[1]Chen Xia

Rapid Strawberry Ripeness Detection And 3D Localization of Picking Point Based on Improved YOLO V8-Pose with RGB-Camera

Journal of Electrical Systems

Abstract: - Accurate identification of strawberries at different growth stages as well as determination of optimal picking points by strawberry picking robots is a key issue in the field of agricultural automation. In this paper, a fast detection method of strawberry ripeness and picking point based on improved YOLO V8-Pose (You Only Look Once) and RGB-D depth camera is proposed to address this problem. By comparing the YOLO v5-Pose, YOLO v7-Pose, and YOLO v8-Pose models, it is determined to use the YOLO v8-Pose model as the fundamental model for strawberry ripeness and picking point detection. For the sake of further improving the accuracy of the model detection, this paper makes targeted improvements: all the Concat modules at the Neck part are replaced with BiFPN richer feature fusion, which enhances the global feature extraction capability of the model; the MobileViTv3 framework is employed to restructure the backbone network, thereby augmenting the model's capacity for contextual feature extraction. Subsequently, the output-side CIoU loss function is supplanted with the SIoU loss function, leading to an acceleration in the model's convergence. The enhanced YOLO v8-Pose demonstrates a 97.85% mAP-kp value, reflecting a 5.49% improvement over the initial model configuration.. For the sake of accurately localizing the three-dimensional information of strawberry picking points, the strawberry picking points are further projected into the corresponding depth information to obtain their three-dimensional information. The experimental results show that the mean absolute error and the mean absolute percentage error of strawberry picking point localization in this paper are 0.63 cm and 1.16%, respectively. In this study, we introduce a method capable of concurrently detecting strawberry maturity and identifying the precise harvesting location while accurately localizing the picking point. This investigation holds considerable theoretical and pragmatic relevance in augmenting the intelligence of strawberry harvesting robots and actualizing automation and smart capabilities in agricultural production.

Keywords: Strawberry identification, Ripeness detection, Picking point localization,Improved YOLO v8-Pose, RGB-D camera

I. INTRODUCTION

Globally esteemed as a crucial fruit produce, strawberries enjoy extensive cultivation and consumption across a wide expanse. [1,2]. However, despite advances in agricultural technology, strawberry harvesting predominantly depends on conventional hand-picking techniques that are both tedious and labor-intensive. The brief ripening window of strawberries exacerbates this challenge, as any delay in harvesting can result in fruit spoilage and significant economic repercussions. [3]. The strawberry ripening period is short, and untimely picking can cause fruit rot, which brings serious economic losses. With the progress and development of science and technology, agricultural picking robots can realize strawberry picking instead of manual labor, which is of great significance and prospect in the field of agricultural production[4–6]. The following is a brief description of the advantages and disadvantages of using a picking robot in agricultural production. Therefore, rapid and accurate identification and detection of strawberries is essential to promote the automated strawberry picking and the intelligent development of agriculture[7].

Yamamoto et al.[8] suggested a strawberry target separation algorithm based on color threshold segmentation. Arefi et al.[9] employed threshold analysis to derive amalgamated features from RGB, HIS, and YIQ domains for ripe tomato localization. However, the detection efficacy was suboptimal for smaller fruits.. Lu et al.[10] developed canny edge detection method for edge detection on color difference maps for further recognition of ripe citrus fruits in complex environments. Wang et al.[11] designed a geometric center based matching method to detect lychee fruits and then used pixel thresholding method for classification. Hayashi et al.[12] designed a strawberry picking robot that also used a color threshold segmentation algorithm for strawberry detection and ripeness estimation. Tang et al.[13] used a modified Otsu algorithm to detect tea leaves by obtaining G and G-B component thresholds. Yang et al.[14] come up with an SVM-based CCL-SVM for disease recognition of tomato leaf images in complex environments by combining color texture features for three common pests and diseases of tomato, and achieved an overall recognition rate of 97.5% while reducing the amount of computation. Although these methods solve the fruit detection problem to a certain extent, the features obtained by these methods need to be manually designed, and it

[1]Haide College, Ocean University of China, Qingdao, China,266100

Email id: 1700066648@qq.com

Copyright © JES 2024 on-line : journal.esrgroups.org

is difficult to automatically extract discriminative information. And targets in complex environments are easily interfered by light, occlusion and other factors, which often hinder the high accuracy and robustness of traditional algorithms.

In contemporary times, the swift advancement and extensive implementation of deep learning methodologies have become increasingly prominent., a series of common target detection algorithms have emerged, such as the R-CNN series[15] , YOLO series[16,17] , SSD, etc.[18]. The extensive application of these algorithms within the realm of fruit target detection has led to substantial advancements in both the academic exploration and practical implementation of this domain.

Yu et al.[19] popped the question that a greenhouse strawberry fruit recognition method based on the R-YOLO model. The backbone network of the model was lightweight and improved by using MobileNet-V1 lightweight network. Liu et al.[20] invoked a method for strawberry detection based on improved YOLOv3. The method improves the fruit recognition efficiency by reducing the number of convolutional layers. The recognition accuracy for ripe and unripe strawberries reached 97.14 % and 96.51 %, respectively. Wang et al.[21] proposed a DSE-YOLO network model for detecting strawberries at different growth stages and designed a DSE module that uses point-by-point convolution and unfolded convolution to extract various details and semantic features in horizontal and vertical dimensions. Zhang et al.[22] projected a new strawberry target detection network RTSD-Net based on YOLOv4-Tiny, which simplifies the network structure of YOLOv4-Tiny, improves the computing speed, and realizes the real-time detection of strawberries. Lamb and Chuah et al.[23] Yu et al. devised an enhanced variant of the SSD model specifically for the purpose of identifying mature strawberry fruits..[24] put in an improved Mask R-CNN[25] method to detect strawberry fruits in the laboratory. Mu et al.[26] combined a faster regional convolutional neural network with transfer learning to detect unripe tomato fruits. The above studies have achieved some success in strawberry target detection and ripeness classification, providing strong support for realizing intelligent strawberry picking. Although the above methods can accurately detect strawberries, they cannot recognize the corresponding picking points, and thus cannot be directly applied in strawberry picking operations.

YU et al.[19] using the fruit attitude estimation rotation YOLO to localize the main axes of fruits, the picking points of strawberries are predicted by the positions of the main axes of the plants. Guo et al.[27] used segmented strawberry images to determine the strawberry pose by calculating the major axis of its binarized image, and used a geometric method to derive the predicted picking point, but in the actual picking work, it would result in the picking point prediction bias resulting in wrong picking and missed picking. In addition, after detecting the strawberry picking point, it is essential to further pinpoint the accurate coordinates of the selection point. Employing either monocular or stereoscopic vision methodologies to achieve object localization has emerged as a focal point in contemporary research pursuits. [28,29]. The target localization has become a research hotspot in the current research. However, because monocular camera is difficult to provide accurate depth information, binocular camera usually adopts parallax method to calculate depth information, which is greatly affected by matching accuracy and has high computational complexity, making it difficult to meet the practical needs. Depth cameras have higher depth measurement accuracy and stronger environment sensing ability, and are suitable for application scenarios that require high precision depth information [30]. Considering that in the natural environment, strawberry targets are small and usually obscured to different degrees, which makes it difficult for the picking robot to recognize them. Picking robots not only need to accurately identify the ripeness of strawberries, but also need to be able to accurately localize the picking point of strawberries. Consequently, this investigation puts forth an approach to ascertain strawberry maturity and 3D pinpointing of the harvesting site, employing a refined YOLO v8-Pose algorithm in conjunction with an RGB-D camera. This technique allows for precise identification of strawberry ripeness, as well as the harvesting site localization, and accurately captures the three-dimensional data of the designated picking point.

II. MATERIALS AND METHODS

2.1Data collection

2.1.1 Image acquisition and analysis

In this paper, the RealSense D435i depth camera made by Intel is used for data acquisition, which can provide images with a resolution of up to 1280×720 at a speed of 30 frames per second, which can be used both indoors and outdoors, and the depth distance is valid from 0.1m-10 m. The research data in this paper were collected in January 2023 at the

strawberry planting base of Ricetime Agricultural Technology Co., Ltd. in Jiaxing City, Zhejiang Province, China, using a depth camera to record video at multiple times throughout the day to collect data on strawberries under different light conditions to ensure the complexity of the light environment. During the recording process, the height of the depth camera was 120 cm, and the distance from the target was 10-70 cm, which was within the permissible range of vision of the depth camera, and the dataset was collected in the way shown in Figure 1, and an external portable computer was used for data storage.

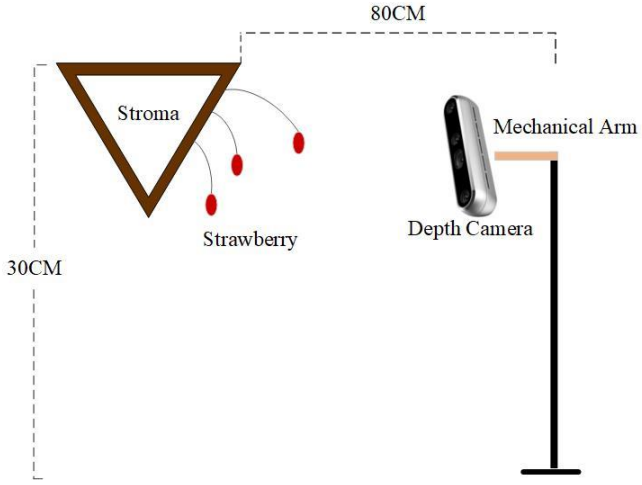


Figure 1: Data collection method

2.1.2 Dataset construction

The recorded video was sliced and segmented to obtain strawberry images using Python script. The original dataset containing 3860 images was obtained in total after screening, and the strawberry dataset contains three kinds of samples: immature, semi-mature and mature strawberries. The fundamental information of the dataset is shown in Table 1, which ensures the completeness and complexity of the dataset. The initial dataset is divided into training set, validation set, and test set according to the division ratio of 7:2:1. In this paper, we use Labelme software to label strawberries and keypoints, and the strawberries with different maturity levels are labeled with the minimum outer rectangle of the target, and the labels are "Maturity", "Medium" and "Immaturity". The key point "Pick" was labeled 2 cm up the fruit stem, and the labeled image is shown in Figure 2. The labeling results are stored in json standard format, and the stored information includes: image path, width and height dimensions, number of channels, and the location information of the strawberry labeling frame and picking point. The test platform is a hardware platform of AMD Ryzen 9 7845HX (CPU) with 3.00GHz main frequency, 16GB of running memory (RAM), and a graphics card (GPU) of NVIDIA GeForce RTX 4070, and all the programs are written in python under the system of Win10, based on the Pytortch deep learning framework.

Table 1 Data capture shooting information

Weather	Counts of Image
Cloudy	1250
Sunny	1230
Rainy	1380

Figure 2: Labelme image annotation

2.2 Strawberry ripeness and picking point detection

2.2.1 Improved YOLO v8-Pose Model

YOLO v8-Pose is a deep convolutional network based on YOLO v8, which not only implements target detection, but also adds estimation of key point locations[31] . Compared with YOLO v5 and YOLO v7, YOLO v8 has improved in accuracy and speed, so we choose to use YOLO v8n model as the base model for strawberry ripening and picking point detection in this paper. The structure of the YOLO v8 model is mainly composed of Backbone, Neck and Head. In the Backbone part of the backbone network, the CSP (Cross Stage Partial) idea is adopted, and feature extraction is performed through the Conv module using a 3×3 convolutional kernel. For the sake of preventing the gradient from vanishing and to ease the training, residual connectivity is introduced and the C2f module is used to convert the feature maps into inputs for the fully connected layers. Finally, the SPPF module receives pooling windows of different scales, pools and splices the feature maps. The Neck part performs bottom-up feature extraction and top-down feature fusion. The Upsample module uses a neighborhood interpolation algorithm to expand the size of the feature maps by a factor of 2. The Concat module joins the feature maps from the Backbone with the feature maps from the upsampled layers to fuse the low-level and high-level semantic information. And the Neck part of the FPN performs the downscaling and channel number adjustment by convolution operation. At the Head end, the detector Detect consists of a series of convolutional and fully connected layers, which is responsible for localizing and identifying the target on the feature map, and generating the location and category prediction of the bounding box. Such a network structure design can effectively accomplish the task of strawberry ripeness and picking point detection while maintaining high accuracy and efficiency.

Typically manifested as diminutive targets with fluctuating shapes and hues within their natural surroundings, strawberries pose an elevated challenge for detection. Furthermore, shifting lighting conditions and potential occlusions can influence the model's efficacy in pinpointing strawberries with precision. Particularly, in terms of key point detection at picking locations, strawberry ripeness and picking points are often tiny targets that are more difficult to capture and localize. To address the above problems, this paper makes the following three improvements to the YOLO v8-Pose model: (1) Replace all the Concat modules at the Neck part with BiFPN richer feature fusion to enhance the global feature extraction capability of the model. (2) Adopting MobileViTv3 to reconstruct the backbone network to enhance the model's contextual feature comprehension capability and further enhance the model's feature extraction capability in complex environments. (3) Replace the CIoU loss function at the output end with the SIoU loss function to accelerate model convergence and improve the regression accuracy of the prediction frame. The improved YOLO v8-Pose model is shown in Figure 3.

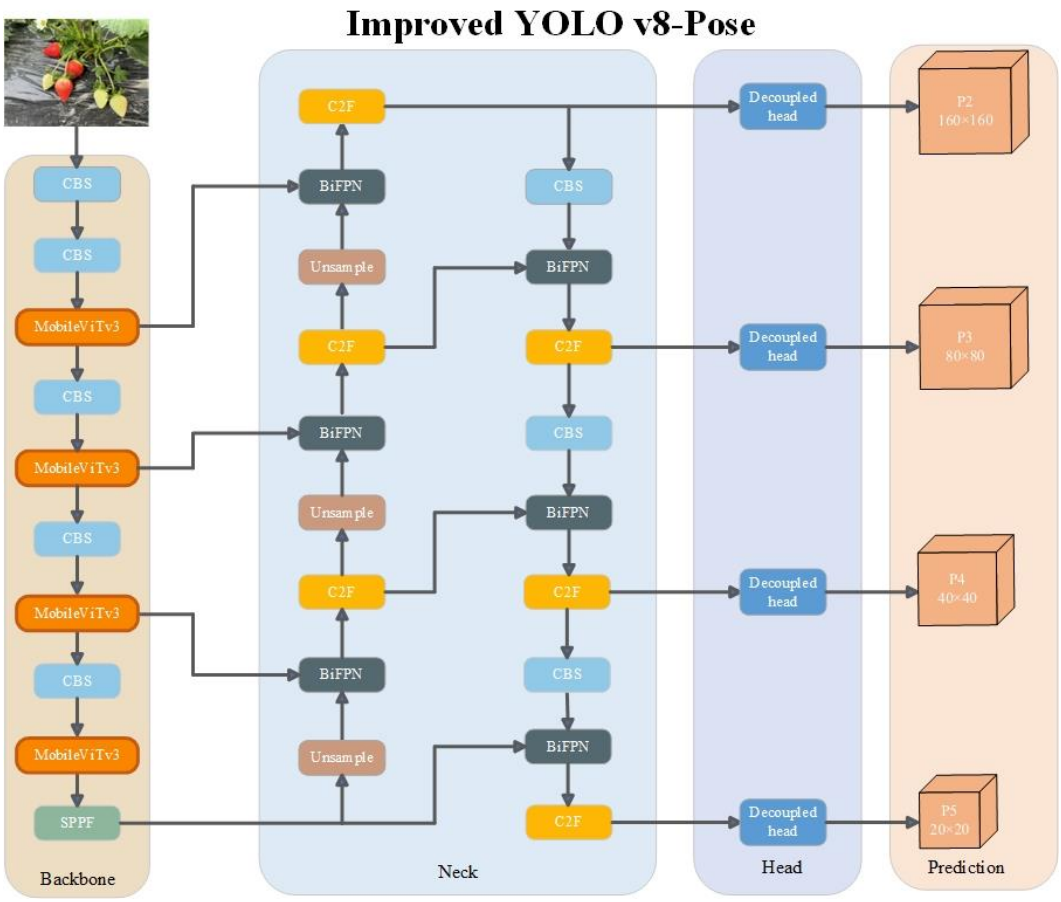


Figure 3: Improved YOLO v8-Pose network structure

2.2.2 Feature Fusion Layer Improvement

The YOLO v8-Pose network structure fuses the FPN and PAN structures, and the target information is mainly concentrated in the top layer semantic information, while the position information of small-sized targets is easily lost, which affects the effectiveness of small target detection. To overcome this problem, we optimize the Neck network of the YOLO v8 network structure to extract the deep features of the image more effectively, thus improving the detection accuracy and efficiency. We introduce the concept of multidimensional feature fusion, which aims at combining features at different resolutions to enhance the characterization of features. Past feature fusion methods usually process all input features in the same way, but the contribution to feature fusion is often unequal due to the different resolutions of different input features. To address this problem, we employ the BiFPN (Bidirectional Weighted Feature Pyramid Network) module, which realizes bidirectional fusion of deep and shallow features from the top layer to the bottom layer and from the bottom layer to the top layer, for the sake of enhancing the transfer of feature information from different network layers[32] .

The BiFPN module, shown in Figure 4, conveys high-level feature semantic information via blue arrows, location information of low-level features via red arrows, and feature fusion at the same level via purple arrows. This bi-directional scale connectivity and weighted feature fusion achieves a better balance between accuracy and efficiency. This study addresses the task of strawberry ripeness and picking point detection, aiming to enable a better balance of the layers of the feature pyramid to provide a more global and semantic feature representation, which helps to improve the detection of targets at different scales in complex environments.

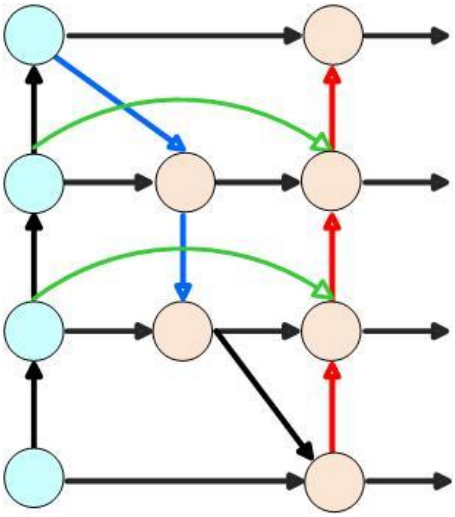


Figure 4: BiFPN network structure

2.2.3 Improvement of backbone feature extraction architecture

The original YOLO V8-Pose model of the CSPDarknet backbone network has a large number of model parameters and requires a large amount of computational information. For the sake of reducing the model parameters and improve the detection speed, MobileViTv3 is used as the backbone network of YOLO v8-Pose in this paper. MobileViT represents a streamlined vision transformer architecture that diminishes the model's dimensions, parameter count, and computational demands, thereby facilitating efficient execution of computer vision tasks on resource-constrained devices.[33,34] The structure of MobileViTv3 is exhibited in Figure 5. The module first takes in local information through convolutional blocks, which enables the model to better recognize the contours, textures, and shapes of objects, and helps to distinguish the differences between the target and the background or other objects. Meanwhile, the MobileViTv3 module leverages the self-attention mechanism in Transformer to perform global correlation computation on the input feature maps, which can capture long-distance dependencies and contextual information, helping the model to understand the visual scene globally, further enhancing the image feature representation, and thus improving the detection accuracy of the model.

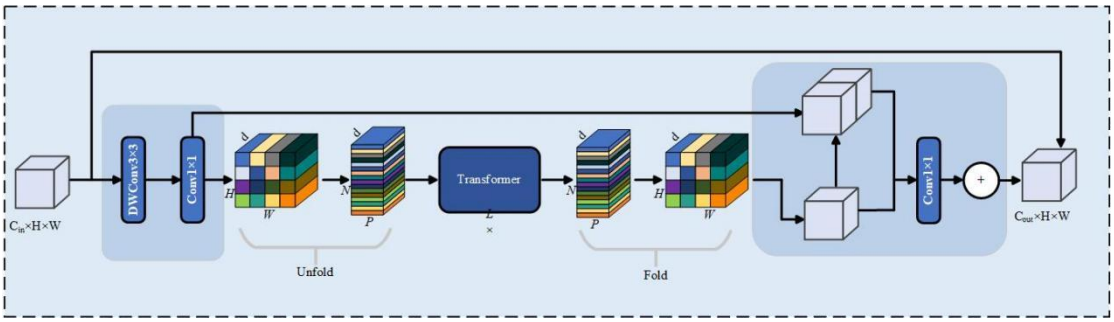


Figure 5: MobileViv3 Block network structure

The challenge in complex environments, such as strawberry ripeness and picking point detection, is the small area and the difficulty of detection. Traditional target detection methods have limited sensory fields and can only acquire local contextual information. MobileViTv3 combines the features of input adaptive weighting and global processing in CNN model and ViT model, uses Transformer as convolution to learn global information, effectively encodes local and global information, and solves the problem of large computation volume and complexity of Transformer-based detection model has the problems of large computation and complexity.

2.2.4 Enhancing the loss function

YOLO v8-Pose uses CIoU as the border regression loss function by default, and evaluates the positional difference between the predicted and target frames by their aspect ratio. However, the full intersection and merger ratio CIoU does not take into account the directional match between the predictor frame and the target frame. Therefore, this paper introduces SIoU[35] as the detection frame regression loss function.

SIoU, on the basis of CIoU loss function, the relative azimuth between the predicted frame and the real frame is added as the cost condition of the distance loss function, and the loss function is redefined to obtain the SIoU loss function:

$$Loss = 1 - IoU + \frac{\Omega + \Delta}{2} \tag{1}$$

Where Δ , IoU and Ω denote the distance loss function, the IoU loss function and the shape loss function, respectively. The SIoU incorporates the angular cost into the distance loss function, which strengthens the constraints of the loss function. When the angle formed by the line connecting the centers of the two frames with the X and Y axes is too large, this loss function makes the prediction frame preferentially move to the nearest coordinate axis. Subsequently, the prediction frame only needs to regress on the X or Y axis where the real frame is located, gradually approaching the real frame, which accelerates the model convergence and improves the precision of the model inference as well as the accuracy of the location of strawberries and picking points.

2.3 Three-dimensional localization of strawberry picking points

The methodology for 3D localization of strawberry picking points based on depth cameras is illustrated in Figure 6. Initially, the RGB camera along with the left and right infrared cameras within the depth camera capture the RGB image and depth image of the strawberry and the picking point, respectively. Subsequently, the refined YOLO v8-Pose object detection algorithm identifies the strawberry and the picking point in the image, procuring the pixel coordinates of the picking point. Following this, depth information is integrated to compute the three-dimensional coordinates of the strawberry picking point within the camera coordinate system. Finally, after a coordinate transformation, the absolute coordinates of the strawberry picking point relative to the robotic arm are acquired. Combined with the corresponding depth information, the three-dimensional coordinates of the picking point in the camera coordinate system are calculated, and the absolute coordinates of the strawberry picking point under the robot arm are obtained after the coordinate transformation to realize the accurate positioning of the strawberry picking point.

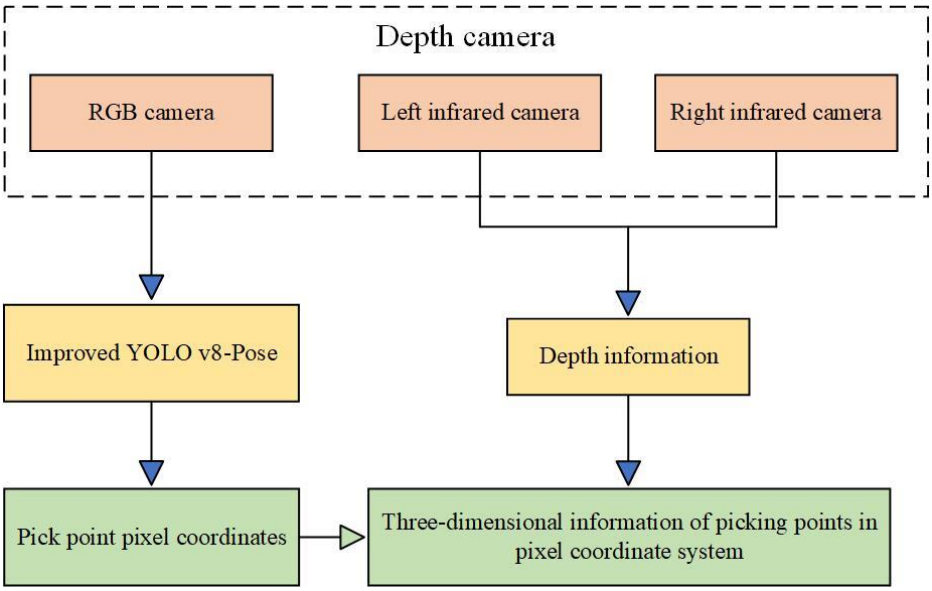


Figure 6: Block diagram of three-dimensional positioning of strawberry picking point

2.3.1 Positioning principles

In the process of harvesting strawberries, it is imperative for the robot to precisely compute the spatial relation between the strawberries and the robotic appendage, guaranteeing accurate manipulation without inflicting harm upon the fruit. This three-dimensional information of the strawberry picking point is crucial for the robot's localization and motion control. The D435i depth camera measures the depth by calculating the distance through the parallax created by the objects on the imaging planes of the left and right infrared cameras. As shown in Figure 7, O_L and O_R represent the optical centers of the left and right infrared cameras respectively, the two cameras are placed horizontally

at a distance of b, with a focal length of f , P_C is a point in the real world, which is mapped to the P_L point on the imaging plane of I_L and the P_R point on the imaging plane of I_R respectively, and Z_C is the distance from P_C to the depth camera. Due to the horizontal placement, the positional deviation only exists in the X-axis, so the deviation of the Y-axis direction is equal, and the deviation of the P_C point in the left and right imaging planes is $C_L P_L = X_L$, $= P_R C_R - X_R (X_R < 0)$ respectively. The similarity can be obtained from $\Delta P_C P_L P_R \sim \Delta P_C O_L O_R$:

$$\frac{Z_C}{Z_C - f} = \frac{b}{b-(X_L - X_R)} \tag{2}$$

where f , b is the camera internal parameter, obtained by calibration, $X_L - X_R$ is the parallax, obtained by pixel-point matching, and the final depth distance Z_C is obtained:

$$Z_C = \frac{fb}{X_L - X_R} \tag{3}$$

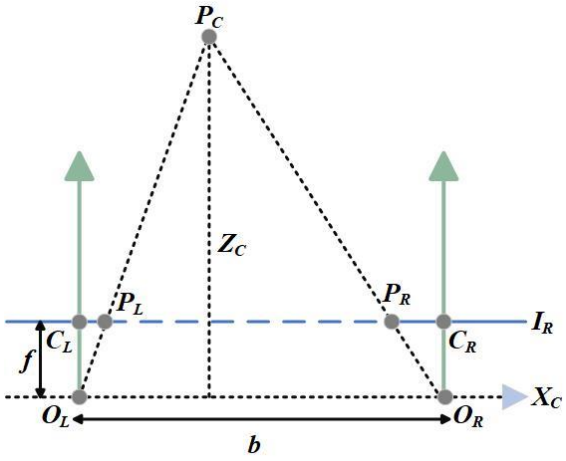


Figure 7: Depth camera ranging schematic

Note: OL and OR are the centers of the left and right cameras, respectively; CL is the origin of the left imaging plane IL , and CR is the origin of the right imaging plane IR ; PC is a point in the real world, and the mapped points on the imaging plane in IL and the imaging plane in IR are the point PL and the point PR , respectively; f is the focal length, and b is the baseline; and ZC is the distance from the point PC to the camera.

Mapping a point in the two-dimensional pixel plane to a position point in the three-dimensional world requires the transformation of the pixel coordinate system, the image coordinate system, the camera coordinate system and the crab pond coordinate system. As shown in Figure 8, the robot arm coordinate system is first established with the origin O_W , the east direction as the X_W axis, the north direction as the Y_W , and the upper direction as the Z_W axis (X_W , Y_W , Z_W). Install the depth camera in front of the robot arm, take the camera as the origin, the front as Z_C axis, the right horizontal direction as X_C axis, and the plumb line over the optical center as Y_C axis to establish the camera coordinate system (X_C , Y_C , Z_C).

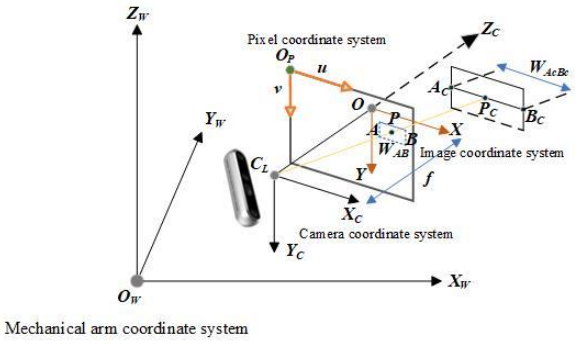


Figure 8: The Sketch Map of transformation relationship between the manipulator and the pixel

Note: {XW , YW , ZW } is the mechanical coordinate system, the origin is OW ; {XC , YC , ZC } is the camera coordinate system, the origin is CL ; {X, Y} is the image coordinate system, the origin is O; {u, v} is the pixel coordinate system, the origin is OP ; the focal length is f; PC is the spatial location of the picking point under the camera coordinate system, and the projected point under the pixel coordinate system is the P point; AB is the connecting line of the midpoints of the two sides of the heights of the projection frame under the pixel coordinate system, and W is the width of the strawberry under the pixel coordinate system; A B is the corresponding projection line of AB under the camera coordinate system, and W is the projection width of the strawberry under the camera coordinate system. AB is the connecting line of the two sides of the projected frame, and WAB denotes the width of the strawberry under the pixel coordinate system; AC BC is the corresponding projected connecting line of AB under the camera coordinate system, and WAcBc denotes the projected width of the strawberry under the camera coordinate system.

As seen in Figure 8, the point P_C is a localization point of the strawberry picking point under the camera coordinate system, after the projection transformation, the localization point P_C is mapped from the three-dimensional space to the point P under the two-dimensional image coordinate system, and after the translation, the point P is then transformed to the pixel coordinate system, and the transformation relationship between the pixel coordinates and the camera coordinate system is as follows:

$$Z_C \begin{bmatrix} \mu \\ \nu \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & \nu_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} \tag{4}$$

Therein: (u,v) is the 2D coordinates of the strawberry picking point under the pixel coordinate system, which is obtained by the improved YOLO v8-Pose model to recognize the RGB image, Z_C is obtained by the depth camera, f_x , f_y , u_0 , and v_0 are the internal references of the depth camera, and the 3D coordinates of the strawberry picking point under the camera coordinate system can be solved by the following equation:

$$\begin{cases} X_C = \dfrac{Z_C(u - u_0)}{f_x} \\ Y_C = \dfrac{Z_C(\nu - \nu_0)}{f_y} \\ Z_C = \dfrac{f \bullet b}{X_L - X_R} \end{cases} \tag{5}$$

It is extremely necessary to convert the 3D coordinates of the picking point under the camera coordinate system to the robot arm coordinate system. The conversion from camera coordinate system to mechanical coordinate system

belongs to rigid body transformation, which can be accomplished by rotation and translation, and the conversion formula is as follows:

$$\begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} \tag{6}$$

Where R is the product of the X, Y, Z 3-direction rotation matrix and T is the translation matrix. The final result is the true 3D coordinates of P_C in the mechanical coordinate system (XPc, YPc, ZPc).

III. RESULTS AND DISCUSSION

3.1 Evaluation indicators description

The evaluation index of the target keypoint detection algorithm is OKS (Object keypoint similarity). The average accuracy AP (Average precision) is obtained from OKS. The mAP-kp (Mean average precision - key point) is calculated from AP as the evaluation index of the key point. Precision P and Recall R were used as the evaluation index of target recognition. Precision and Recall were used as evaluation indexes of target detection results. OKSp is calculated as follows:

$$OKSp = \frac{\sum_i \exp\{-d_{pi}^2 / 2S_p{}^2\sigma_i{}^2\}(\nu_{pi} > 0)}{\sum_i \delta(\nu_{pi} > 0)} \tag{7}$$

$$\delta = \begin{cases} =1(\nu_{pi} > 0) \\ =0(\nu_{pi} \le 0) \end{cases} \tag{8}$$

where dpi represents that the Euclidean distance between the i-th keypoint detected and the corresponding keypoint in the target, Sp is the scale factor of p-points, ν_{pi} is the picking point visibility, 0 is unlabeled, 1 is labeled occluded, and 2 is labeled unseen, and σ_i is the normalization factor of picking points of type i.

The AP mathematical calculation expression is as follows:

$$AP = \frac{\sum_m \sum_p \beta(OKSp > T)}{\sum_m \sum_p 1} \tag{9}$$

$$\beta = \begin{cases} OKSp \quad (OKSp > T) \\ 0 \quad (OKSp > T) \end{cases} \tag{10}$$

3.2 Analysis of model training results

Figure 9 shows the changes in the predicted bounding box loss function of the benchmark model and the improved YOLO v8-Pose model after training. From the figure, it can be observed that compared to the benchmark model, the improved YOLO v8-Pose has a faster convergence speed at the beginning of training and is able to accurately detect the ripeness of strawberries and the location of the picking point while having a lower loss value. As the training proceeds, after about 300 iterations, the model's convergence rate exhibits marked stability, and the enhanced version demonstrates superior convergent characteristics.

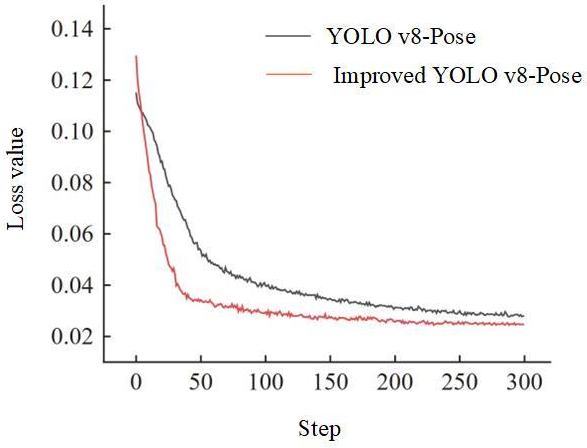


Figure 9: Training process loss function change diagram

The enhanced model underwent evaluation with a dedicated test set, resulting in a normalized confusion matrix that highlights the categorization of varied obstructions by the refined model, as depicted in Figure 10. Each column signifies the accurate classification of the sample, while each row corresponds to the anticipated classification of the sample. Observation of the confusion matrix reveals that the elements on the main diagonal are significantly larger than the other non-diagonal elements. This indicates that the improved YOLO v8-Pose has a high discrimination and recognition rate in strawberry ripeness and picking point detection. 5% of the strawberries were misidentified as Medium strawberries; 2% of Medium strawberries were misidentified as Medium strawberries; 2% of Medium strawberries were misidentified as Immaturity strawberries; and 5% of Immaturity strawberries were misidentified as Medium strawberries. The analysis was due to the fact that individual Medium strawberries were closer in appearance to Maturity strawberries and individual Medium strawberries were closer in appearance to Immaturity strawberries. In addition, 6% of the strawberry picks were misidentified as Background, mainly because the strawberry picks are small and difficult to distinguish from the background when there is occlusion. Overall, the improved YOLO v8-Pose used in this section has a high recognition rate and differentiation between the three different maturity strawberries and picking points.

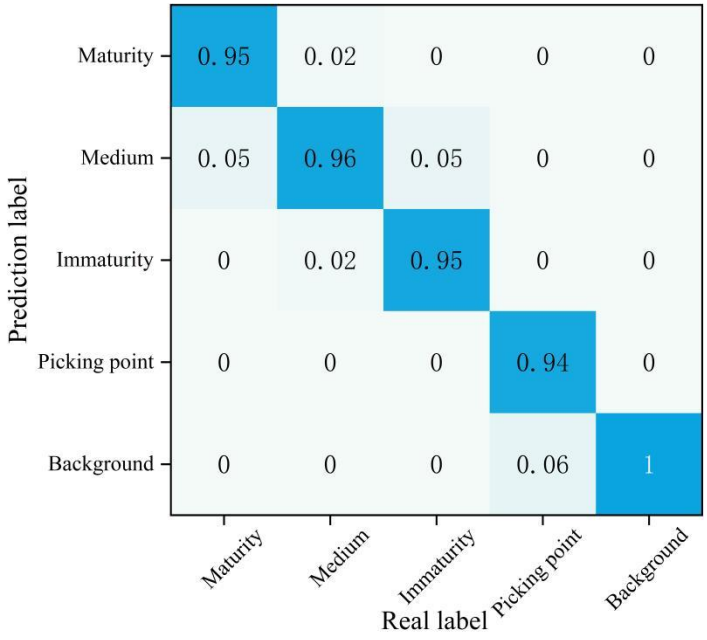


Figure 10: Improved YOLO v8-Pose detection result confusion matrix

3.3 Analysis of ablation experiment

For the purpose of verifying the effectiveness of the three improvement strategies of BiFPN feature fusion network, MobileViTv3 reconfiguration backbone network, and SIoU loss function in improving the performance of the model,

the ablation analysis is conducted utilizing an identical dataset, with progressive incorporation of enhancement techniques based on the initial model. The results of the ablation study are displayed in Table 2.

It can be evidently seen from Table 2, compared with the original YOLO v8-Pose model, Model 2 increases the Precision, Recall and mAP − kpvalues by 1.24%, 0.88%, and 1.92%, respectively, without a significant increase in the number of parameters and computation. It shows that replacing all the Concat modules with BiFPN at the Neck part effectively increases the feature extraction capability and improves the detection accuracy. Model 3 is based on Model 2, the MobileViTv3 module is used to reconfigure and optimize the backbone network, and the Precision, Recall and mAP-kp values are increased by 0.89%, 1.36% and 1.19%, respectively. The reason for analyzing this is mainly because the MobileViTv3 module is able to extract global long-distance dependency relationships through the self-attention in Transformer. mechanism to extract full Director Sequence Modeling Dependency and contextual information while fusing the local information obtained by the convolutional block to improve the detection accuracy of the model. On the basement of model 3, the border loss function of the original model is changed to the SIoU loss function to obtain the final model 4, in which the number of parameters and computation amount remain basically unchanged, the Precision, Recall and mAP-kp values are increased by 1.58, 2.11 and 2.38 percentage points, respectively. The detection time of a single frame is 17.1 ms. In contrast to the the past model, the number of parameters, computation and detection time of the improved YOLO v8-Pose are basically unchanged from that of the initial model and meet the real-time requirements, but the Precision, Recall and mAP-kp values are increased by 3.71%, 4.35%, and 5.49%, respectively. For picking robots, improving the accuracy of strawberry ripeness, and picking point detection while keeping the model parameters basically unchanged is an important prerequisite to ensure the subsequent precise picking.

Table 2. Ablation test results

No.	Model	Parameter Quantity/M	Calculation Quantity/(× 109•s-1)	Precision/%	Recall /%	mAP-kp/%	Detection Time/ms
1	YOLO v8-Pose	3.01	3.79	93.41	92.54	92.36	16.2
2	YOLO v8-Pose-BiFPN	3.25	4.09	94.65	93.42	94.28	16.4
3	YOLO v8-Pose-BiFPN-MobileViTv3	3.65	4.59	95.54	94.78	95.47	17.1
4	YOLO v8-Pose-BiFPN-MobileViTv3-SIoU	3.65	4.59	97.12	96.89	97.85	17.1

3.4 Analysis of the results of experiments comparing different models

This study carries out comparative investigations employing YOLO v5-Pose, YOLO v7-Pose, YOLO v8-Pose, and the refined YOLO v8-Pose models. The deep learning training outcomes for each respective model are illustrated in Table 3, and the analysis of the computer calculation results exhibits that the YOLO v8-Pose base model selected in this paper has the smallest number of parameters, and the number of parameters, Precision, Recall , mAP-kp values and the single-frame image detection time of the improved YOLO v8-Pose are 3.65M, 97.12%, 96.89%, respectively, 97.85% and 17.1 ms. The improved YOLO v8-Pose in this paper. where Precision, Recall, mAP-kp value and Detection time are the best, the YOLO v8-Pose model is overall better than other common models. This demonstrates that the foundational model selected for this study is not merely superior, but the employed enhancement approach also proves effectual, consequently augmenting the model's detection precision.

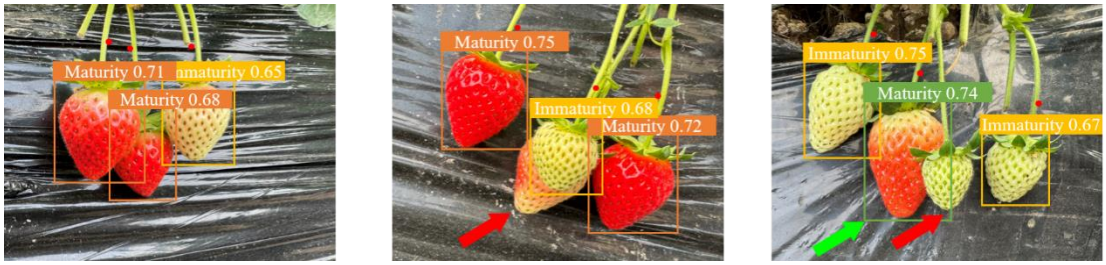
Table 3. Red ripe stage strawberry identification and picking point detection results of different models

Model	Parameter Quantity/M	Precision/%	Recall/%	mAP-kp/%	Detection Time/ms
YOLO v5-Pose	16.45	89.35	90.18	90.16	32
YOLO v7-Pose	13.76	90.89	91.24	91.32	24

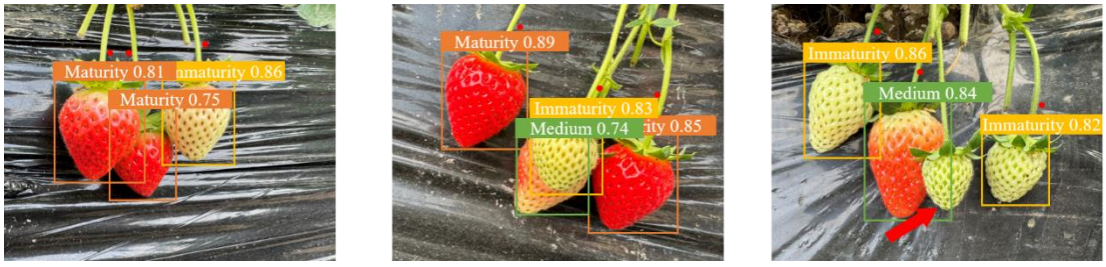
YOLO v8-Pose	3.01	93.41	92.54	92.36	16.2
Improved YOLO v8-Pose	3.65	97.12	96.89	97.85	17.1

Comparison of the results of red ripening strawberry identification and picking point prediction for YOLO v5-Pose, YOLO v7-Pose, YOLO v8-Pose and improved YOLO v8-Pose are shown in Figure 11. As can be seen from Figure 11a, the YOLO v5-Pose detection results showed omissions (marked by red arrows), and its overall confidence was low, and its prediction of picking points was relatively poor, with a large deviation in the location of the predicted "Pick" point, and the "Medium" point was incorrectly predicted as "Medium". Medium" was incorrectly predicted as "Maturity" (marked with a green arrow). As can be seen from Figure 11b, YOLO v7-Pose also misses the detection of shaded strawberries (marked by red arrows), and the overall confidence level is low; the picking points predicted by YOLO v7-Pose have the same bias, which is not able to satisfy the accurate detection of stalk picking points.

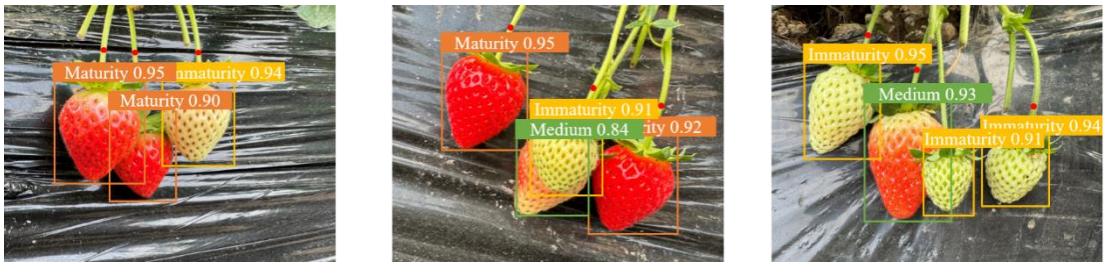
The recognition effects of YOLO v8-Pose model and the improved YOLO v8-Pose model are shown in Figure 11c and Figure 11d, and the comparison shows that YOLO v8-Pose and the improved YOLO v8-Pose can accurately recognize strawberries at different periods of time, but the confidence level of the prediction of YOLO v8-Pose is lower; in the aspect of the detection of the picking point, both models can detect the "pick" key point for detection, but there is the same deviation in the location of the key point predicted by YOLO v8-Pose. In terms of picking point detection, both models are able to detect "pick" keypoints for detection, but YOLO v8-Pose predicts the location of keypoints with the same bias. The improvement of YOLO v8-Pose can solve the problem of low confidence of strawberry detection in YOLO v8-Pose and make more accurate prediction of the key point location, which is more suitable for the picking work of the robot.



(b) YOLO v5-Pose



(c) YOLO v7-Pose



(d) YOLO v8-Pose

Figure 11: Comparison of strawberry ripeness and picking point detection with different network models

3.5 Picking point localization accuracy analysis

During the test, the strawberries in the greenhouse were detected and located within the distance range of 10~70 cm, and after the strawberries were successfully detected, the three-dimensional coordinates of the center point of the strawberries were recorded (XPci , YPci , ZPci,), and the depth distance ZPci was used as the measurement data for evaluation and analysis, and the distance from the center point of the strawberries to the mid-point of the camera baseline, ZdPci , which was obtained with a high-precision laser distance meter, was used as the real distance, and the measurement accuracy of the high-precision laser distance meter was + (2.0 mm + 5 × 10-5 D) (D denotes the distance, km), and the maximum measurement error within 10 m is ± 2.5 mm. The absolute and relative errors of the two sets of data were finally analyzed, and the validation results are shown in Table 4.

The validation results showed that the mean absolute error and the mean absolute percentage error of the measured distances were 0.63cm and 1.16% in the range of 10-70cm, respectively. The maximum absolute error and maximum relative error of the measured distance were 1.37cm and 2.22% respectively. It can be approximately believed that the error steadily increases as the distance increases, mainly due to the limitation of the camera resolution. The depth information of the camera is not sufficient when capturing a long-distance target, and it is also affected by the illumination, which leads to the loss of target details, thus affecting the accuracy of distance measurement. Nevertheless, the overall distance error is still in line with the actual acquisition requirements.

Table 4 Test results of locating accuracy of picking point

Number	Pick			
	ZPci /cm	ZdPci /cm	EZ /cm	EZr /%
1	10.12	10.32	0.20	0.02
2	13.31	13.36	0.05	0.00
3	16.43	16.48	0.05	0.00
4	19.23	19.19	0.04	0.00
5	22.56	22.48	0.08	0.00
6	25.72	25.57	0.15	0.01
7	28.01	27.78	0.23	0.01
8	31.45	31.21	0.24	0.01
9	34.98	34.62	0.36	0.01
10	37.51	38.08	0.57	0.01
11	40.87	41.56	0.69	0.02
12	43.93	44.64	0.71	0.02
13	46.36	47.28	0.92	0.02
14	49.81	50.94	1.13	0.02
15	52.92	54.11	1.19	0.02
16	55.94	54.82	1.12	0.02
17	58.69	59.83	1.14	0.02
18	61.78	62.91	1.13	0.02
19	64.85	66.12	1.27	0.02
20	67.68	69.05	1.37	0.02
Mean			0.63	1.36
Maximum			1.37	2.22

The detection research outputs are shown in Figure 12. In this paper, In accordance with the RGB-D depth camera and using the improved YOLO v8-Pose algorithm can not only accurately detect the ripeness of strawberries, but also accurately recognize the exact location information of the picking point.

Figure 12: Strawberry ripeness detection and three-dimensional location of picking point

Note: The information above the prediction box is the strawberry category and confidence level, and the information above the picking point is the three-dimensional coordinate information of the picking (unit:m)

IV. CONCLUSIONS

This paper presents an accurate detection of strawberry ripeness and picking point based on an RGB-D depth camera with an improved YOLO v8-Pose algorithm. The foremost advancements presented in this study encompass the following:

(1) In response to the difficulty of detecting strawberry ripeness and picking point at the same time in the current research, this paper adopts the improved YOLO v8-Pose that can accurately detect strawberry ripeness and picking point in complex environments simultaneously.

(2) In order to improve the detection performance of the model in complex background environments, this paper has targeted the YOLO v8-Pose model to improve the Neck part by replacing all the Concat modules with BiFPN richer feature fusion; adopting MobileViTv3 to reconfigure the backbone network to cement the ability of the model's contextual feature comprehension; and replacing the CIoU loss function at the output end with an SIoU loss function to accelerate the model convergence. The mAP-kp value of the improved YOLO v8-Pose is 97.85%, which is 5.49% higher than that of the initial model, and the single-frame image detection time is 17.1ms, which is basically unchanged from that of the original model and meets the real-time requirements.

(3) In this paper, we utilize the three-dimensional perception characteristics of the depth camera to accurately detect the strawberry picking point while accurately obtaining its three-dimensional coordinate information. The mean absolute error and the mean absolute percentage error in depth estimation amount to 0.63cm and 1.16%, respectively, which are of high accuracy. This study provides effective technical support for the picking robot to accurately localize the strawberry position.

REFERENCE

[1] Kim B, Han Y K, Park J H, et al. Improved Vision-Based Detection of Strawberry Diseases Using a Deep Neural Network[J]. Frontiers in Plant Science, 2021, 11.

[2] Martin R R, Tzanetakis I E. Characterization and Recent Advances in Detection of Strawberry Viruses[J]. Plant Disease, 2006, 90(4): 384-396.

[3] Shin J, Chang Y K, Heung B, et al. Effect of directional augmentation using supervised machine learning technologies: a case study of strawberry powdery mildew detection[J]. Biosystems Engineering, 2020, 194: 49-60.

[4] Van Henten E J, Hemming J, Van Tuijl B A J, et al. Collision-free Motion Planning for a Cucumber Picking Robot[J]. Biosystems Engineering, 2003, 86(2): 135-144.

[5] Bloss R. Robot innovation brings to agriculture efficiency, safety, labor savings and accuracy by plowing, milking, harvesting, crop tending/ picking and monitoring[J]. Industrial Robot: an International Journal, 2014, 41(6): 493-499.

[6] Luo L, Zou X, Xiong J, et al. Automatic positioning for picking point of grape picking robot in natural environment[J]. Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(2): 14-21.

[7] An Q, Wang K, Li Z, et al. Real-Time Monitoring Method of Strawberry Fruit Growth State Based on YOLO Improved Model[J]. IEEE Access, 2022, 10: 124363-124372.

[8] Yamamoto S, Hayashi S, Yoshida H, et al. Development of a Stationary Robotic Strawberry Harvester with a Picking Mechanism that Approaches the Target Fruit from Below[J]. Japan Agricultural Research Quarterly: JARQ, 2014, 48(3): 261-269.

[9] Arefi A, Motlagh A M, Mollazade K, et al. Recognition and localization of ripen tomato based on machine vision[J]. Australian Journal of Crop Science, 2011, 5(10): 1144-1149.

[10] Lu J, Sang N. Detecting citrus fruits and occlusion recovery under natural illumination conditions[J]. Computers and Electronics in Agriculture, 2015, 110: 121-130.

[11] Wang C, Tang Y, Zou X, et al. Recognition and Matching of Clustered Mature Litchi Fruits Using Binocular Charge-Coupled Device (CCD) Color Cameras[J]. Sensors, 2017, 17(11): 2564.

[12] Hayashi S, Yamamoto S, Saito S, et al. Field Operation of a Movable Strawberry-harvesting Robot using a Travel Platform[J]. Japan Agricultural Research Quarterly: JARQ, 2014, 48(3): 307-316.

[13] Tang Yiping H W. Design and Experiment of Intelligentized Tea-plucking Machine for Human Riding Based on Machine Vision[J]. Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery, 2016, 47(7).

[14] Yingru Y, Huarui W, Yan Z, et al. Tomato disease recognition using leaf image based on complex environment[J]. Journal of Chinese Agricultural Mechanization, 2021, 42(9): 177.

[15] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]//Advances in Neural Information Processing Systems: volume 28. Curran Associates, Inc. 2015.

[16] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779- 788.

[17] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[EB/OL]. (2020-04-23)[2024-03-05]. https://arxiv.dosf.top/abs/2004.10934v1.

[18] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]//Computer Vision - ECCV 2016. Springer, Cham, 2016: 21-37.

[19] Yu Y, Zhang K, Liu H, et al. Real-Time Visual Localization of the Picking Points for a Ridge-Planting Strawberry Harvesting Robot[J]. IEEE Access, 2020, 8: 116556-116568.

[20] Liu X. Liu X, Fan C, Li J, et al. Identification method of strawberry based on convolutional neural network[EB/OL].(2020)[2024-03-05]. https://so1.typicalgame.com/scholar?q=Identification+method+of+strawberry+based+on+convolutional+neural+network

[21] Wang Y, Yan G, Meng Q, et al. DSE-YOLO: Detail semantics enhancement YOLO for multi-stage strawberry detection[J]. Computers and Electronics in Agriculture, 2022, 198: 107057.

[22] Zhang Y, Yu J, Chen Y, et al. Real-time strawberry detection using deep neural networks on embedded system (rtsd-net): an edge AI application[J]. Computers and Electronics in Agriculture, 2022, 192: 106586.

[23] Lamb N, Chuah M C. A Strawberry Detection System Using Convolutional Neural Networks[C]//2018 IEEE International Conference on Big Data (Big Data) . 2018: 2515-2520.

[24] Yu Y, Zhang K, Yang L, et al. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN[J]. Computers and Electronics in Agriculture, 2019, 163: 104846.

[25] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV). 2017: 2980-2988.

[26] Mu Y, Chen T S, Ninomiya S, et al. Intact Detection of Highly Occluded Immature Tomatoes on Plants Using Deep Learning Techniques[J]. Sensors, 2020, 20(10): 2984.

[27] Feng G, Qixin C, Yongjie C, et al. Fruit location and stem detection method for strawberry harvesting robot[J]. 2008.

[28] Yamaguchi K, Kato T, Ninomiya Y. Moving Obstacle Detection using Monocular Vision[C]//2006 IEEE Intelligent Vehicles Symposium. Meguro-Ku,. Japan: IEEE, 2006: 288-293.

[29] Wang Q, Meng Z, Liu H. Review on Application of Binocular Vision Technology in Field Obstacle Detection[J]. IOP Conference Series: Materials Science and Engineering, 2020, 806(1): 012025.

[30] Skoczeń M, Ochman M, Spyra K, et al. Obstacle Detection System for Agricultural Mobile Robot Application Using RGB-D Cameras[J]. Sensors, 2021, 21(16): 5292.

[31] Maji D, Nagori S, Mathew M, et al. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss[C]//2022 IEEE/ CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, LA, USA: IEEE, 2022: 2636-2645.

[32] Tan M, Pang R, Le Q V. EfficientDet: Scalable and Efficient Object Detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 10778-10787.

[33] Mehta S, Rastegari M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer[A]. arXiv, 2022.

[34] Wadekar S N, Chaurasia A. MobileViTv3: Mobile-Friendly Vision Transformer with Simple and Effective Fusion of Local, Global and Input Features[A]. . arXiv, 2022.

[35] Gevorgyan Z. SIoU Loss: More Powerful Learning for Bounding Box Regression[A]. arXiv, 2022.