agronomy

MDPI

Article

Realtime Picking Point Decision Algorithm of Trellis Grape for High-Speed Robotic Cut-and-Catch Harvesting

Zhujie Xu [1], Jizhan Liu [1,2,3,*], Jie Wang [1], Lianjiang Cai [1], Yucheng Jin [1,2,3], Shengyi Zhao [1,2,3] and Binbin Xie [1]

[1] Key Laboratory of Modern Agricultural Equipment and Technology, Ministry of Education, Jiangsu University, Zhenjiang 212013, China; 2212016051@stmail.ujs.edu.cn (Z.X.); 2222116026@stmail.ujs.edu.cn (J.W.); 2222016002@stmail.ujs.edu.cn (L.C.); 1000003026@ujs.edu.cn (Y.J.); 2111916017@stmail.ujs.edu.cn (S.Z.); 2111816004@stmail.ujs.edu.cn (B.X.)

[2] National Digital Agricultural Equipment (Artificial Intelligence and Agricultural Robotics) Innovation Sub-Centre, Jiangsu University, Zhenjiang 212013, China

[3] Key Laboratory for Theory and Technology of Intelligent Agricultural Machinery and Equipment, Jiangsu University, Zhenjiang 212013, China

* Correspondence: liujizhan@163.com; Tel.: +86-0511-8879-7338

Abstract: For high-speed robotic cut-and-catch harvesting, efficient trellis grape recognition and picking point positioning are crucial factors. In this study, a new method for the rapid positioning of picking points based on synchronous inference for multi-grapes was proposed. Firstly, a three-dimensional region of interest for a finite number of grapes was constructed according to the "eye to hand" configuration. Then, a feature-enhanced recognition deep learning model called YOLO v4-SE combined with multi-channel inputs of RGB and depth images was put forward to identify occluded or overlapping grapes and synchronously infer picking points upwards of the prediction boxes of the multi-grapes imaged completely in the three-dimensional region of interest (ROI). Finally, the accuracy of each dimension of the picking points was corrected, and the global continuous picking sequence was planned in the three-dimensional ROI. The recognition experiment in the field showed that YOLO v4-SE has good detection performance in various samples with different interference. The positioning experiment, using a different number of grape bunches from the field, demonstrated that the average recognition success rate is 97% and the average positioning success rate is 93.5%; the average recognition time is 0.0864 s; and the average positioning time is 0.0842 s. The average positioning errors of the x, y, and z directions are 2.598, 2.012, and 1.378 mm, respectively. The average positioning error of the Euclidean distance between the true picking point and the predicted picking point is 7.69 mm. In field synchronous harvesting experiments with different fruiting densities, the average recognition success rate is 97%; the average positioning success rate is 93.606%; and the average picking success rate is 92.78%. The average picking speed is 6.18 s × bunch^{-1}, which meets the harvesting requirements for high-speed cut-and-catch harvesting robots. This method is promising for overcoming time-consuming harvesting caused by the problematic positioning of the grape stem.

Keywords: trellis grape; cut-and-catch; YOLO v4; picking point; positioning

check for updates

Citation: Xu, Z.; Liu, J.; Wang, J.; Cai, L.; Jin, Y.; Zhao, S.; Xie, B. Realtime Picking Point Decision Algorithm of Trellis Grape for High-Speed Robotic Cut-and-Catch Harvesting. *Agronomy* **2023**, *13*, 1618. https://doi.org/10.3390/agronomy13061618

Academic Editor: Chao Chen

Received: 8 May 2023
Revised: 3 June 2023
Accepted: 13 June 2023
Published: 15 June 2023

1. Introduction

Grapes are one of the most important fruits in the world [1], and China is the main producer of grapes. Due to the intensification of grape industrialization and the increase in demand, it has become a realistic remand to realize the automation of harvesting [2]. Various prototypes of grape harvesting robots have been developed in China, Greece, Canada, and other countries, which all adopt classic hand–eye calibration [3–5]. They all pick grapes by stem clipping and cutting, heavily relying on visual positioning accuracy, and then place the grapes into bins. Inadequate efficiency has become the key bottleneck standing in the way of meeting the actual needs of grape production. Significantly improving work efficiency has become an urgent task in technical development. Inspired by the high-efficiency mechanized harvesting of grain and leaf vegetables, a vision-based high-speed cutting-and-catching harvesting robot is proposed by our team, which transforms the traditional precise positioning and picking method into a continuous harvesting-by-vision method. It provides new possibilities to break through the efficiency ceiling of robotic harvesting. Different from the working process of combine harvesters, vision is still necessary for fresh grape harvesting. Nevertheless, the need of vision technology for this cut-and-catch scheme is greatly different from traditional harvesting robots.

At present, scholars from all over the world have made some progress in research on the positioning of trellis grape picking points. Computer vision and deep learning methods based on a convolutional neural network are widely used in the precise recognition of ripe grape picking points due to their strong robustness, strong adaptability, and high accuracy [6–10]. Fusing depth information to filter the foreground and background in fruit recognition is an effective means to improve the accuracy of fruit recognition and the robustness of the picking point location [11–13]. There are two main methods to accurately position the picking point coordinates: traditional image processing models [14–17] and deep learning models [18–21]; these methods directly detect grape stems and infer the location of picking points after identifying the grapes. Jin. et al. [22] constructed a far–close range stereoscopic vision system to identify grape ears and stems through image segmentation and a morphological operation. Luo. et al. [23] proposed a method based on binocular stereo vision to detect and extract the three-dimensional space coordinates of the cutting point on the grape stem. Xiong. et al. [15] analyzed the shape of grapes and used the Ostu threshold to segment grapes and stems to realize the positioning of disturbed grape picking points. Lei. et al. [24] proposed a method based on image segmentation and minimum angle constraints to determine the best grape picking points via corner detection and K-means clustering. Yin. et al. [25] used a Mask R-CNN model to detect a mask image of the target grapes for the collision-free picking problem and extracted a grape point cloud to estimate the picking position, with an average accuracy of 89.53%, a recall of 95.33%, and a time of 1.7 s. Kalampokas, T. et al. [26] proposed a regression convolutional neural network (RegCNN) model to implement grape stem detection via images. Li. et al. [27] proposed a grape and picking point detection model, YOLO-Grape, by adding an attention mechanism and introducing depth separable convolution. However, the diversity of fruit types and appearance, the variability of lighting conditions, and the complexity of natural backgrounds often lead to poor recognition results. Additionally, grape stems are frequently submerged by complex backgrounds and interference due to occlusion caused by branches and leaves, as well as the similar color of grape stems and vines, making it challenging to locate the picking point. As a result, research on identifying grape clusters and subsequently inferring the picking point has been conducted often to try and overcome these issues. Zhang et al. [17] used a machine vision reverse recognition algorithm combined with a template-matching algorithm to infer the two-dimensional positioning of grape picking points. Zhao. et al. [28] proposed a lightweight end-to-end model of YOLO-GP to simultaneously detect grape clusters and predict the picking point based on the grape bounding boxes. Although the interference of the background was reduced, the impact of different end effectors on the estimation and calculation of the picking point was not considered, which may easily lead to the inability of the end effector to accurately clamp the grape stem.

The grape picking point detection algorithms currently being researched utilize the method of accurately identifying fruit stems to achieve the precise positioning of the picking point. However, this approach suffers from the issue of time-consuming visual positioning, heavily relying on its accuracy. Moreover, the complex trellis environment introduces numerous interference factors that hinder the recognition of grape stems. Firstly, thinner fruit stems located behind branches and leaves cannot be detected due to occlusion. Secondly, the similarity in color between grape stems and grape vines makes it difficult to extract them from complex backgrounds. Furthermore, it is necessary to accurately detect the grape stem from a wide field of view through the method of close-range recognition to precisely clip the fruit stem with the end effector. Close-range recognition can be achieved by placing the camera close enough to the grapes, but this approach reduces the camera's field of view and makes global continuous picking sequence planning impossible. Therefore, in this study a specific visual algorithm was proposed for high-speed robotic cut-and-catch harvesting, which eliminates the need to directly locate the picking point of the grape stem. This vision algorithm is based on the improved YOLO v4-SE model, enabling the inference of picking points and avoiding interference from the complex background of a trellis environment in fruit stem recognition.

The main contributions of this paper are summarized as follows:

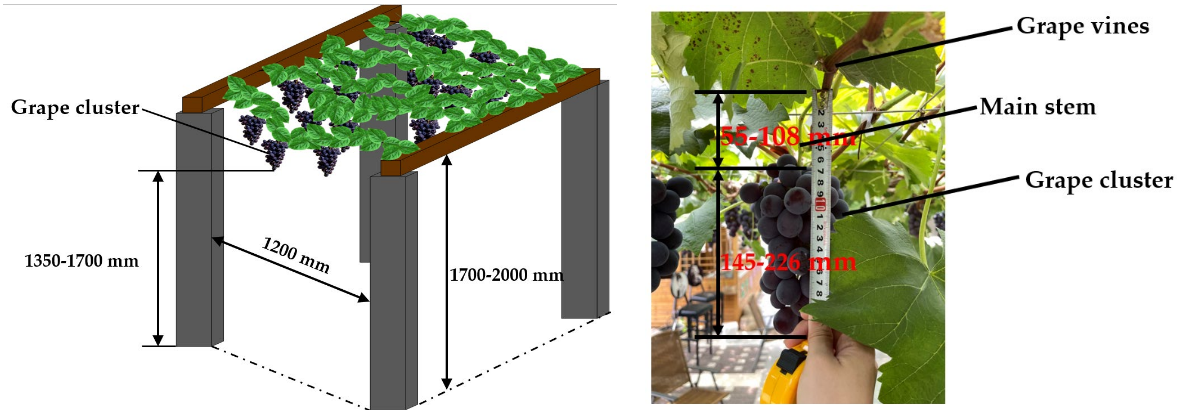
- This paper proposed a three-dimensional ROI (ROI means the region of interest) containing a finite number of grapes divided for subsequent harvesting work using depth images and RGB images obtained using Realsense D455, with the model being developed according to the structure of the robot body and working parameters.
- In this paper, the SE-Net attention mechanism module and the distance threshold segmentation module were deeply integrated into the feature-enhanced YOLO v4-SE model with multi-channel inputs to realize the synchronous recognition of multi-target grapes in the three-dimensional ROI, including overlapping or occluded grapes and grapes imaged completely.
- After the synchronous recognition of multi-grapes, a method based on the inference from the center point of the prediction boxes was proposed for the rapid positioning of the grape stem picking point for the first time, combining a disc knife cutting end effector and the physical properties of the trellis grapes. It relied on Gaussian distance weights to plan the picking sequence in the three-dimensional ROI.

2. Materials and Methods

2.1. System Analysis

2.1.1. Technical Solution of the High-Speed Cut-and-Catch Robot

The horizontal trellis cultivation of table grapes, shown Figure 1, is performed by drawing metal wires on the beams connected to the vertical cement columns, so as to form a grid surface parallel to the ground. Trellis grapes are structurally adjacent and hang vertically under the barbed wire in a vertical posture. The horizontal trellis cultivation of table grapes has the following characteristics.



(**a**) Horizontal trellis cultivation environment structure. (**b**) Ripe grape grown on horizontal trellis.

Figure 1. Relative parameters of horizontal trellis cultivated grapes.

(1) The distribution of ripe grapes is random and discrete; only relying on high-speed cut-and-catch machinery harvesting will affect work efficiency. It is necessary to plan

the picking sequence first and then combine the high-speed cut-and-catch operation method to really improve work efficiency.

(2) The length of a main stem is generally between 5.5 cm and 10.8 cm after field investigation and measurement, which is ideal for high-speed cut-and-catch harvesting with a disc knife end effector. However, it is necessary to position the picking point of the stem to avoid damaging the grapes or colliding with the trellis in the height direction during the cutting process.

(3) The terrain of the trellis vineyard is uneven, which will affect the working height of the disc knife cutting end effector. Thus, it is necessary to adjust the working height of the end effector by positioning the picking point, which can reduce working errors caused by uneven terrain.

Due to the cultivation characteristics of trellis grapes, Figure 2 shows the high-speed cut-and-catch harvesting robotic solution. It consists of a vision system, an operational harvest system, a chassis self-propelled system, and a harvesting fruit bin-lifting mechanism. In order to meet the vision technology requirements of this robot solution, a rapid rough positioning technology for picking points based on the disc knife cutting end effector is proposed for the first time in this paper. Rough positioning refers to a picking point positioning method with certain precision based on the upward inference achieved with the vision system, which is suitable for the disc knife cutting end effector with a continuous cutting method, certain working diameter, good error tolerance, and efficient catching after cutting.

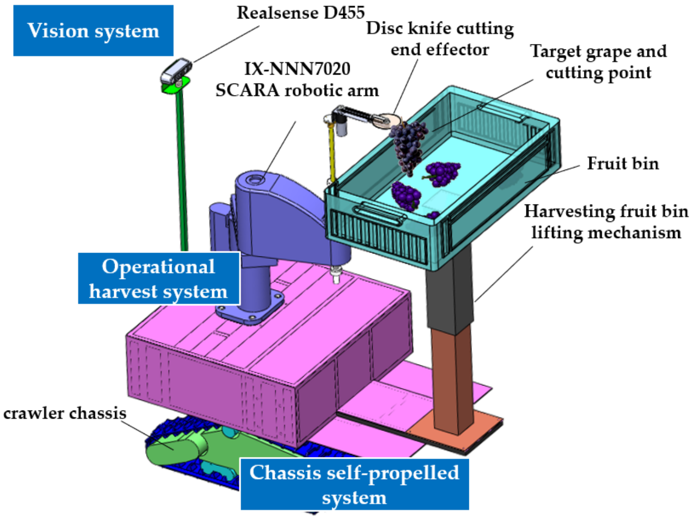

Vision system
Realsense D455
Disc knife cutting end effector
IX-NNN7020 SCARA robotic arm
Target grape and cutting point
Fruit bin
Harvesting fruit bin lifting mechanism
Operational harvest system
crawler chassis
Chassis self-propelled system

Figure 2. Various functional systems of the robot.

The Realsense D455 depth camera in the vision system is mounted on the middle of the robot tail and can obtain full and sufficient RGB images and depth images of trellis grapes. The operational harvesting system consists of an IX-NNN7020 SCARA robotic arm and a disc knife cutting end effector. A SCARA robotic arm is installed in the center of the robot body. The disc knife cutting end effector is attached to the end effector of the SCARA robotic arm and always faces the front of the robot. The output of the picking point coordinates and the planned picking sequence of the vision system drive the robotic arm for movement and the end effector for continuous cutting. The chassis self-propelled system realizes the walking. Through the position information of the picking points that is fed back by the vision system, the height of the fruit bin is adjusted to catch the cut grapes smoothly. The high efficiency and continuous harvesting of the robot are finally realized through the high cooperation between the vision scheme and the robotic solution.

2.1.2. Overall Visual Scheme

Trellis vineyards have many unique structural features. To realize the robot solution for continuous cutting and catching and achieve the high-speed advantage of the disc knife cutting end effector, the formulation of the special vision scheme should meet the following requirements in this special working environment.

(1) Considering the working width limitations imposed by the trellis vineyard, the vertical profiling that avoids colliding with the trellis, and the elimination of interference of irrelevant and complex fruits, foreground, and background in the walking operation, this paper designs a three-dimensional ROI to achieve a robotic solution for continuous cutting and catching. This restricted harvest region containing only a finite quantity of grapes is the three-dimensional ROI for robotic harvesting.

(2) The vision algorithm mainly includes the synchronous recognition of a finite quantity of grapes, the inference of picking points upwards along the fruit, global continuous picking sequence planning, and visual feedback, which all serve for the three-dimensional ROI, so as to realize high-speed continuous cutting and catching for the disc knife cutting end effector.

For the high-speed cut-and-catch robotic harvesting of the trellis grape, the robot will cut all the grapes that meet the harvesting requirements after one recognition and make them fall into the fruit box. Aiming to deliver high-speed continuous harvesting using the robot, the field of view of the far-range field of view and the positioning information of the picking point determine the overall vision scheme design of the robot (Table 1).

Table 1. Key hand-eye positions and parameters of the high-speed cut-and-catch trellis grape harvesting robot.

Key Position	Related Hardware Parameters	Related Parameters of Horizontal Trellis Cultivated Grape	Vision Information
Far-range field of view	• Workspace of SCARA robotic arm; • Fruit bin installation position and size; • Work features and methods of disc knife cutting end effector; • Work width of robot	• Fruiting height • Distribution regularity • Stem length • Fruit shape • Fruit size	RGB and depth image information of grapes
Picking point	• Disc knife working parameters • Height of SCARA robotic arm end effector • Height of fruit bin-lifting mechanism		Pixel and spatial information of picking point

The overall framework of the visual scheme for the high-speed cutting and catching harvest robot is shown in Figure 3.

(1) Firstly, the regularity of fruit distribution, the structure of the robot body, and working parameters are determined using the trellis grape viticulture environment and the purpose of high-speed cutting and catching. The structure of the robot body is formed by mutual cooperation between the maximum working area of SCARA, the installation position and size of the fruit bin, and the features and working methods of the disc knife, which all help to clarify the relationship between Realsense D455 and the end effector.

(2) Secondly, a three-dimensional ROI of a finite number of harvest grapes is marked out according to the above information. Then, a feature perception-enhanced model is used for the synchronous recognition of the trellis grapes in the three-dimensional ROI. The picking points of the multi-grapes are synchronously inferred upwards along the fruit recognition boxes.

(3) Thirdly, the rough positioning of the picking points using the fusion of multi-dimensional information is confined to the cutting area with differences in three-dimensional

directions, and then the picking points are calculated reliably and quickly based on the corner information.

(4) Finally, picking points are sent to the PLC controller continuously on the basis of the global continuous picking sequence planned in the ROI area, and occluded and overlapping grapes are separated with the fusion of depth information during the process, which significantly helps to improve the efficiency of the cutting and catching robot.

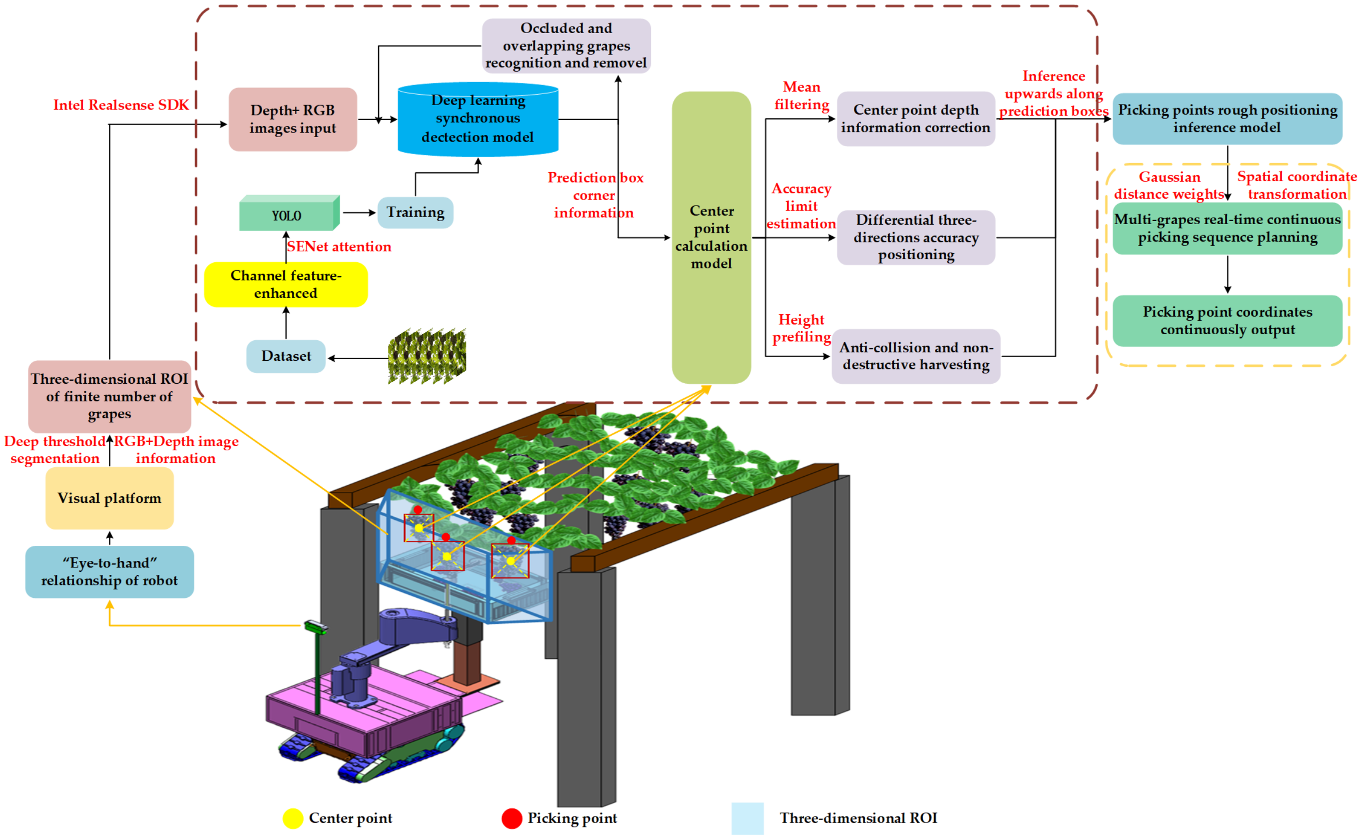

Intel Realsense SDK
Depth+ RGB images input
Deep learning synchronous dectection model
Occluded and overlapping grapes recognition and removel
YOLO
Training
SENet attention
Channel feature-enhanced
Dataset
Prediction box corner information
Center point calculation model
Mean filtering
Center point depth information correction
Accuracy limit estimation
Differential three-directions accuracy positioning
Height prefiling
Anti-collision and non-destructive harvesting
Inference upwards along prediction boxes
Picking points rough positioning inference model
Gaussian distance weights
Spatial coordinate transformation
Multi-grapes real-time continuous picking sequence planning
Picking point coordinates continuously output
Three-dimensional ROI of finite number of grapes
Deep threshold segmentation
RGB+Depth image information
Visual platform
"Eye-to-hand" relationship of robot
Center point
Picking point
Three-dimensional ROI

Figure 3. Overall framework of visual scheme.

2.2. System

2.2.1. System Architecture

As shown in Figure 4, the software and hardware composition of this robot includes using Intel Realsense SDK 2.0 to call Realsense D455, using ROS to deploy each functional module in Jetson Nano, and using Pytorch and YOLO v4-SE to realize the grape detection and picking point positioning. By using the trellis grape information collected using the Intel Realsense D455 camera as a feedback signal, the signal is input into a Jetson Nano board for image processing and model recognition, relying on setting distance thresholds to process the three-dimensional coordinate information of a limited number of mature grapes picking points within the obtained area. The SCARA X-SEL controller is utilized to perceive the relative position information between the current position of the robotic arm and the target ripe grapes and finally to guide the disc cutter to reach the specified position area. By comparing the real-world target grape pose information with the expected pose information given by using the deep learning model, the processed control signals Signal 1 and Signal 2 are used to drive the robotic arm and the cutting end effector, realize the positioning and cutting of the target grape using the robot, and establish closed-loop control of the system.

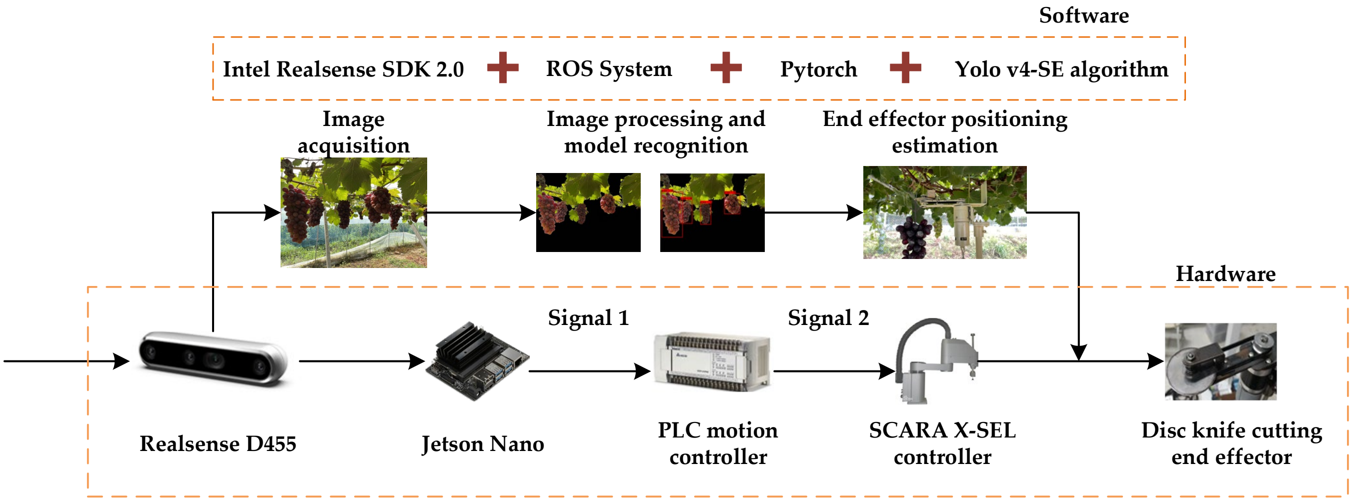


Figure 4. Software and hardware composition for high-speed cutting.

2.2.2. "Eye-to-Hand" Configuration

According to the field research, we drew a space size map of 1.4 m × 2.2 m in the trellis vineyard, as shown in Figure 5. The optimal working width of the picking robot is 1 m by getting rid of the weed area around the working corridor of the robot and the supporting column area, which can avoid collisions with the trellis and the fruit.

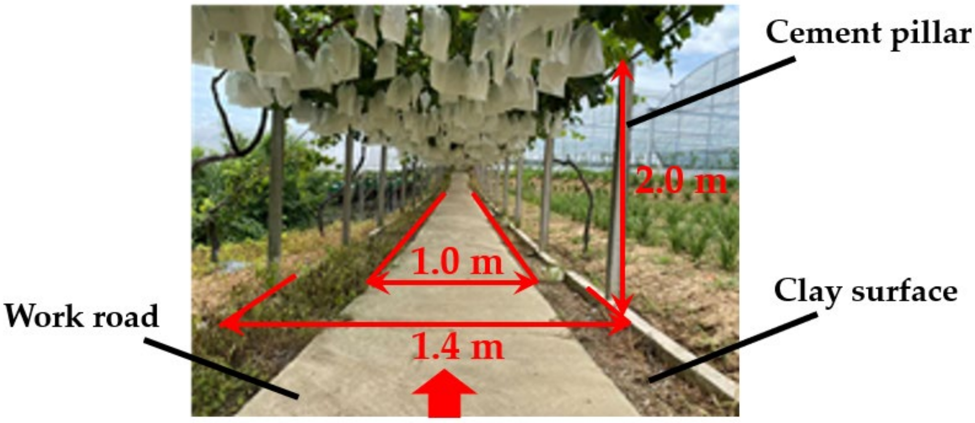


Figure 5. Picking robot working environment.

The Realsense D455 camera installed at the midpoint of the tail of the harvest robot is placed 350 mm behind the base center of the SCARA robotic arm at a height of 185 cm above ground, therefore avoiding collisions with the trellis and acquiring images of multiple bunches of whole grapes. The camera always faces the front of the robot, which is consistent with the working orientation of the disc knife cutting end effector. The color field of view of Realsense D455 is 90° × 65°, and the depth field of view is 87° × 58°. In order to obtain reliable depth information and RGB information and accurately align the depth image with the RGB image, the field of view of Realsense D455 is set to 87° × 58°, operating as the working field of view of the robot's vision (Figure 6).

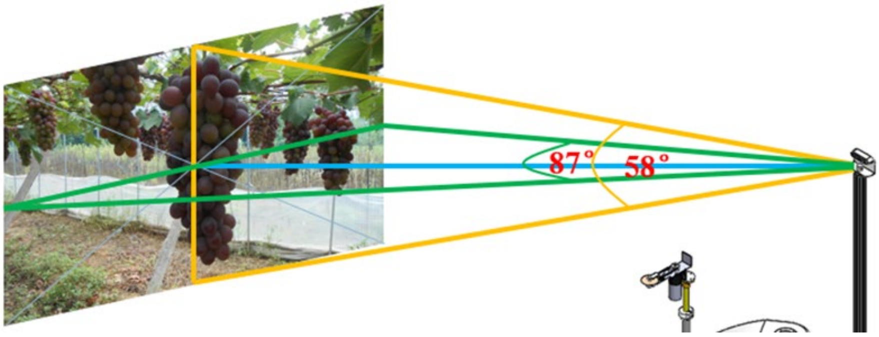


Figure 6. Realsense D455' s field of view.

2.2.3. Three-Dimensional ROI

The length of the large arm of the SCARA robotic arm is 350 mm, and the farthest extension distance of the robotic arm is 700 mm. In order to prevent the large arm of the robotic arm from being over-folded and affecting work efficiency, the optimal working range of the robotic arm is 150 mm–700 mm. The fruit bin is installed at the midpoint of the front end of the harvest robot, and the distance from the base of the SCARA robotic arm is 400 mm. In order to prevent the collision between the robotic arm and the fruit bin and ensure that the target grapes are caught within the fruit bin smoothly, the distance from Realsense D455 to the inside of the fruit bin is set as the minimum depth, and the distance from Realsense D455 to the center of the fruit bin is the maximum depth. According to the minimum depth and maximum depth, this paper divides an ROI area for limited bunches of harvest grapes, as shown in Figure 7.

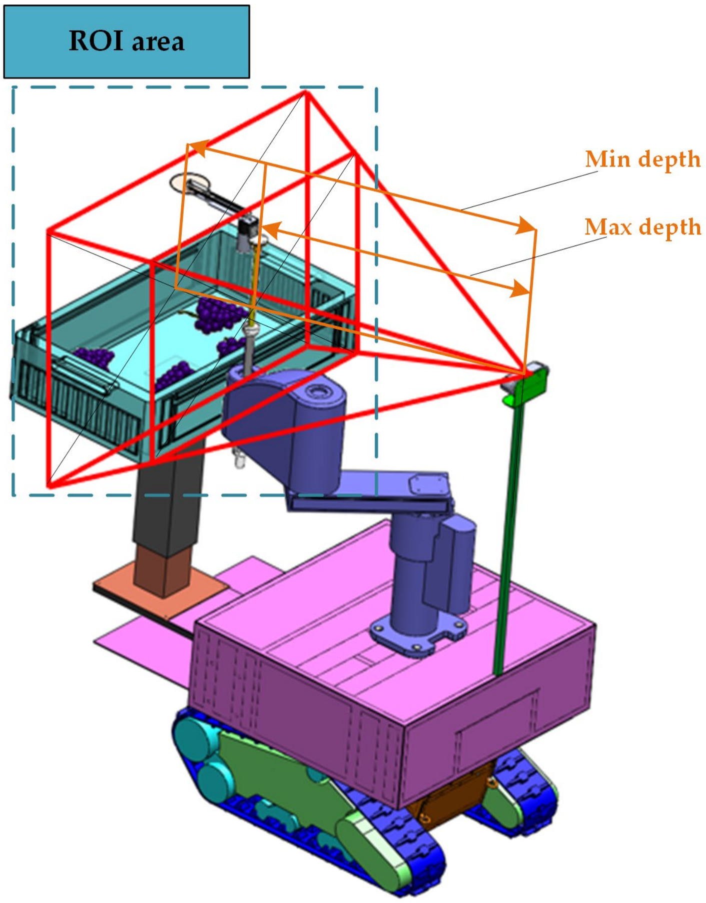
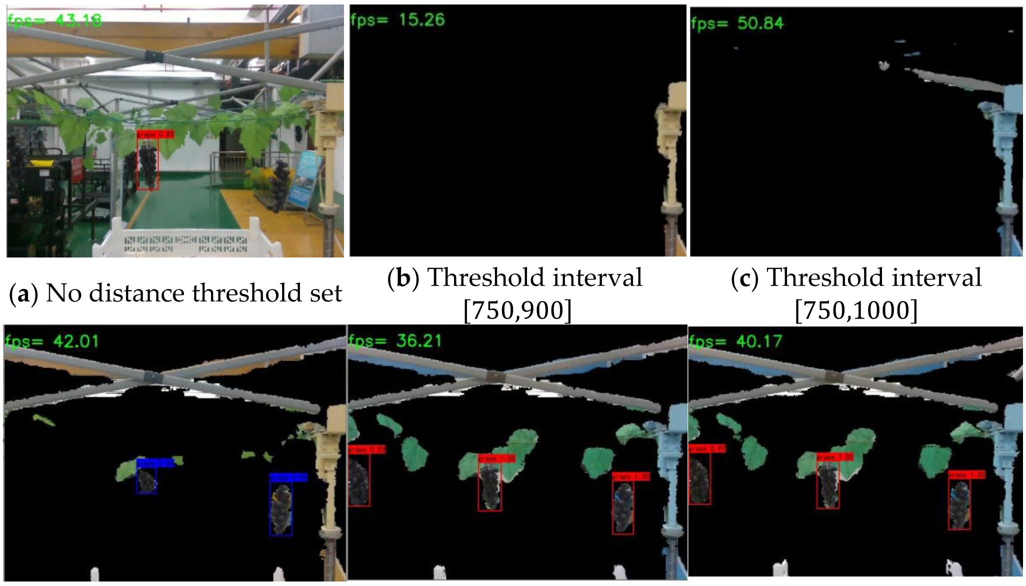


Figure 7. ROI area formed using maximum depth and minimum depth.

In the actual growing environment of the grapes, there are interferences from different factors, such as the background trellis and the vineyard ground. The distance threshold is used to segment the depth information fed back by the depth camera, and the optimal segmentation threshold is determined according to the optimal working range of the SCARA robotic arm and working width of the robot. In this way, the complex background and unstable foreground of the target grapes are eliminated.

In order to ensure the working width of 1 m and avoid too many grapes outside the working space of the robotic arm appearing in the field of view of Realsense D455, the vertical distance between Realsense D455 and the plane behind the fruit bin is set as the minimum depth limit, which is 750 mm. The vertical distance between Realsense D455 and the front plane of the fruit bin is taken as the maximum depth limit, which is 1350 mm. Considering the diameter range of ripe grapes is 6.8–17.2 cm, the maximum depth distance threshold setting range of Realsense D455 is determined according to the camera and the relative position of the SCARA robotic arm, and it is between 900 mm and 1300 mm.

As shown in Figure 8, according to the imaging and recognition effects of different distance thresholds of Realsense D455, it is found that the model cannot recognize all grapes within the camera range due to complex background interference when the distance threshold is not set; within the distance threshold range of $[750, 900]$–$[750, 1000]$, the camera can only accept some of the end effector images and cannot fetch other grape images that meet harvest requirements. At the $[750, 1100]$ distance threshold, some grape images can be obtained, but missing detection is still apparent. At the distance threshold interval of $[750, 1200]$–$[750, 1300]$, the camera has the best imaging and detection effect, which can eliminate most interference from complex foliage foregrounds and irrelevant backgrounds and realize the recognition within the camera's field of view of all grapes that meet the harvest requirements. However, if $[750, 1300]$ is used as the optimal threshold segmentation interval, this may result in large grapes falling out of the fruit bin after cutting.

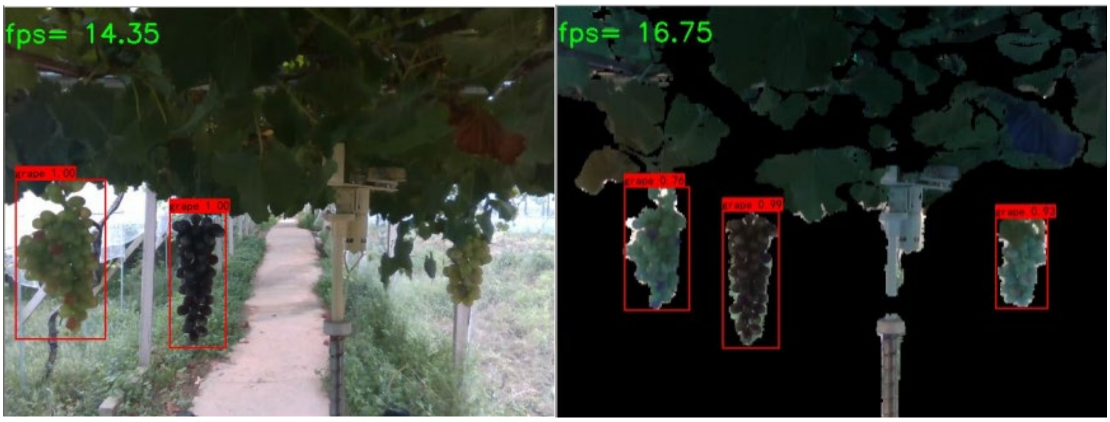


(**a**) No distance threshold set (**b**) Threshold interval [750,900] (**c**) Threshold interval [750,1000]

(**d**) Threshold interval [750,1100] (**e**) Threshold interval [750,1200] (**f**) Distance threshold [750,1300]

Figure 8. Imaging and recognition effect pictures of Realsense D455 with different distance thresholds.

According to the relevant parameters of the high-speed operation area, the distance threshold interval of [750, 1200] is applied to the trellis environment where the robot practically works, as shown in Figure 9. The recognition effect before and after threshold segmentation in the actual harvesting process of the trellis vineyard environment is compared. It is found that at the optimal depth threshold interval, not only is the interference from the complex foliage foreground and irrelevant background in the field of view of the camera eliminated, but the recognition effect of the model is also good at this time. The detection of all target grapes that meet the harvest requirements in the field of view is realized.



(**a**) Recognition effect diagram without threshold segmentation

(**b**) Recognition effect diagram using threshold segmentation

Figure 9. Comparison of recognition effects before and after threshold segmentation in the actual picking process of the trellis vineyard environment.

According to the relevant parameters of the high-speed operation area, the distance threshold interval of is applied to the trellis environment where the robot practically works, as shown in Figure 9. The recognition effect before and after threshold segmentation in the actual harvesting process of the trellis vineyard environment is compared. It is found that at the optimal depth threshold interval, not only is the interference from the complex foliage foreground and irrelevant background in the field of view of the camera eliminated, but the recognition effect of the model is also good at this time. The detection of all target grapes that meet the harvest requirements in the field of view is realized.

2.3. Algorithm

2.3.1. Algorithm Structure

Multivariate information including RGB images and depth images is acquired, which is then input to the feature-enhanced deep learning model. At the same time, samples with different variables are collected with the same device for the training of the feature-enhanced deep learning model. As Figure 3 shows, the vision algorithm for high-speed cut-and-catch includes the following functions.

(1) The multi-target grapes in the three-dimensional ROI can be synchronously identified, which is beneficial to the global continuous picking sequence established in the three-dimensional ROI.

(2) The pixel coordinates of the picking point can be inferred by using the corner information of the prediction box output from the feature-enhanced deep learning model.

(3) After the spatial coordinate transformation, according to the Gaussian distance weight and dual-indicator spatial coordinates that sort the picking point in the three-dimensional ROI, the global continuous picking sequence is planned.

2.3.2. Multivariate Image Acquisition

The trellis grape images used for training and testing in this experiment were collected at Erya Vineyard, Jurong Grape Technology Comprehensive Demonstration Base, Jurong City, Jiangsu Province. The image acquisition device is an Intel Realsense D455 camera. An RGB image and corresponding depth image are captured using Intel Realsense SDK 2.0 simultaneously, as shown in Figure 10, which will be aligned for further deep learning detection.

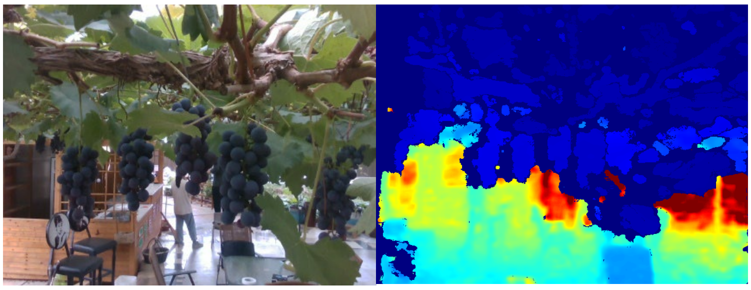
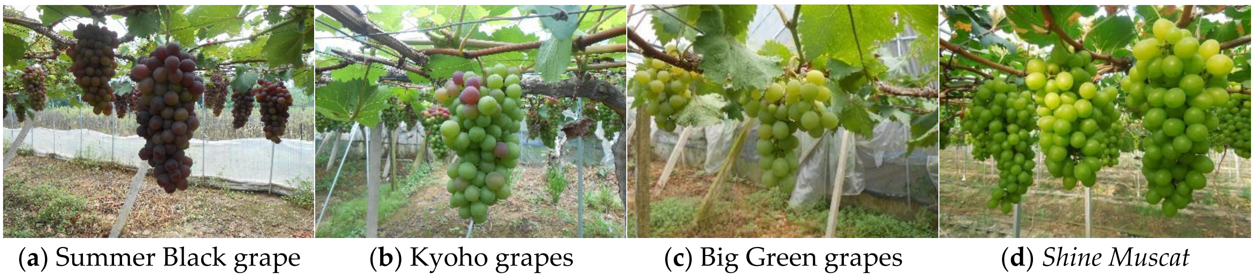


Figure 10. RGB image and corresponding depth image.

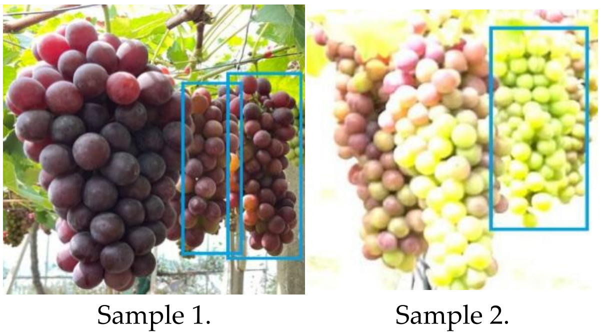
There are 200 images of each variety of trellis grape with different colors, with a total of 800 images, as shown in Figure 11. The collection time focuses only on the period between July and September 2021. The camera is installed at a height of 185 cm from the ground, and the view distance ranges from 30 cm to 150 cm.



(**a**) Summer Black grape (**b**) Kyoho grapes (**c**) Big Green grapes (**d**) *Shine Muscat*

Figure 11. Four different varieties of trellis grapes.

Due to the complexity of the actual harvesting trellis environment, the samples of fruit overlapping, fruit occluded, and strong light-affected color changes in the captured images are added, which are indicated by blue boxes, as shown in Figure 12.



Sample 1. Sample 2.

Figure 12. Samples that are overlapping, occluding, and strongly affected by light. Note: Sample 1. is a sample of ripe grapes overlapping or occluding each other. Sample 2. is a sample of grapes affected by strong light.

The labeling method of the trellis grape depth image is the same as that of the RGB image, and the grape labeling in the image should be as consistent as possible. The labeling method for overlapping and occluded grapes is shown in Figure 13. The overlapping and occluded grapes are marked as grape 2 and grape 3, respectively, and input into the deep learning model for training, so that they can be distinguished in the recognition and positioning process of subsequent harvesting.

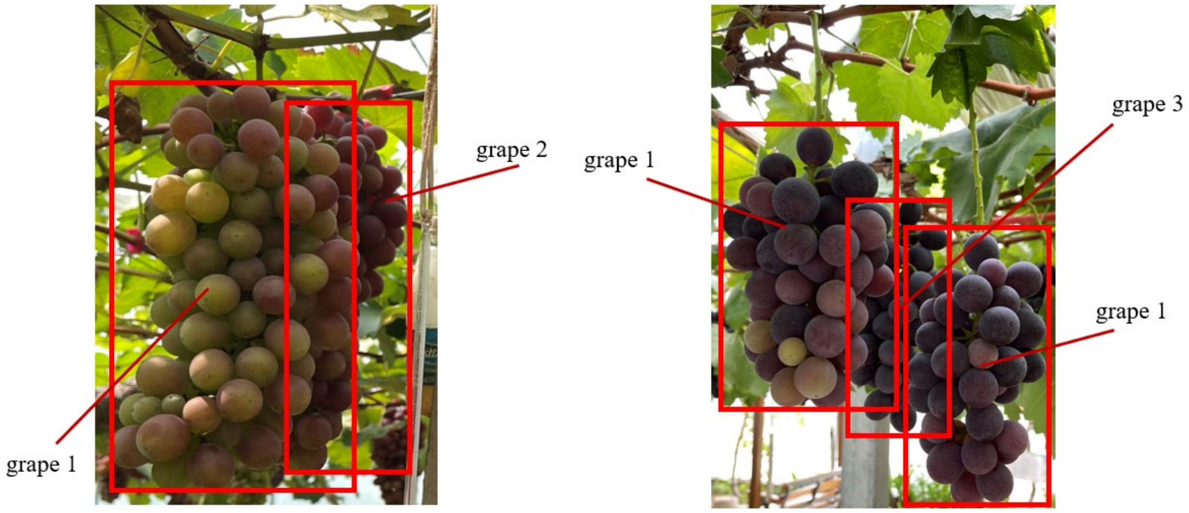


Figure 13. Labeling of overlapping and occluded grapes. Note: Grape 1 indicates the complete grape imaged within the camera's field of view; grape 2 indicates the overlapping grape in the image; grape 3 indicates the occluded grape in the image.

During the training process for YOLO v4-SE, three types of labels are used for training including grape 1 (complete grapes), grape 2 (overlapping grapes) and grape 3 (occluded grapes). The generalization ability and robustness of the model are improved by increasing the sample size of overlapping and occluded grapes in the dataset.

2.3.3. Feature-Enhanced Model for Synchronous Recognition of Trellis Grapes

For high-speed cut-and-catch harvesting, the accurate recognition of grapes and the positioning of picking points in three-dimensional ROI are key prerequisites. YOLO v4, as one of the most advanced real-time detection models, can simultaneously output category anchor boxes and probabilities during the detection process. It is further optimized on the basis of YOLO v3, which improves the overall performance of the model significantly [29].

The attention mechanism is a biological imitation vision mechanism. The area of interest is screened out by quickly scanning the global image, so as to invest more attention resources and suppress other useless information [30]. In a dense trellis environment, each grape can have a complex background, and it becomes a small target. SENet (Squeeze-and-Excitation Networks) [31] is an efficient attention mechanism, with its structure shown in Figure 14. It can rapidly select and focus on salient objects in complex scenarios. Its core SE module (SE block) is mainly composed of two parts, squeeze and excitation, running a channel attention mechanism in tandem.

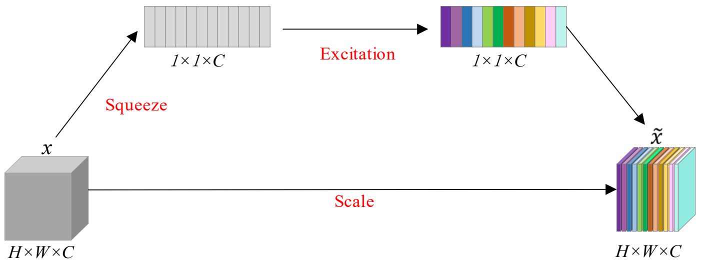


Figure 14. Structure of SENet.

Since the actual harvesting environment of trellis grapes is relatively complex, the collected images were easily interfered with by the overlap of vines, leaves, fruits, as well as the intensity of light, which cases an excessive background depth, thus affecting the accuracy of recognition and causing difficulty for the subsequent harvest performed by robot. The feature perception ability for trellis grapes in the YOLO v4 recognition model is enhanced by integrating the SENet attention mechanism so that it can focus more on the target area of a trellis grape, which can help to achieve efficient and accurate grape detection. The new recognition model for trellis grapes is called YOLO v4-SE.

The YOLO v4-SE model, an improvement over YOLO v4, incorporates the SENet attention mechanism after the feature fusion layer, as illustrated in Figure 15. The YOLO v4-SE model automatically calculates the importance of each feature map based on the correlation among the feature maps extracted from the input trellis grape image. The calculated importance values are then utilized to enhance the discriminative features of the trellis grapes themselves during the YOLO v4-SE detection process. Additionally, they help suppress interference caused by complex environmental factors, improving the accuracy of detection, which can realize efficient detection in the three-dimensional ROI.

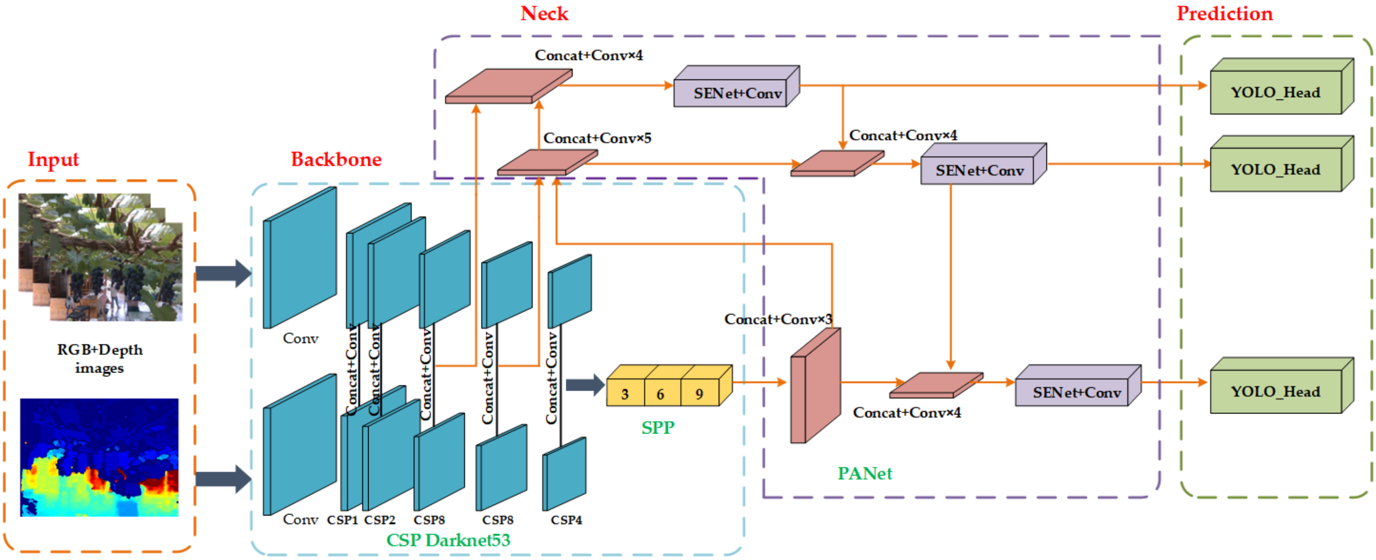


Figure 15. Feature-enhanced YOLO v4 model with SENet.

2.3.4. Rough Positioning of Picking Points with Fusion of Depth Images and RGB Images

During the fruit harvesting process, the accuracy of picking point positioning is primarily influenced by errors in the depth as well as in the horizontal and vertical directions. Errors in the depth direction and left–right directions will cause the disc knife to fail to cut the fruit. At the same time, it is easy to collide with the fruit or the trellis if the error in height direction is too large. Therefore, the accuracy requirements in the three directions of x, y, and z are given by this paper to limit the rough positioning of the picking points, which can aid the disc knife cutting end effector in successfully cutting the grape without damage as well as in avoiding collisions.

This paper presents a three-dimensional positioning method that incorporates differential rough positioning accuracy to optimize efficiency. It was found during field research that the diameter of the circumscribed circle at the widest point of mature grapes ranges from 6.8 cm to 17.2 cm. Therefore, the cutting end effector used by the harvesting robot proposed in this paper is a disc knife with a diameter of 60 mm. The cutting speed of the disc knife is fast. At the same time, it has enough error tolerance for the positioning of the trellis grape picking points during the picking process. Figure 16 illustrates the coordinates of the true picking point (m, n, p) and the predicted picking point (x_1, y_1, z_1). A circular plane picking area with a radius of 30 mm is formed using the working diameter of the disc knife cutting end effector to allow enough error tolerance, which is centered on the x and y coordinates of the true picking point.

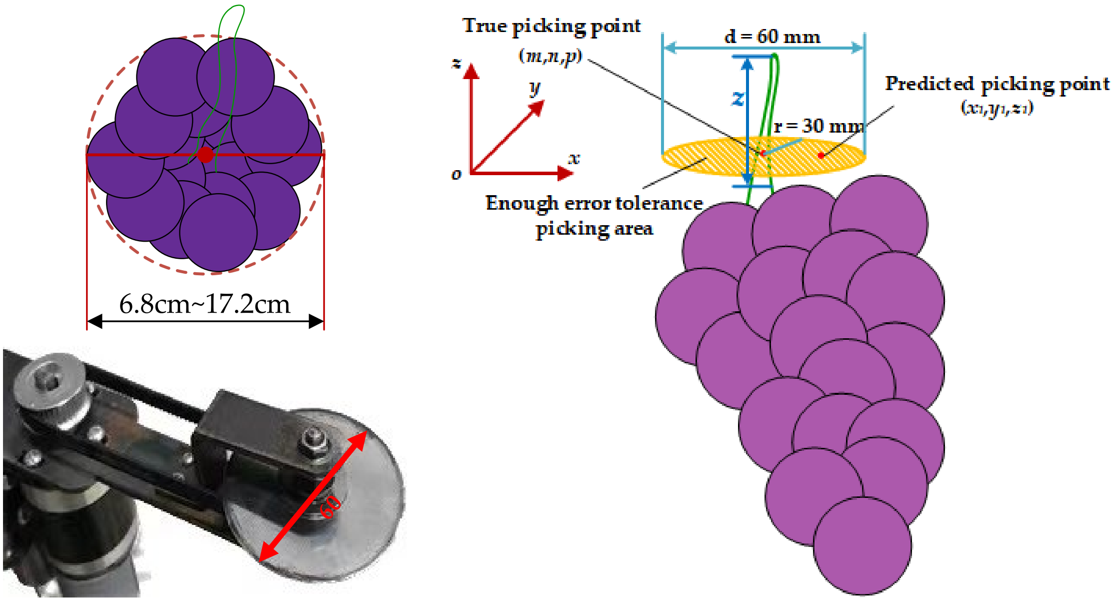


Figure 16. Enough error tolerance picking area built based on the disc knife cutting end effector.

(1) Accuracy requirements in the x and y directions:

$$\sqrt{(x_1 - m)^2 - \quad (y_1 - n)^2} \leq 30\text{ mm} \tag{1}$$

(2) Accuracy requirements in the z direction:

$$z_1 \leq l \quad l \in (55\text{ mm}, 108\text{ mm}) \tag{2}$$

where l is the measured main stem length of the grape.

After the grapes on the trellis are ripe, they can often fall vertically in the air due to the support of the grape clusters, and the fruit stems are naturally and vertically downward. The disc knife cutting end effector has a large radius, a high cutting speed, and enough error tolerance. The method of directly locating the grape picking point [32] cannot meet the high-speed cutting and catching work requirements of the disc knife because of its long-time consumption. Therefore, this paper proposes for the first time a rough positioning method for trellis grape picking points based on prediction boxes from YOLO v4-SE. Compared with the complex traditional contour extraction algorithm, a center point calculation based on the corner information of the prediction boxes and the synchronous inference of picking points upwards along the prediction boxes of multi-grapes demonstrate good reliability and stability, which can significantly improve efficiency. The positioning flowchart of this method is shown in Figure 17.

A frame of image randomly read from the real-time video stream of trellis grapes is captured using Realsense D455; then, this image is divided into n × n grid cells. If the center point of the grape (object) falls in the grid, then the grid is responsible for predicting this target grape. Each layer network predicts the position information and confidence information of the bounding box, as well as the position information of the four corner points (top, left, bottom, right) that correspond to the bounding box. The target fruit is obtained via depth distance threshold segmentation. The interference of complex foliage foreground and irrelevant background outside the distance threshold range is eliminated. Then, the multi-target grapes can be synchronously identified. The prediction boxes obtained using YOLO v4-SE are retained for subsequent information processing.

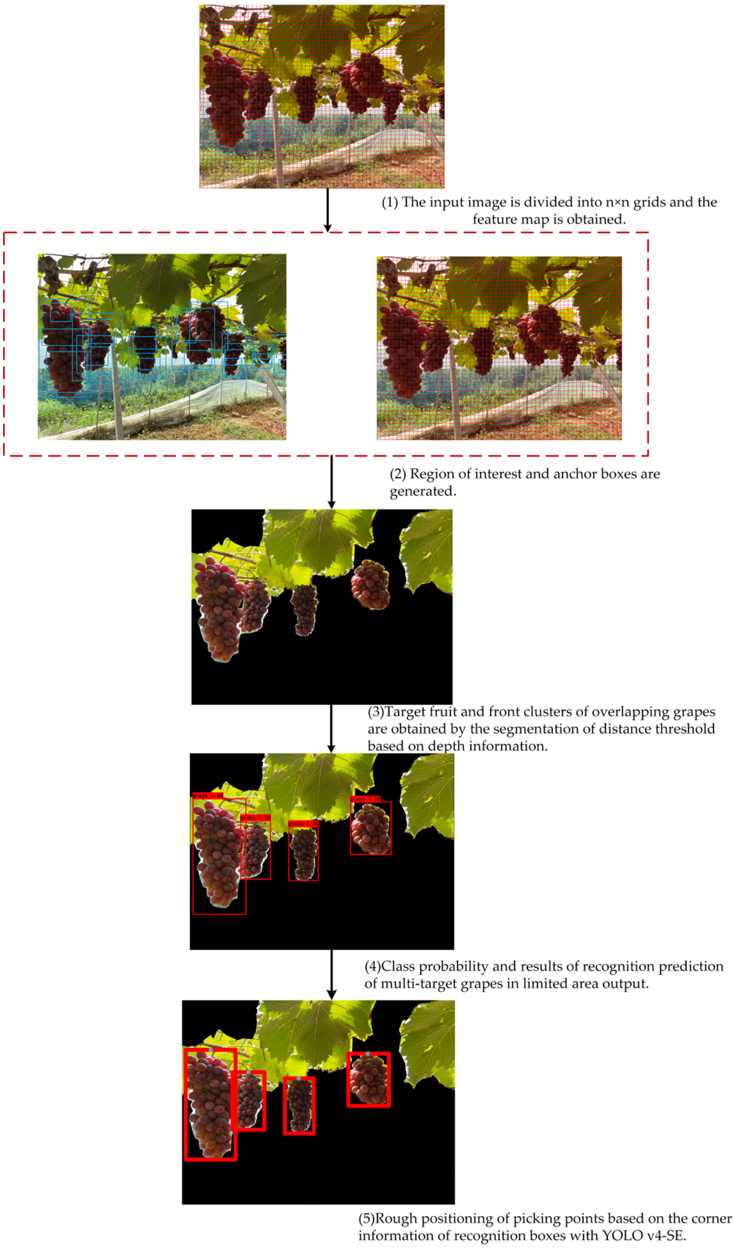

(1) The input image is divided into n×n grids and the feature map is obtained.
(2) Region of interest and anchor boxes are generated.
(3)Target fruit and front clusters of overlapping grapes are obtained by the segmentation of distance threshold based on depth information.
(4)Class probability and results of recognition prediction of multi-target grapes in limited area output.
(5)Rough positioning of picking points based on the corner information of recognition boxes with YOLO v4-SE.

Figure 17. Trellis grape picking point positioning process.

In order to meet the working requirements of a high-speed harvest with the cutting end effector and avoid the time-consuming drawback of precise positioning, the prediction box can be regarded as the minimum bounding rectangle of the target grape. As shown in Figure 18, the center point based on the corner information of the prediction box and the predicted picking point are on the same straight line. Point 1. on the grape stem is the true picking point, and Point 2. at the center of the upper edge of the prediction box is the predicted picking point. The pixel coordinates of the picking point (Point 2.) at this time can be calculated by relying on the coordinate information of the center point of the prediction box. Due to the working characteristics and physical parameters of the cutting end effector, Point 2. can replace Point 1. as the picking point, thus avoiding a time-consuming process and realizing the rapid positioning of the picking points.

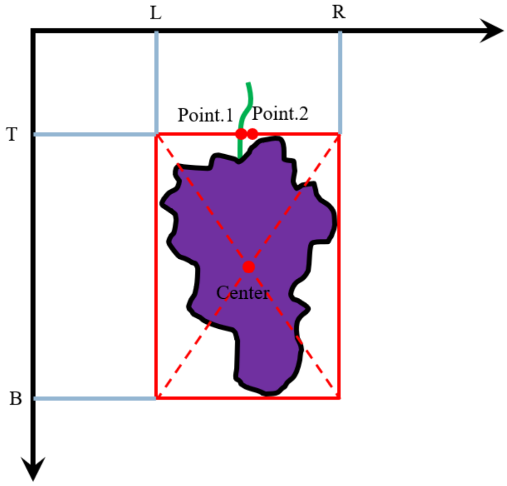


Figure 18. Acquisition of the two-dimensional pixel coordinates of the picking point.

The pixel coordinates of the center point of the prediction box:

$$(x, y) = \left(\frac{R+L}{2}, \frac{T+B}{2}\right) \tag{3}$$

The pixel coordinates of predicted picking point (Point 2.):

$$(x, y) = \left(\frac{R+L}{2}, T\right) \tag{4}$$

where R, L, T, and B are the right value, left value, top value, and bottom value of the four corner points of the prediction box, respectively.

Sufficient error tolerance of the picking area is formed using the working diameter of the disc knife cutting end effector. However, this paper still uses part of the calculation in the model to reduce the error between the true picking point and the predicted picking point. In this way, the sacrifice of the precision of the model to directly predict the picking point is avoided while improving the positioning speed of the grape picking point.

The depth information used for model calculation is obtained from the center point of the grape prediction box. Since the surface of the imaged grape is not smooth and there are gaps between the fruit grains, the depth information of the center point of the prediction box is corrected using the mean filter algorithm. As is shown in Figure 19, this paper takes the center point of the prediction box as the center and draws a purple area with a fixed size of $\frac{w}{4} \times \frac{w}{4}$. Outliers in the depth values of all pixels in the purple area are eliminated, then the average is used as the depth value of the center point of the target grape, and the depth values of all remaining pixels in the area are calculated. This method can be expressed as Equation (5).

$$g(i, j) = \frac{1}{\left(\frac{w}{4}\right)^2} \sum_{s=-\frac{w}{4}}^{\frac{w}{4}} \sum_{t=-\frac{w}{4}}^{\frac{w}{4}} f(i+s, j+t) \tag{5}$$

where *f(i,j)* represents the depth value of the pixel of row *i* and column *j* in the depth image, and *g(i,j)* represents the new depth value of the pixel of row *i* and column *j* in the depth image after mean filtering.

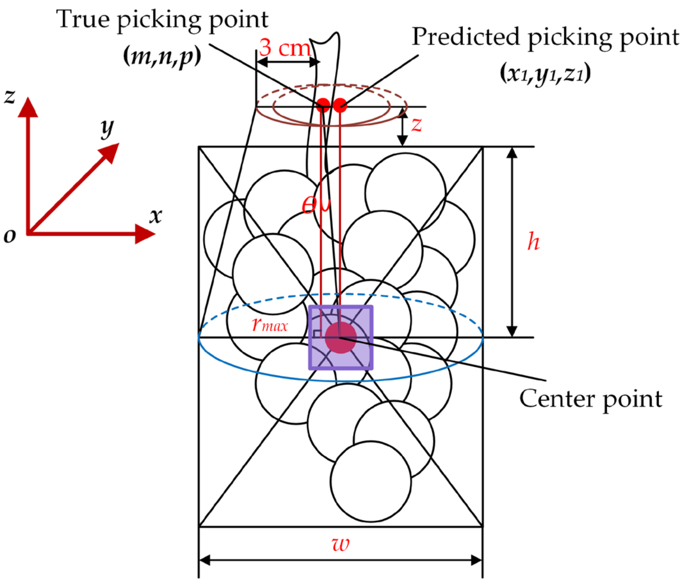


Figure 19. Picking point positioning error compensation calculation.

At the same time, this paper uses a calculation in three directions to improve the accuracy of the inference of the grape picking points upwards along the prediction boxes. As shown in Figure 19, the prediction box can be used as the minimum bounding rectangle of the target grape. Since the grape itself is center-symmetrically distributed along the fruit stem, half the width of the prediction box is used as the radius, and the center point is moved by a radius distance along the y direction and taken as the center of the circle, which is then used to draw a blue circle in the figure.

The error in the horizontal direction is denoted as n, and its relationship with the deviation angle θ of the true and predicted picking point and the compensation parameter z of the height direction should satisfy Equation (6).

$$\frac{n}{z+h} = tan\theta \tag{6}$$

The measured length of the main stem of the mature grapes falls within the range of 5.5 cm to 10.8 cm. A parameter compensation in height direction is added to avoid damage to the fruit after the spatial coordinate conversion. Therefore, the parameter is set to 1.8 cm. By incorporating a compensation in the height direction, the picking point's spatial coordinates are adjusted to ensure that it is positioned above the target grape, thus preventing unnecessary harm during the picking process. After the three-dimensional coordinate conversion, the error compensation of the picking point space coordinates can be expressed as Equation (7).

$$\begin{cases} (m,n,p) = (x_1 + (z+h)tan\theta, y_1 + r_{max}, z_1 + z) \\ r_{max} = \frac{w}{2} \\ z = 1.8\text{ cm} \end{cases} \tag{7}$$

2.3.5. Real-Time Continuous Picking Sequence Planning in Three-Dimensional ROI for Multi-Grapes

Previously, our deep learning model identified and removed overlapping and occluded samples of grapes. Therefore, to enable high-speed cut-and-catch robotic harvesting and ensure real-time updates of the picking sequence, it is necessary to plan the picking sequence only for mature grapes that are completely imaged in the three-dimensional ROI. This ensures that the picking sequence is refreshed in real-time. The overlapping grapes become complete grapes after the previous grapes are picked; the robot still picks the complete grapes imaged in the previous field of view. The robot picks according to the sequence of coordinates of picking points stored in the PLC controller. After picking the current complete grape, the image will be refreshed to identify the subsequent overlapping grapes and re-plan their picking sequence.

As shown in Figure 20, in order to ensure the picking sequence refreshes and continuously sends the spatial coordinates of the picking points to the PLC motion controller in real-time, the pixel coordinates of the picking points obtained using the rough positioning method after the above-mentioned picking sequence planning need to be transferred into the robot base coordinate system through an intrinsic matrix and extrinsic matrix. The spatial coordinates of the picking points can be obtained at this point to guide the disc knife to approach and cut the grapes. Additionally, the Z value of the space coordinates of the picking point can adjust the height of the fruit bin-lifting mechanism, so that the fruit bin can catch the cut grapes smoothly.

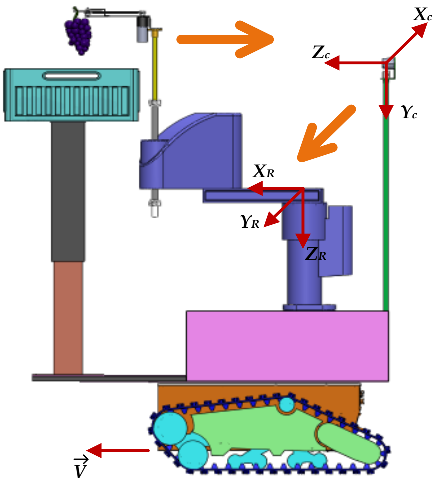


Figure 20. Spatial coordinate of picking point acquisition.

Taking the base coordinate system of the SCARA robotic arm as the center of the coordinate system, a Gaussian kernel function is used to assign distance weights to each grape in the three-dimensional ROI:

$$G(x, y, z) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2+z^2}{2\sigma^2}} \tag{8}$$

where *G(x, y, z)* is the Gaussian kernel function, *(x, y, z)* are the spatial coordinates of the picking points, and σ^2 is the mean squared deviation. The distance weights of the grapes determine how far they are from the disc knife cutting end effector. The less distance, the higher its weight. Each grape is then sorted by its weight.

3. Experimental Results and Discussion

3.1. Experimental Environment and Conditions

Model experiments are conducted with Windows 10 running a deep learning workstation graphics card, 2× Quadro RTX 5000, with 64 GB of video memory; CPU model Intel$^{(R)}$ Xeon$^{(R)}$ Gold 6248, 3.00 GHz, 2.99 GHz; and a deep learning framework Pytorch (Python 3.6, Pytorch 1.2, torchvision 0.4). During training, the batch size is set to 8. The initial learning rate is set to 0.001. The epoch is set to 100. In order to prevent the over-fitting phenomenon of the model in the training process, the weight decay (also called L2 regularization) is set to 5×10^{-4}.

The field experimental site of the grape-harvesting robot is at Erya Vineyard, Jurong Grape Science and Technology Integrated Demonstration Base, Jurong City, Jiangsu Province.

3.2. Feature-Enhanced Deep Learning Model Field Application

In order to verify the effect of the feature-enhanced deep learning model in this paper and present the difference in the recognition results more intuitively, under the same test set collected at Erya Vineyard, focusing on the occlusion caused by stems and leaves, the overlapping grapes, green grapes, purple grapes, and grapes affected by strong light, the application effects of the initial model, YOLO v7, YOLO X, Faster R-CNN, and the improved YOLO v4-SE model are compared, as shown in Figure 21 and Table 2.

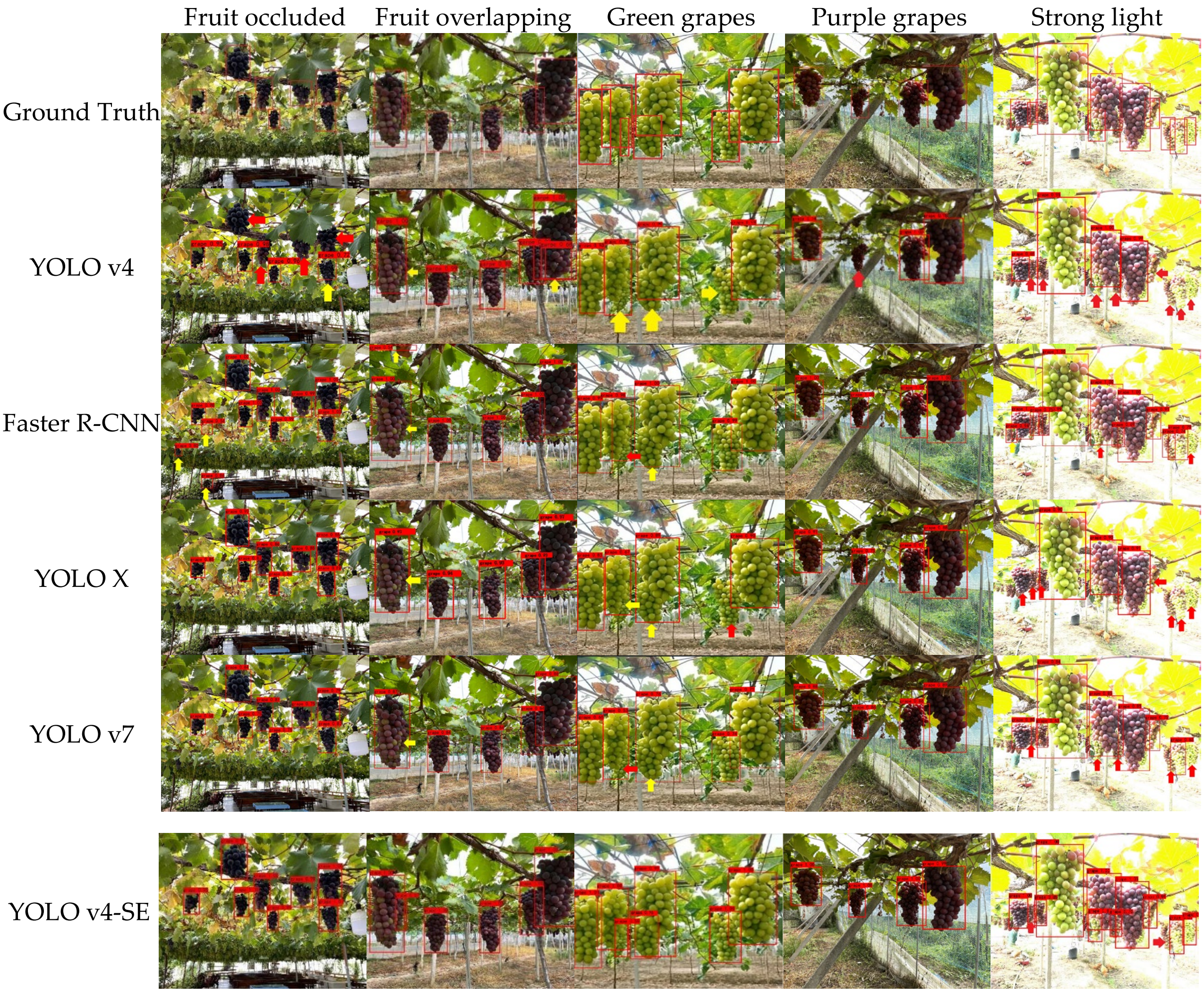


Figure 21. Comparison of model recognition effect for different samples.

Table 2. Performance of different models for trellis grapes (score_threshold = 0.5).

	mAP/%	Precision/%	Recall/%	F1
YOLO v4	93.87	93.43	93.58	0.93
Faster R-CNN (resnet50)	94.76	69.61	94.61	0.82
YOLO X	94.88	87.80	92.31	0.90
YOLO v7	94.27	93.06	89.33	0.91
YOLO v4-SE (Our)	95.21	95.75	95.83	0.95

The improved YOLO v4-SE model, with the same dataset, has the highest precision rate for grape recognition, with a precision rate of 95.75% (Table 2). The mAP rate of YOLO v4-SE is 95.21%, which is 0.94%, 0.33%, 0.45%, and 1.34% than that of YOLO v7, YOLO X, Faster R-CNN, and YOLO v4, respectively. The YOLO v4-SE model proposed here achieved the highest score, reaching 0.95, which indicates that it has better robustness and can meet the requirements of efficient trellis grape detection.

As shown in Figure 21, the red rectangular box is the prediction result, the red arrow indicates the missed detection of the target, and the yellow arrow indicates the false detection of the target. It can be seen from Figure 21 that the improved YOLO v4-SE model proposed in this paper has good detection performance for samples with occluded grapes, overlapping grapes, green grapes, purple grapes, and grapes affected by strong light, and the improved model can mostly identify the target trellis grapes that are missed or occluded in the image. At the same time, it also can be applied to the detection of grapes of different colors and the working environment affected by strong light, which is ideal for robotic harvest. Although the improved YOLO v4-SE also has missed detection under strong light conditions, the grapes eligible for robot harvesting are all correctly identified. Therefore, YOLO v4-SE with high recognition accuracy is suitable for the complex trellis grape cultivation environment and trellis grape harvesting.

Interestingly, we found that YOLO v4-SE with feature enhancement can still accurately identify most of the grapes in a scene where green grapes and purple grapes are mixed under strong lighting conditions. It shows that YOLO v4-SE can be applied for recognition in the purple and green grape picking task under complex conditions.

3.3. Field Positioning Accuracy Verification Test

The PLC motion controller was used to send the 3D spatial coordinate information after pixel coordinate conversion to the SCARA robotic arm, realizing communication between the coordinates of the model-identified fruit and the robotic arm as well as conducting the overall prototype picking experiments. The field work process of the prototype is shown in Figure 22.

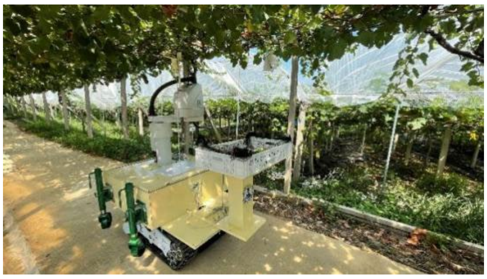


Figure 22. Prototype field work.

In order to analyze the overall working performance of YOLO v4-SE in more detail, and to test the accuracy of the spatial coordinate information of the target fruit fed back via the model to the prototype, a target fruit model predicted and recorded true picking points in 3D during the field experiment. The spatial coordinate information, the relative position of the disc knife cutting end effector, and the picking point are shown in Figure 23.

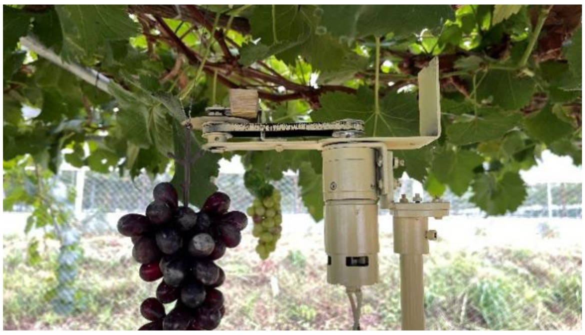


Figure 23. Relative position between the end of the end effector and the picking point of the fruit.

The verification test of the picking point positioning accuracy in the three-dimensional method takes the true picking point of the grapes on the trellis as the target point, moves the disc knife cutting end effector to reach the true picking point of the trellis grapes, and records the position coordinates of the disc knife cutting end effector at this time; then, the algorithm in this paper is used to identify the trellis grapes in the three-dimensional ROI using the high-speed cut-and-catch trellis grape picking robot, and then the trellis grape picking points are estimated quickly and roughly based on the prediction boxes. When the relative distance between the actual picking point and the predicted picking point is within 30 mm, the positioning is considered successful.

$$\begin{gathered}E_x = \frac{\sum_{i=1}^{N} \left| x_p - x_t \right|}{N} \\ E_y = \frac{\sum_{i=1}^{N} \left| y_p - y_t \right|}{N} \\ E_z = \frac{\sum_{i=1}^{N} \left| z_p - z_t \right|}{N} \\ E_r = \sqrt{\left(x_p - x_t\right)^2 + \left(y_p - y_t\right)^2 + \left(z_p - z_t\right)^2}\end{gathered} \tag{9}$$

As shown in Equation (9), E_x, E_y, and E_z indicate the average positioning errors of the *x*,*y*, and *z* directions in the base coordinate system of the SCARA manipulator. E_r indicates the average positioning error of the Euclidean distance between the true picking point and the predicted picking point. (x_p, y_p, z_p) indicates the spatial coordinate of the predicted picking point of the target grape. The spatial coordinates of the predicted picking points are calculated by the positioning model. (x_t, y_t, z_t) indicates the spatial coordinate of the true picking point of the target grape. The spatial coordinates of the true picking point make the outer edge of the maximum cutting force of the disc knife close to the grape stem by controlling the SCARA robotic arm. The spatial coordinates of the center of the disc cutter read in the X-SEL controller of the robotic arm at this time are recorded, which is the true picking point.

In this paper, trellis grape picking point positioning experiments with five kinds of different numbers of bunches in a three-dimensional ROI were recorded. All of the bunch experiments were carried out three times, with a total of fifteen groups of experiments. The test results are recorded in Table 3.

Table 3. Picking point positioning results.

Grape Bunches	Recognition Success Rate/%	Positioning Success Rate/%	Average Recognition Time/s	Average Positioning Time/s	E_x/mm	E_y/mm	E_z/mm	E_r/mm
1	100	100	0.085	0.075	2.58	2.06	1.32	7.16
2	100	100	0.091	0.081	2.55	1.83	1.09	7.85
3	100	88.9	0.082	0.095	2.36	2.02	1.44	7.59
4	91.7	91.7	0.093	0.086	2.83	2.08	1.67	8.85
5	93.3	86.7	0.081	0.084	2.67	2.07	1.37	7.00
average	97	93.5	0.0864	0.0842	2.598	2.012	1.378	7.69

The picking point positioning test species tested 15 times with different numbers of grape bunches show that the average recognition success rate is 97%; the average positioning success rate is 97%; the average recognition time after 15 tests with different numbers of grape bunches is 0.0864 s; the average positioning time after 15 tests with different numbers of grape bunches is 0.0842 s. The average positioning errors of the *x*, *y*, and *z* directions are 2.598, 2.012, and 1.378 mm, respectively. The average positioning error of the Euclidean distance between the true picking point and the predicted picking point is 7.69 mm, which is far smaller than the 30-mm working radius of the disc knife cutting end effector.

3.4. Synchronous Harvesting Experiment in Field

In the trellis environment, the size of the three-dimensional ROI of the robot is used as the standard, and the grape fruiting density within the area is investigated. It is found that the density of the sparsest part of the grape fruiting is 2 bunches × mm^{-2}, and the density of the densest part of grape fruiting is 11 bunches × mm^{-2}. Therefore, this paper designs a synchronous harvesting experiment using a high-speed cut-and-catch harvesting robot in the trellis environment with a gradient of 2 bunches × mm^{-2}. When the target grapes are in the three-dimensional ROI of the robot, the robot stops walking and harvests, recording the relevant indexes at this point in time. Three synchronous harvesting experiments are carried out for different fruiting densities, with a total of fifteen groups of experiments. The average values are taken and recorded in Table 4.

Table 4. Synchronous harvesting experiment in field results.

Fruiting Density/ *Bunches*×mm^{-2}	Recognition Success Rate/%	Positioning Success Rate/%	Collision	Harvesting Time/s	Picking Success Rate/%	Picking Speed/*s*×*Bunch*$^{-1}$
2	100	100	×	12.4	100	6.2
4	100	100	×	23.33	100	5.833
6	100	88.87	×	31.93	88.87	5.95
8	91.7	95.83	√	44.97	91.67	6.14
10	93.3	83.33	√	56.7	83.33	6.80
average	97	93.606	-	-	92.78	6.18

As is shown in Table 4, under the different fruiting densities, the average recognition success rate is 97%; the average positioning success rate is 93.606%; and the average picking success rate is 92.78%. Under the fruiting densities of 2–6 bunches × mm^{-2}, the harvesting times are 12.4, 23.33, 31.93 s, respectively, and no collision occurred. Under the fruiting densities for 8–10 bunches × mm^{-2}, the harvesting times are 44.97 and 56.7 s, and some collision occurred. The main reason for the missed harvesting phenomenon in the two groups of experiments (8–10 bunches × mm^{-2}) is the positioning error of the target grape picking point itself. Although in the identification process, overlapping or occluded grapes have been eliminated in the process of the picking sequence planning, when they are close to the imaged complete grapes, collisions will still occur. However, such collisions rarely take place. Picking speed refers to the time required to complete a single picking task per unit time. The average picking speed is 6.18 s × bunch^{-1}, which meets the harvesting requirements for high-speed cut-and-catch harvesting robots.

3.5. Discussion

In the feature-enhanced deep learning recognition model field application, the feature-enhanced YOLO v4-SE model with multi-channel inputs has good detection performance for samples with occluded grapes, overlapping grapes, green grapes, purple grapes, and grapes affected by strong light. In the field positioning accuracy verification test, from the single-bunch and multi-bunch grape picking point positioning experiments, it can be found that the number of grapes that require a positioned picking point has almost no impact on the recognition time and positioning time of the visual recognition and positioning model. However, during the recognition process for grapes, the shaking caused by wind speed often affects the accuracy of picking point positioning. Strong lighting will cause the edge outline of the grapes to be blurred, resulting in inaccurate position information for the acquired grape bounding box. The acquired bounding box is smaller than the complete bounding box, and accurate center point information cannot be extracted from the acquired bounding box.

In the synchronous harvesting experiment in the field, although in the recognition process overlapping or occluded grapes have been eliminated in the process of picking sequence planning, when they are close to the imaged complete grapes, their distance weight values are similar or the same, which will still lead to collisions. However, such collisions rarely take place. The disc knife rotates too fast, which will cause the grapes with a light weight to be thrown behind the knife disc, which produces picking failures.

Therefore, YOLO v4-SE can realize precise recognition for grapes and the rapid positioning of picking points in the high-speed cutting and catching operation designed for trellis grapes.

4. Conclusions and Future Works

In this study, we successfully developed a rough positioning method for high-speed cutting and catching harvesting robots. The method utilized the depth information from the structure of the robot body to build a three-dimensional region of interest where grapes were synchronously identified via a feature perception-enhanced model combined with multi-channel inputs of RGB and depth images. At the same time, synchronous inference based on the recognition boxes of the multi-grapes provided corner information for reliable and fast picking point calculation, which was confined to the cutting area with differences in three-dimensional directions. Additionally, this method planned the global continuous sequence in the three-dimensional region of interest and also solved the problem of sequence planning for the picking of occluded and overlapping grapes.

Next, we will attempt to verify and further optimize our method through field experiments. At the same time, we will optimize the model structure and reduce the size of the model so as to lower its weight and accelerate the detection speed. We will also build a dataset for trellis grapes in a real trellis vineyard during the night to meet the night work requirements for high-speed cutting and catching harvesting robots. In future research, we consider using inertial unit, electronic compass, or gyroscope to measure the roll angle, pitch angle, heading angle, and altitude (ground fluctuation height) of the prototype and introduce these parameters into the algorithm model to reduce the errors in the depth data acquired by the Realsense D455 caused by the tilt of the prototype, so as to improve robot scene adaptability and anti-interference ability. Thus, we will help fruit growers to realize rapid and efficient harvesting during the ripening period of trellis grapes as well as promote the modernization and intelligence of agriculture.

Author Contributions: Conceptualization, Z.X. and J.L.; methodology, Z.X. and J.L. and J.W.; software, Z.X., S.Z. and B.X.; validation, Z.X. and J.L.; formal analysis, Z.X., J.L. and J.W.; investigation, Z.X. and L.C.; resources, J.L. and Y.J.; data curation, Z.X. and J.W.; writing—original draft preparation, Z.X.; writing—review and editing, Z.X. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (NSFC) (No. 31971795), Project of Jiangsu Modern Agricultural Machinery Equipment & Technology Demonstration and Promotion (No. NJ2021-13), Project of Faculty of Agricultural Equipment of Jiangsu University (No. 5543000001), and Project of Faculty of Agricultural Equipment of Jiangsu University (No. 4111680002).

Data Availability Statement: Not applicable.

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

References

1. Liu, L.S.; Liu, H.Y. Discussion on the management and storage technology of grapes before and after picking. *Rural. Econ. Sci. Technol.* **2017**, *28*, 32.
2. Liu, J.Z. Research Progress Analysis of Robotic Harvesting Technologies in Greenhouse. *Trans. Chin. Soc. Agric. Mach.* **2017**, *48*, 1–18.
3. Xiong, Y.; Peng, C.; Grimstad, L. Development and field evaluation of a strawberry harvesting robot with a cable-driven gripper. *Comput. Electron. Agric.* **2019**, *157*, 392–402. [CrossRef]
4. Rong, J.; Wang, P.; Wang, T. Fruit pose recognition and directional orderly grasping strategies for tomato harvesting robots. *Comput. Electron. Agric.* **2022**, *202*, 107430. [CrossRef]
5. Williams, H.A.; Jones, M.H.; Nejati, M.; Seabright, M.J.; Bell, J.; Penhall, N.D.; Barnett, J.J.; Duke, M.D.; Scarfe, A.J.; Ahn, H.S.; et al. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms. *Biosyst. Eng.* **2019**, *181*, 140–156. [CrossRef]

6. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 3065386. [CrossRef]
8. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
10. Pereira, C.S.; Morais, R.; Reis, M.J.C.S. Pixel-based leaf segmentation from natural vineyard images using color model and threshold techniques. In Proceedings of the International Conference Image Analysis and Recognition, Waterloo, ON, Canada, 27–29 August 2019; pp. 96–106. [CrossRef]
11. Gong, L.; Wang, W.; Wang, T.; Liu, C. Robotic harvesting of the occluded fruits with a precise shape and position reconstruction approach. *J. Field Robot.* **2022**, *39*, 69–84. [CrossRef]
12. Rong, J.; Zhou, H.; Zhang, F.; Yuan, T.; Wang, P. Tomato cluster detection and counting using improved YOLOv5 based on RGB-D fusion. *Comput. Electron. Agric.* **2023**, *207*, 107741. [CrossRef]
13. Sun, Q.; Chai, X.; Zeng, Z.; Zhou, G.; Sun, T. Noise-tolerant RGB-D feature fusion network for outdoor fruit detection. *Comput. Electron. Agric.* **2022**, *198*, 107034. [CrossRef]
14. Liu, S.; Whitty, M. Automatic grape bunch detection in vineyards with an SVM classifier. *J. Appl. Log.* **2015**, *13*, 643–653. [CrossRef]
15. Xiong, J.; He, Z.; Tang, L.; Lin, R.; Liu, Z. Visual localization of disturbed grape. Picking Point in Non-structural Environment. *Nongye Jixie Xuebao/Trans. Chin. Soc. Agric. Mach.* **2017**, *48*, 29–33. [CrossRef]
16. Luo, L.; Tang, Y.; Lu, Q.; Chen, X.; Zhang, P.; Zou, X. A vision methodology for harvesting robot to detect cutting points on peduncles of double overlapping grape clusters in a vineyard. *Comput. Ind.* **2018**, *99*, 130–139. [CrossRef]
17. Zhang, T.; Liu, P. A Fast and Efficient Recognition Method for Grape Picking Point. *J. Agric. Mech. Res.* **2020**, *42*, 189–193. [CrossRef]
18. Peng, H.; Huang, B.; Shao, Y.; Li, Z.; Zhang, C.; Chen, Y. General improved SSD model for picking object recognition of multiple fruits in natural environment. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 155–162. [CrossRef]
19. Zhao, D.; Wu, R.; Liu, X.; Zhao, Y. Apple positioning based on YOLO deep convolutional neural network for picking robot in complex background. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 172–181. [CrossRef]
20. Liu, F.; Liu, Y.; Lin, S.; Guo, W.; Xu, F.; Zhang, B. Fast recognition method for tomatoes under complex environments based on improved YOLO. *Trans. CSAM* **2020**, *51*, 229–237. [CrossRef]
21. Yan, B.; Fan, P.; Lei, X.; Liu, Z.; Yang, F. A real-time apple targets detection method for picking robot based on improved YOLOv5. *Remote Sens.* **2021**, *13*, 1619. [CrossRef]
22. Jin, Y.C.; Yu, C.C.; Yin, J.J.; Simon, X.Y. Detection method for table grape ears and stems based on a far-close-range combined vision system and hand-eye-coordinated picking test. *Comput. Electron. Agric.* **2022**, *202*, 107364. [CrossRef]
23. Luo, L.; Tang, Y.; Zou, X.; Ye, M.; Feng, W.; Li, G. Vision-based extraction of spatial information in grape clusters for harvesting robots. *Biosyst. Eng.* **2016**, *151*, 90–104. [CrossRef]
24. Lei, W.; Lu, J. Visual positioning method for picking point of grape picking robot. *Jiangsu J. Agric. Sci.* **2020**, *36*, 29–33+81.
25. Yin, W.; Wen, H.J.; Ning, Z.T.; Ye, J.; Dong, Z.Q.; Luo, L.F. Fruit detection and pose Estimation for Grape Cluster-Harvesting Robot Using Binocular Imagery Based on Deep Neural Networks. *Front. Robot. AI* **2021**, *8*, 626989. [CrossRef] [PubMed]
26. Kalampokas, T.; Vrochidou, E.; Papakostas, G.A. Grape stem detection using regression convolutional neural networks. *Comput. Electron. Agric.* **2021**, *186*, 106220. [CrossRef]
27. Li, H.; Li, C.; Li, G.; Chen, L.X. A real-time table grape detection method based on improved YOLO v4-tiny network in complex background. *Biosyst. Eng.* **2021**, *212*, 347–359. [CrossRef]
28. Zhao, R.; Zhu, Y.; Li, Y. An end-to-end lightweight model for grape and picking point simultaneous detection. *Biosyst. Eng.* **2022**, *223*, 174–188. [CrossRef]
29. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
30. Su, B.F.; Shen, L.; Chen, S.; Mi, Z.W. Multi-features Identification of Grape Cultivars Based on Attention Mechanism. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 226–233+252.
31. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
32. Ning, Z.T.; Luo, L.F.; Liao, J.X. Recognition and the optimal picking point location of grape stems based on deep learning. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 222–229.

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