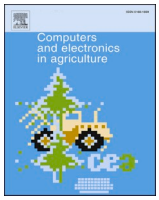
Computers and Electronics in Agriculture 212 (2023) 108056

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Tomato 3D pose detection algorithm based on keypoint detection and point cloud processing

Xiaoqiang Du [a,b,c,*], Zhichao Meng [a], Zenghong Ma [a,b,c], Wenwu Lu [a], Hongchao Cheng [a]

[a] School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
[b] Key Laboratory of Transplanting Equipment and Technology of Zhejiang Province, Hangzhou 310018, China
[c] Key Laboratory of Agricultural Equipment for Hilly and Mountainous Areas in Southeastern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Hangzhou 310018, China

ARTICLE INFO

Keywords:
YOLO v5
Point cloud
3D pose
Keypoint detection
Tomato

ABSTRACT

Tomatoes are widely grown all over the world and are one of the favourite vegetables of humanity. Tomato harvesting requires a lot of labor. With the aging population and the increasing demand for tomatoes, it is imminent to develop tomato picking robots. In the picking environment with a complex background, a 3D pose of the target is essential. It can provide guidance for robotic arm attitude and obstacle avoidance during picking. In this paper, a tomato pose detection algorithm (TPD) is proposed for 3D pose detection of a single fruit of clustered tomato. The TPD algorithm is divided into two modules: YOLO-lmk model and the point cloud processing module. By analyzing the network structure, optimizing parameters, optimizing loss function, adding attention mechanism and adding keypoint prediction, a YOLO-lmk model with YOLO v5s as the core is proposed to realize tomato bounding box and keypoint detection. The point cloud processing module is composed of point cloud segmentation, voxel downsampling, removing outliers, Euclidean color clustering, RANSAC sphere fitting, and keypoint index. As the experimental results show, YOLO-lmk model bounding box mAP is 92.9%, d_{lmk} is 7.9, FLoating point Operations (FLOPs) is 16.6B, and speed is 0.062 s/sheet. The d_{lmk} represents the Euclidean distance between true keypoints and predicted keypoints. Compared with YOLO v5s, mAP is increased 1.8%, and FLOPs is increased only 0.1B. The point cloud processing module only takes 0.028 s to complete a tomato 3D pose detection, which is fast. The accuracy of the TPD algorithm in detecting tomatoes is 93.4%, and the time cost is 0.09 s for detecting one tomato. The TPD algorithm can provide a theoretical basis for tomato 3D pose detection, and provide a reference for other fruits (pears, citrus, apples) 3D pose detection.

1. Introduction

Tomato harvesting belongs to a labour-intensive industry. With the expansion of the tomato planting area and the increase of labour costs year by year, robotic picking is the future development direction (Tang et al., 2020). Mechanical massive harvesting has been widely used for tomatoes planted in a large field. However, non-destructive harvesting robots for fresh tomatoes have yet to be used commercially (Zhang et al., 2022). In recent years, many researchers have developed various tomato picking robots for non-destructive picking of tomatoes (Jun et al., 2021; Lili et al., 2017; Fujinaga et al., 2021). The above researches on picking robots only determine the 3D position of the tomato, not the 3D pose of the tomato. In the picking environment with a complex background, a 3D pose is essential. It can provide guidance for robotic arm attitude and obstacle avoidance during picking. Vision technology is also used in various agricultural engineering fields, such as target detection and location (Wu et al., 2022; Tang et al., 2023), target 3D reconstruction (Lin et al., 2021), target counting (He et al., 2022), etc. Vision technology is a research highlight in agricultural engineering applications. Many types of tomatoes can be roughly divided into single fruit growth and cluster growth. For tomatoes growing in clusters, the whole bundle picking is unreasonable because a cluster of tomatoes is not mature simultaneously. So the method to detect the 3D pose of a single mature fruit in clustered tomatoes is studied in this paper.

With the development of artificial intelligence and the expansion of datasets, researchers are committed to using deep learning technology to solve the target detection task in agriculture (Sozzi et al., 2022; Yao et al., 2021; Zhao et al., 2021). Wang et al. (2022a) proposed DSE-YOLO

* Corresponding author.
E-mail address: xqiangdu@zstu.edu.cn (X. Du).

https://doi.org/10.1016/j.compag.2023.108056
Received 19 August 2022; Received in revised form 26 May 2023; Accepted 10 July 2023
Available online 16 July 2023

(DetailSemantics Enhancement You Only Look One) to detect multi-stage strawberries. The detection results showed that the mAP value was 86.58 %, and the F1 score was 81.59 %. Li et al. (2021a) proposed an efficient grape detection model to solve the problem of decreased detection accuracy caused by complex growing environments, shadows of branches and leaves, and overlapping grapes. Experiments showed that the YOLO-Grape model had high detection accuracy for occluded grapes. The above research shows that the target detection technology based on deep learning could achieve the target detection task in complex agricultural environments, whether a tiny target or a serious occlusion.

The keypoint detection technology is widely used in human bone and joint extraction and is less applied in agriculture (Cao et al., 2017; Liang et al., 2014; Wu et al., 2020). Du et al. (2022) proposed a method that relies on keypoints to achieve body measurements for cattle and pigs. Nasirahmadi et al. (2017) implemented the classification of sweet and bitter almonds using a Bag-of-Feature (BoF) model composed of keypoint detectors and SIFT descriptors. The research above shows that the keypoint detection technology can extract the required keypoints.

In recent years, research on fruit pose detection has become a hot topic. Zhang et al. (2022) proposed a 3D pose detection method for tomato bunch, named Tomato Pose Method (TPM). By cascading target and keypoint detection networks, the bounding box and the 11 keypoints of tomato bunches were detected. The pose detection of the tomato bunches was detected according to multiple keypoints. Since the image needed to be input into two networks, the speed was slow. It took 0.93 s to realize the pose detection of a bunch of tomatoes, and 11 keypoints were easily blocked. Luo et al. (2022)relied on Mask R-CNN and point cloud technology to detect the pose of grapes. Since Mask R-CNN belonged to the two-stage network, the speed was slow. It took 1.786 s to realize the pose detection of a bunch of grapes. GUO et al. (2020) completed the pose detection of orange, cucumber, pineapple, and cabbage in the laboratory environment through point cloud matching. Cucumber is the fastest among the four fruits, and SAC-IA ICP takes 12.898 s. In summary, it is meaningful to propose a fast tomato 3D pose detection algorithm. This paper proposes an end-to-end YOLO-lmk model, which simultaneously completes tomato bounding box and keypoint detection tasks in one model, and combines point cloud processing to complete tomato 3D pose detection. Since the image is only input to one model, the speed is fast.

In agriculture, spatial information must be obtained for fruit phenotypic detection, spatial positioning, and path recognition. Due to the imaging method, 2D images cannot provide more depth information. A RGB-D camera obtains the color and depth images to synthesize the point cloud, providing a new direction for machine vision. Li et al. (2021b) used point cloud to realize 3D positioning of tea picking points. Miao et al. (2021) realized the automatic segmentation of corn buds, stems, and leaves based on the point cloud. The above research shows that point clouds can effectively provide the required 3D information.

In this paper, a tomato keypoint dataset is constructed, and a tomato pose detection algorithm (TPD) is designed for 3D pose detection of a single fruit of clustered tomato. TPD is divided into two modules: YOLO-lmk and the point cloud processing module. Specifically, the tomato's bounding box and the calyx's center are output through the YOLO-lmk model. The point cloud in the bounding box is segmented, and the 3D position of the tomato's centroid and the calyx's center point is obtained through a series of point cloud processing operations. Connecting two 3D points is the 3D pose of a tomato. The main contributions of this paper are as follows:

1) By analyzing the network structure, optimizing parameters, optimizing loss function, adding attention mechanism and adding keypoint prediction, a YOLO-lmk model with YOLO v5s as the core is proposed to realize tomato bounding box and keypoint detection.
2) The point cloud processing module only needs to process the point cloud in the bounding box according to the bounding box information, avoiding processing the entire point cloud. Combined with the voxel downsampling and the Euclidean color clustering, the speed is faster, and the accuracy is high.
3) Combining YOLO-lmk model with the point cloud processing module, a fast tomato pose detection algorithm (TPD) is proposed, which only needs to pass in a set of RGB-D images to obtain the 3D pose of the tomato.
4) By referring to the concept of coaxiality in mechanical design, an evaluation method of whether 3D pose detection is correct is proposed. This method is more intuitive and effective.

The rest of the paper is structured as follows: Section 2 describes the dataset construction and illustrates the data augmentation method used for training and the details of the method proposed in this paper; Section 3 evaluates the performance of the TPD algorithm for bounding box, keypoint, and 3D pose detection; Section 4 discusses, and Section 5 presents the conclusion.

2. Materials and methods

2.1. Data acquisition

This paper takes tomatoes as the research object. The images used in this research are all from the Haining Yangdu Base in Jiaxing, Zhejiang Province, China. The shooting time is June 17, 2022. Fig. 1 shows a schematic diagram of the image acquisition process during this study. The experimenter takes images at a position 0.5 m horizontally away from the tomato plant, keeping the mobile phone or Realsense L515 camera at the height of 1.5 m and an inclination of 45 degrees. Table 1 shows 3354 tomatoes with visible calyx and 146 tomatoes with invisible calyx in dataset 1, totaling 3500. Shooting in this position can better

Fig. 1. Schematic diagram of the image acquisition process.

Table 1
Dataset 1.

Name	The number of images in the training set	The number of images in the test set	Number of dataset images	Number of tomatoes with visible calyx	Number of tomatoes with invisible calyx	Total number of tomatoes
Dataset 1	895	224	1119	3354	146	3500

obtain tomato fruit with visible calyx. A total of two datasets are used in this study. Dataset 1 was photographed by a mobile phone for training the YOLO-lmk model. The pixels are 1080 × 1920, and the saving format is.jpg. There are a total of 1119 images divided into the training set and test set by 4:1. Dataset 2 consists of 15 samples taken with Realsense L515 camera, each sample includes a RGB image and a depth image, used to evaluate the 3D pose of tomatoes. The pixels of the RGB image are 1280 × 720, the pixels of the depth images are 1280 × 720, and the saving format is png.

2.2. Dataset preparation

2.2.1. Color point cloud acquisition

This study uses RGB and depth images to synthesize colored point clouds. The relationship between depth images and camera space can be calculated using Realsense L515 camera intrinsics. As shown in Fig. 2, assuming that A is the pixel point of the tomato, and the corresponding 3D point in the camera is A1, the relationship between A and A1 is shown in Eq. (1). The RGB image and the depth image are aligned and have the same pixel size. Through Eq. (1), each pixel in the depth image is converted into a three-dimensional point in the camera space, and then the color information of the same pixel position in the RGB image is integrated to form a color point cloud.

$$Z_1\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} \tag{1}$$

where (u, v) and (x, y) are A's pixel coordinates and image coordinates, respectively. (X_1, Y_1, Z_1) are the 3D coordinates of A1 in camera space. (c_x, c_y) is the pixel coordinate of the center of the image. f is the focal length of the camera.

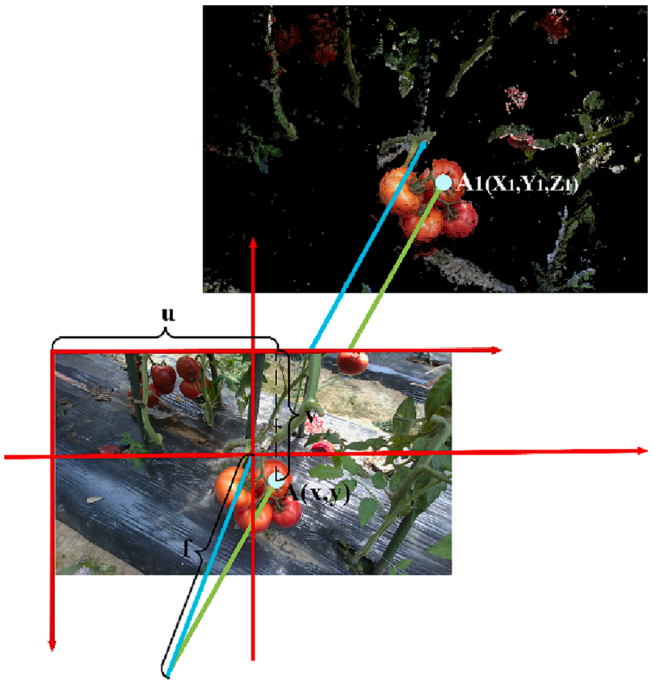


Fig. 2. Schematic diagram of point cloud synthesis.

2.2.2. Labelling bounding boxes and labelling keypoints

The labelme annotation tool marks the bounding box and the keypoint in this study. An annotation file recorded the coordinates of the center of the bounding box, its width and height, and the keypoint coordinates. As shown in Fig. 3, if the calyx is visible, the keypoint is marked at the center of the calyx. If the calyx is not visible due to the tomato posture, it is marked at the junction of the tomato border and the fruit axis. A bounding box contains only one keypoint, because too many keypoints are easily affected by leaf occlusion, and it will take more time to predict multiple keypoints. From Table 1, due to the optimal shooting angle, the proportion of tomatoes with invisible calyx to the total number of tomatoes is extremely low, so the tomatoes with visible and invisible calyx are divided into one category for training.

2.2.3. Data augmentation

To improve the richness of experimental data and the generalization of the model, random data augmentation operations are performed on the training set in Dataset 1, and no data augmentation operations are used on the test set. As shown in Table 2, data augmentation method includes mosaic, translation, scale, and color space conversion. Compared with the RGB space, the HSV color space is closer to human's subjective perception of color, and it is very intuitive to show the main tone of the color, the degree of vividness, and the degree of lightness and darkness. In the tomato image, the background, tomato, stem, and leaves have obvious differences in color characteristics. By adjusting the HSV during training, the adaptability of the model under various lighting conditions can be effectively improved. When the YOLO-lmk model is trained, each image is processed through a random combination of 4 data augmentation methods, and the corresponding annotation files for each image are simultaneously transformed.

2.3. Real 3D pose of tomato

When creating Dataset 2, as shown in Fig. 4(a), the colored stick was inserted into the center of the tomato calyx and the tomato tail respectively, and the 3D pose of the stick was considered as the tomato 3D pose. Fig. 4(b) showsthe marker segmentation obtained by dividing the color image by HSV color and area screening. Fig. 4(c) shows that the mark center is obtained by averaging. Finally, as shown in Fig. 4(d), two central positions in the point cloud are indexed to obtain its 3D coordinate points. Two 3D coordinate points are connected to obtain the real 3D pose of the tomato. Coaxiality is a position constraint on the point features in the diameter direction of the rotating body relative to the datum center feature. As shown in Fig. 4(e), this study proposes to use the coaxiality method to evaluate the accuracy of 3D pose detection, which is to use the real 3D pose of a tomato to construct a cylinder with a radius of 1.5 cm. The radius of the tomato pedicle is slightly smaller than that of the tomato calyx, and the radius of the tomato calyx is about 2.2 cm (Liu et al., 2015). Because the radius of the cylinder is designed according to the tomato pedicle, so the radius of the cylinder is set to 1.5 cm. If the tomato 3D pose is inside the cylinder, the detection is correct, and if it is outside the cylinder, the detection fails.

2.4. Overall technical route

In order to solve the problem of tomato 3D pose detection, this study

Fig. 3. Dataset labelling methods.

Table 2
Data Augmentation probability.

Name	Mosaic	Translation	Scale	Color space conversion(HSV)
Dataset 1	0.5	0.1	0.5	hsv_h:0.015, hsv_s:0.7, hsv_v:0.4

proposes a tomato 3D pose detection algorithm, which is mainly composed of YOLO-lmk model and point cloud processing module. The overall flowchart of the proposed method is shown in Fig. 5. Firstly, a tomato keypoint detection dataset is created for training the YOLO-lmk model and then is tuned. Deploy the model in the ROS framework and use actions to communicate between the YOLO-lmk model and the point cloud processing module. The YOLO-lmk model consists of YOLO v5s adding keypoint predictions. Perform both bounding and keypoint detection tasks in one model. The YOLO-lmk model outputs bounding boxes and keypoints mapped on the point cloud. The point cloud image is cropped according to the bounding box information, avoiding the use of the entire point cloud and speeding up the operation of the algorithm. The point cloud in the bounding box is subjected to voxel down-sampling, outlier removal, Euclidean color clustering and RANSAC sphere fitting algorithm to obtain the tomato centroid. Finally, the point cloud position of the keypoint is indexed, and the sphere center and the keypoint are connected to obtain the 3D pose of the tomato.

2.5. YOLO v5

To solve the problem of the slow detection speed of the two-stage deep learning model, Redmon et al. (2016) proposed a one-stage target detection algorithm YOLO. Since it was proposed, YOLO has been widely used in the agricultural field with its simple network structure and fast detection speed (Fan et al., 2022; Jintasuttisak et al., 2022; Liu et al., 2021). After many improvements, the low detection accuracy and weak detection of small targets in the YOLO series have

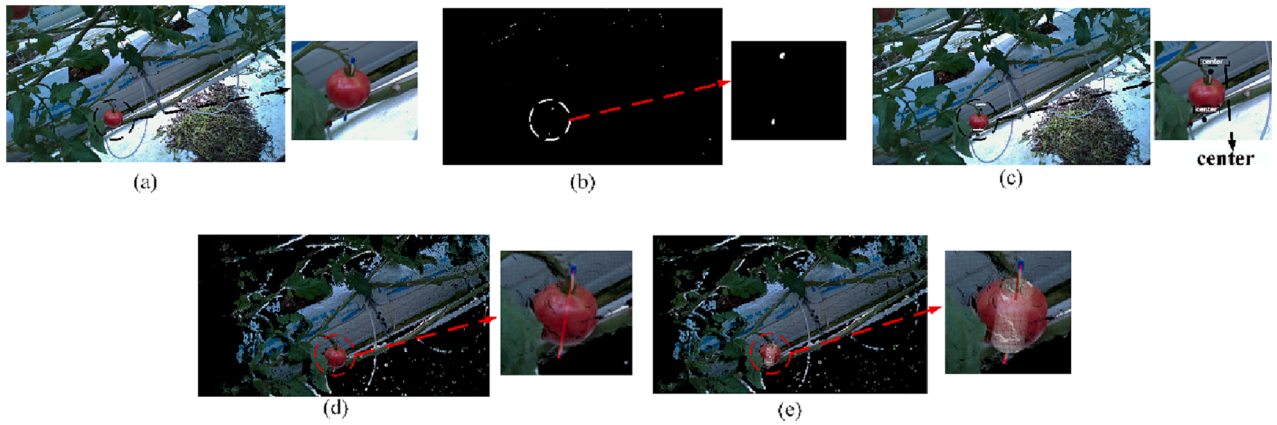


Fig. 4. Real 3D pose of tomato. (a) The original image. (b)The marker segmentation. (c) The marker center. (d) Real 3D pose of tomato. (e) The coaxiality method.

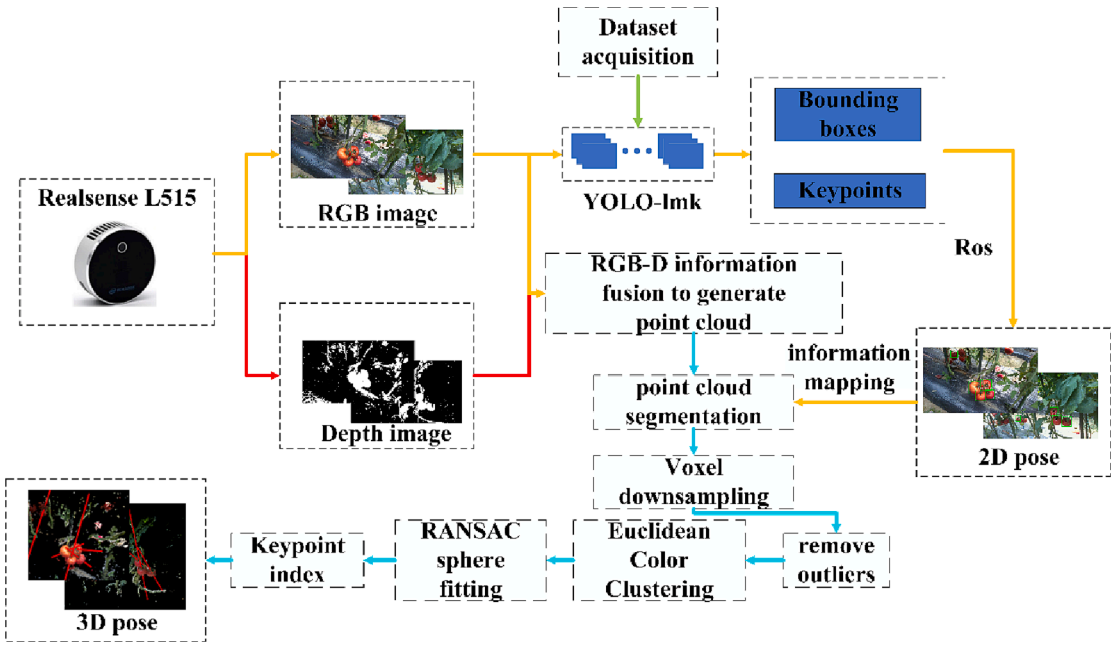


Fig. 5. Overall flowchart of the proposed method.

been greatly improved (Wang et al., 2022b). Fig. 6 shows the model architecture of YOLO v5s. YOLO v5 is mainly improved by adding Focus, BottleneckCSP, SPP, and PANet modules based on YOLO v3. YOLO v5 is mainly composed of a backbone, neck, and three detection heads. The backbone is used to extract image features at different scales. The neck plays the role of feature fusion. The three feature maps obtained by detecting heads are used to predict small, medium, and large objects from large to small. These feature maps are divided into grids. For each grid, 3 prior boxes (anchors) will be used to predict the target bounding box. Finally, each bounding box will output a feature vector, including the predicted bounding box coordinate offset, object confidence and classification probability. According to the depth and width of the network, it is divided into four models: YOLO v5s, YOLO v5m, YOLO v5l, and YOLO v5x.

Different from other versions, YOLO v5 uses an adaptive anchor strategy (Gao et al., 2019), the backbone part uses the Focus and C3 (Wang et al., 2020), and the bounding box regression loss uses GIoU Loss.

2.6. *Bounding box loss function*

YOLO v5 bounding box regression uses GIoU loss, but there are two problems. 1) When the prediction box contains the ground truth box, GioU loss degenerates to IoU loss, and the prediction cannot be evaluated. 2) Slowly converge in the horizontal and vertical directions when the prediction box and the ground truth box intersect. Hence some variants of IoU loss functions have emerged, such as DIoU loss, CIoU loss, and SIoU loss. To correct the shortcomings of GIoU loss, Zheng et al. (2020) proposed the DIoU loss and the CIoU loss. The DIoU loss sums up the Euclidean distance between the center point of the prediction box and its corresponding ground truth box to speed up the regression process. The CIoU loss reduces the aspect ratio gap between the prediction box and the ground truth box by adding new items, thus further optimizing the DIoU loss.

So far, the proposed methods do not take into account the direction of the mismatch between the prediction box and the ground truth box. This deficiency results in slower and less efficient convergence, resulting in a worse model. SIoU loss (Gevorgyan, 2022) is a newly proposed loss function in 2022, which redefines the penalty metric considering the vector angle between the required regressions. The SIoU loss consists of

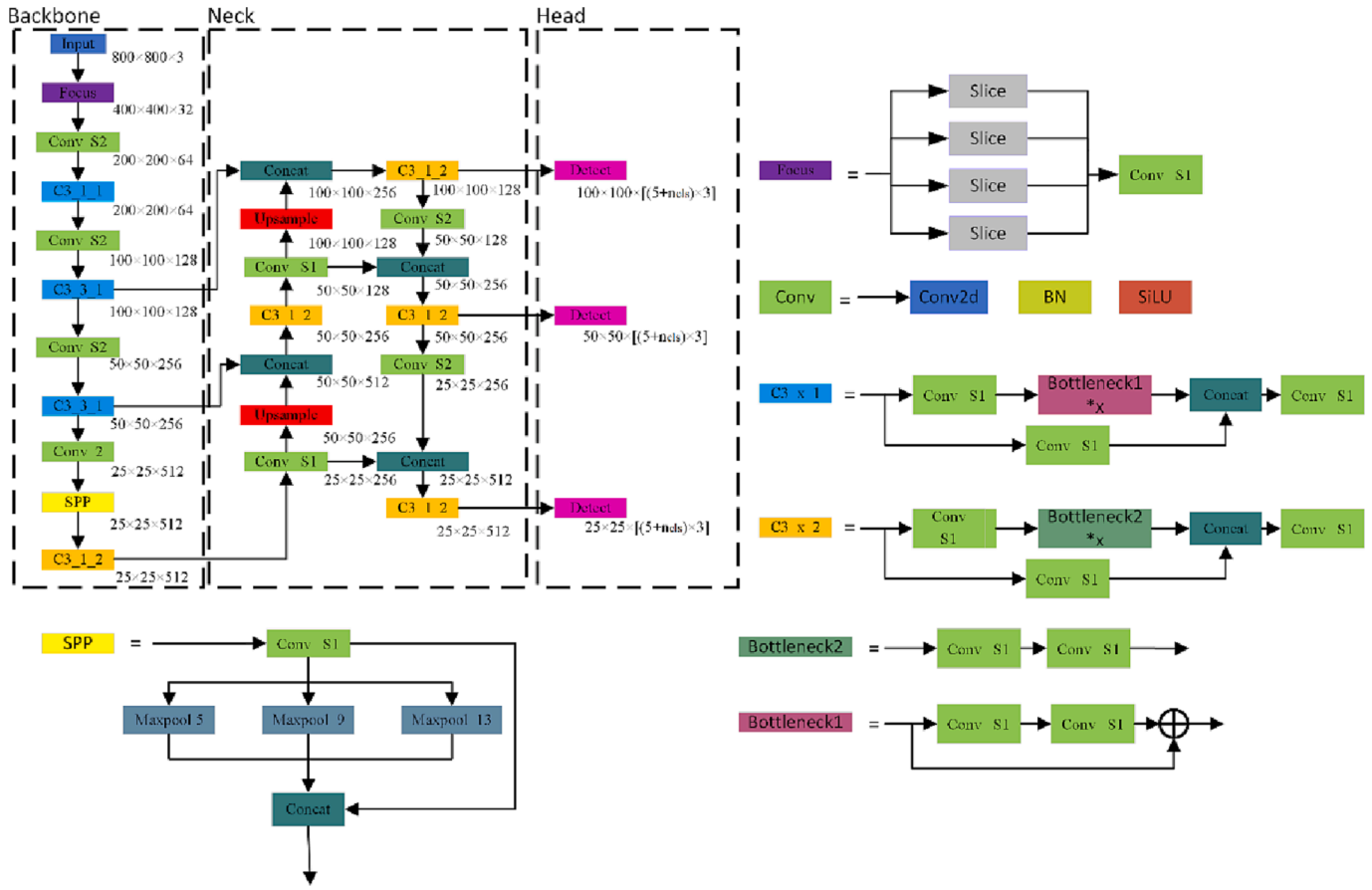


Fig. 6. YOLO v5s model architecture.

4 Cost functions: Angle cost, Distance cost, Shape cost, and IoU cost.

2.7. Keypoint loss function

Keypoint detection is widely used in face detection. This paper draws on the keypoint loss function in face detection. The training image is I, the network structure is Φ, the prediction result is $S^{'} = \Phi(I)$, and the loss function is shown in Eq. (2).

$$loss(\boldsymbol{s}, \boldsymbol{s}^{'}) = \sum_{i=1}^{2L} f\left(s_i - s_i^{'}\right) \tag{2}$$

where s represents the real position of the keypoint, and $f(x)$ uses the most loss functions: L_1 loss, L_2 loss and smooth$_{L1}$ loss.

In this regard, Feng et al. (2018) proposed Wing loss, which is shown as Eq. (3). For small errors, logarithmic function with offset is used, and L_1 loss is used for large errors, which balances the impact of large and small errors on keypoints in the early and late training stages.

$$\text{Wing}\ (x) = \begin{cases} w\ln(1 + |x|/\varepsilon) \ \text{if}\ |x| < w \\ |x| - C \ \text{otherwise} \end{cases} \tag{3}$$

where w is an integer to constrain the range of the nonlinear part in the interval $[-w, w]$. ϵ is the curvature of the nonlinear bound region, and $C = w - w\ln(1 + x/\epsilon)$ is a constant that can smoothly connect the linear and nonlinear parts of the segment.

2.8. Attention mechanism

The attention mechanism enhances the representation of important features in the feature map and suppresses unimportant features (She et al., 2022). Existing research shows that adding attention mechanisms can bring performance leaps to the network. This study added Coordinate Attention (CA) to the YOLO-lmk model for ablation experiments to obtain better results. Fig. 7 shows the network architecture of CA attention mechanism modules. CA attention mechanism modules embed position information into the channel attention mechanism and decompose the global pool into one-to-one feature encoding operations, so that the attention mechanism modules can capture precise information.

2.9. Evaluation metrics

2.9.1. Evaluation metrics for bounding-box detection

In order to verify the accuracy of YOLO-lmk in tomato bounding box regression, this paper uses precision (Eq. (4)), Recall (Eq. (5)) and mAP (Eq. (6)) as evaluation metrics. At the same time, FLoating point Operations (FLOPs) is used to measure the complexity of the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

where TP (true positive) represents the number of correctly detected tomatoes, FP (false positive) represents the number of falsely detected tomatoes, and FN (false negative) represents the number of missed tomatoes.

Models are tested on the test set, and all samples are ordered by their confidence scores. The corresponding Precision and Recall are calculated when the sample threshold is determined. Changing the threshold affects Precision and Recall, and the confidence in this study is set to 0.4.

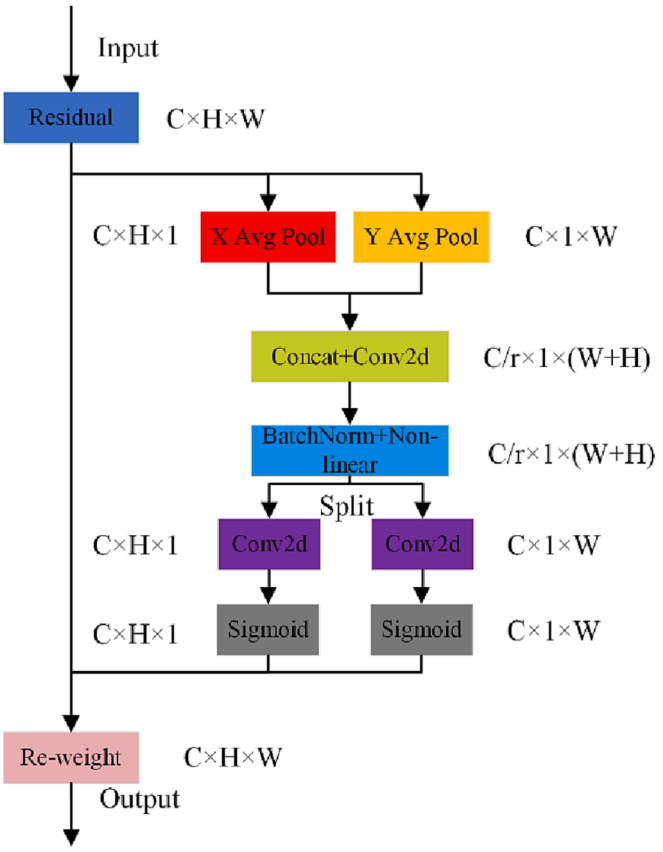


Fig. 7. The network architecture of CA attention mechanism modules.

The Precision-Recall curve is a 2D curve with the *x*-coordinate representing Recall and the *y*-coordinate representing Precision. AP (Eq (21)) is the area enclosed by the curve and the coordinate axis.

$$\mathrm{AP} = \int_0^1 P(R)d(R) \tag{6}$$

There is only one category of tomatoes in this study. Therefore, the AP value is consistent with the mAP value.

2.9.2. Evaluation metrics for keypoint detection

The Euclidean distance(Eq. (7)) between the predicted and real keypoint is selected as the evaluation metric for keypoint detection.

$$d_{lmk} = \frac{\sum_{i=1}^{n}\sqrt{\left(lmkx_{pi} - lmkx_{ti}\right)^2 + \left(lmky_{pi} - lmky_{ti}\right)^2}}{n} \tag{7}$$

where d_{lmk} represents the average keypoint error, n represents the number of keypoints, $lmkx_{\mathrm{pi}}$ represents the x-axis coordinate of the i-th predicted keypoint, $lmky_{\mathrm{pi}}$ represents the y-axis coordinate of the i-th predicted keypoint, and $lmkx_{\mathrm{ti}}$ represents the x-axis coordinate of the i-th real keypoint, $lmky_{\mathrm{ti}}$ represents the y-axis coordinate of the i-th real keypoint.

2.10. Experimental setting

The software and hardware configurations for model training and testing in this study are listed in Table 3. The training epochs are set to 700, the batch size is 32, the optimizer uses Adam, and the cosine return fire algorithm is used to update the learning rate.

Table 3
Software and hardware configuration.

Accessories	Model
CPU	Intel Xeon E5-2680 v3
RAM	32G
Operating system	Ubuntu20.04
GPU	NVIDIA GeForce RTX 2080Ti *2
Development Environments	Python3.7,Pytorch1.8.1 CUDA11.1

2.11. YOLO-lmk

In this research, the parameters of YOLO v5s and the loss function are optimized, and an attention mechanism is added, which is the proposed YOLO-lmk. Fig. 8 is the model architecture of YOLO-lmk.

The bounding box loss function of YOLO-lmk is changed from GIoU to SIoU. SIoU redefines the penalty metric considering the vector angle between the desired regressions. SIoU is one of the best loss functions out there. The keypoint loss function is Wing loss, which can well balance the influence of the size error before and after training on the prediction of keypoints. The parameters in the SPP module of YOLO v5 are designed according to the objectives in the MSCOCO dataset. Analysis of the tomato dataset shows that the tomato samples are small targets, which are very different from the MSCOCO dataset. Therefore, the parameters in SPP module modified from 5, 9, 13 to 5, 7, 9, and 13, which adds more scale features to improve the detection accuracy of tomato. Since the shape and color of the tomato are different from the background, an attention mechanism is introduced to improve the accuracy of the YOLO-lmk model. A CA attention mechanism is added after the third C3 module to improve the model's ability to extract important features. The detect layer is mainly used for the final reasoning and detection of the model. The network applies the Anchor box to the feature map output by the previous layer of the Neck network. Finally, it outputs the category probability, object score, bounding box position vector, and keypoint position vector containing the target object. Compared with YOLO v5, YOLO-lmk has more keypoint position vectors in the detect layer to complete the detection task of keypoints and bounding boxes in one model.

2.12. Point cloud processing module

The point cloud processing module is composed of six parts: point cloud segmentation, voxel downsampling, removing outliers, Euclidean color clustering, RANSAC sphere fitting, and keypoint index. It will obtain the 3D position of the centroid of the tomato and the 3D position of the keypoints, which can be used to determine the 3D pose of the tomato fruit.

Point cloud segmentation: Input a color image to the YOLO-lmk model to output the target bounding box and tomato keypoints. Map bounding boxes and keypoints onto the point cloud. Segment the tomato target from the point cloud according to the bounding box, reducing the number of point clouds processed and speeding up the point cloud processing time.

Voxel downsampling: With the improvement of the accuracy of point cloud acquisition instruments, the number of point clouds that need to be processed and calculated continues to increase. Traditional point cloud feature algorithms are not able to process the point cloud with numerous point data in real time. Voxel downsampling creates a three-dimensional voxel grid based on the original point cloud data. Calculate the centroid of all points in each cube to replace all points in the three-dimensional grid, reduce the number of point clouds, maintain the morphological characteristics of tomato point clouds, and increase the algorithm's speed.

Euclidean color clustering: In this study, the 3D pose of the tomato is obtained by finding the tomato centroid and a keypoint. The

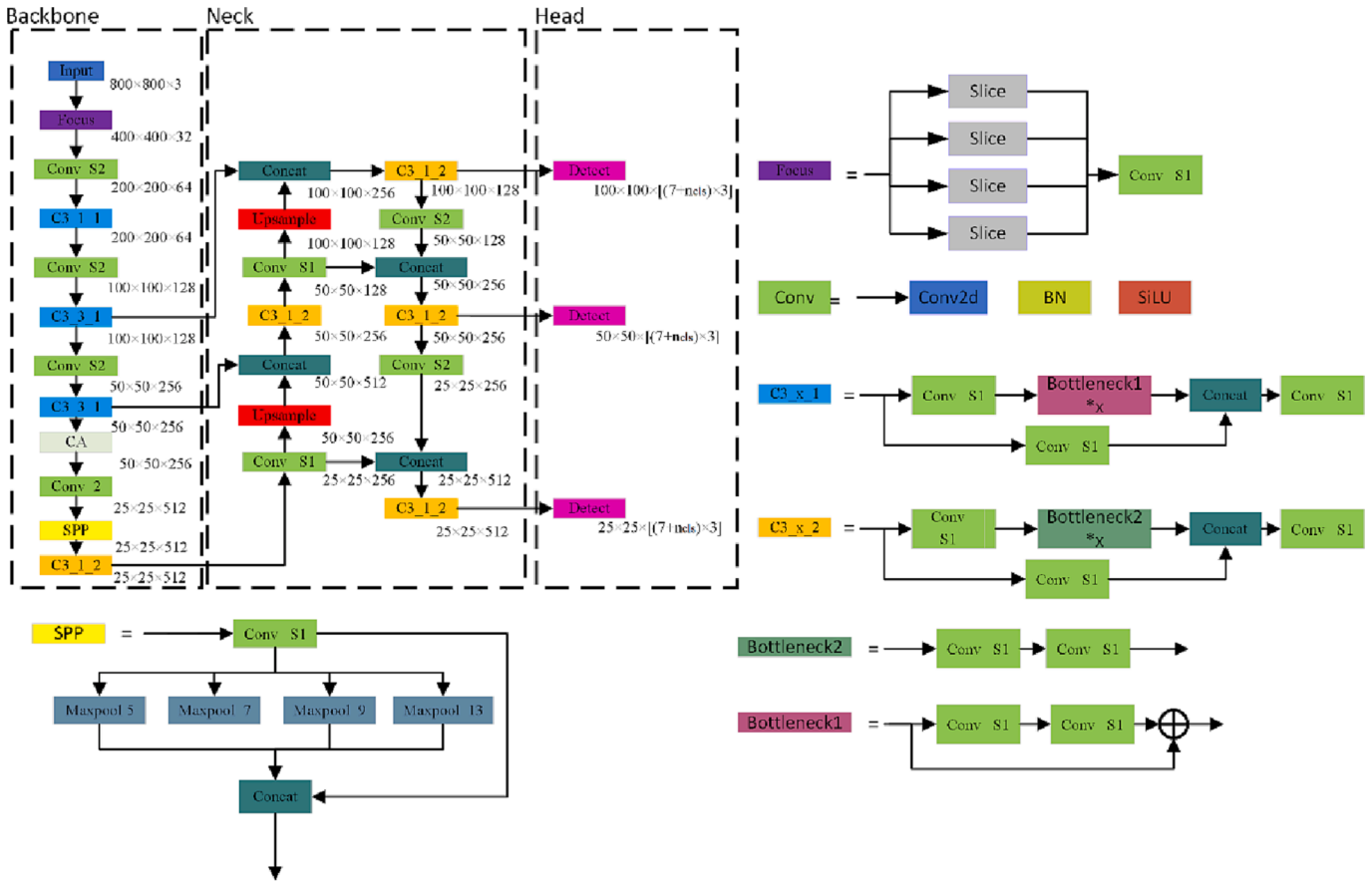


Fig. 8. YOLO-lmk model architecture.

biological morphology of the tomato is oblate, and the point cloud of the calyx has an adverse effect on the tomato centroid obtained by fitting the tomato point cloud with a RANSAC sphere. The point cloud of the tomato calyx is green, and the tomato fruit is red, with a big color difference. Tomato calyx can be separated from tomato fruit by Euclidean color clustering. The tomato point cloud is more spherical after removing the calyx, which is beneficial to obtaining a more accurate tomato centroid.

RANSAC sphere fitting: The tomato centroid is obtained by RANSAC sphere fitting algorithm. The principle of RANSAC sphere fitting algorithm is as follows:

1) Randomly select 4 points and calculate their corresponding spherical surface.
2) The distance d from all points to the sphere is calculated, and the threshold d_r is selected. If $d < d_r$, it is considered as an inner point of the model. Otherwise, it is considered as an outer point of the model, recording the number of points in the current model.
3) Iterate the above two steps, and select the model parameters with the largest number of inner points. Each iteration computes an iterative judging factor based on the desired error rate, number of inliers, total samples, and the current number of iterations. It is determined whether to stop the iteration according to the evaluation factor.

Keypoint index: The keypoints provided by the YOLO-lmk model are 2D points, which need to be indexed in the point cloud to obtain 3D points. The depth image obtained by Realsense L515 cannot obtain the depth of all pixels. Therefore, there will be cases where the tomato keypoint cannot be indexed. The algorithm traverses the left, right, up and down 10 pixels of the keypoints until the 3D position is obtained. Because the position of the 10 pixels in the 3D space has no obvious change, the keypoint position of the index is accurate.

Finally, the 3D position of the keypoint and the 3D position of the tomato centroid are connected to obtain the 3D pose of the tomato.

3. Results

YOLO v5 only implements bounding box regression and cannot detect keypoints. In this study, the model structure of YOLO v5 is analyzed, and a keypoint is regressed in the detection layer based on YOLO v5. The best YOLO v5 model for the tomato keypoint detection task is selected. YOLO v5x is not considered due to excessive computation. The experimental results are shown in Table 4, the mAP, d_{lmk} and FLOPs of YOLO v5s are 91.1%, 7.08, 16.5B, respectively. Although the mAP of YOLO v5s is 0.1% and 0.8% lower than YOLO v5m and YOLO v5l, respectively, but has the lowest d_{lmk} and FLOPs. YOLO v5s can predict bounding boxes and keypoints faster and is easier to deploy in embedded devices, so YOLO v5s is chosen as the base network.

Table 5 shows the impact of different bounding box loss functions on the YOLO model. It can be seen from the table that using DIoU, CIoU and SIoU loss functions in the YOLO model, mAP has a certain degree of improvement compared to GIoU. The highest mAP obtained using SIoU loss is 92%, which is 0.9% higher than GIoU loss. Because the SIoU loss function is the bounding box loss, changing the loss function does not

Table 4
Comparison of different YOLO v5 models.

Model	Precision	Recall	mAP	d_{lmk}	FLOPs/B
YOLO v5s + keypoint	90.7	93	91.1	7.08	16.5
YOLO v5m + keypoint	90.2	93.1	91.2	7.18	50.6
YOLO v5l + keypoint	90.8	93.4	91.9	7.47	114.6

Table 5
Comparison of different loss functions.

Model	Precision	Recall	mAP	d_{lmk}
YOLO v5s + keypoint + GIoU	90.7	93	91.1	7.08
YOLO v5s + keypoint + DIoU	91.3	93.1	91.6	7.27
YOLO v5s + keypoint + CIoU	91.7	93.5	91.7	7.48
YOLO v5s + keypoint + SIoU	91.5	93.8	92	7.21

help improve d_{lmk}. So d_{lmk} only increases by 0.13, which has little impact on 3D pose detection. The experimental results show that using SIoU loss in YOLO-lmk can better regress the bounding box. The bounding box loss function considers the vector angle between the required regressions, which can effectively improve the detection accuracy compared to only considering the Euclidean distance between the bounding box centers, the overlapping area and the aspect ratio.

It can be seen from Table 6 that if three Maxpool layers are used in SPP when the kernel is 5, 7, and 9, the highest mAP and the lowest d_{lmk} are obtained, which are 91.7% and 7.04, respectively. The mAP and d_{lmk} increased by 0.6% and decreased by 0.04, compared with the original kernel parameters. It proves that tomato is not a big target, and the original parameters of YOLO v5 are not suitable and need to be reduced. When the kernel is 3, 5 and 7, the effect is poor. It proves that tomato is not a small target, and the parameters of the kernel cannot be too small. In this study, a Maxpool layer is added on the basis of the three Maxpool layers. The kernel is 5, 7, 9, and 13. The mAP is 92%. The d_{lmk} is 7.68. Compared with the best effect among the three Maxpools, the mAP is improved by 0.3%. Although d_{lmk} is increased by 0.64, it has little effect on the real tomato 3D pose. It proves that adding a branch in spp can obtain more characteristic information. Therefore, the SPP in YOLO-lmk uses four maxpool layers, and the kernel is 5, 7, 9 and 13.

As shown in Table 7, the CA attention mechanism is added to different positions of YOLO v5s. The first is to add CA after the first three C3 modules of YOLO v5s, the second is to add CA after the first and second C3 modules, and the third is to add CA after the third C3 module. From the experimental results, the third method has the best effect. The mAP is 92%, which is increased by 0.9% compared with the original Yolo v5s. Therefore, YOLO-lmk adds CA attention mechanism after the third C3 module. From the first method, we can see that the CA attention mechanism is not that adding more is better. From the second and third methods, it can be seen that adding CA attention mechanism in appropriate positions can effectively enhance model performance.

As shown in Table 8, ablation experiments with data augmentation are performed. The experimental results show that when NO mosaic, No translation, No scale, and No color space conversion (HSV) are not used, the mAPs are 84.88%, 91%, 86.08%, and 90.5%, respectively. Compared with our method of using all data enhancements, No mosaic decreased the mAP the most, reaching 8.02%, and No translation decreased the mAP the least, reaching 1.9%. Mosaic and scale data enhancement is crucial to the improvement of model performance. It enables more detection targets in an image and enriches the image background. All four data augmentation methods can effectively enhance model performance. In the experiment, it is found that when no mosaic and no scale are used, the recall is low but the precision is high, indicating that many label bounding box have not been detected, but the correct rate of the detected bounding box is high.

Table 9 shows the performance comparison between the YOLO-lmk model and YOLO v5s. It can be seen from the experimental results

Table 6
Comparison of different kernels in SPP.

Model	Precision	Recall	mAP	d_{lmk}
YOLO v5s + keypoint + SPP(5,9,13)	90.7	93	91.1	7.08
YOLO v5s + keypoint + SPP(5,7,9)	89.7	94.1	91.7	7.04
YOLO v5s + keypoint + SPP(3,5,7)	91.5	92.6	90.9	7.23
YOLO v5s + keypoint + SPP(5,7,9,13)	91	93.8	92	7.68

Table 7
Comparison of different positions of CA attention mechanism.

Model	Precision	Recall	mAP	d_{lmk}
YOLO v5s + keypoint	90.7	93	91.1	7.08
YOLO v5s + keypoint + CA(1,2,3)	91.6	91.9	90.7	7.46
YOLO v5s + keypoint + CA(1,2)	90.3	92.9	91.3	7.50
YOLO v5s + keypoint + CA(3)	92.1	93.8	92	7.56

Table 8
Data augmented ablation experimental results.

Model	Precision	Recall	mAP
No mosaic	95.39	86.27	84.88
No translation	92.5	93	91
No scale	94.08	87.75	86.08
No color space conversion(HSV)	91.8	92.1	90.5
All data enhancements	89.4	95.2	92.9

Table 9
Comparison between YOLO-lmk and YOLO v5s.

Model	Precision	Recall	mAP	d_{lmk}	FLOPs/B
YOLO v5s + keypoint	90.7	93	91.1	7.08	16.5
YOLO-lmk	89.4	95.2	92.9	7.9	16.6

that the mAP of YOLO-lmk is increased by 1.8% compared to the original YOLO v5s, and the FLOPs is only increased by 0.1B. Although d_{lmk} increased by 0.82, it has little effect on the real tomato 3D pose.

Fig. 9 shows the change process of the image when the TPD algorithm detects the tomato 3D pose, (a) is the original image, (b) is the image after detection by the YOLO-lmk model, (c) is to segment the point cloud according to the bounding box mapping information, (d) is the image after voxel downsampling, (e) is the image after Euclidean color clustering, (f) is the point cloud required for RANSAC sphere fitting, (g) is the image of tomato to achieve 3D pose detection. It can be seen from (e) that the Euclidean color clustering is affected by lighting and other reasons. The clustered tomato point cloud still has a green calyx part, but in (f), it can be seen that RANSAC sphere fitting removes the calyx part of the point cloud.

Table 10 shows the experimental results of the TPD algorithm. The TPD algorithm is used to detect dataset 2. Among the 15 samples, 14 samples are correctly detected, with an accuracy of 93.4%. The YOLO-lmk model takes an average of 0.062 s to detect a color image, and the point cloud processing module takes an average of 0.028 s to detect a tomato. Fig. 10 shows the tomato 3D pose detection evaluation diagram. When the detected 3d pose is within the cylinder, the detection is correct.

Table 11 shows the comparison between the TPD algorithm and other research algorithms. It can be seen from the table that the TPD algorithm is the fastest, 18.9 times faster than the algorithm proposed by Yin et al. and 11.1 times faster than the algorithm proposed by Zhang et al. The method proposed by Yin et al uses a two-stage instance segmentation model, resulting in slower speed. The method proposed by Zhang et al cascades multiple networks and requires the prediction of 11 keypoints, resulting in slower speed. The method proposed by Guo et al requires a lot of point cloud data, and the calculation of coarse registration is large, so the speed is slow. As far as I know, in the field of agriculture, the TPD algorithm is currently the fastest 3D pose detection algorithm.

4. Discussion

The proposed method to complete keypoint and bounding box detection through one model is innovative and beneficial for tomato 3D pose detection. Compared with the method of cascading object detection

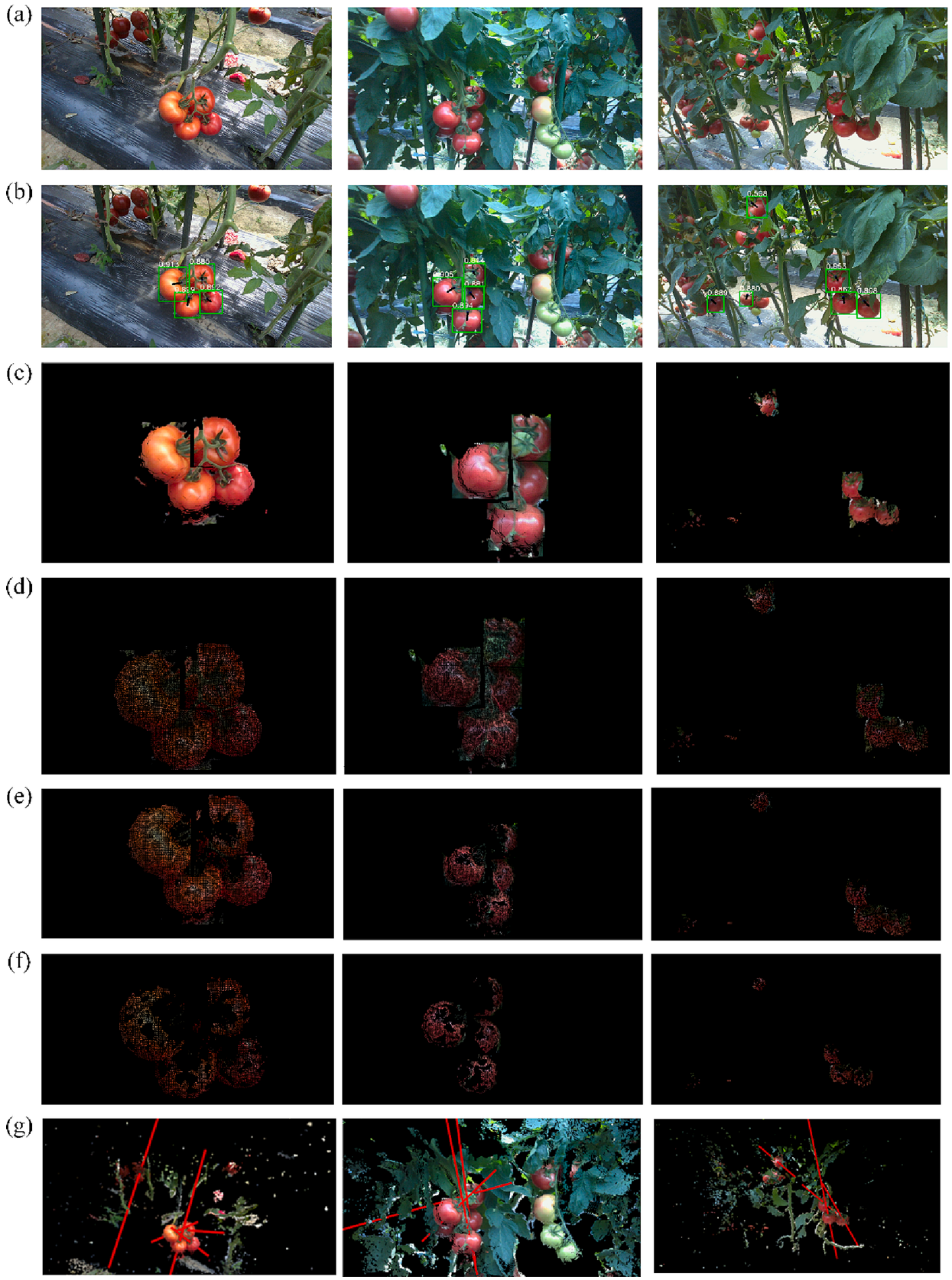


Fig. 9. The change process of the image when the TPD algorithm detects the tomato 3D pose. (a) The original image. (b) The image after detection by the YOLO-lmk model. (c) Segment the point cloud according to the bounding box mapping information. (d) The image after voxel downsampling. (e) The image after Euclidean color clustering. (f) The point cloud required for RANSAC sphere fitting. (g) The image of tomato to achieve 3D pose detection.

Table 10
The experimental results of the TPD algorithm.

Method	Number of samples	Number of samples that the tomato 3D pose is correctly detected	Accuracy	YOLO-lmk model speed	The time that the point cloud processing module takes to detect a tomato
TPD algorithm	15	14	93.4%	0.062 s/sheet	0.028 s

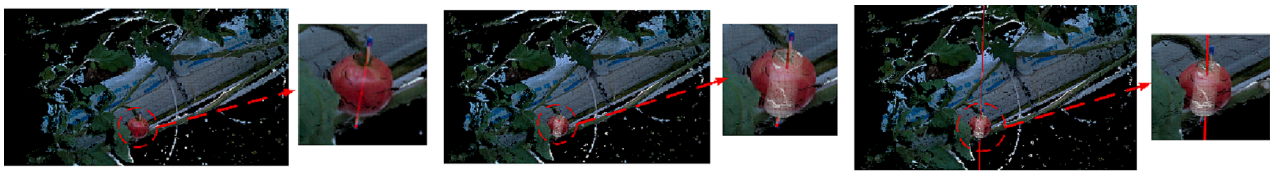


Fig. 10. The tomato 3D pose detection evaluation diagram. (a) Tomato real 3D pose. (b) The coaxiality method. (C) The tomato 3D pose detection image.

Table 11
Performance comparison between TPD algorithm and other research algorithms.

Author	Research object	Research objectives	Method	Time
Yin et al., 2021	Grape	3D pose	Mask R-CNN + RANSAC algorithm	1.7 s/string
Guo et al., 2020	Cucumber, Pineapple, Orange, Cabbage	3D pose	SA-IA + ICP	Cucumber (12.898 s), Pineapple (34.723 s), Orange (29.115 s), Cabbage (50.024 s)
Zhang et al., 2022	Tomato bunch	3D pose	TPM	A single-bunch scenario in 1.0 s a multi-bunch scenario in 2.0 s
	Tomato	3D pose	TPD	0.09 s

models and keypoint detection models, our model has a smaller weight space, is faster, and is easier to deploy into embedded devices. In the tomato keypoint detection, we only detect one keypoint, which is the center of the calyx. Compared with those algorithms that detect multiple keypoints, our algorithm is less affected by stem and leaf occlusion and is faster. Based on YOLO v5s, we added keypoint prediction in the detection layer, modified the original bounding box loss function GIoU loss to SIoU loss, added attention mechanism and modified SPP so that the bounding box mAP reached 92.9%. The keypoint detection was only 7.9 pixels error. Compared with other models of YOLO v5, YOLO v5s has the fastest speed and the least amount of calculation and has more advantages in practical engineering. Due to its self-learning and easy-to-train properties, our proposed TPD algorithm is easy to transplant to other crops, such as apples, pears, citrus, etc.The TPD algorithm only needs an RGB-D image to realize the 3D pose detection of the tomato, and the running speed is also the fastest we know so far. It can run in real-time in the equipment and complete the 3D pose detection task well, meeting the needs of tomato picking robots. Section 3 shows that we only use 895 training sets, and the detection accuracy of the tomato 3D pose reaches 93.4%. Compared with other keypoint datasets (such as MPII and MSCOCO datasets) containing tens of thousands of images, 895 images is a small training set. However, it can satisfy the training of YOLO-lmk. The method we proposed to evaluate the accuracy of tomato 3D pose detection by referring to the coaxiality concept in mechanical design is also innovative. This method is more intuitive and does not need to calculate the deviation between the real 3D pose and the detected 3D pose on each projection plane. There is no single fruit pose detection algorithm for clustered tomato in the existing research, and our research fills this field. Regarding the generality of the model, the model trained on this dataset can be used in other greenhouses, but the accuracy will be reduced to some extent. The model can only be used in a greenhouse with tomato varieties and environments similar to the training set. In order to enhance the generality of the model, in future research, we will build tomato datasets of different varieties in different greenhouses. In our research, the point cloud information obtained by Realsese L515 is not complete, and some point clouds are missing, but Realsese L515 is already one of the best RGB-D cameras. In future research, we will consider using the triangulation depth method or consider using multi-line lidar and RGB camera to make RGB-D camera. The Robotic arm picking tomatoes needs to obtain the position of obstacles. In future research, we will combine semantic segmentation in the algorithm to realize the three-dimensional position detection of stems.

5. Conclusions

In this research, a new TPD algorithm is proposed to detect the 3D pose of tomatoes. The TPD method is mainly composed of YOLO-lmk model and point cloud processing module. The YOLO-lmk model is based on the YOLO v5s selected from the YOLO v5 series models, which adds keypoint prediction in the detection layer, modifies the original bounding box loss function GIoU loss to SIoU loss, adds attention mechanism and modifies SPP. The bounding box and keypoint detection tasks are done in one model. The YOLO-lmk model bounding box mAP is 92.9 %, d_{lmk} is 7.9, FLOPs is 16.6B, and the speed is 0.062 s/sheet. Compared with the original model, mAP is increased 1.8%, and FLOPs is increased only 0.1B. The experimental results show that the YOLO-lmk model can complete the bounding box and keypoint detection tasks well. The model has high accuracy and a small amount of calculation. Our proposed point cloud processing module consists of point cloud segmentation, voxel downsampling, Euclidean color clustering, RANSAC sphere fitting and keypoint index. The point cloud processing module only takes 0.028 s to complete a tomato 3D pose detection, which is fast. The accuracy of the TPD algorithm in detecting tomatoes is 93.4%, and the speed is 0.09 s for detecting one tomato. It is the fastest algorithm we currently know for 3D pose detection. The TPD algorithm can provide a theoretical basis for tomato 3D pose detection, and provide a reference for other fruits (pears, citrus, apples) 3D pose detection.

CRediT authorship contribution statement

Xiaoqiang Du: Conceptualization, Methodology, Writing – review & editing. **Zhichao Meng:** Visualization, Software, Writing – original draft. **Zenghong Ma:** Software, Investigation. **Wenwu Lu:**

Investigation. **Hongchao Cheng:** Investigation.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFD2202103), the National Natural Science Foundation of China (Grant No. 31971798), the Zhejiang Provincial Key Research & Development Program (Grant No. 2023C02049), the SNJF Science and Technology Collaborative Program of Zhejiang Province (Grant No. 2022SNJF017), the 521 Talent Plan of Zhejiang Sci-Tech University, and the Cultivation Project for Youth Discipline Leader in Zhejiang Provincial Institute.

References

Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299.

Du, A., Guo, H., Lu, J., Su, Y., Ma, Q., Ruchay, A., Pezzuolo, A., 2022. Automatic livestock body measurement based on keypoint detection with multiple depth cameras. Comput. Electron. Agric. 198, 107059 https://doi.org/10.1016/j.compag.2022.107059.

Fan, Y., Zhang, S., Feng, K., Qian, K., Wang, Y., Qin, S., 2022. Strawberry maturity recognition algorithm combining dark channel enhancement and YOLOv5. Sensors-Basel. 22 (2), 419. https://doi.org/10.3390/s22020419.

Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J., 2018. Wing loss for robust facial landmark localisation with convolutional neural networks. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245. https://doi.org/10.1109/CVPR.2018.00238.

Fujinaga, T., Yasukawa, S., Ishii, K., 2021. Evaluation of tomato fruit harvestability for robotic harvesting. In: In: 2021 IEEE/SICE International Symposium on System Integration, pp. 35–39. https://doi.org/10.1109/IEEECONF49454.2021.9382603.

Gao, M., Du, Y., Yang, Y., Zhang, J., 2019. Adaptive anchor box mechanism to improve the accuracy in the object detection system. Multimed. Tools. Appl. 78 (19), 27383–27402. https://doi.org/10.1007/s11042-019-07858-w.

Gevorgyan, Z. 2022. SIoU Loss: More Powerful Learning for Bounding Box Regression. arXiv preprint arXiv:2205.12740. https://doi.org/10.48550/arXiv.2205.12740.

Guo, N., Zhang, B., Zhou, J., Zhan, K., Lai, S., 2020. Pose estimation and adaptable grasp configuration with point cloud registration and geometry understanding for fruit grasp planning. Comput. Electron. Agric. 179, 105818 https://doi.org/10.1016/j.compag.2020.105818.

He, L., Wu, F., Du, X., Zhang, G., 2022. Cascade-SORT: A robust fruit counting approach using multiple features cascade matching. Comput. Electron. Agric. 200, 107223 https://doi.org/10.1016/j.compag.2022.107223.

Jintasuttisak, T., Edirisinghe, E., Elbattay, A., 2022. Deep neural network based date palm tree detection in drone imagery. Comput. Electron. Agric. 192, 106560 https://doi.org/10.1016/j.compag.2021.106560.

Jun, J., Kim, J., Seol, J., Kim, J., Son, H.I., 2021. Towards an efficient tomato harvesting robot: 3D perception, manipulation, and end-effector. IEEE Access. 9, 17631–17640. https://doi.org/10.1109/ACCESS.2021.3052240.

Li, Y., He, L., Jia, J., Lv, J., Chen, J., Qiao, X., Wu, C., 2021b. In-field tea shoot detection and 3D localization using an RGB-D camera. Comput. Electron. Agric. 185, 106149 https://doi.org/10.1016/j.compag.2021.106149.

Li, H., Li, C., Li, G., Chen, L., 2021a. A real-time table grape detection method based on improved YOLOv4-tiny network in complex background. Biosyst. Eng. 212, 347–359. https://doi.org/10.1016/j.biosystemseng.2021.11.011.

Liang, Z., Wang, X., Huang, R., Lin, L., 2014. An expressive deep model for human action parsing from a single image. In: In: 2014 IEEE International Conference on Multimedia and Expo, pp. 1–6. https://doi.org/10.1109/ICME.2014.6890158.

Lili, W., Bo, Z., Jinwei, F., Xiaoan, H., Shu, W., Yashuo, L., Chongfeng, W., 2017. Development of a tomato harvesting robot used in greenhouse. Int. J. Agr. Biol. Eng. 10 (4), 140–149. https://doi.org/10.25165/j.ijabe.20171004.3204.

Lin, G., Tang, Y., Zou, X., Wang, C., 2021. Three-dimensional reconstruction of guava fruits and branches using instance segmentation and geometry analysis. Comput. Electron. Agric. 184, 106107 https://doi.org/10.1016/j.compag.2021.106107.

Liu, J., Yang, J., Qin, L., Li, C., Liang, Y., 2015. Development and diversity of calyx morphology from bud stage to fruit maturity in tomato. J. Plant Genetic Res. 16 (02), 300–306. https://doi.org/10.13430/j.cnki.jpgr.2015.02.014.

Liu, C., Zhu, H., Guo, W., Han, X., Chen, C., Wu, H., 2021. EFDet: an efficient detection method for cucumber disease under natural complex environments. Comput. Electron. Agric. 189, 106378 https://doi.org/10.1016/j.compag.2021.106378.

Luo, L., Yin, W., Ning, Z., Wang, J., Wei, H., Chen, W., Lu, Q., 2022. In-field pose estimation of grape clusters with combined point cloud segmentation and geometric analysis. Comput. Electron. Agric. 200, 107197 https://doi.org/10.1016/j.compag.2022.107197.

Miao, T., Zhu, C., Xu, T., Yang, T., Li, N., Zhou, Y., Deng, H., 2021. Automatic stem-leaf segmentation of maize shoots using three-dimensional point cloud. Comput. Electron. Agric. 187, 106310 https://doi.org/10.1016/j.compag.2021.106310.

Nasirahmadi, A., Ashtiani, S.H.M., 2017. Bag-of-Feature model for sweet and bitter almond classification. Biosyst. Eng. 156, 51–60. https://doi.org/10.1016/j.biosystemseng.2017.01.008.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.

She, J., Zhan, W., Hong, S., Min, C., Dong, T., Huang, H., He, Z., 2022. A method for automatic real-time detection and counting of fruit fly pests in orchards by trap bottles via convolutional neural network with attention mechanism added. Ecol. Inform., 101690 https://doi.org/10.1016/j.ecoinf.2022.101690.

Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., Marinello, F., 2022. Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms. Agronomy-Basel. 12 (2), 319. https://doi.org/10.3390/agronomy12020319.

Tang, Y., Chen, M., Wang, C., Luo, L., Li, J., Lian, G., Zou, X., 2020. Recognition and localization methods for vision-based fruit picking robots: a review. Front. Plant. Sci. 11, 510. https://doi.org/10.3389/fpls.2020.00510.

Tang, Y., Zhou, H., Wang, H., Zhang, Y., 2023. Fruit detection and positioning technology for a Camellia oleifera C. Abel orchard based on improved YOLOv4-tiny model and binocular stereo vision. Expert Syst. Appl. 211, 118573 https://doi.org/10.1016/j.eswa.2022.118573.

Wang, Z., Jin, L., Wang, S., Xu, H., 2022b. Apple stem/calyx real-time recognition using YOLO-v5 algorithm for fruit automatic loading system. Postharvest. Biol. Tec. 185, 111808 https://doi.org/10.1016/j.postharvbio.2021.111808.

Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H., 2020. CSPNet: a new backbone that can enhance learning capability of CNN. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391.

Wang, Y., Yan, G., Meng, Q., Yao, T., Han, J., Zhang, B., 2022a. DSE-YOLO: detail semantics enhancement YOLO for multi-stage strawberry detection. Comput. Electron. Agric. 198, 107057 https://doi.org/10.1016/j.compag.2022.107057.

Wu, F., Duan, J., Ai, P., Chen, Z., Yang, Z., Zou, X., 2022. Rachis detection and three-dimensional localization of cut off point for vision-based banana robot. Comput. Electron. Agric. 198, 107079 https://doi.org/10.1016/j.compag.2022.107079.

Wu, Q., Xu, G., Zhang, S., Li, Y., Wei, F., 2020. Human 3D pose estimation in a lying position by RGB-D images for medical diagnosis and rehabilitation. In: in 2020 42nd Annual international Conference of the IEEE Engineering in Medicine & Biology Society, pp. 5802–5805. https://doi.org/10.1109/EMBC44109.2020.9176407.

Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X., 2021. A real-time detection algorithm for Kiwifruit defects based on YOLOv5. Electronics-Switz. 10 (14), 1711. https://doi.org/10.3390/electronics10141711.

Yin, W., Wen, H., Ning, Z., Ye, J., Dong, Z., Luo, L., 2021. Fruit detection and pose estimation for grape cluster–harvesting robot using binocular imagery based on deep neural networks. Front. Robot. Ai. 163 https://doi.org/10.3389/frobt.2021.626989.

Zhang, F., Gao, J., Zhou, H., Zhang, J., Zou, K., Yuan, T., 2022. Three-dimensional pose detection method based on keypoints detection network for tomato bunch. Comput. Electron. Agric. 195, 106824 https://doi.org/10.1016/j.compag.2022.106824.

Zhao, J., Zhang, X., Yan, J., Qiu, X., Yao, X., Tian, Y., Cao, W., 2021. A wheat spike detection method in UAV images based on improved YOLOv5. Remote Sens-Basel. 13 (16), 3095. https://doi.org/10.3390/rs13163095.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2020. Distance-IoU loss: faster and better learning for bounding box regression. In: In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12993–13000. https://doi.org/10.1609/aaai.v34i07.6999.