

Real-Time Fruit Recognition and Grasping Estimation for Autonomous Apple Harvesting

1st Hanwen Kang

Monash University

Department of Mechanical and Aerospace Engineering
Melbourne, Australia
hanwen.kang@monash.edu

2nd Chao Chen

Monash University

Department of Mechanical and Aerospace Engineering
Melbourne, Australia
chao.chen@monash.edu

Abstract—In this paper, a fully neural network based visual perception framework for autonomous apple harvesting is proposed. The proposed framework includes a multi-function neural network for fruit recognition and a Pointnet network to determine the proper grasp pose to guide the robotic execution. Fruit recognition takes raw input of RGB images from the RGB-D camera to perform fruit detection and instance segmentation. The Pointnet grasping estimation takes the point cloud of each fruit as input, and predict the grasp pose for each fruit as output. The proposed framework is validated by using RGB-D images collected from laboratory and orchard environments, robotic grasping experiments in a controlled environment are also included. Experimental results shows that the proposed framework can accurately localise fruits and estimate the grasp pose for robotic grasping.

Index Terms—Fruit recognition; Grasping estimation; Pose estimation; Pointnet; Autonomous harvesting.

I. INTRODUCTION

Autonomous harvesting plays a significant role in the recent development of the agricultural industry [1]. Vision is one of the essential tasks in autonomous harvesting, as it can detect and localise the crop, and guide the robotic arm to perform detachment [2]. Vision tasks in orchard environments are challenging as there are many factors influencing the performance of the system, such as variances in illumination, appearance, and occlusion between crop and other items within the environment. Meanwhile, occlusion between fruits and other items can also decrease the success rate of autonomous harvesting [3]. In order to increase the efficiency of harvesting, the vision system should be capable of guiding the robotic arm to detach the crop from a proper approach pose. Overall, an efficient vision algorithm which can robustly perform crop recognition and grasp pose estimation is the key to the success of autonomous harvesting [4].

In this work, a fully deep-learning based vision algorithm which can perform real-time fruit recognition and grasping estimation for autonomous apple harvesting by using sensory data from the RGB-D camera is proposed. The proposed method includes two function blocks: fruit recognition and grasping estimation. Fruit recognition applies a one-stage multi-task neural network to perform fruit detection and instance segmentation on colour images. Grasp pose estimation processes the information from the fruit recognition together

with depth information to estimate the proper grasp pose for each fruit by using the Pointnet. The following highlights are presented in the paper:

- Applying a multi-task neural network to perform fruit detection and instance segmentation on input colour images from RGB-D camera.
- Proposing a modified Pointnet-based network to perform fruit modelling and grasp pose estimation by using point clouds from RGB-D camera.
- Realising and combining the aforementioned two features to guide the robot to perform autonomous harvesting.

The rest of the paper is organised as follows. Section II reviews the related works on fruit recognition and grasp pose estimation. Section III introduces the methods of the proposed vision processing algorithm. The experimental setup and results are included in Section IV. In Section V, conclusion and future works are presented.

II. LITERATURE REVIEW

A. Fruit Recognition

Fruit recognition is an essential task in the autonomous agricultural applications [5]. There are many methods which have been studied in decades, including the traditional method [6]–[8] and deep-learning based method. Traditional method applies hand-crafted feature descriptors to describe the appearances of objects within images, and uses machine-learning algorithm to perform classification, detection, or segmentation by using extracted feature descriptors [9]. The performance of the traditional method is limited by the express ability of the feature descriptor, which required to be adjusted before applying in different conditions [10]. Deep-learning based method applies deep convolution neural network to perform automatic image feature extraction, which has shown the good performance and generalisation in many core tasks of the computer vision [11]. Deep-learning based detection method can be divided into two classes: two-stage detection and one-stage detection [12]. Two-stage detection applies a Region Proposal Network (RPN) to search the Region of Interest (RoI) from the image, and a classification branch is applied to perform bounding box regression and classification [13], [14]. One-stage detection combines the RPN and classification

into a single architecture, which speeds up the processing of the images [15], [16]. Both two-stage detection and one-stage detection have been widely studied in autonomous harvesting [17]. Bargoti and Underwood [18] applied Faster Region Convolution Neural Network (Faster-RCNN) to perform multi-class fruit detection in orchard environments. Yu et al. [19] applied Mask-RCNN [20] to perform strawberry detection and instance segmentation in the non-structural environment. Liu et al. [21] applied a modified Faster-RCNN on kiwifruit detection by combining the information from RGB and NIR images, an accurate detection performance was reported in this work. Tian et al. [22] applied an improved Dense-YOLO to perform monitoring of apple growth in different stages. Koirala et al. [23] applied a light-weight YOLO-V2 model which named as 'Mongo-YOLO' to perform fruit load estimation. Kang and Chen [24], [25] develop a multi-task network based on YOLO, which combines the semantic, instance segmentation, and detection in a one-stage network. To efficiently perform the robotic harvesting, the grasping estimation which can guide the accurate robotic harvesting is also required [26]. The aforementioned studies only applied detection network to perform fruit recognition while lack the ability of the grasping estimation.

B. grasping estimation

Grasp pose estimation is one of the key techniques in the robotic grasp [27]. Similar to the methods developed for fruit recognition, the grasp pose estimation methods can be divided into two categories: traditional analytical approaches and deep-learning based approaches [28]. Traditional analytical approaches extract feature/key points from the point clouds and then perform matching between sensory data and template from the database to estimate the object pose [29]. The pre-defined grasp pose can be applied in this condition. For the unknown objects, some assumption can be made, such as grasp the object along the principle axis [27]. The performance of the traditional analytical approaches is limited when being performed in the real world, the noise or partial point cloud can severely influence the accuracy of the estimation [30]. In the following development, deep-learning based methods recast the grasp pose estimation as an object detection task, which can directly produce grasp pose from the images [31]. Recently, with the development of the deep-learning architecture for 3D point cloud processing [32], [33], some studies focus on performing grasp pose estimation by using the 3D point clouds. These methods apply convolution neural network architectures to process the 3D point clouds and estimate the grasp pose to guide the grasping, such as Grasp Pose Detection (GPD) [34] and Pointnet GPD [35], which showed accurate performance in the specific conditions. In the robotic harvesting case, Lehnert et al. [36] modeled the sweep pepper as the super-ellipsoid and estimated the grasp pose by performing shape matching between the super-ellipsoid and fruit. In their following work [37], surface normal orientation of fruits were applied as grasp candidates and ranked by the an utility function, which is time consuming and not robust

to the outdoor environments. Some other studies [38]–[40] performed the grasping by translating towards the fruits, which can not secure the success rate of harvesting in unstructured environments. The aforementioned studies are limited to be applied in the specific conditions or not accurate and robust to the orchard environments. In this study, a Pointnet based grasping estimation is proposed to perform fruit modelling and grasp pose estimation by combining with the fruit recognition, which shows the accurate and robust performance in the experiments.

III. METHODS AND MATERIALS

A. System Design

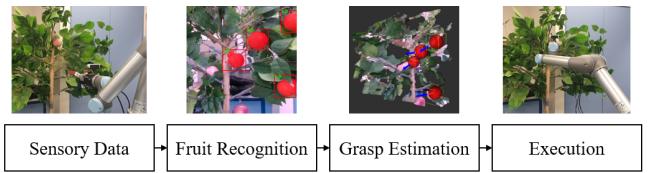


Fig. 1. Two-stage vision perception and grasping estimation for autonomous harvesting.

The proposed method include two-stages: fruit recognition and grasp pose estimation. The workflow of the proposed vision processing algorithm is shown in Figure 1. In the first step, the fruit recognition block performs fruit detection and segmentation on input RGB images from the RGB-D camera. The outputs of the fruit recognition are projected to the depth images, and the point clouds of each detected fruit are extracted and sent to the grasp pose estimation block for further processing. In the second step, the Pointnet architecture is applied to estimate the geometry and grasp pose of fruits by using the point clouds from the previous steps. The method of the fruit recognition block and grasp pose estimation are presented in Section III-B and III-C, respectively. The implementation details of the proposed method are introduced in Section III-D.

B. Fruit Recognition

1) *Network Architecture*: A one-stage neural network Dasnet [41] is applied to perform fruit detection and instance segmentation tasks. Dasnet applies a 50 layers residual network (resnet-50) [42] as the backbone to extract features from the input image. A three levels Feature Pyramid Network (FPN) is used to fuse feature maps from the C3, C4, and C5 level of the backbone (as shown in Figure 2). That is, the feature maps from the higher level are fused into the feature maps from the lower level since feature maps in higher level include more semantic information which can increase the classification accuracy [43].

On each level of the FPN, an instance segmentation (includes detection and instance segmentation) branch is applied, as shown in Figure 3. Before the instance segmentation branch, an Atrous Spatial Pyramid Pooling (ASPP) [44] is used to process the multi-scale features within the feature maps. ASPP

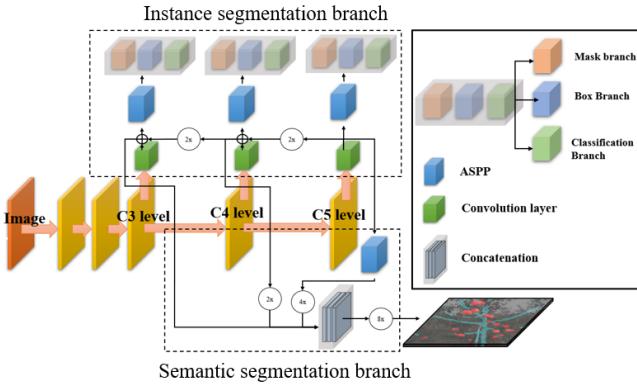


Fig. 2. Network architecture of the Dasnet [41], Dasnet is a one-stage detection network which combines detection, instance segmentation and semantic segmentation.

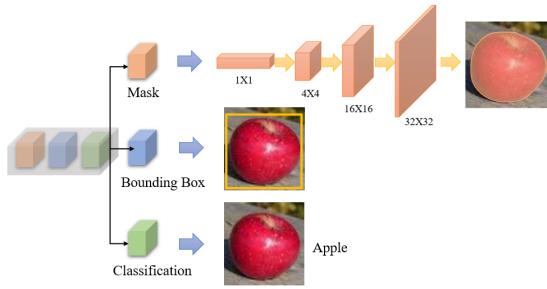


Fig. 3. Architecture of the instance segmentation branch, which can perform instance segmentation, bounding box regression, and classification.

applies dilation convolution with different rates (e.g. 1, 2, 4 in this work) to process the feature, which can process the features of different scale separately. The instance segmentation branch includes three sub-branches, which are mask generation branch, bounding box branch, and classification branch. Mask generation branch follows the architecture design proposed in Single Pixel Reconstruction Network (SPRNet) [45], which can predict a binary mask for objects from a single pixel within the feature maps. Bounding box branch includes the prediction on confidence score and the bounding box shape. We apply one anchor bounding box on each level of FPN (size of anchor box of instance segmentation branch on C3, C4 and C5 level are 32 x 32 (pixels), 80 x 80, and 160 x 160, respectively.). Classification branch predicts the class of the object within the bounding box. The combined outputs from the instance segmentation branch form the results of the fruit recognition on colour images. Dasnet also has a semantic segmentation branch for environment semantic modelling, which is not applied in this research.

2) *Network Training:* More than 1000 images are collected from apple orchards located in Qingdao, China and Melbourne, Australia. Types of apples, includes Fuji, Gala, Pink Lady, and so on. The images are labelled by using LabelImage tool from Github [46]. We applied 600 images as the training set, 100 images as the validation set, and 400 images as the test set. We introduce multiple image augmentations in the

network training, including random crop, random scaling (0.8-1.2), flip (horizontal only), random rotation ($\pm 10^\circ$), randomly adjust on saturation (0.8-1.2) and brightness (0.8-1.2) in HSV colour space. We apply focal loss [47] in the training and Adam-optimiser is used to optimise the network parameters. The learning rate and decay rate of the optimiser are 0.001 and 0.75 per epoch. We train the instance segmentation branch for 100 epochs and train the whole network for another 50 epochs.

3) *Post Processing:* The results of the fruit recognition are projected into the depth image. That is, the mask region of each apple on depth image is extracted. Then, the 3D position of each point in the point clouds of each apple is calculated and obtained. The generated point clouds are the visible part of the apple from the current view-angle of the RGB-D camera. These point clouds are further processed by grasp pose estimation block to estimate the grasp pose, which is introduced in the following section.

C. grasping estimation

1) *Grasp Planning:* Since most of the apples are presented in sphere or ellipsoid, we modelling the apple as sphere shape for simplified expression. In the natural environments, apples can be blocked by branches or other items within the environments from the view-angle of the RGB-D camera. Therefore, the visible part of the apple from the current view-angle of the RGB-D camera indicates the potential grasp pose, which is proper for the robotic arm to pick the fruit. Unlike GPD [34] or Pointnet GPD [35] which generates multiple grasp candidates and uses the network to determine the best grasp pose , we formulate the grasp pose estimation as an object pose estimation task which is similar to the Frustum PointNets [48]. We select the centre of the visible part and orientation from the centre of the apple to this centre as the position and orientation of the grasp pose (as shown in Figure 4). The Pointnet takes 1-viewed point cloud of each fruit as input and estimates the grasp pose for the robotic arm to perform detachment.

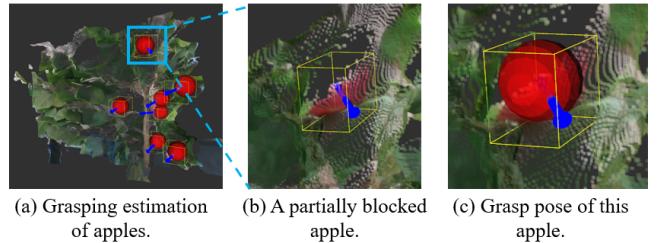


Fig. 4. Our method select orientation from the fruit centre to visible part centre as grasp pose.

2) *Grasp Representation:* The pose of an object in 3D space has 6 Degree of Freedom (DoF), includes three positions (x , y , and z), and three rotations (θ , ϕ , and ω , along Z-axis, Y-axis, and X-axis, respectively). We apply Euler-ZYX angle to represent the orientation of the grasp pose, as shown in Figure 5. The value of ω is set as zero since we can always assume that fruit will not rotate along its X-axis (since apples

are presented in a spherical shape). The grasp pose (GP) of an apple can be formulated as follow:

$$T_{GP} = \begin{bmatrix} \cos \theta \cos \phi & -\sin \theta & \cos \theta \sin \phi & x \\ \sin \theta \cos \phi & \cos \theta & \sin \theta \sin \phi & y \\ -\sin \phi & 0 & \cos \phi & z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

Therefore, a parameter list $[x, y, z, \theta, \phi]$ is used to represent the grasp pose of the fruit.

3) *Data Annotation*: Grasp pose block use point clouds as input and predicts the 3D Oriented Bounding Box (3D-OBB) (oriented in grasp orientation) for each fruit. Each 3D-OBB includes six parameters, which are x, y, z, r, θ, ϕ . The position (x, y, z) represents the offsets on X-, Y-, Z-axis from the centre of point clouds to the centre of the apple, respectively. The parameter r represents the radius of the apple, as the apples is modelled as sphere. The length, width, and height can be derived by radius. θ and ϕ represent the grasp pose of the fruit, as described in Section III-C2.

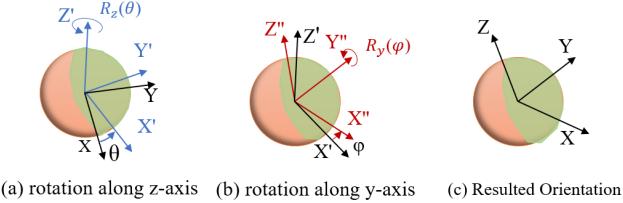


Fig. 5. Euler-ZYX angle is applied to represent the orientation of the grasp pose.

Since the values of the parameters x, y, z , and r may have large variances when dealing with prediction in different situations, a scale parameters S is introduced. We apply S to represent the mean scale (radius) of the apple, which equals 30 (cm) in our case. The parameters x, y, z , and r are divided by S to obtain the united offset and radius (x_u, y_u, z_u, r_u) . After remapping, the range of the x_u, y_u, z_u is reduced to $[-\infty, \infty]$, and the range of r_u are in $[0, \infty]$. To keep the grasp pose in the range of motion of the robotic arm, the θ and ϕ are limited in the range of $[-\frac{1}{4}\pi, \frac{1}{4}\pi]$. We divide the θ and ϕ by $\frac{1}{4}\pi$ to map the range of grasp pose into the range of $[-1, 1]$. The united θ and ϕ are denoted as θ_u and ϕ_u . In total, we have six united parameters to predict the 3D-OBB for each fruit, which are $[x_u, y_u, z_u, r_u, \theta_u, \phi_u]$. Among these parameters, $[x_u, y_u, z_u, \theta_u, \phi_u]$ represent the grasp pose of the fruit, r_u controls the shape of 3D-OBB.

4) *Pointnet Architecture*: Pointnet [32] is a deep neural network architecture which can perform classification, segmentation, or other tasks on point clouds. Pointnet can use raw point clouds of the object as input and does not requires any pre-processing. The architecture of the Pointnet is shown in Figure 6 and 7. Pointnet uses an $n \times 3$ (n is the number of points) unordered point clouds as input. Firstly, Pointnet applies convolution operations to extract a multiple dimensional feature vector on each point. Then, a symmetric function

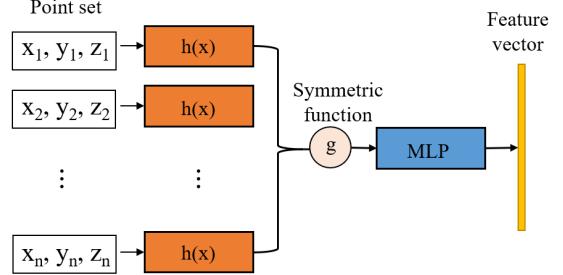


Fig. 6. Pointnet applies symmetric function to extract features from the unordered point cloud.

is used to extract the features of the point clouds on each dimension of the feature vector.

$$f(x_1, x_2, \dots, x_n) = g(h(x_1), h(x_2), \dots, h(x_n)) \quad (2)$$

In Eq. 2, g is a symmetric function and f is the extracted features from the set. Pointnet applies max-pooling as the symmetric function. In this manner, Pointnet can learn numbers of features from point set and invariant to input permutation. The generated feature vectors are further processed by Multi-Layer Perception (MLP) (fully-connected layer in Pointnet), to perform classification of the input point clouds. Batch-norm layer is applied after each convolution layer or fully-connection layer. Drop-out is applied in the fully-connected layer during the training.

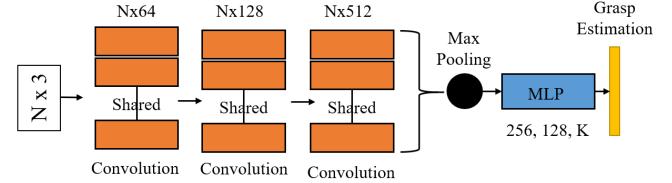


Fig. 7. Network architecture of the Pointnet applied in grasping estimation.

In this work, the output of the Pointnet is changed to the 3D-OBB prediction, which includes prediction on six parameters $[x_u, y_u, z_u, r_u, \theta_u, \phi_u]$. The range of the parameters x_u, y_u , and z_u are in $[-\infty, \infty]$, hence we do not applies an activation function on these three parameters. The range of the r_u are from 0 to ∞ , the exponential function is used as activation. The range of the θ_u, ϕ_u is from -1 to 1, hence a tanh activation function is applied. The Pointnet output before activation are denoted as $[x_p, y_p, z_p, r_p, \theta_p, \phi_p]$. Therefore, we have

$$\begin{cases} x_u, y_u, z_u = x_p, y_p, z_p, \\ r_u = \exp(r_p), \\ \theta_u, \phi_u = \tanh(\theta_p), \tanh(\phi_p). \end{cases} \quad (3)$$

The output of the Pointnet can be remapped to their original value by following the description in Section III-C3.

5) *Network Training*: The data labelling is performed on our own developed labelling tool, as shown in Figure 8. Our labelling tool records the six parameters of the 3D-OBB and

all the points within the point clouds. The training of the Pointnet for 3D-OBB prediction is independent of the fruit recognition network training. There are 570 1-viewed point clouds of apples labelled in total (250 are collected in lab, 250 are collected in orchards). We apply 300 point sets as the training set (150 in each data set), 50 samples as validation set (25 in each data set), and the rest 220 samples as test set (110 in each data set). We introduce scaling (0.8 to 1.2), translation (-15 cm to 15 cm on each axis), rotation (-10° to 10° on θ and ϕ), adding Gaussian noise (mean equals 0, variance equals 2cm), and adding outliers (1% to 5% in total number of point clouds) in the data augmentation. One should notice that the orientation of samples after augmentation should still in the range between $-\frac{1}{4}\pi$ and $\frac{1}{4}\pi$.

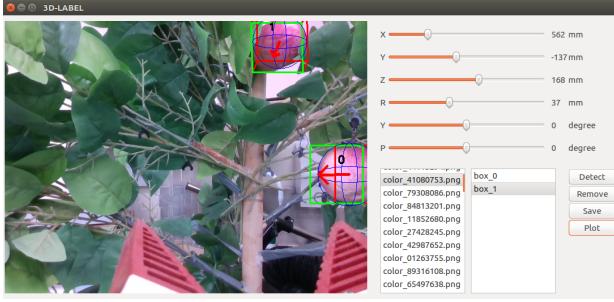


Fig. 8. The developed labelling tool for RGB-D images.

The square error between prediction and ground truth is applied as the training loss. The Adam-optimizer in Tensorflow is used to perform the optimisation. The learning rate, decay rate, and total training epoch of the applied optimiser are 0.0001, 0.6 /epoch, and 100 epochs, respectively.

D. Implementation Details

1) System Configuration and Software: The Intel-D435 RGB-D camera is applied in this research, a laptop (DELL-INSPRITION) with Nvidia-GPU GTX-980M and Intel-CPU i7-6700 is used to control the RGB-D camera and perform the test. The connection between RGB-D camera and laptop is achieved by using the RealSense communication package in the Robot Operation System (ROS) in kinetic version [49] on the Linux Ubuntu 16.04. The calibration between the colour image and the depth image of the RGB-D camera is included in the realsense-ros. The implementation code of the Pointnet (in Tensorflow) is from the Github [50], and it is trained on the Nvidia-GPU GTX-980M. The implementation code of the Dasnet is achieved by using Tensorflow. The training of the Dasnet is performed on the Nvidia-GPU GTX-1080Ti. In the autonomous harvesting experiment, an industry robotic arm Universal Robot UR5 is applied. The communication between UR5 and the laptop is performed by using universal-robot-ROS. MoveIt! [51] with TackIK inverse kinematic solver [52] is used in the motion planning of the robotic arm.

2) Point Clouds Pre-processing: An Euclidean distance based outlier rejection algorithm is applied to filter out outliers within point clouds before it is processed by Pointnet. When

the distance between a point and point clouds centre is two times larger than the mean distance between the points and centre, we consider this point as an outlier and reject it. This step is repeated three times to ensure the efficiency of rejection. To improve the inference efficiency, a voxel downsampling function (resolution 3 mm) from the 3D data processing library open3D is used. Then we randomly pick 200 points from the downsampled point sets as the input of the Pointnet grasping estimation. The point set with the number of points less than 200 after voxel downsampling will be rejected since the insufficient number of points are presented.

IV. EXPERIMENT AND DISCUSSION

A. Experiment Setup

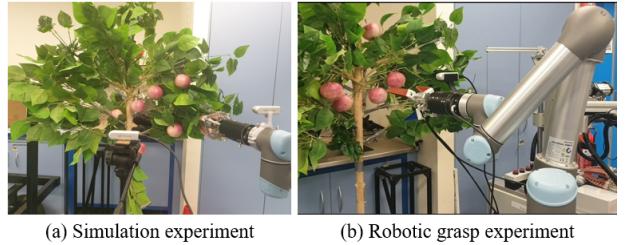


Fig. 9. Experiment setup in laboratory scenario.

We evaluate our proposed fruit recognition and grasping estimation algorithm in both simulation and the robotic hardware. In the simulation experiment, we perform the proposed method in the RGB-D data on the test set, which includes 110 point sets respectively in the laboratory environment and orchard environment. In the robotic harvesting experiment, we apply the proposed method to guide the robotic arm to perform the grasp of apples on the artificial plant in the lab. We apply IoU between predicted and ground-truth bounding box to evaluate the accuracy of 3D localisation and shape estimation of the fruits. We use 3D Axis Aligned Bounding Boxes (3D-AABB) to simplify the IoU calculation of 3D bounding box [53]. The IoU between 3D-AABB is denoted as IoU_{3D} . We set 0.75 ($\text{thres}_{\text{IoU}}$) as the threshold value for IoU_{3D} to determine the accuracy of fruit shape prediction. In terms of the evaluation of the grasp pose estimation, we apply absolute error between the predicted value and ground truth value of grasp pose, as it can intuitively show the accuracy of predicted grasp pose. The maximum accepted error of grasp pose estimation for the robot to perform a successful grasp is 8°, which is set as the threshold value in the grasp pose evaluation. This experiment is conducted in several scenarios, including noise and outlier presented conditions, and also dense clutter condition.

B. Simulation Experiments

In the simulation experiment, we compare our method with traditional shape fitting methods, which include sphere Random Sample Consensus (sphere-RANSAC) [54] and sphere Hough Transform (sphere-HT) [55], in terms of accuracy on

fruit localisation and shape estimation. Both RANSAC and HT based algorithms take point clouds as input and generate the prediction of the fruit shape. The 3D bounding box of predicted shapes are then used to perform accuracy evaluation and compared with our method. This comparison are conducted on RGB-D images collected from both laboratory and orchard scenarios.

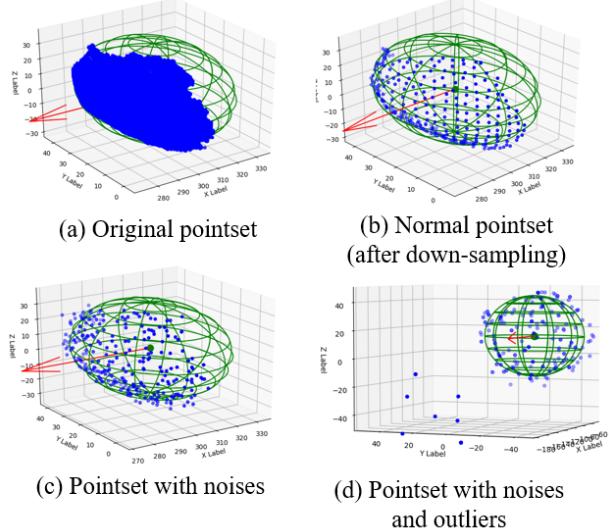


Fig. 10. Pointset under different conditions, green sphere is the ground truth of the fruit shape.

TABLE I

ACCURACY OF THE FRUIT SHAPE ESTIMATION BY USING POINTNET, RANSAC, AND HT IN DIFFERENT TESTS.

| | Normal | Noise | Outlier | Dense clutter | Combined |
|----------|--------|-------|---------|---------------|----------|
| Pointnet | 0.94 | 0.92 | 0.93 | 0.91 | 0.89 |
| RANSAC | 0.82 | 0.71 | 0.81 | 0.74 | 0.61 |
| HT | 0.81 | 0.67 | 0.79 | 0.73 | 0.63 |

1) Experiments in laboratory Environments: We performed Pointnet grasping estimation, RANSAC, and HT on the collected RGB-D images from the laboratory environment. The experimental results of three methods in different tests are shown in Table I. From the experimental results, Pointnet grasping estimation significantly increases the localisation accuracy of the 3D bounding box of the fruits. Pointnet grasping estimation achieves 0.94 on IoU_{3D} , which is higher than the RANSAC and HT methods, respectively. To evaluate the robustness of different methods when dealing with noisy and outlier conditions, we randomly add Gaussian noise (mean equals 0, variance equals 2cm) and outlier (1% to 5% in the total number of point clouds) to the point clouds, as shown in Figure 10. Three methods show similar robustness when dealing with outliers. Since both RANSAC and HT apply vote framework to estimate the primitives of the shape, which is robust to the outlier. However, when dealing with the noisy environment, Pointnet grasping estimation achieves better robustness, as compared to the RANSAC and HT. Since noisy point

clouds can influence the accuracy of vote framework to a large extent. We also tested Pointnet grasping estimation, RANSAC, and HT in dense clutter condition. grasping estimation in dense clutter condition is challenging since the point clouds of objects can be influenced by other neighbouring objects. Pointnet grasping estimation can robustly perform accurate localisation and shape fitting of apples in this condition, which shows a significant improvement, as compared to the performance of the RANSAC and HT algorithms. The experimental results obtained by using Pointnet grasping estimation are presented in Figure 11, and the 3D-OBBs are projected into image space by using the method applied in the work of Novak [56].

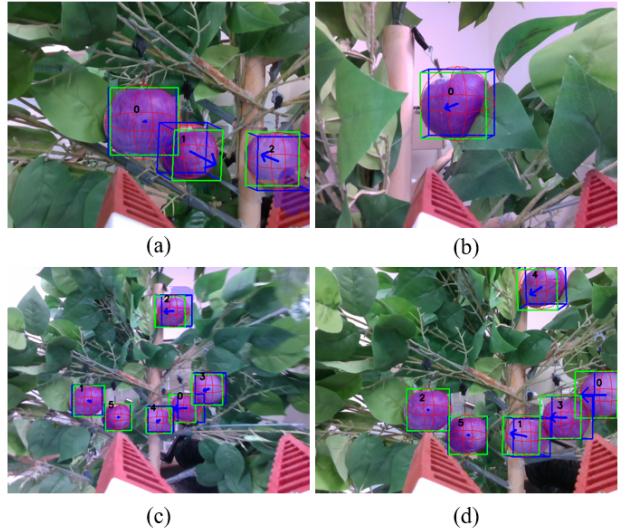


Fig. 11. grasping estimation by using Pointnet. The green box are the front of the 3D-OBB, blue arrows are the predicted grasp pose, red sphere are the predicted shape of the fruits.

TABLE II

MEAN ERROR OF GRASP ORIENTATION ESTIMATION BY USING POINTNET IN DIFFERENT TESTS.

| | Normal | Noise | Outlier | Dense clutter | Combined |
|----------|-------------|-------------|-------------|---------------|-------------|
| Pointnet | 3.2° | 5.4° | 4.6° | 4.8° | 5.5° |

In terms of the evaluation of the grasp orientation estimation, Pointnet grasping estimation shows accurate performance in the experimental results, as shown in Table II. The mean error between predicted grasp pose and ground truth grasp pose is 3.2° . Experimental results also show that Pointnet grasping estimation can accurately and robustly determine the grasp orientation of the objects in noisy, outlier presented, and dense clutter conditions.

TABLE III

PERFORMANCE EVALUATION OF FRUIT RECOGNITION IN RGB-D IMAGES COLLECTED IN ORCHARD SCENARIOS

| | F ₁ score | Recall | Accuracy | IoU _{mask} |
|--------|----------------------|--------|----------|---------------------|
| Dasnet | 0.873 | 0.868 | 0.88 | 0.873 |

2) *Experiments in Orchards Environments*: In this experiment, we performed the fruit recognition (Dasnet) and Pointnet grasping estimation on the collected RGB-D images from apple orchards. The performance of the Dasnet is evaluated by using the RGB images in test set. We apply F_1 score and IoU as the evaluation metric of the fruit recognition. IoU_{mask} stands the IoU value of instance mask of fruits in colour images. Table III show the performance of the Dasnet (in terms of the detection accuracy and recall) and Pointnet grasping estimation, Figure 12 shows fruit recognition results by using Dasnet on test set. Experimental results show that Dasnet performs well on fruit recognition in orchard environment, having 0.88 and 0.868 on accuracy and recall, respectively. The accuracy of the instance segmentation on apples is 0.873. The inaccuracy of the fruit recognition is due to the illumination and fruit appearance variances. From the experiments, we found that Dasnet can accurately detect and segment the apples in the most conditions.

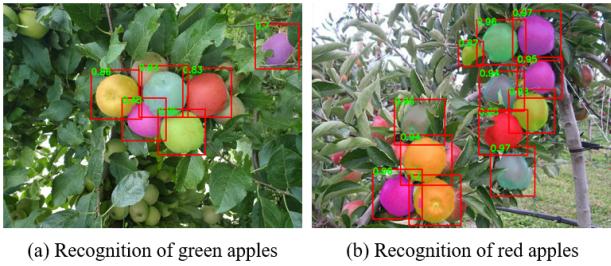


Fig. 12. Detection and instance segmentation performed by using Dasnet on collected RGB images.

TABLE IV

EVALUATION ON GRASP POSE ESTIMATION BY USING POINTNET IN DIFFERENT TESTS IN THE ORCHARD SCENARIO.

| | Pointnet | RANSAC | HT |
|-------------------|----------|--------|------|
| Accuracy | 0.88 | 0.76 | 0.78 |
| Grasp Orientation | 5.2° | - | - |

Table IV shows the performance comparison between Pointnet grasping estimation, RANSAC, and HT. In the orchard environments, grasp pose estimation is more challenging compared to the indoor environments. The sensory depth data can be affected by the various environmental factors, as shown in Figure 14. In this condition, the performance of the RANSAC and HT show the significant decrease from the indoor experiment while Pointnet grasping estimation shows better robustness. The IoU_{3D} achieved by Pointnet grasping estimation, RANSAC, and HT in orchard scenario are 0.88, 0.76, and 0.78, respectively. In terms of the grasp orientation estimation, Pointnet grasping estimations show robust performance in dealing with flawed sensory data. The mean error of orientation estimation by using Pointnet grasping estimation is 5.2°, which is still within the accepted range of orientation error. The experimental results of grasp pose estimation by using Pointnet grasping estimation in orchard scenario is shown in Figure 13.

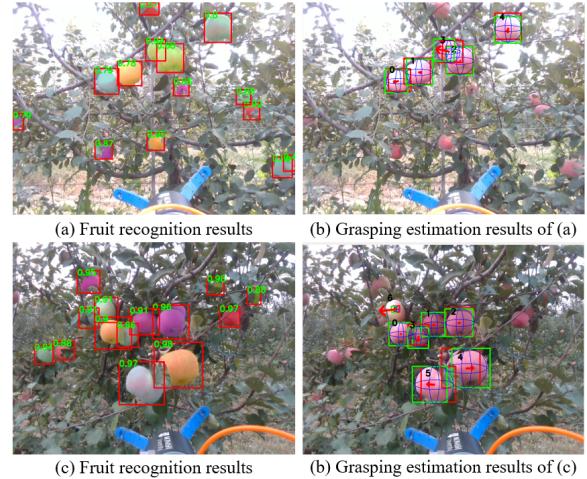


Fig. 13. Fruit recognition and grasping estimation experiments in orchard scenario.

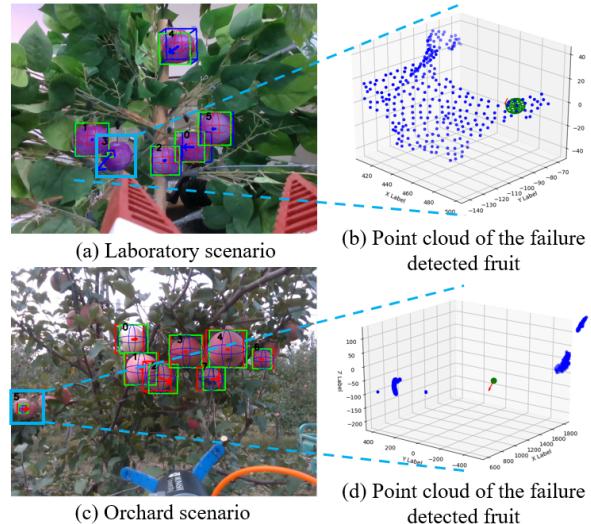


Fig. 14. Failure grasping estimation in laboratory and orchard scenarios.

3) *Common Failures in grasping estimation*: The major reason leading to the grasping estimation failure by using Pointnet grasping estimation is due to the sensory data defect, as shown in Figure 14. When under this condition, the results of Pointnet grasping estimation will always predicts a sphere with a very small value of radius. We can apply a radius value threshold to filter out this kind of failure during the operation.

C. Experiments of Robotic Harvesting

The Pointnet grasping estimation was tested by using a UR5 robotic arm to validate its performance in the real working scenario. We arranged apples on a fake plant in the laboratory environment, which is shown in Figure 9. We conducted multiple trials (each trial contains three to seven apples on the fake plant) to evaluate the success rate of the grasp. The success rate records a fraction of success grasps in the total number of grasp attempts. The operational procedures follow

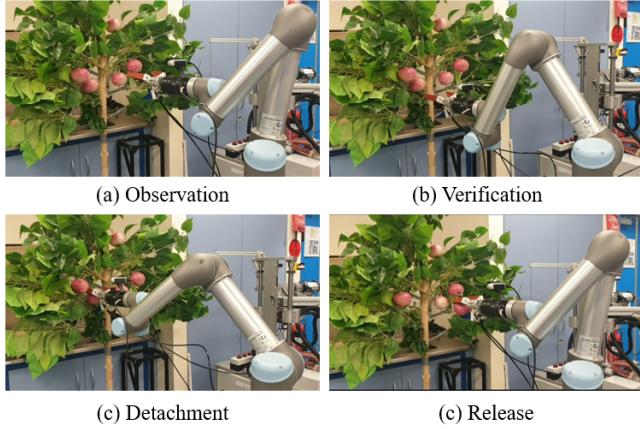


Fig. 15. Autonomous harvesting experiment in the laboratory scenario.

the design of our previous work [57], as shown in Figure 15. We simulate the real outdoor environments of autonomous harvesting by adding noises and outliers into the depth data. We also tested our system in dense clutter condition. The experimental results are shown in Table V.

TABLE V
EXPERIMENTAL RESULTS ON ROBOTIC GRASP BY USING POINTNET GRASPING ESTIMATION IN LABORATORY SCENARIO

| | Normal | Noise | Outlier | Dense clutter | Combined |
|--------------|--------|-------|---------|---------------|----------|
| success rate | 0.91 | 0.87 | 0.90 | 0.84 | 0.837 |

From the experimental results presented in Table V, Pointnet grasping estimation performs efficiently in the robotic grasp tests. Pointnet grasping estimation achieves accurate grasp results on normal, noise, and outlier conditions, which are 0.91, 0.87, and 0.9, respectively. In dense clutter condition, the success rate shows a decrease compared to the previous conditions. The reason for the success rate decreasing in dense clutter condition is due to the collision between gripper and fruits side by side. When collision presented in the grasp, it will cause the shift of the target fruit and lead to the failure of the grasp. This defect can be either improved by re-design the gripper or propose multiple grasp candidates to avoid the collision. The collision between gripper and branches can also lead to grasping failure in the other three conditions. Although such defect can affect the success rate of robotic grasp, it still achieves good performance in experiments. The success rate of robotic grasp under dense clutter condition and that all of factors combined condition are 0.84 and 0.837, respectively. The average running time of the fruit recognition and grasping estimation for one frame RGB-D image (5-7 apples included) is about 0.32 seconds on GTX-980M, showing a real-time ability to be performed in the robotic harvesting.

V. CONCLUSION AND FUTURE WORK

In this work, a fully deep-learning neural network based fruit recognition and grasping estimation method were proposed and validated. The proposed method includes a multi-

functional network for fruit detection and instance segmentation, and a Pointnet grasping estimation to determine the proper grasp pose of each fruit. The proposed multi-function fruit recognition network and Pointnet grasping estimation network was validated in RGB-D images taken from the laboratory and orchard environments. Experimental results showed that the proposed method could accurately perform visual perception and grasp pose estimation. The Pointnet grasping estimation was also tested with a robotic arm in a controlled environment, which achieved a high grasping success rate (0.847 in all factor combined condition). Future works will focus on optimising the design of the end-effector and validating the developed robotic system in the coming harvest season.

ACKNOWLEDGEMENT

This research is supported by ARC ITRH IH150100006 and THOR TECH PTY LTD. We also acknowledge Zhuo Chen for her assistance in preparation of this work.

REFERENCES

- [1] J. P. Vasconez, G. A. Kantor, and F. A. A. Cheein, "Human–robot interaction in agriculture: A survey and current challenges," *Biosystems engineering*, vol. 179, pp. 35–48, 2019.
- [2] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "Deepfruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, p. 1222, 2016.
- [3] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan, "Harvesting robots for high-value crops: State-of-the-art review and challenges ahead," *Journal of Field Robotics*, vol. 31, no. 6, pp. 888–911, 2014.
- [4] Y. Zhao, L. Gong, Y. Huang, and C. Liu, "A review of key techniques of vision-based control for harvesting robot," *Computers and Electronics in Agriculture*, vol. 127, pp. 311–323, 2016.
- [5] A. Vibhute and S. Bodhe, "Applications of image processing in agriculture: a survey," *International Journal of Computer Applications*, vol. 52, no. 2, 2012.
- [6] G. Lin, Y. Tang, X. Zou, J. Xiong, and Y. Fang, "Color-, depth-, and shape-based 3d fruit detection," *Precision Agriculture*, pp. 1–17, 2019.
- [7] G. Lin, Y. Tang, X. Zou, J. Cheng, and J. Xiong, "Fruit detection in natural environment using partial shape matching and probabilistic hough transform," *Precision Agriculture*, pp. 1–18, 2019.
- [8] L. Fu, E. Tola, A. Al-Mallahi, R. Li, and Y. Cui, "A novel image processing algorithm to separate linearly clustered kiwifruits," *Biosystems engineering*, vol. 183, pp. 184–195, 2019.
- [9] K. Kapach, E. Barnea, R. Mairon, Y. Edan, and O. Ben-Shahar, "Computer vision for fruit harvesting robots—state of the art and challenges ahead," *International Journal of Computational Vision and Robotics*, vol. 3, no. 1/2, pp. 4–34, 2012.
- [10] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.
- [11] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.
- [12] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.
- [13] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.
- [15] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

- [17] H. Kang and C. Chen, "Fast implementation of real-time fruit detection in apple orchards using deep learning," *Computers and Electronics in Agriculture*, vol. 168, p. 105108, 2020.
- [18] S. Bargoti and J. Underwood, "Deep fruit detection in orchards," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3626–3633, IEEE, 2017.
- [19] Y. Yu, K. Zhang, L. Yang, and D. Zhang, "Fruit detection for strawberry harvesting robot in non-structural environment based on mask-rcnn," *Computers and Electronics in Agriculture*, vol. 163, p. 104846, 2019.
- [20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.
- [21] Z. Liu, J. Wu, L. Fu, Y. Majeed, Y. Feng, R. Li, and Y. Cui, "Improved kiwifruit detection using pre-trained vgg16 with rgb and nir information fusion," *IEEE Access*, 2019.
- [22] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved yolo-v3 model," *Computers and electronics in agriculture*, vol. 157, pp. 417–426, 2019.
- [23] A. Koirala, K. Walsh, Z. Wang, and C. McCarthy, "Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of mangoyolo," *Precision Agriculture*, pp. 1–29, 2019.
- [24] H. Kang and C. Chen, "Fruit detection and segmentation for apple harvesting using visual sensor in orchards," *Sensors*, vol. 19, no. 20, p. 4599, 2019.
- [25] H. Kang and C. Chen, "Fast implementation of real-time fruit detection in apple orchards using deep learning," *Computers and Electronics in Agriculture*, vol. 168, p. 105108, 2020.
- [26] G. Lin, Y. Tang, X. Zou, J. Xiong, and J. Li, "Guava detection and pose estimation using a low-cost rgb-d sensor in the field," *Sensors*, vol. 19, no. 2, p. 428, 2019.
- [27] S. Chitta, E. G. Jones, M. Ciocarlie, and K. Hsiao, "Perception, planning, and execution for mobile manipulation in unstructured environments," *IEEE Robotics and Automation Magazine, Special Issue on Mobile Manipulation*, vol. 19, no. 2, pp. 58–71, 2012.
- [28] S. Caldera, A. Rassau, and D. Chai, "Review of deep learning methods in robotic grasp detection," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 57, 2018.
- [29] A. Aldoma, Z. Marton, F. Tombolini, W. Wohlkinger, C. Potthast, B. Zeisl, and M. Vincze, "Three-dimensional object recognition and 6 dof pose estimation," *IEEE Robotics & Automation Magazine*, pp. 80–91, 2012.
- [30] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.
- [31] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4–5, pp. 705–724, 2015.
- [32] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.
- [33] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in neural information processing systems*, pp. 5099–5108, 2017.
- [34] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 598–605, IEEE, 2016.
- [35] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3629–3635, IEEE, 2019.
- [36] C. Lehnert, I. Sa, C. McCool, B. Upcroft, and T. Perez, "Sweet pepper pose detection and grasping for automated crop harvesting," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2428–2434, IEEE, 2016.
- [37] C. Lehnert, A. English, C. McCool, A. W. Tow, and T. Perez, "Autonomous sweet pepper harvesting for protected cropping systems," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 872–879, 2017.
- [38] Y. Si, G. Liu, and J. Feng, "Location of apples in trees using stereoscopic vision," *Computers and Electronics in Agriculture*, vol. 112, pp. 68–74, 2015.
- [39] H. Yaguchi, K. Nagahama, T. Hasegawa, and M. Inaba, "Development of an autonomous tomato harvesting robot with rotational plucking gripper," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 652–657, IEEE, 2016.
- [40] Y. Onishi, T. Yoshida, H. Kurita, T. Fukao, H. Arihara, and A. Iwai, "An automated fruit harvesting robot by using deep learning," *ROBOMECH Journal*, vol. 6, no. 1, p. 13, 2019.
- [41] H. Kang and C. Chen, "Fruit detection, segmentation and 3d visualisation of environments in apple orchards," *Computers and Electronics in Agriculture*, vol. 171, p. 105302, 2020.
- [42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [43] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.
- [44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [45] J. Yu, J. Yao, J. Zhang, Z. Yu, and D. Tao, "Sprnet: Single-pixel reconstruction for one-stage instance segmentation," *IEEE Transactions on Cybernetics*, pp. 1–12, 2020.
- [46] Tzutalin, "LabelImg," <https://github.com/tzutalin/labelImg>, 2015. Git code (2015).
- [47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.
- [48] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 918–927, 2018.
- [49] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, p. 5, Kobe, Japan, 2009.
- [50] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," <https://github.com/charlesq34/pointnet>, 2016.
- [51] I. A. Sucan and S. Chitta, "Moveit!," <http://moveit.ros.org>, 2016.
- [52] P. Beeson and B. Ames, "Trac-ik: An open-source library for improved solving of generic inverse kinematics," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 928–935, IEEE, 2015.
- [53] J. Xu, Y. Ma, S. He, and J. Zhu, "3d-giou: 3d generalized intersection over union for object detection in point cloud," *Sensors*, vol. 19, no. 19, p. 4093, 2019.
- [54] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," in *Computer graphics forum*, vol. 26, pp. 214–226, Wiley Online Library, 2007.
- [55] A. Torii and A. Imiya, "The randomized-hough-transform-based method for great-circle detection on sphere," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1186–1192, 2007.
- [56] L. Novak, *Vehicle detection and pose estimation for autonomous driving*. PhD thesis, Masters thesis, Czech Technical University in Prague, 2017.
- [57] H. Kang, H. Zhou, and C. Chen, "Visual perception and modelling for autonomous apple harvesting," *IEEE Access*, pp. 1–1, 2020.