variety, uncertainty, changes in background characteristics and changes in lighting conditions. (2) Only using depth images is susceptible to external influence, resulting in more noise in depth images, thus affecting detection results. (3) Occlusion or aggregation can lead to missed detection and false detection, which is a major challenge to fruit detection and positioning accuracy. (4) Although the segmentation accuracy of point cloud is high, the detection time of point cloud of the whole fruit tree is very long, which does not meet the fast detection and positioning algorithm required by fruit picking robot. There are few studies on mature pomegranate fruit detection and localization based on machine vision at present.

The main objective of this paper is to realize accurate detection and location for mature pomegranate fruit combining improved F-PointNet and 3D point cloud clustering, which is used to detect and locate ripe pomegranate in orchard from 2D image to 3D point cloud. The specific

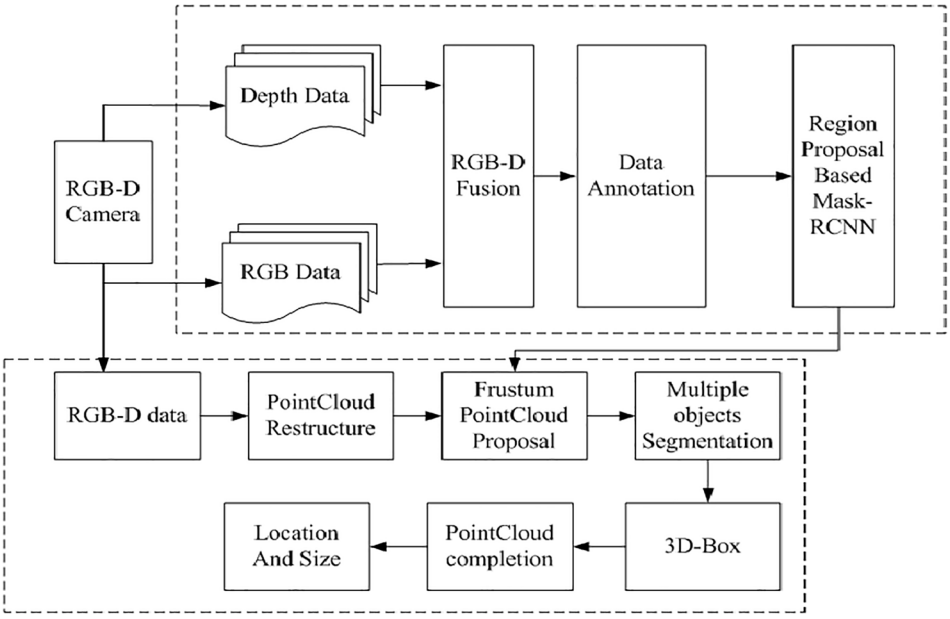


Fig. 1. Framework of the pomegranate detection and location.

Fig. 2. Data acquisition site.

objectives are (1) to segment the RGB and depth region of ripe pomegranates from RGB-D pictures based on Mask R-CNN; (2) to further extract the 3D point cloud of a single ripe pomegranate using PointNet combined with OPTICS clustering algorithm based on manifold distance; (3)to obtain the accurate 3D bounding box of a single pomegranate using PointFusion; (4)to obtain the position and size by fitting a 3D sphere for the target fruit, and (5)to validate the performance of the proposed method by carrying out a lot of different comparative experiments.

2. Materials and methods

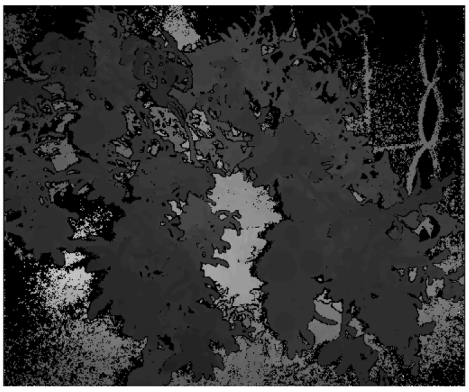
2.1. Overall framework

In order to accurately detect and locate a mature pomegranate, we put forward a robust algorithm for fruit detection and location based on F-PointNet using RGB-D data. We used Kinect V2 camera to obtain 1000 pomegranate images, including RGB images, depth data and RGB-D fusion data. The regions proposed by the Mask R-CNN trained using the fusion of RGB-D data are used to obtain the frustum-shaped regional point cloud according to the two-dimensional square box and transformation relations. To ensure that the point clouds fed into the PointNet network contain only one target, we use an Ordering points to identify the clustering structure (OPTICS) clustering algorithm based on manifold distance to separate the individual fruit point clouds from the overlapping fruit point clouds. Subsequently, we use trained PointNet networks to predict and segment point clouds in conical regions. At the same time, a 3D prediction box is placed on the target point cloud to obtain the data of the fruit employing PointFusion. In order to get the size and the position of the target fruit accurately, a 3D sphere fitting algorithm was introduced to complete a single pomegranate, and the sizes and center position of the target pomegranate were obtained. The framework of this paper is shown in Fig. 1.

2.2. RGB-D data acquisition and preprocessing

This study takes mature pomegranate fruit trees as the research object and uses Microsoft Kinect 2.0 camera (as shown in Fig. 2) to acquire color images and depth images simultaneously. The experimental site is located in Hugang Pomegranate Garden, Nanjing City, Jiangsu Province. The research objects were mature pomegranate fruit trees. The first experiment was conducted on October 11, 2020, on cloudy days, and the second experiment was conducted on October 12, 2020, on sunny days. As the RGB-D camera needs to avoid direct sunlight, it is necessary to

(a) RGB data



(b) Depth data

(c) RGB-D fusion data

(d) Point clouds

Fig. 3. The collected samples. (a) RGB data. (b) Depth data. (c) RGB-D fusion data. (d) Point clouds.

avoid direct sunlight when collecting data. The experiment lasted from 10:00 a.m. to 5:00p.m. We randomly acquired images from different angles and different distances of 1 m,1.2 m, and 1.5 m. We collected 1000 images from 48 pomegranate trees for this experiment.

In this study, we mainly collected RGB data, depth data and RGB-D fusion data, and we generated 3D point clouds from the RGB-D fusion data as shown in Fig. 3. A large number of samples are the basis for high-precision training. Therefore, in order to avoid over-fitting, it is necessary to train the neural network with as much training data as possible. In each epoch, based on the original training dataset, the data was rotated at random angles and vertical axes. At the same time, each point is slightly offset along a random vector. Gaussian noise with zero mean and small standard deviation (range) was used to make the position of each point in each training sample jitter. Data enhancement effectively provides a large amount of training data without changing the amount of data in the training set and improves the accuracy and robustness of the model.

2.3. Ripe pomegranate segmentation based on improved F-PointNet

2D detection models have made great progress in recent years, although 3D sensors are increasingly used, accurate three-dimensional fruit localization still needs further research. The previous 3D detection framework converts the 3D point cloud into images or voxels and then adopts CNN, which affects the invariance of the original 3D data. The emergence of PointNet proposes a scheme to directly process point cloud data, but this method still faces some challenges in the field of 3D detection, such as how to effectively locate the possible position of the target in 3D space, that is, how to generate 3D candidate box, if the global search will cost a lot of computing power and time. F-PointNet (Qi et al. 2018) conducts 3D detection based on RGB and point cloud data. Instead of MV3D data-level fusion, F-PointNet uses decision-level fusion. In this paper, Mask R-CNN is first utilized to generate candidate regions. In the results, 0.5 was used as the threshold to screen out fruit regions, and each detected fruit was painted with different colors to separate different fruits. The Frustum between the camera shooting point and the fruit area detected by Mask R-CNN was retained by the principle of projection. Then the subsequent 3D box regression network based on point cloud is adopted to detect the target pomegranate, which greatly improves the detection efficiency.

The network is divided into three parts as shown in Fig. 4. The first part uses images to detect objects and generate frustums. The second part is the segmentation of point cloud instances inside the frustums. And then we generates a 3D frame based on the point cloud.

2.3.1. Frustum Proposal

Mask R-CNN (He et al., 2017) was used to extract 2D object regions in RGB images and classify objects. For input images, The network segmented the input image at the pixel level. The operation is shown in Fig. 5, architecturally, it consists of two parts: one is to extract the backbone of features, and the other is to carry out classification regression and mask prediction for each ROI.

First, the input image was obtained through VGG16 feature extraction backbone network. Then, Region Suggestion Network (RPN) can generate all anchors with a fixed size, then apply the regression offset obtained from the network to the Anchor to make it closer to the truth value, clip the Proposal beyond the image size, and get the initial suggested area. Then sort Anchors according to the scores output from the classified network and retain the first 12,000 high-scoring Anchors. Since an object may have more than one Anchors overlapping correspondence, apply non-maximum suppression (NMS) to remove the overlapping boxes and choose the first 2000 in the remaining Proposal again based on the predicted RPN score. Next, the RoIAlign process will align the extracted features correctly with the input image. Through RoI Align, the feature graph corresponding to the original image is transferred to the full connection layer for classification and boundary box

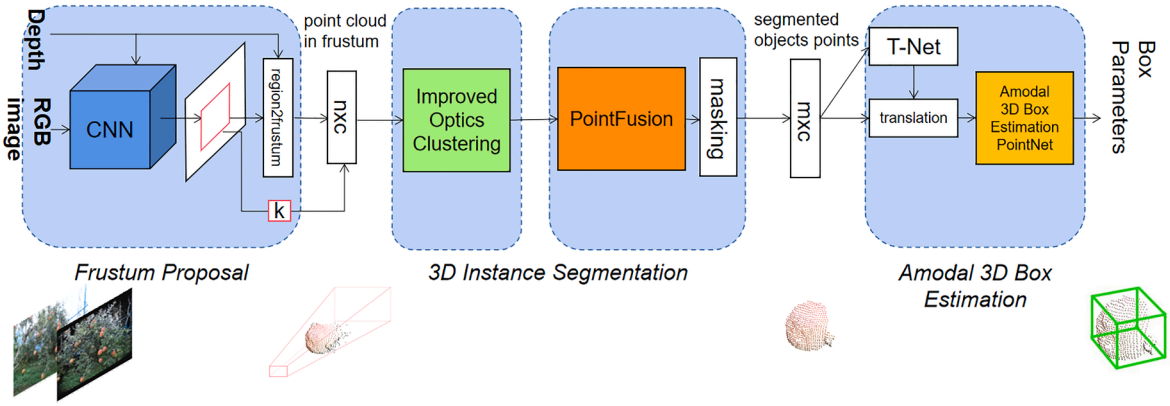


Fig. 4. Diagram of F-PointNet architecture.

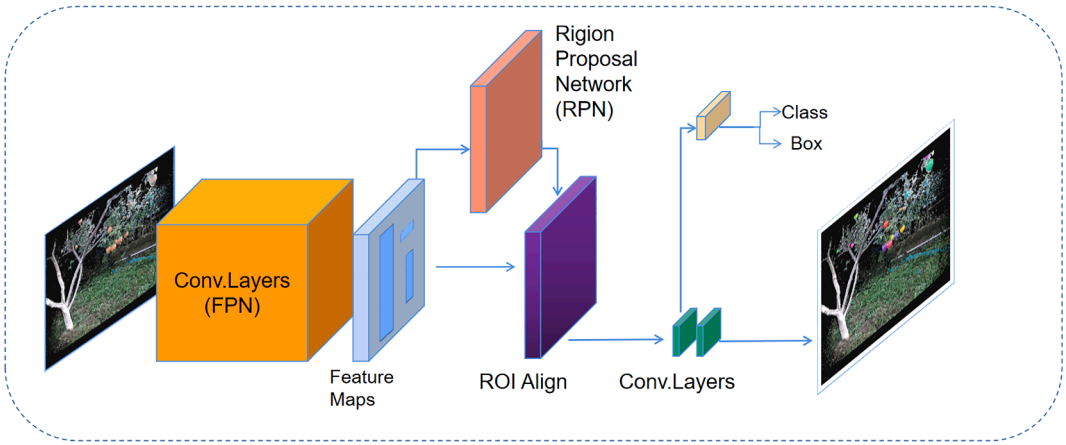


Fig. 5. The architecture of Mask R-CNN.

regression, and transferred to the convolution layer for mask task. In the fruit detection task, the detection effect of overlapping fruits in 2D images is relatively poor. Depth information can better reflect the difference of spatial position than RGB information. In order to improve the detection rate of overlapping fruits, a multi-feature fusion method, RGB-D information fusion method is adopted (Hazirbas et al., 2016). Features are extracted from RGB and depth images respectively, and then the depth feature graph is fused into the RGB branch.

In the point cloud generated by using RGB-D information, the point cloud of the target region provided by the 2D detector is segmented into frustum point cloud. Frustum may be many different directions, different directions of the point cloud is different. Therefore, by rotating them to center the view, we make the center of the frustum axis orthogonal to the image plane to standardize on them. The normalized helps to improve the accuracy of the algorithm. The whole process is to generate the frustum point cloud of the target region from RGB-D data.

2.3.2. 3D instance segmentation

The network takes a point cloud in frustum and uses the PointNet (Qi et al.,2017) predicts a probability score for each point that indicates how likely the point belongs to the object of interest (Chen et al.,2021).

PointNet is the first deep neural network that directly processes out-of-order point cloud data. The PointNet has three core building blocks, the transformation networks (T-Net), the max pooling layer as a symmetric function to aggregate information from all the voxels and the multi-layer perceptron (MLP) network.

Mask R-CNN may produce false detections. The false result area may contain leaves or multiple fruits. In classical F-PointNet algorithm, the point cloud input into the PointNet network is required to contain only one target. For highly overlapping fruit targets, the point cloud contained in the obtained cone area is not the point cloud of a single target. In order to separate most overlapping fruits individually, DBSCAN is a density-based point cloud clustering method, which is usually used to segment point cloud. DBSCAN does not require pre-setting the number of clusters and anti-noise points. DBSCAN does not require pre-setting the number of clusters and anti-noise points. We adopt a more effective multi-objective point cloud segmentation method, combining OPTICS with manifold distance clustering algorithm. Traditional methods cluster according to Euclidean distance, however, this can only reflect the local similarity and does not consider the global similarity. Using Euclidean distance only focuses on the distance between two points, while using manifold distance takes all points between two points into account, taking into account the local relationships of all points and reflecting whether two points belong to the same fruit. OPTICS is an improvement of DBSCAN algorithm, so their parameters look very similar However, the OPTICS algorithm does not require a minimum number of points to become a cluster. By calculating the sort, a dynamic minimum number of points can be obtained to achieve a more flexible and adaptable clustering process.

1. Define two queues, the ordered queue and the result queue. The ordered queue is used to store core points and their density points, and is arranged in ascending order according to the reachable distance; The result queue is used to store the output order of sample points. Points in the ordered queue are the samples to be processed, and points in the result queue are the samples after processing. 2. Select an unprocessed core point and put it into the result queue. Meanwhile, calculate the reachable distance of the sample points in the neighborhood and put the sample points into the ordered queue in ascending order according to

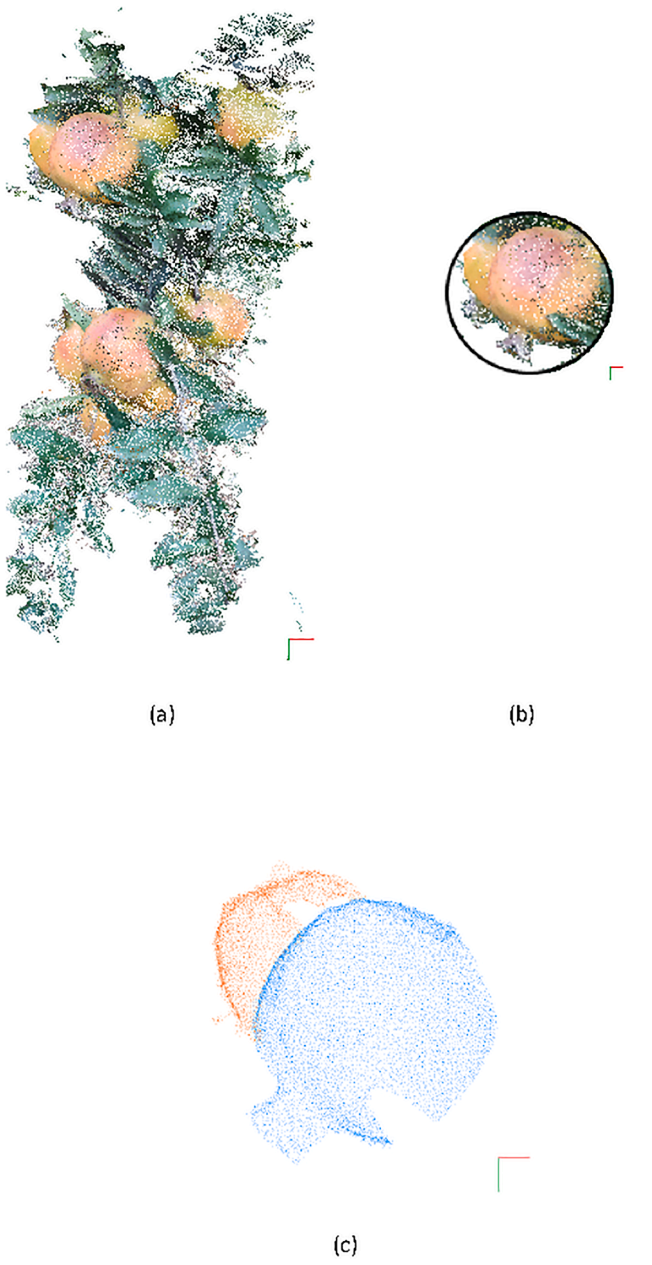


Fig. 6. 3D point cloud segmentation of pomegranate. (a) Primordial point clouds. (b) Overlapped pomegranate point cloud separated using the optics clustering method. (c) OPTICS clustering results.

the reachable distance; 3. Extract the first sample from the ordered queue. If it is core point, calculate the reachable distance and put the point with the smallest reachable distance into the result queue. If it is not core point, skip the point and select a new core point. 4. Iterate the second and third steps until all sample points are processed, and then output the sample points in the result queue and their reachable distance.

In an orchard, many fruits get in the way of each other. The clustering segmentation method based on Euclidean distance cannot segment point cloud accurately. Based on this, an overlapping fruit segmentation method combining OPTICS with manifold distance is proposed. The algorithm only needs to input radius and minPts (The minimum number of points) to get clustering results. Combined with manifold distance, the algorithm can effectively segment overlapping fruit point clouds, as shown in Fig. 6.

In 3D detection, some location clouds are sparse and have no color information, The reason why color information is not used in 3D detection is that in the F-PointNet algorithm, color information is only used when the fruit area is detected by Mask R-CNN. After the point cloud behind the fruit area is obtained, the PointNet algorithm is used to detect the point cloud and only extracts the spatial location features of the point cloud, but does not extract RGB features. Our RGB and depth cameras have been registered, and the reason for the sparse point clouds is the distant and obscured fruits. There will be a certain amount of missed detection and false detection. Although the 2D proposal was detected, PointNet failed to detect the presence of an object in Frustum because there were some sparse point clouds (a single fruit point cloud is less than 200). The fusion network corresponds the image features extracted by CNN to the corresponding point cloud features generated by PointNet.

In order to solve the point cloud sparsity and lack of color information will lead to the problem of missed detection and misdetection. A point cloud and RGB information fusion method PointFusion (Xu et al., 2018) is adopted. PointFusion uses PointNet to fuse the features of the original point cloud data with the color features extracted from the input image by CNN (See Fig. 7).

2.3.3. Training with Multi-task losses

Our overall losses were mainly in three areas: PointNet 3D segmentation, T-Net, and box estimation.

$$L_{mulit-task} = L_{seg} + \lambda\left(L_{c1-reg} + L_{c2-reg} + L_{h-cls} + L_{h-reg} + L_{s-cls} + L_{s-reg} + \gamma L_{corner}\right) \tag{1}$$

The loss of T-NET is L_{c1-reg}, and the loss of central regression is L_{c2-reg}. L_{h-cls} and L_{h-cls} are losses for heading angle prediction.L_{s-cls} and L_{s-reg} are for box size.

2.4. Fruit localization

Since the source of our point cloud data is recovered according to the data collected by the RGB-D camera, our point cloud is incomplete. In order to better obtain the parameters of fruits, we need to fit the incomplete fruit point cloud and then conduct the 3D reconstruction. The positioning of the fruit center also helps us to eliminate repeated detection targets. Because the shape of pomegranate fruit is spherical, we used a spherical model to reconstruct our pomegranate fruit. We used Random Sample Congruence (RANSAC) algorithm to complete the spherical model fitting of pomegranate fruit.

The parametric model for spherical fitting is expressed as follows:

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = R^2 \tag{2}$$

where (x_0,y_0,z_0,R) are the parameters that need at least four points to solve. Given 4 points, a solution to these parameters can be obtained by the linear least square (LLS).

2.5. Evaluations

We use precision-recall curve, mean accuracy and F1 score to evaluate our segmentation writing results. The following indicators were selected as evaluation criteria :

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

TP, FP and FN respectively represent true positive, false positive and false negative in the test results. TP represents the number of pomegranate fruits correctly identified, FP represents the number of pomegranate fruits incorrectly identified, and FN represents the number of errors identified as background. We take 0.5 as the threshold,TP was the

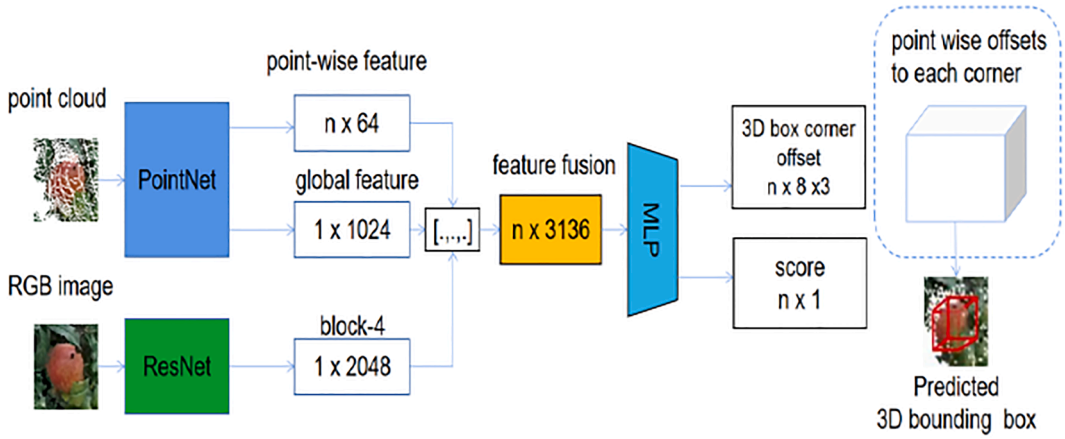


Fig. 7. An overview of the dense PointFusion architecture.

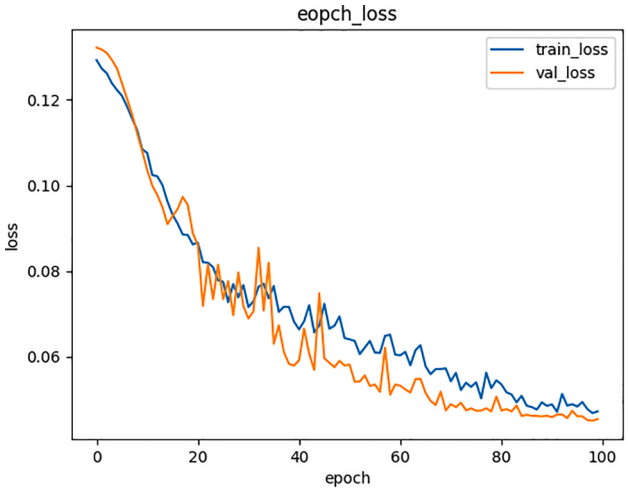


Fig. 8. Train-loss and val-loss.

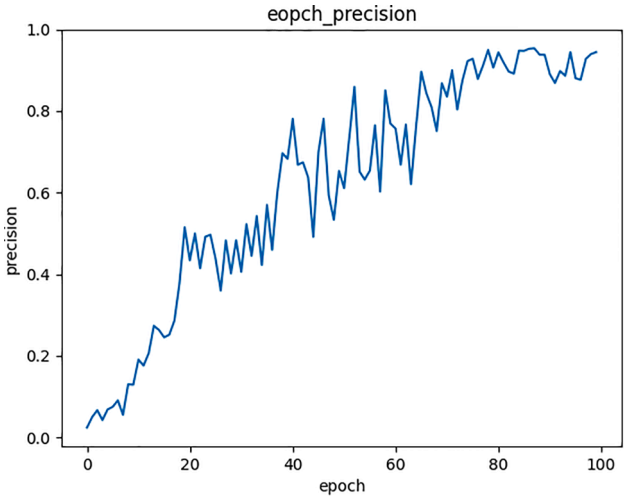


Fig. 9. Train-precision.

number of pomegranate fruits with correct recognition score greater than 0.5,FP was the number of pomegranate fruits with false recognition score greater than 0.5, and FN was the number of pomegranate fruits with false recognition score less than 0.5 as the background.

F1 score is the harmonic mean of accuracy and recall and is defined as:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

3. Results

With the exception of the section on deep learning, these experiments were conducted on a Windows 10 64-bit PC equipped with an Intel (R) Core (TM) i7-8700 CPU @ 2.80 GHz processor (Intel, Santa Clara, CA, USA) and 32 GB-RAM. Since deep learning involves automated computer systems to study large amounts of training data and requires high computing power, we used the NVIDIA RTX 1080TI GPU (NVIDIA Inc., Santa Clara, CA, USA) instead of the CPU to reduce our training time. After the collected data is preprocessed, the number of images in the training, validation, and test set are 600, 200, and 200, respectively.

In order to analyze the reliability of the improved the proposed algorithm in this paper, the accuracy of the pomegranate fruit regions segmented by Mask R-CNN is firstly analyzed, and then the segmentation results of point cloud are analyzed, and the radius and position of the target are obtained by fitting the point cloud.

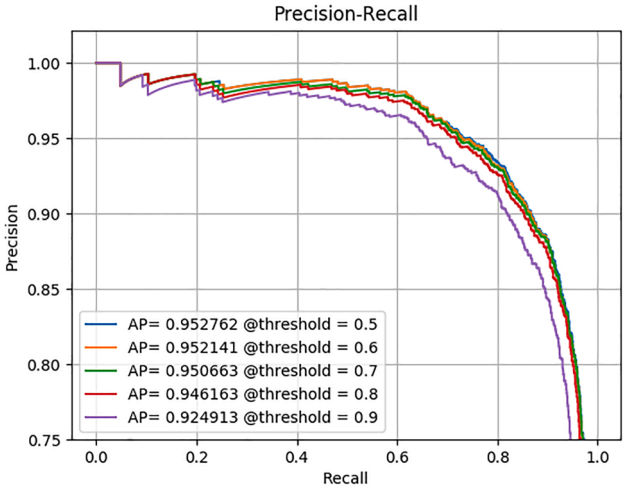


Fig. 10. P-R curves of our Mask R-CNN.

3.1. Training assessment

In the F-PointNet model, the learning rate is 0.0001 and the epoch count is 100. Training losses and training accuracy are plotted in Fig. 8 and Fig. 9.

Table 1
Detection results of different network structures and preprocessing methods.

Network	AP	F1 score
the classic Mask R-CNN	0.872	0.793
the proposed network without jet colormap	0.897	0.801
the proposed network with jet colormap	0.952	0.845

(a)

(b)

(c)

Fig. 11. The segmentation results of RGB-D samples. The red box represents the fruit of successful detection and the blue box represents the fruit of missed detection. (a) No missed detection (b) Missed detection due to fruit occlusion (c) Missed detection due to leaf occlusion. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

With the delay of the learning process, the training accuracy of the training samples showed an upward trend, while the training loss showed a downward trend, indicating that our F-PointNet is a global optimization process. Training precision increased and Training loss down. After 60 epochs, train loss values tend to be stable, while val loss did not increase, indicating that F-PointNet has a strong fitting ability.

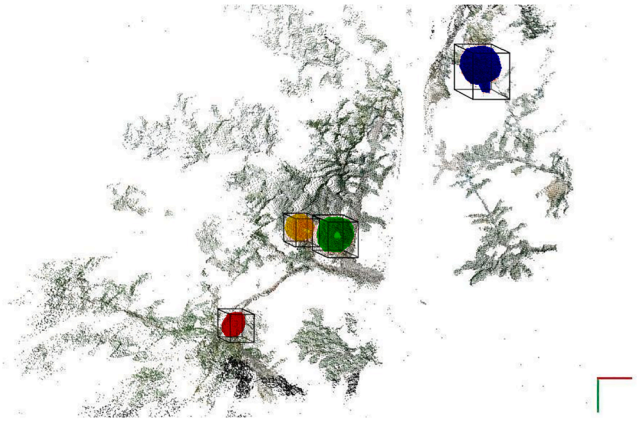


Fig. 12. Results of F-PointNet. Segmentation result of F-PointNet.

3.2. Results of Mask R-CNN

The P-R curves of our Mask R-CNN in different thresholds are shown in Fig. 10. In the orchard environment, there are many factors that can affect the detection results, such as different lighting conditions, overlapping fruits, and changes in appearance. These factors will seriously affect the accuracy of detection and segmentation.

Table 1 shows the detection results of different pre-training networks including the classic Mask R-CNN, the proposed network without jet colormap (Eitel et al., 2015) and the proposed network with jet colormap. The results of all networks with the same confidence degree of 0.7 are evaluated by AP and F1 scores. The F1 score and AP score of the classic Mask R-CNN network are 0.793 and 0.872 respectively. However, the F1 score and AP score of the proposed network without jet colormap are increasing. The proposed network with jet colormap is one of the most accurate networks with the F1 score of 0.845 and the AP score of 0.952 respectively. The results showed that our improved Mask R-CNN network had better detection performance for fruits in the orchard with complex environment.

Fig. 11 shows the detection results of 3 images selected from the validation data set under different distance and different sunlight. It is observed that most separated fruits are successfully detected, and some fruits with less occlusion can also be successfully segmented. In addition, Mask R-CNN correctly masks the pixels belonging to a fruit, even though the fruit is visually separated by branches or leaves, The experimental results showed that the Mask R-CNN combined with depth information had a high recall rate and accuracy in the detection of pomegranate in the orchard environment. There are many environmental factors leading

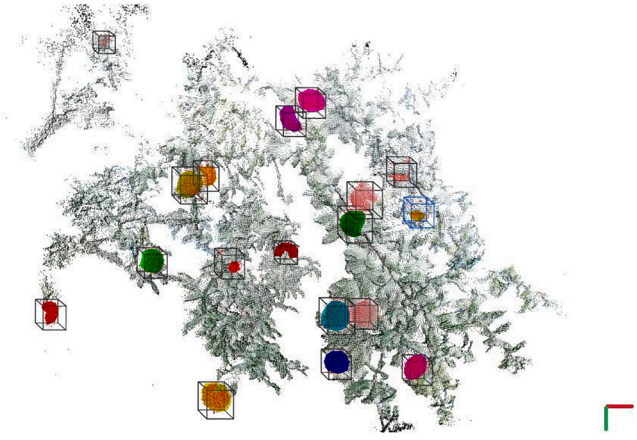
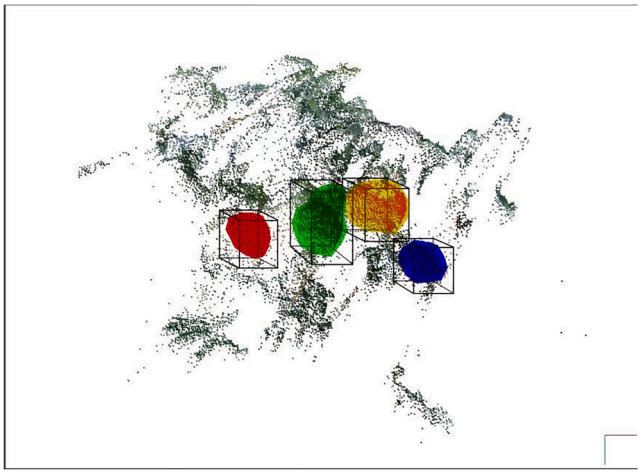
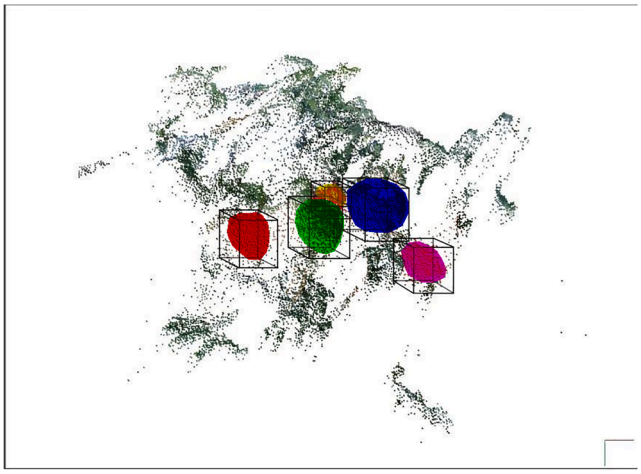


Fig. 13. Segmentation boxes of F-PointNet. Segmentation result of improved F-PointNet.



(a)



(b)

Fig. 14. Result of overlapping parts. (a) Segmentation result of classical F-PointNet. (b) Segmentation result of improved F-PointNet.

to false negative detection, such as strong sunlight reflection shadow, fruit appearance, color, shape occlusion or changes in perspective. The experimental results in Fig. 10 show that our method achieves high accuracy in pomegranate detection in orchard environment and provides relatively accurate region of interest.

3.3. Results of multi-object 3D instance segmentation

The PointNet model is used to identify the point cloud data provided by the Frustum Proposal. We visualized some simple cases of F-PointNet output for reasonably distant unobstructed objects, and our model outputs 3D instance segmentation and 3D bounding boxes (See Fig. 12).

A common mistake is because of the point cloud is too sparse to accurate recognition and segmentation of target point cloud, we think the image characteristics are very helpful for the solution to this problem, so the improved network for recognition and segmentation of sparse point cloud is put forward combining with an RGB characteristics and characteristics of point cloud feature. As shown in Fig. 13, the segmentation boxes of the point cloud of pomegranate fruit regions. It can be found that the point cloud in the blue rectangular box in Fig. 13 is too sparse, and the classcial F-PointNet failed to identify the fruit point cloud. F-PointNet uses PointNet network to extract features from the point cloud of fruits, a global feature is finally extracted by global maximum pooling. When the point cloud is too sparse (the number of points is less than 400), few global features are extracted, and the PointNet network classifies the fruit point cloud as non-fruit. Improved F-PointNet successfully identifies the fruit point cloud.

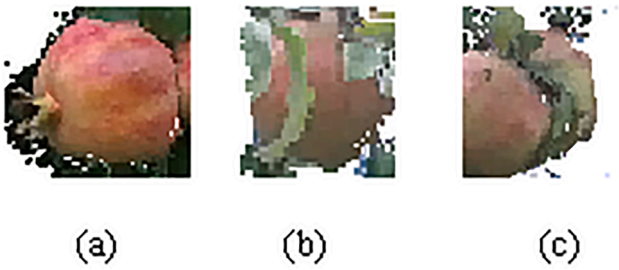


Fig.15. The fruits were classified according to the degree of occlusion. (a) not being occluded is defined as easy, (b) the occluded part is less than 60% and is defined as common, (c) occlusion of more than 60% is defined as hard.

Table 2
Comparison of segmentation results of different models

model	precision
F-PointNet	0.657
Improved F-PointNet	0.826

Table 3
Segmentation performance.

Method		Precision
traditional F-PointNet	Easy	0.847
	Common	0.744
	Hard	0.565
improved F-PointNet	Easy	0.957
	Common	0.917
	Hard	0.838

Another common mistake is that classical F-PointNet assumes that there is only a single object of interest in each frustum, which relies on 2D detector results. Although our improvement of 2D detector has effectively increased the detection capability of overlapping images, in front of the partial overlap rate is too high, and there is still appearing error segmentation with multiple instances mistakenly identified as a single example. However, multi-target three-dimensional segmentation algorithm can effectively solve this problem. Firstly, the OPTICS algorithm is used first to segment the multi-instance target point cloud that may exist in the frustum into a single instance target point cloud. Then, the single instance target point cloud is input into PointNet for classification and segmentation. The results are shown in the Fig. 14. In Fig. 14, there are two pomegranate fruits that are identified as one pomegranate because they largely overlap. In Fig. 14a, because there were two fruits in the input point cloud, PointNet mistakenly detected them as one target and output the wrong result. Fig. 14b visualizes the output result of the improved F-PointNet, successfully splitting the highly overlapping fruits and producing the correct result.

To further verify the effectiveness of our segmentation method, we have analyzed evaluations of the comparison experiment between the traditional F-PointNet and our improved F-PointNet. The results show that the improved F-PointNet segmentation's F1-score is increased from 0.823 to 0.912. We randomly selected 100 RGB-D images from the test set, and divided the target pomegranate fruits into three categories as shown in Fig. 15. The pomegranate fruits without occlusion were classified as easy, the pomegranate fruits with occlusion degree less than 60% were classified as common, and those with occlusion degree greater than 60% were defined as hard. The results indicate that the improved F-PointNet can achieve higher segmentation effect regardless of easy

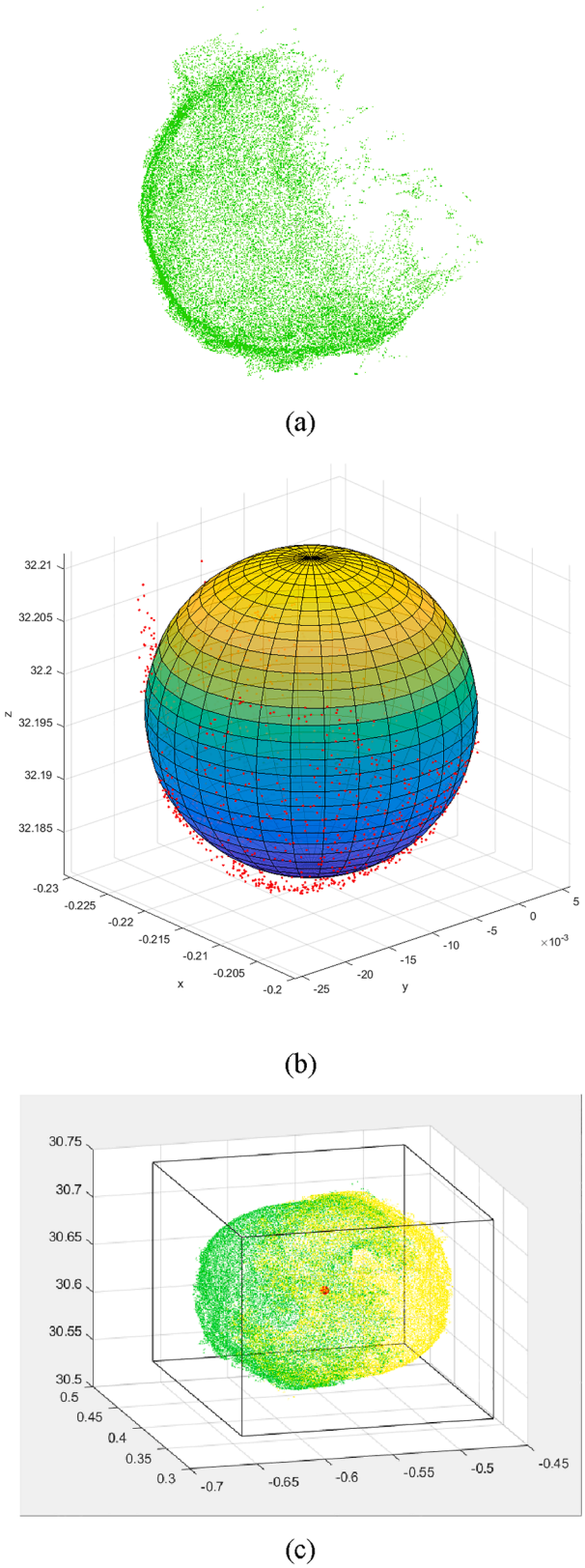


Fig. 16. Fitting and location of fruit point cloud (a) Original point cloud (b) Spherical completion based on RANSAC (c) Complete the fruit point cloud and fruit center.

samples or hard samples than the traditional F-PointNet (See Table 2 and 3).

3.4. Fruit localization

The fruit point cloud extracted by the improved F-PointNet is shown in Fig. 16a, and Fig. 16b is the completion point cloud generated by fruit spherical fitting. The red point (shown in Fig. 16c) is the center of the sphere fitting, which is defined by us as the center of the fruit. The spherical center was defined as the center point of the fruit, and the three-dimensional spatial coordinates of the fruit could be obtained by combining the radius information of the sphere obtained by RANSAC algorithm. The spherical center was defined as the center point of the fruit, so the position of the fruit in space could be obtained by using the measured radius information. Accurate segmentation of pomegranate fruit point cloud can make the fitting sphere and fruit shape error smaller, which helps us to accurately locate and pick pomegranate fruit. In order to verify the rationality of the algorithm proposed by us, we randomly selected 100 pomegranate fruits to estimate their location and radius. The results of comparison between the actual radius and the measured ones with the proposed method are shown in the Fig. 17. The regression RMSE is 0.235 cm, and R^2 is 0.826. The Mean Absolute Error (MAE) of the position information measured by the algorithm proposed in this paper is less than 5 mm. The MAE in a similar three-dimensional fruit localization study was 7 mm, which can provide reliable data for automatic robot harvesting (Gené-Mola et al., 2021).

4. Conclusion

In this paper, an OPTICS algorithm based on manifold distance is proposed to improve the fruit detection and 3D localization methods of F-PointNet. The improved Mask R-CNN (F1-Score = 0.845) achieved good results in unstructured orchards with complex environments. Due to the complex environment of orchards, most fruits have occlusion problems in 2D images, but the occlusion problem is easier to solve in 3D space. Point network in 3D space is helpful to improve the precision of segmentation and location. Although the detection speed of Mask R-CNN was slower than that of other networks, it could provide us with accurate fruit regions. This can greatly reduce the resource consumption when searching the target fruit point cloud in 3D space. PointNet was used to classify and segment point clouds preprocessed using the OPTICS algorithm based on manifold distance, avoiding the problem of unobtainable obscured fruits. The improved F-PointNet network has better segmentation result of point cloud with 82.6% than the traditional F-PointNet network.

This paper proposes an improved F-PointNet fruit detection and 3D localization method combining OPTICS algorithm based on manifold distance. Due to the uncertainty of the fruit location, the highly overlapping fruit may be mistaken as an object of interest by our two-dimensional segmentation results. The improved Mask R-CNN utilizes depth information more effectively and improves the detection rate of overlapping targets. Due to distance or illumination reasons, classical F-PointNet algorithms may misestimate the category and size of point clouds due to the sparsity of point clouds. Therefore, we use the fusion method of RGB information and point cloud features to effectively increase the detection ability of sparse point clouds. In the classical F-PointNet default truncated frustum, there is only one object of interest. When multiple instances appear, it may be confused and output mixed segmentation results. To solve this problem, we segment the point clouds in the truncated frustum before inputting the truncated frustum point cloud into the point cloud segmentation network to ensure that there is only one object of interest in each truncated frustum. F-PointNet outputs the point cloud and 3D Box of the target fruit. By integrating depth features into RPN, the F1-score of 2D segmentation result was improved 6.55% (from 0.793 to 0.845). Combining with OPTICS algorithm based on manifold distance and FusionNet, multi-objective

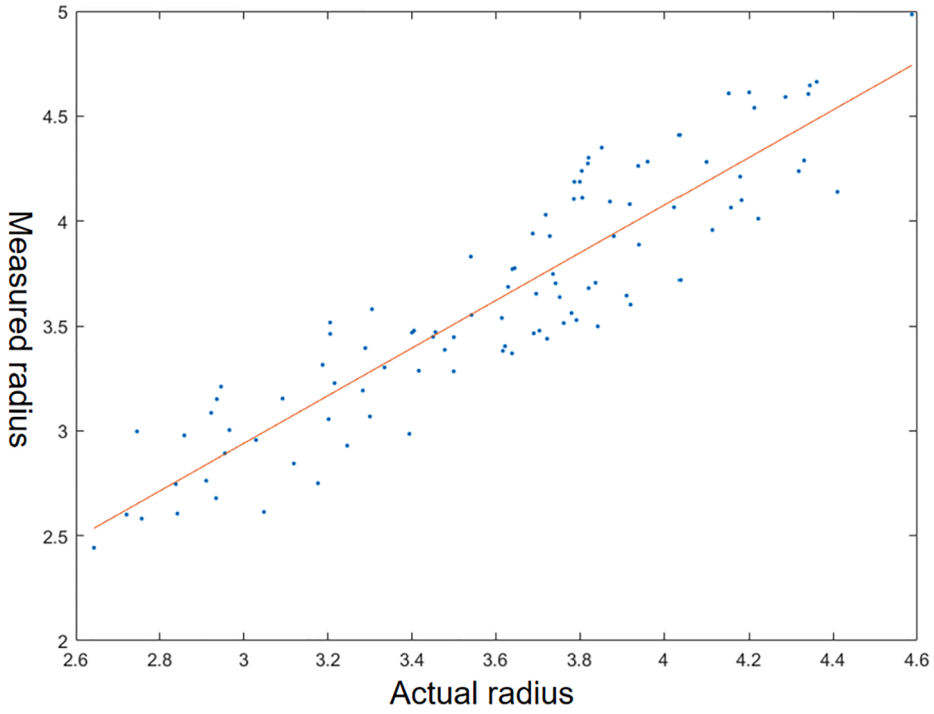


Fig. 17. Fruit size compare actual size with measured size.

segmentation and positioning accuracy was improved by 25.7% (from 0.657 to 0.826). And the error of the position information measured by the algorithm proposed is less than 5 mm. In order to better localize the fruit, we perform sphere fitting on the fruit point cloud to locate the center of the fruit, because the shape of pomegranate fruit is very close to the sphere. The experimental results show that the RMSE is 0.235, and the location effect is good. Due to the higher detection rate and spatial accuracy achieved by the improved F-PointNet and providing more accurate positioning for fruit picking, future work should extend this method to the measurement of fruit growth, and thus to the estimation of fruit yield.

CRediT authorship contribution statement

Tao Yu: Data curation, Writing – original draft. **Chunhua Hu:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Yuning Xie:** Visualization, Investigation. **Jizhan Liu:** Resources, Supervision. **Pingping Li:** Software, Validation.

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.

References

Auat Cheein, F., Torres-Torriti, M., Hopfenblatt, N.B., Prado, Á.J., Calabi, D., 2017. Agricultural service unit motion planning under harvesting scheduling and terrain constraints. J. Field Rob. 34 (8), 1531–1542. https://doi.org/10.1002/rob.21738.

Arad, B., Balendonck, J., Barth, R., Ben-Shahar, O., Edan, Y., Hellström, T., Hemming, J., Kurtser, P., Ringdahl, O., Tielen, T., Tuijl, B., 2020. Development of a sweet pepper harvesting robot. J. F. Robot. 37 (6), 1027–1039. https://doi.org/10.1002/rob.21937.

Chen, X., Jiang, K., Zhu, Y., Wang, X., Yun, T., 2021. Individual tree crown segmentation directly from uav-borne lidar data using the pointnet of deep learning. Forests 12, 1–22. https://doi.org/10.3390/f12020131.

Chiu, C.-Y., Thelwell, M., Senior, T., Choppin, S., Hart, J., Wheat, J., 2019. Comparison of depth cameras for threedimensional reconstruction in medicine. Proc. Instit. Mech. Eng. H-J. Eng. Med. 233 (9), 938–947.

Dey, D., Mummert, L., Sukthankar, R., 2012. Classification of plant structures from uncalibrated image sequences. IEEE.

Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W., 2015. Multimodal deep learning for robust RGB-D object recognition. Int. Conf. Intell. Robot. Syst. 681–687. https://doi.org/10.1109/IROS.2015.7353446.

Fu, L., Feng, Y., Elkamil, T.T., et al., 2018. Image recognition method of multi-cluster kiwifruit in field based on convolutional neural networks. Trans. Chin. Soc. Agric. Eng. 34 (2), 205–211.

Fu, L., Majeed, Y., Zhang, X., Karkee, M., Zhang, Q., 2020. Faster R-CNN based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. Biosyst. Eng. 197, 245–256. https://doi.org/10.1016/j.biosystemseng.2020.07.007.

Feng, J., Zeng, L., He, L., 2019. Apple fruit recognition algorithm based on multi-spectral dynamic image analysis. Sensors (Switzerland) 19, 1–13. https://doi.org/10.3390/s19040949.

Gené-Mola, J., Vilaplana, V., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Gregorio, E., 2019. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometriccapabilities. Comput. Electron. Agric. 162, 689–698.

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Morros, J.-R., Ruiz-Hidalgo, J., Vilaplana, V., Gregorio, E., 2020. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. Comput Electron Agric 169, 105165. https://doi.org/10.1016/j.compag.2019.105165.

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J.R., Escolà, A., Gregorio, E., 2021. In-field apple size estimation using photogrammetry-derived 3D point clouds: Comparison of 4 different methods considering fruit occlusions[J]. Comput. Electron. Agric. 188, 106343. https://doi.org/10.1016/j.compag.2021.106343.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In Asian Conf. Comput. Vis. 213–228. https://doi.org/10.1007/978-3-319-54181-5_14.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In Proc. IEEE Int. Conf. Comput. Vis. 2961–2969. https://doi.org/10.1109/ICCV.2017.322.

Jay, S., Rabatel, G., Hadoux, X., Moura, D., Gorretta, N., 2015. In-field crop row phenotyping from 3D modeling performed using Structure from Motion. Comput. Electron. Agric. 110, 70–77.

Kapach, K., Barnea, E., Mairon, R., Edan, Y., Shahar, O.B., 2012. Computer vision for fruit harvesting robots-state of the art and challenges ahead. Int. J. Comput. Vis. Robot. 3 (1/2), 4–34.

Liu, X.Y., Jia, W.K., Ruan, C.Z., et al., 2018. The recognition of apple fruits in plastic bags based on block classifification. Precis. Agric. 19 (4), 735–749.

Mai, C., Zheng, L., Sun, H., Yang, W., 2015. Research on 3D reconstruction of fruit tree and fruit recognition and location method based on RGB-D camera. Trans. Chinese Soc. Agric. Mach. 46, 35–40. https://doi.org/10.6041/j.issn.1000-1298.2015.S0.006.

Nguyen, T.T., Vandevoorde, K., Wouters, N., Kayacan, E., De Baerdemaeker, J.G., Saeys, W., 2016. Detection of red and bicoloured apples on tree with an RGB-D

camera. Biosyst. Eng. 146, 33–44. https://doi.org/10.1016/j.biosystemseng.2016.01.007.

Qi CR,Su H,Mo K,Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. Proc-30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017;2017-January: 77 - 85. https://doi.org/10.1109/CVPR.2017.16.

Qi, Charles R:; Liu, Wei; Wu, Chenxia;Su, Hao; Guibas, Leonidas J. Frustum PointNets for 3D Object Detection from RGB-D Data. [IEEE 2018 /EEE/CVF Conference on Computer Vision and PatternRecognition (CVPR) Salt Lake City, UT,USA (2018.6. 18-2018.6.23)] 0, 918-927.doi:10.1 109/CVPR.2018.00102.

Syazwani, R.W.N., Asraf, H.H., Amin, M.A.M.S., Dalil, K.A.N., 2022. Automated image identification, detection and fruit counting of top-view pineapple crown using machine learning. Alexandria Eng. J. 61 (2), 1265–1276. https://doi.org/10.1016/j.aej.2021.06.053.

Song, J., Teng, D., Wang, K., 2016. Segmentation and localization method of greenhouse cucumber based on image fusion technology. Int. J. Simul. Syst. Sci. Technol. 17, 11–14. https://doi.org/10.5013/IJSSST.a.17.25.07.

Tan, K., Lee, W.S., Gan, H., Wang, S., 2018. Recognising blueberry fruit of different maturity using histogram oriented gradients and colour features in outdoor scenes. Biosyst. Eng. 176, 59–72. https://doi.org/10.1016/j.biosystemseng.2018.08.011.

Tsouvaltzis, P., Babellahi, F., Amodio, M.L., Colelli, G., 2020. Early detection of eggplant fruit stored at chilling temperature using different non-destructive optical techniques and supervised classification algorithms. Postharvest biology and technology. 159, 111001. https://doi.org/10.1016/j.postharvbio.2019.111001.

Tu, S., Pang, J., Liu, H., Zhuang, N., Chen, Y., Zheng, C., Wan, H., Xue, Y., 2020. Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-D images. Precis. Agric. 21 (5), 1072–1091. https://doi.org/10.1007/s11119-020-09709-3.

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., Liang, Z., 2019a. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Comput. Electron. Agric. 157, 417–426. https://doi.org/10.1016/j.compag.2019.01.012.

Tian, Y., Duan, H., Luo, R., Zhang, Y., Jia, W., Lian, J., Zheng, Y., Ruan, C., Li, C., 2019b. Fast recognition and location of target fruit based on depth information. IEEE Access 7, 170553–170563. https://doi.org/10.1109/ACCESS.2019.2955566.

Tang, Y., Chen, M., Wang, C., Luo, L., Li, J., Lian, G., Zou, X., 2020. Recognition and Localization Methods for Vision-Based Fruit Picking Robots: A Review[J]. Front. Plant Sci. 11 https://doi.org/10.3389/fpls.2020.00510.

Wu, G., Li, B., Zhu, Q., Huang, M., Guo, Y.a., 2020. Using color and 3D geometry features to segment fruit point cloud and improve fruit recognition accuracy. Comput Electron Agric 174, 105475. https://doi.org/10.1016/j.compag.2020.105475.

Wang, Z.H., Xun, Y., Wang, Y.K., Yang, Q.H., 2022. Review of smart robots for fruit and vegetable picking in agriculture. Int. J. agricult. biological eng. 15 (1), 33–54. https://doi.org/10.25165/j.ijabe.20221501.7232.

Wang, C., Tang, Y., Zou, X., SiTu, W., Feng, W., 2017. A robust fruit image segmentation algorithm against varying illumination for vision system of fruitharvesting robot. Optik-Int. J. Light Electron Opt. 131, 626–631.

Wang, Z., Walsh, K., Verma, B., 2017. On-tree mango fruit size 745 estimation using RGB-D images. Sensors 17 (12), 2738.

Xu, D., Anguelov, D., Jain A, A., 2018. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 244–253.

Yang, C., Liu, Y., Wang, Y., Xiong, L., Xu, H., Zhao, W., 2019. Research and experiment on recognition and location system for citrus picking robot in natural environment. Trans. Chinese Soc. Agric. Mach. 50 (14–22), 72. https://doi.org/10.6041/j.issn.1000-1298, 2019.12.002.

Yu, Y., Zhang, K., Yang, L.i., Zhang, D., 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Comput. Electron. Agric. 163, 104846. https://doi.org/10.1016/j.compag.2019.06.001.

Zhang, Y., Tian, Y.e., Zheng, C., Zhao, D., Gao, P.o., Duan, K.e., 2019. Segmentation of apple point clouds based on ROI in RGB images. Inmateh - Agric. Eng. 59 (3), 209–218.

Zhao, Y., Gong, L., Huang, Y., & Liu, C. (2016a). A review of key techniques of vision-based control for harvesting robot. Computers and Electronics in Agriculture, 127, 311e323. https:// doi.org/10.1016/j.compag.2016.06.022.

Zhou, J., Fu, X., Zhou, S., Zhou, J., Ye, H., Nguyen, H.T., 2019. Automated segmentation of soybean plants from 3D point cloud using machine learning. Comput. Electron. Agric. 162, 143–153.