TYPE Original Research
PUBLISHED 29 July 2022
DOI 10.3389/fpls.2022.972445

Detection and localization of citrus fruit based on improved You Only Look Once v5s and binocular vision in the orchard

Chaojun Hou[1†], Xiaodi Zhang[1†], Yu Tang[2*], Jiajun Zhuang[1], Zhiping Tan[2], Huasheng Huang[2], Weilin Chen[3], Sheng Wei[4], Yong He[5] and Shaoming Luo[3]

[1]Academy of Contemporary Agriculture Engineering Innovations, Zhongkai University of Agriculture and Engineering, Guangzhou, China, [2]Academy of Interdisciplinary Studies, Guangdong Polytechnic Normal University, Guangzhou, China, [3]School of Mechatronics Engineering and Automation, Foshan University, Foshan, China, [4]Engineering Research Center for Intelligent Robotics, Jihua Laboratory, Foshan, China, [5]College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou, China

Intelligent detection and localization of mature citrus fruits is a critical challenge in developing an automatic harvesting robot. Variable illumination conditions and different occlusion states are some of the essential issues that must be addressed for the accurate detection and localization of citrus in the orchard environment. In this paper, a novel method for the detection and localization of mature citrus using improved You Only Look Once (YOLO) v5s with binocular vision is proposed. First, a new loss function (polarity binary cross-entropy with logit loss) for YOLO v5s is designed to calculate the loss value of class probability and objectness score, so that a large penalty for false and missing detection is applied during the training process. Second, to recover the missing depth information caused by randomly overlapping background participants, Cr-Cb chromatic mapping, the Otsu thresholding algorithm, and morphological processing are successively used to extract the complete shape of the citrus, and the kriging method is applied to obtain the best linear unbiased estimator for the missing depth value. Finally, the citrus spatial position and posture information are obtained according to the camera imaging model and the geometric features of the citrus. The experimental results show that the recall rates of citrus detection under non-uniform illumination conditions, weak illumination, and well illumination are 99.55%, 98.47%, and 98.48%, respectively, approximately 2–9% higher than those of the original YOLO v5s network. The average error of the distance between the citrus fruit and the camera is 3.98 mm, and the average errors of the citrus diameters in the 3D direction are less than 2.75 mm. The average detection time per frame is 78.96 ms. The results indicate that our method can detect and localize citrus fruits in the complex environment of orchards with high accuracy and speed. Our dataset and codes are available at https://github.com/AshesBen/citrus-detection-localization.

KEYWORDS

citrus detection, citrus localization, binocular vision, YOLO v5s, loss function

OPEN ACCESS

EDITED BY
Gregorio Egea,
University of Seville, Spain

REVIEWED BY
Thomas Burks,
University of Florida, United States
Jun Lu,
Huazhong Agricultural University,
China

*CORRESPONDENCE
Yu Tang
yutang@gpnu.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

RECEIVED 18 June 2022
ACCEPTED 12 July 2022
PUBLISHED 29 July 2022

CITATION
Hou C, Zhang X, Tang Y, Zhuang J, Tan Z, Huang H, Chen W, Wei S, He Y and Luo S (2022) Detection and localization of citrus fruit based on improved You Only Look Once v5s and binocular vision in the orchard. *Front. Plant Sci.* 13:972445. doi: 10.3389/fpls.2022.972445

Introduction

Citrus plays an essential role in the fruit industry around the world, with an annual production of approximately 140 million tons (Zheng et al., 2021; Noorizadeh et al., 2022). As the cost of fruit harvesting increases and the availability of skilled labor decreases in China, the traditional manual harvesting method is no longer practical (Gongal et al., 2015; Tang et al., 2021). Presently, fruit harvesting has become increasingly automated for labor-saving and large-scale agriculture (Onishi et al., 2019). The development of an automated citrus picking robot is an inevitable trend for fruit harvesting (Zhuang et al., 2018). In recent work, the development of automatic fruit picking with a robot involves two main tasks: (1) fruit detection and (2) fruit localization via computer vision. The accuracy of fruit detection and fruit localization directly determines the picking efficiency of the robot.

Fruit detection using computer vision has been investigated in numerous recent studies, and most have applied deep learning methods to achieve good performance and robustness (Yang et al., 2020; Chen et al., 2021; Yan et al., 2021). Wan and Goudos (2020) integrated multiclass classification into Faster R-CNN to detect oranges, apples, and mangoes. The improved model achieved a 90.72% mAP. Kang and Chen (2020) proposed a LedNet network with a feature pyramid network and an atrial space pyramid pool for mature apple detection; the recall rate and precision were 0.821 and 0.853, respectively. Chu et al. (2021) improved mask R-CNN by adopting a suppression branch to suppress the generation of nonapple fruit features. However, their method has poor detection performance under backlight conditions. He et al. (2020) developed a deep bounding box regression forest to describe the characteristics of immature citrus on three levels, which is beneficial for differentiating an object from the background. However, the detection speed is slow (0.759 s per frame), making it challenging to apply in real-time applications. For the real-time application of fruit harvesting, the detection speed should be at least 10–15 frames per second (Tu et al., 2020). YOLO series models have been used in various applications for fast detection speed with high accuracy (Jiang et al., 2020; Wang et al., 2021). Xiong et al. (2020) used a YOLO v2 model to detect green mango and reported a recall of 89.0%, a precision of 96.1%, and an average detection time of 0.08 s per frame. Liang et al. (2020) combined YOLO v3 and U-Net to detect litchi fruits and litchi stems at night for picking robots under different illuminations; 96.1% precision and 89.0% recall were achieved. However, the method has not yet been assessed in the daytime. Wang and He (2021) developed an improved YOLO v5 model to detect apple fruitlets using the channel pruning method. However, the network architecture must be manually adjusted during detection. Notably, the target-background class imbalance is typically the main obstacle encountered in training convolutional neural networks (Buda et al., 2018). To address such class imbalance, Lin et al. (2020) designed a focal loss function to make the network pay more attention to hard samples in training, but the approach cannot push the object further from the background. Rahman et al. (2020) proposed polarity loss to improve focal loss. In the above studies, various deep learning methods have been proposed to detect fruit targets and have achieved good results. However, the detection performance deteriorates in unstructured growing environments with variable illumination conditions. For better accuracy, the disparity between citrus and background under variable illumination conditions and different occlusion states should be incorporated into the network structure.

The purpose of fruit localization is to determine the spatial coordinates of the detected fruit and its location information, such as posture and shape (Huang et al., 2019). Many fruit localization methods require a binocular stereo vision system. The depth map or point cloud image is captured to obtain three-dimensional (3D) localization of fruit. Yang et al. (2020) employed a mask R-CNN to detect citrus objects and branches and matched the color and depth maps to locate fruits and branches. The average error in the diameter of the fruit and the branch was less than 4 mm. However, the distance from the fruit to the camera was not provided in their work. Nguyen et al. (2016) used a Euclidean clustering algorithm to segment a single apple using a point cloud image. The results showed that the errors in the spatial coordinates and the diameter of the fruit were slightly less than 10 mm, but the 3D location information about apples was not the aim of their work. Xu et al. (2018) proposed the PointFusion structure to estimate the 3D object bounding box and its confidence from RGB image and point cloud information. The approach produces good results in the KITTI and SUN-RGBD datasets, with 78% AP. Since the information of the depth map or point cloud is incomplete, fruit localization often requires the use of empirical knowledge (Liu et al., 2017). Wang et al. (2017) adopted Otsu’s method and a one-dimensional filter to remove occluded objects (leaves, branches, fruit particles, etc.) and employed ellipse fitting to extract a well-separated mango region. Finally, mango dimensions were calculated using depth information. Ge et al. (2020) developed a shape completion method to reconstruct the point clouds of strawberries; the average error of the center point of strawberries was 5.7 mm. However, the reconstructed error is larger in the case of the neighboring overlapping fruits. Note that an incomplete depth map makes it difficult to recover the missing depth value lost by variable illumination or the fruit region being occluded by randomly overlapping participants, such as neighboring fruits and other background objects. Therefore, this paper aims to restore the depth map with high accuracy for locating fruits in unstructured orchard environments.

The objective of this work is to develop a novel method for the detection and localization of mature citrus fruits in natural orchards using a binocular camera. The pipelines of the study are to (1) design a new loss function to enhance the detection performance of the YOLO v5s network architecture under variable illumination conditions, (2) extract the fruit region in the RGB image and recover the missing value in the depth map under different occlusion states of citrus fruit, and (3) estimate the 3D localization of citrus fruits using the camera imaging model and the geometric features of citrus fruits. Our method can provide 3D localization information of citrus fruits, such as the diameters of citrus fruits in the 3D direction, the spatial coordinates of citrus fruits, the distance between citrus fruits and the camera, and the 3D bounding box of citrus fruits.

Materials and methods

Datasets

A variety of citrus named "Shantanju" was investigated in the hillside orchard of the Guangzhou Conghua Hualong Fruit and Vegetable Freshness Co. Ltd., located in Guangzhou, China (113°39'2.38'E, 23°33'12.48'N). A total of 4855 groups of images were captured in December 2020 and December 2021 before harvest. Image acquisition was performed using a binocular camera (Model ZED 2, Stereolab's Co. Ltd, USA) with a 1920 × 1080 pixel resolution under sunny and cloudy conditions. The distance between the camera and citrus was set to approximately 30∼150 cm. Each group of images contains a left view (RGB image) and a depth map (grayscale image). Note that the right view images were also captured and used only to generate the depth map with the left view images. The depth map is provided with a Z value for every pixel (X, Y) in the left view image. According to the illumination of the citrus surface, images are divided into three groups: non-uniform illumination (non), weak illumination (weak), and well illumination (well). In total, 2913 images were randomly selected as the training dataset (train), 971 images were selected as the validation dataset (validation), and 971 images were selected as the test dataset (test), the number of citrus samples in each group is shown in **Table 1**.

The hand of the harvesting robot is designed to pick citrus fruits that are in the correct position in front of the camera. In each left view image, the citrus fruits located near the center of the image were manually labeled with bounding boxes using Labelme software. **Figures 1A–C** provides examples of citrus images from each illumination group. The bounding boxes of labeled citrus are annotated with red rectangles. The corresponding depth maps with labeled citrus are shown in **Figures 1D–F**, where the grayscale of color is based on distance from the camera, i.e., closer objects are darker; further objects are lighter.

Detection and localization of citrus

An overview of our proposed method for citrus detection and localization is presented in **Figure 2**. The main procedure involves the following steps: Firstly, an improved YOLO v5s is developed to detect citrus in the 2D bounding box. Secondly, Cr-Cb chromatic mapping, Otsu threshold algorithm, and morphology processing are used to extract citrus shape. The missing depth values are recovered by the kriging method. Finally, the 3D localization of citrus fruit is realized by geometric imaging model. Each procedure is described in detail in the following subsections.

Detection of the 2D bounding box of citrus fruit

YOLO (You Only Look Once) is a one-stage detection network that converts object detection into a regression problem using convolutional neural networks (Wang et al., 2021, 2022). YOLO v5, the latest version of the YOLO model (Jocher and Stoken, 2021), has a faster detection speed and higher accuracy than the previous version. The release of YOLO v5 consists of four different model sizes: YOLO v5s (smallest), YOLO v5m, YOLOv5l, and YOLO v5x (largest). The network structures of these four models are basically the same, but the numbers of modules and convolution kernels are different. Considering that the application scenario of this paper requires fast detection efficiency, the YOLO v5s model is selected as the basic network, and its structure is shown in **Figure 3A**. The YOLO v5s network is divided into three parts. The first part is the backbone network, which is responsible for the feature extraction of the target. The second part is PANet, which generates feature pyramids for object scaling. The third part is the head network, which conducts the final detection.

In YOLO v5s, binary cross-entropy with a logit loss function ($Loss_B$) is used to calculate the class probability and objectness score for each sample, as follows:

$$Loss_B\left(x_i, y_i\right) = -y_i \log\left(\sigma\left(x_i\right)\right) - \left(1 - y_i\right) \log\left(1 - \sigma\left(x_i\right)\right), \tag{1}$$

where i is the sample index, x_i is the predicted likelihood, y_i stands for the ground truth, and $\sigma\left(\cdot\right)$ is the sigmoid function that maps the prediction x_i to the probability for the ground truth. In object detection tasks, the problem of unbalanced training sets is considerable (Lin et al., 2020), i.e., the background information in the dataset used for training is overrepresented compared to that of the target class. The sum of $Loss_B$ from the easy samples over the entire images can overwhelm the overall $Loss_B$ from the hard samples. Moreover, the training is inefficient, as most

TABLE 1 Dataset distribution.

	Non		Weak		Well		Total	
	Images	Samples	Images	Samples	Images	Samples	Images	Samples
Train	923	4569	814	2503	1176	4016	2913	11088
Validation	333	1636	255	809	383	1435	971	3880
Test	307	892	269	655	395	1052	971	2599

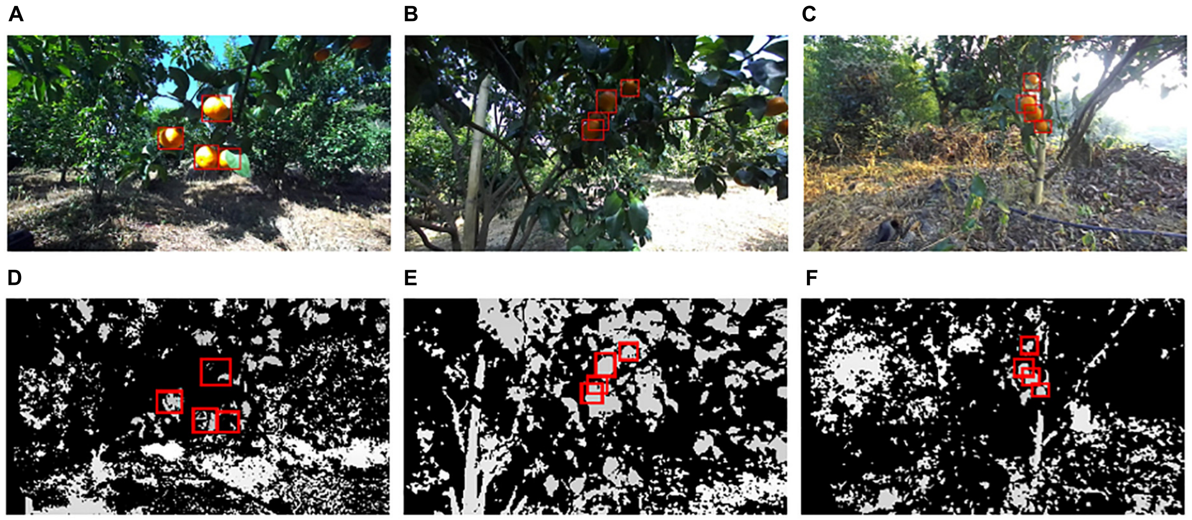

FIGURE 1
Examples of citrus images captured in three illumination conditions: **(A)** non, **(B)** weak, **(C)** well, **(D)** depth map of **(A)**, **(E)** depth map of **(B)**, and **(F)** depth map of **(C)**.

locations are easy samples that do not contribute to learning. Furthermore, in our trial-and-error experiments, the hard negative samples, i.e., the citrus misclassified as background, are difficult to distinguish from the background under weak illumination or obvious occlusion. On the other hand, the hard positive samples, i.e., the background misclassified as a citrus target, exhibit similar characteristics to mature citrus due to the uncontrolled factors in the orchard environment.

To better differentiate citrus from the background under variable illumination conditions and different occlusion states, we design a new loss function, the polarity binary cross-entropy with logit loss ($Loss_{PB}$), to calculate the class probability and objectness score to penalize the hard samples. In particular, a penalty function f_p (Rahman et al., 2020) is developed to represent the disparity between the prediction for citrus and background. $Loss_{PB}$ is defined as follows:

$$\begin{cases} Loss_{PB}\left(x_i, y_i\right) = f_p\left(\sigma\left(x_i\right)\right) Loss_B\left(x_i, y_i\right) \\ f_p\left(z_i\right) = \frac{2}{1+\exp(-\gamma(\overline{z}_i - z_i))} \end{cases} \tag{2}$$

where z_i is the probability of sample i being predicted as the true class, such as citrus target or background, $\overline{z}_i = 1 - z_i$ is the probability of sample i being misclassified as the incorrect class, and γ is a slope parameter of the sigmoid function f_p (Figure 3B). f_p is used to calculate the disparity between the prediction for the true class and false class based on the value of $\overline{z}_i - z_i$. If the citrus target is misclassified as background, the prediction probability $\overline{z}_i$ is greater than, such that a large value of $\overline{z}_i - z_i$ is obtained, and a large penalty will be assigned by f_p. In this case, the penalty value of $Loss_{PB}$ is larger than that of $Loss_B$, which helps to suppress the missed detection of citrus. Similarly, if the background is misclassified as citrus, a large penalty will be assigned by f_p due to the large value of $\overline{z}_i - z_i$, which will improve the false detection of citrus. On the other hand, if a citrus target or the background is predicted with a more reliable probability of z_i, the penalty value applied by f_p will be closer to 0 due to the small value of $\overline{z}_i - z_i$. In such a case, the penalty value of $Loss_{PB}$ is smaller than that of $Loss_B$ and is pushed toward zero. In general, a large penalty is applied to missed detection and false detection of citrus targets. Thus, f_p enforces a large margin to push predictions z_i and $\overline{z}_i$ further apart.

Recall rate (R), precision (P), and F_β-score (F_β) are selected to evaluate the performance of the improved YOLO v5s in the test dataset:

$$R = \frac{TP}{TP + FN}, \tag{3}$$

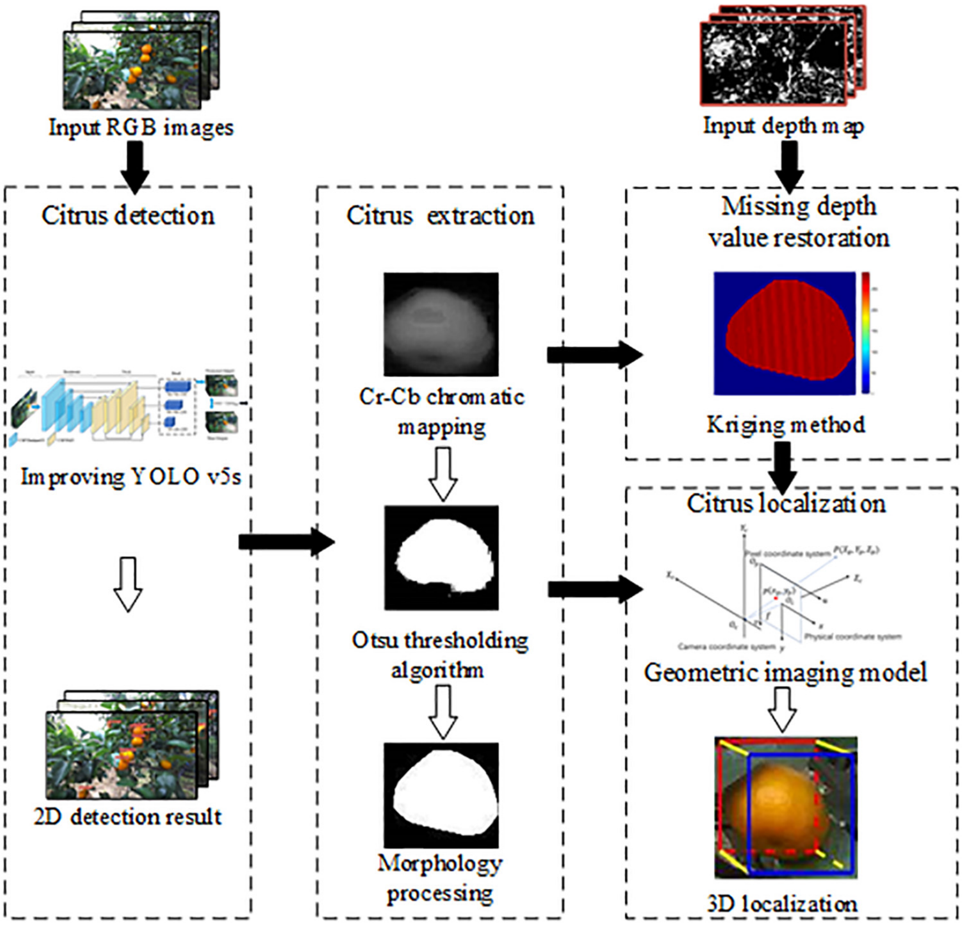

FIGURE 2
Flow diagram of our proposed method.

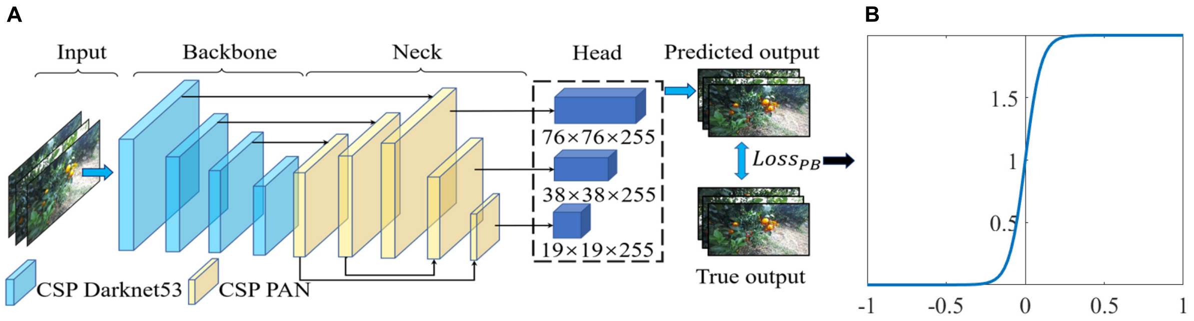

FIGURE 3
Citrus detection model based on You Only Look Once (YOLO) v5s: **(A)** network structure of improved YOLO v5s and **(B)** function graph of penalty function f_p.

$$P = \frac{TP}{TP + FP}, \tag{4}$$

$$F_\beta = (1 + \beta^2)\frac{P \times R}{\beta^2 \times P + R}, \tag{5}$$

where *FN* is the number of false negatives for the false detection of citrus samples, *FP* is the number of false positives for the missed detection of citrus samples, and *TP* is the number of true positives for the detected citrus samples. F_β uses a positive real number β to weigh the importance between *R* and *P*. In

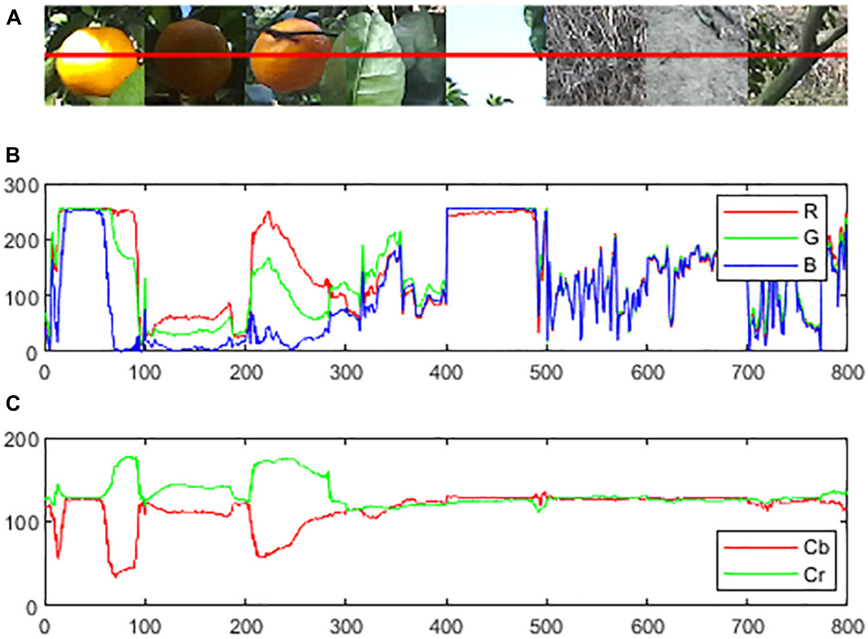

FIGURE 4
Examples of color curves of the citrus and background in different color spaces: **(A)** original RGB image, **(B)** color intensity on the line on R, G, and B elements in RGB color space, and **(C)** color intensity on the line on Cb and Cr elements in YCrCb color space.

this paper, β is set to 1 as F_1 by regarding R and P are equally necessary for our experiment.

Extraction of the citrus fruit region from the 2D bounding box

Image data captured in a natural orchard always contain multiple participants, e.g., grass, soil, lawn, leaves, branches, trunks, and sky. The citrus fruit region is difficult to extract exactly from the 2D bounding box predicted from the improved YOLO v5s. Fortunately, these participants have different color characteristics, so the different targets can be extracted based on their color information. Here, the proper color space is beneficial to robustly extract the citrus fruit region from the background. Zhuang et al. (2018) and Zhuang et al. (2019) adopted improved R-G chromatic mapping to extract fruit regions. In this paper, the input images are converted into the YCbCr color space for better contrast enhancement between the citrus fruit region and background.

As shown in **Figure 4A**, a horizontal red line was drawn across citrus fruits and the background. The color intensities of the pixels of the line are represented with the R curve (the red element of RGB), the G curve (the green element), and the B curve (the blue element) in **Figure 4B**. The Cr curve (the Cr element of YCbCr) and the Cb curve (the Cb element) are represented in **Figure 4C**. The intensity difference between the R curve and G curve is small in both the citrus region and background, and there are no obvious rules exhibited in the B curve among the citrus fruit regions and backgrounds. However, the intensity difference between the Cr curve and Cb curve values within the citrus region is obviously greater than that of the background. Thus, Cr-Cb chromatic mapping is suitable to enhance the disparity between the citrus region and the background participants.

The Otsu thresholding algorithm is an appropriate method to segment the potential citrus regions from the background, where the best threshold value is selected by maximizing the variance between foreground and background. As shown in **Figure 5**, the Cr-Cb chromatic mapping has prominent bimodal characteristics in the intensity histogram under variable illumination, where the citrus fruit region contributes to the high value and background contributes to the lower value. Therefore, the best threshold value from Otsu is suitable to segment the citrus fruit region from the background.

The fruit region segmented by Otsu thresholding will not be complete in terms of shape due to the irregular growth situations of citrus fruit that are occluded by adjacent fruits, branches and leaves. To address this problem, the mathematical morphology operations of erosion, dilation, and hole filling are subsequently adopted to fill the gaps between detected regions, remove noise, fill small holes, and smooth the region's boundary. Then, the mathematical morphology operation of convex hull is used to estimate the occluded regions of the fruit from the partially compact region. In this way, the citrus fruit can be almost completely segmented from its corresponding 2D bounding box.

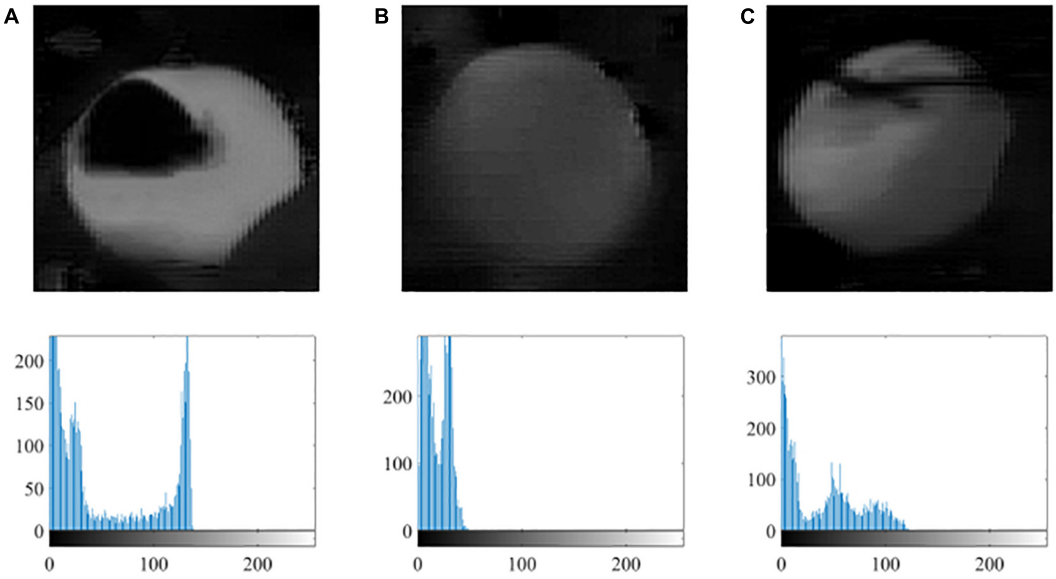

FIGURE 5
Examples of the citrus image after Cr-Cb chromatic mapping and its gray histogram under variable illumination: **(A)** non, **(B)** weak, and **(C)** well.

Recovery of missing depth values

To achieve the 3D localization result of citrus, it is essential to obtain a complete depth map of the whole citrus fruit region; however, an incomplete depth map is always obtained for two main reasons. First, the depth map is sparse in the case of binocular stereo conditions. The depth value is missing and set to zero for pixels where no depth information is sensed by the ZED camera, which may be caused by variable illumination, camera performance limitation, and shooting angle (Liu et al., 2017). Second, the depth values can be missing due to the occluded region estimated from the morphological processing. To restore the complete depth map of the citrus region, the kriging method is adopted to predict the missing depth value by adding the weight of the observed depth value.

Let I_O be the citrus region segmented by Otsu thresholding and I_C be the citrus fruit region extracted via the convex hull operation. We denote by I_{in} the set of pixels whose depth value is missing in I_C, such that the depth value is zero or the pixel is located outside of I_O. Let I_V be the set of pixels whose observed depth value is available in I_O. Therefore, the missing depth value in I_{in} can be obtained as follows:

$$\hat{Z}(s) = \sum_{p\in I_V} \lambda_p(s)\, Z(p), \quad \forall s \in \mathrm{I}_{in}, \tag{6}$$

where $Z(p)$ is the observed depth value at pixel p and $\lambda_p(s)$ is the weight of $Z(p)$, which depends not only on the distance between the depth values but also on the position and overall spatial arrangement of the observed depth value around pixel s. Note that the kriging method is the best linear unbiased estimator to restore the missing depth value using observed depth values from the incomplete depth map. Therefore, all the missing depth values in I_C will be restored completely.

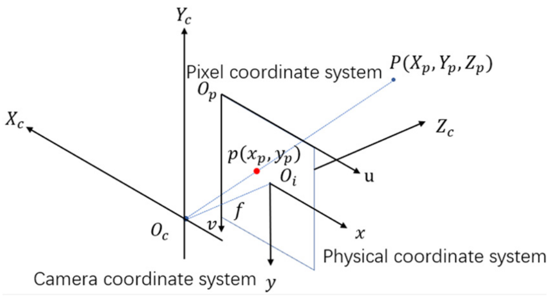

FIGURE 6
Coordinate system transformation diagram.

3D localization of citrus fruit

The 3D localization of citrus determines the spatial position and posture information, such as citrus diameter in the 3D direction d_x, d_y, and d_z, the spatial coordinates of citrus $Q_0(X_q,\ Y_q,\ Z_q)$, the distance between the citrus and camera d, the spatial coordinates of the citrus 3D bounding box $P_1,\ P_2,\ \ldots,\ P_8$, and the corresponding 2D coordinates of the 3D bounding box in the image plane $p_1,\ p_2,\ \ldots,\ p_8$. The 3D coordinates of a point in the real world must be precisely mapped to the 2D coordinates of a pixel in the imaging plane. Here, the transformation relation among the camera coordinate system O_c, the physical coordinate system O_i, and the pixel coordinate system O_p should be analyzed.

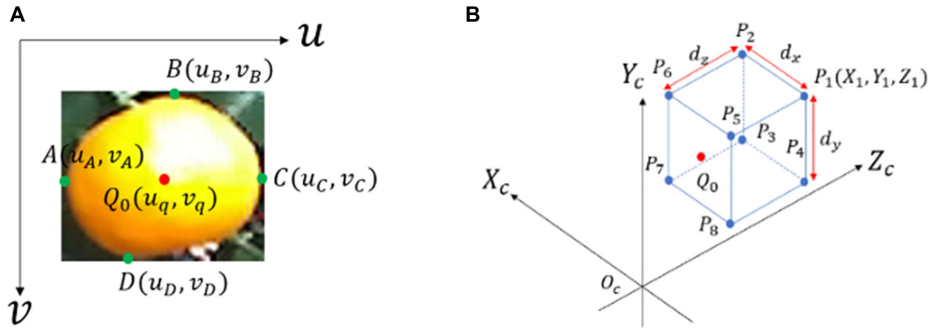

FIGURE 7
Example of 2D and 3D location information of citrus fruit: **(A)** 2D information of citrus fruit with four endpoints (green points) and center points Q_0 (red points) in O_p. **(B)** Citrus 3D bounding box in O_c with eight vertices.

As illustrated in Figure 6, a physical coordinate system O_i is depicted with the origin in the imaging plane (unit: millimeter). The camera coordinate system O_c is created with the optical center as the coordinate origin. Note that the coordinates of the object in the real world are represented relative to O_c, and O_c reaches O_i through perspective projection transformation. Suppose the coordinates of point P in O_c are (X_p, Y_p, Z_p), and the corresponding coordinates projected onto O_i are (x_p, y_p). The relationship of point P between O_c and O_i is given by

$$\begin{cases} x_p = f\frac{X_p}{Z_p} \\ y_p = f\frac{Y_p}{Z_p} \end{cases} \tag{7}$$

where f is the camera focal length. As demonstrated in Figure 6, a pixel coordinate system O_p is depicted with the origin on the top-left vertex of the image (unit: pixel). The u- and v-axes are parallel to the x- and y-axes of O_i. Let the point (u_p, v_p) in O_p corresponding to the point (x_p, y_p) in O_i. The two coordinate values can be obtained as follows:

$$\begin{cases} u_p = \frac{x_p}{d_u} + u_0 \\ v_p = \frac{y_p}{d_v} + v_0 \end{cases} \tag{8}$$

where (u_0, v_0) represents the translation of the origin of O_i relative to the origin of O_p and d_u and d_v represent the actual size of the pixels in the u-axis and v-axis directions, respectively. According to Eqn. (7) and (8), the transformed relationship between O_p and O_c is given as

$$Z_p \begin{bmatrix} u_p \\ v_p \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_p \\ Y_p \\ Z_p \end{bmatrix}, \tag{9}$$

where $f_x = f/d_u$, $f_y = f/d_v$. Note that f, d_u, d_v, u_0, and v_0 are the intrinsic camera parameters that can be provided from the factory parameters of the ZED camera, and Z_p is the observed depth value of the depth map.

As shown in Figure 7A, let A, B, C, and D be the leftmost, topmost, rightmost, and bottom-most endpoints of the citrus fruit region projected in O_i, respectively, which have coordinates (u_A, v_A), (u_B, v_B), (u_C, v_C), and (u_D, v_D). Denote (X_A, Y_A, Z_A), (X_B, Y_B, Z_B), (X_C, Y_C, Z_C), and (X_D, Y_D, Z_D) as the corresponding spatial coordinates of points A, B, C, and D in O_c. According to Eqn. (9), the spatial coordinates of A, B, C, and D are given by

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \end{bmatrix} \begin{bmatrix} u_i \\ v_i \\ Z_i \end{bmatrix}, \tag{10}$$

where i is A, B, C and D. Let d_x, d_y and d_z be the fruit diameter in the X_c-, Y_c-, and Z_c-axes, respectively. d_x and d_y are obtained according to the spatial coordinates of A, B, C, and D,

$$\begin{cases} d_x = X_C - X_A \\ d_y = Y_D - Y_B \end{cases} \tag{11}$$

In 3D perspective projection, the citrus fruit diameter d_z cannot be obtained directly from the image. Fortunately, the shape of a citrus fruit is similar to an ellipsoid; thus, the magnitudes of d_x, d_y, and d_z will be highly correlated. In this paper, d_z can be estimated by fitting a quadratic polynomial function of d_x and d_y:

$$\hat{d_z} = \beta_0 + \beta_1 d_x^2 + \beta_2 d_y^2 + \beta_3 d_x + \beta_4 d_y, \tag{12}$$

where β_0, β_1, ..., β_4 are the regression coefficients of a polynomial that can be determined using the least-squares method.

Let $Q_0(u_q, v_q)$ be the center point of the citrus 2D bounding box (Figure 7), which indeed corresponds to the center of the citrus surface. The spatial coordinates (X_q, Y_q, Z_q) of Q_0 in O_c are obtained using Eqn. (10). Denote by d the Euclidean distance from Q_0 to the origin point, i.e., the distance between the citrus and camera,

$$d = \sqrt{X_q^2 + Y_q^2 + Z_q^2}. \tag{13}$$

The position and posture information for detected targets can usually be determined by the 3D bounding box (Xu et al., 2018). Let $P_1, P_2, \ldots, P_8$ be the eight vertices of the citrus 3D bounding box (**Figure 7B**), which have coordinates of (X_i, Y_i, Z_i) for $i = 1, 2, \ldots, 8$. In particular, (X_i, Y_i, Z_i) can be obtained from the relative geometrical position of P_i to Q_0, e.g., (X_1, Y_1, Z_1) is inferred as follows:

$$\begin{cases} X_1 = X_q - d_x/2 \\ Y_1 = Y_q + d_y/2 \\ Z_1 = Z_q + d_z \end{cases} \tag{14}$$

To visualize the 3D bounding box of citrus in the image, the corresponding projected 2D coordinates are calculated. Let the eight vertex points $p_1, p_2, \ldots, p_8$ be the corresponding $P_1, P_2, \ldots, P_8$ projected on O_p, which have coordinates (u_i, v_i), $i = 1, 2, \ldots, 8$. They can be deduced by Eqn. (9). Therefore, the 3D localization for each citrus is summarized in **Algorithm 1**.

Algorithm 1 - Calculation of 3D localization for a citrus fruit.

```
Input: Citrus fruit region $I_C$ and depth map $I_d$.
Output:$d_x$, $d_y$, $d_z$, $Q_0(X_q, Y_q, Z_q)$, $d$, $(u_i, v_i)$ and $(X_i, Y_i, Z_i)$ for $i = 1, 2, \ldots, 8$.
S1: According to $I_C$, 2D coordinates of citrus region extreme points $A(u_A, v_A)$, $B(u_B, v_B)$, $C(u_C, v_C)$, and $D(u_D, v_D)$ are obtained.
S2: The spatial coordinates of $(X_A, Y_A, Z_A)$, $(X_B, Y_B, Z_B)$, $(X_C, Y_C, Z_C)$, and $(X_D, Y_D, Z_D)$ are calculated by Eqn. (10).
S3: Citrus fruit diameter$d_x$ and $d_y$ are calculated by Eqn. (11), and $d_z$ is estimated by Eqn. (12).
S4: According to $I_d$, the spatial coordinates of citrus $Q_0(X_q, Y_q, Z_q)$ are determined by Eqn. (10).
S5: The distance $d$ between $Q_0$ and the origin point in $O_c$ is obtained by Eqn. (13).
S6: The spatial coordinates $(X_i, Y_i, Z_i)$ of citrus 3D bounding box are calculated by Eqn. (14).
S7: The 2D coordinates $(u_i, v_i)$ of citrus 3D bounding box are calculated from $(X_i, Y_i, Z_i)$ using Eqn. (10).
```

Results and discussion

The performance of the proposed method was evaluated on a workstation with an Intel Core i9-9920X processor with 3.50 GHz, 32 GB RAM, and an NVIDIA GeForce RTX 2080 GPU with 8 GB RAM. The operating system is Windows 10, and the software framework is PyTorch 1.8. All the algorithms were developed in Visual Studio Code 1.63 and MATLAB R2020a software.

Performance evaluation of citrus 2D detection

To evaluate the performance of citrus 2D detection using our proposed loss function, ($Loss_{PB}$), on YOLO v5s, three loss functions, $Loss_B$, focal loss ($Loss_F$) (Lin et al., 2020), and polarity loss ($Loss_P$) (Rahman et al., 2020), were used for comparison. The YOLO v5s models were trained using the training dataset, and the hyperparameters of the model were fine-tuned using the validation dataset. The performance of the final model was evaluated using the test dataset. After several trial-and-error training runs, the learning rate was set to 0.0032, the batch size was set to 32, the IoU threshold was set to 0.5, the training epoch was 200 and γ was set to 20. All the input images were resized to 640 × 640 pixels. The network weights of YOLO v5s were initialized with the weights of the model pretrained on the COCO image dataset.

The detection results under three illumination conditions on the test dataset are provided in **Table 2**. With our proposed loss function, $Loss_{PB}$, we achieves the best improvement on the non-uniform illumination than weak illumination and well illumination, compared to $Loss_B$, $Loss_F$, and $Loss_P$. Specifically, under non-uniform illumination, the recall of our loss is 99.55%, which is an average improvement of 9.08% over $Loss_B$, 7.17% over $Loss_F$, and 5.38% over $Loss_P$. The precision of our loss is 95.79%, which is almost the same result as that of the other three loss functions, while the highest precision of 95.93% is obtained by $Loss_F$. The F_1-score of our loss is 0.98, which is the highest.

Under weak illumination, the precision of our loss is 96.13%, which is 1.33% higher than that of $Loss_B$ and 1.04% higher than that of $Loss_F$ and $Loss_P$. The recall of our loss is 98.47%, and the F_1-score is 0.97, both of which are better than those of the other loss functions. Under well illumination, the F_1-score of our loss is 0.98, an average of 3%, 4%, and 2% higher than that of $Loss_B$, $Loss_F$ and $Loss_P$, respectively. The precision and recall of our loss are 96.64% and 98.48%, respectively, which are both the best highest.

Overall, for our loss, the recall is 98.85%, the precision is 96.22%, and the F_1-score is 0.98, on average, under the three illumination conditions, values that are approximately 2–9% higher than those of $Loss_B$, about 1–6% higher than those of $Loss_F$, and approximately 1–4% higher than those of $Loss_P$. In terms of other metrics, the detection time per image (T) is similar for all loss functions and is consistent with the requirements of the picking robot (Tu et al., 2020).

Figure 8 shows the citrus samples detected by our loss function $Loss_{PB}$ but not $Loss_B$ under different illumination conditions. As listed in **Table 2**, the recall rate of $Loss_B$ under non-uniform illumination is the lowest at 90.47% than other illumination conditions. On the other hand, the recall rate of $Loss_{PB}$ performed the best at 99.55% over other illumination conditions. The reason may be twofold: (1) As shown in **Figure 8**, the illumination component is uniform on the surface of a citrus fruit under weak or well illumination conditions. Therefore, the total number of samples is larger under weak and well illumination than under non-uniform

TABLE 2 Detection results of You Only Look Once (YOLO) v5s using different loss functions in the test dataset.

Loss function	Illumination	P (%)	R (%)	F_1	T (ms)	TP	FP	FN
Our Loss $Loss_{PB}$	Non	95.79	**99.55**	0.98	79.31	888	39	4
	Weak	96.13	98.47	0.97	**75.49**	645	26	10
	Well	**96.64**	98.48	0.98	81.04	1036	36	16
	Total	96.22	98.85	**0.98**	78.96	2569	101	30
$Loss_B$	Non	95.50	90.47	0.93	81.34	807	38	85
	Weak	94.80	91.91	0.93	78.63	602	33	53
	Well	96.07	93.06	0.95	83.16	979	40	73
	Total	95.56	91.88	0.94	81.33	2388	111	211
$Loss_F$	Non	95.93	92.38	0.94	79.91	824	35	68
	Weak	95.09	91.76	0.93	75.38	601	31	54
	Well	96.25	92.78	0.94	82.59	976	38	76
	Total	95.85	92.38	0.94	79.75	2401	104	198
$Loss_P$	Non	95.67	94.17	0.95	79.33	840	38	52
	Weak	95.09	94.66	0.95	75.53	620	32	35
	Well	96.06	95.06	0.96	81.73	1000	41	52
	Total	95.68	94.65	0.95	79.25	2460	111	139

The bold values means the best result on each metrics.

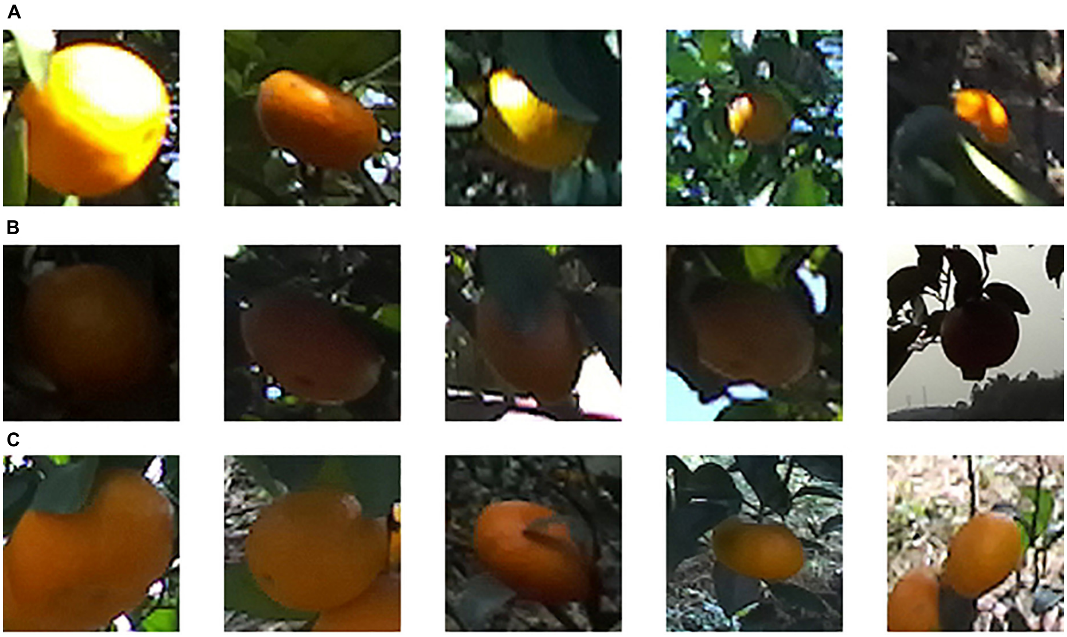

FIGURE 8
The missed detection of citrus samples of You Only Look Once (YOLO) v5s but detected by our proposed loss in different illumination on test data: **(A)** non, **(B)** weak, and **(C)** well.

illumination, making the YOLO v5s with $Loss_B$ more likely to learn citrus with uniform color features. (2) It is likely that, compared with weak and well illumination, the color features of a citrus fruit under non-uniform will be hard to extract by the Yolo v5s with $Loss_B$, such that the most citrus sample cannot be detected. Using our loss function, the citrus target under non-uniform illumination will be further pushed from the background. A large penalty is applied to missed detection from the penalty function f_P in the training process.

Figure 9 shows the detection results for different loss functions. Specifically, the red bounding box represents

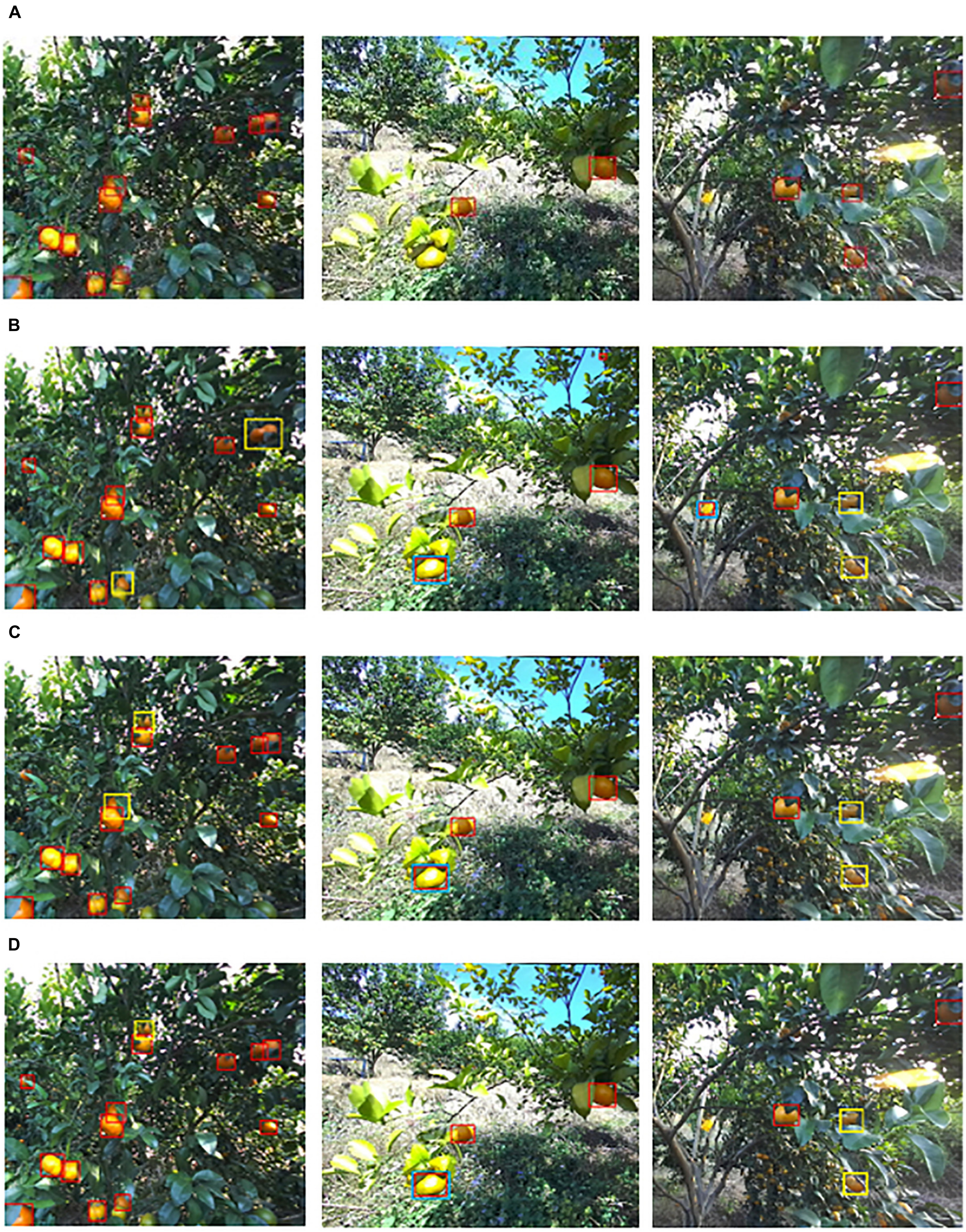

FIGURE 9
Comparison of detection results using different loss functions: **(A)** Our loss, **(B)** $Loss_B$, **(C)** $Loss_F$, and **(D)** $Loss_P$.

the predicted output by models, the yellow bounding box represents the missed detection, and the blue bounding box represents the false detection. Figure 9A indicates that the YOLO v5s model with our loss function achieves the best citrus detection performance under all illumination conditions, reducing the occurrence of both missed detection and false detection.

There are several examples of missed detection or false detection by other loss functions, as presented in Figures 9B–D. With such loss functions, some background objects, such as

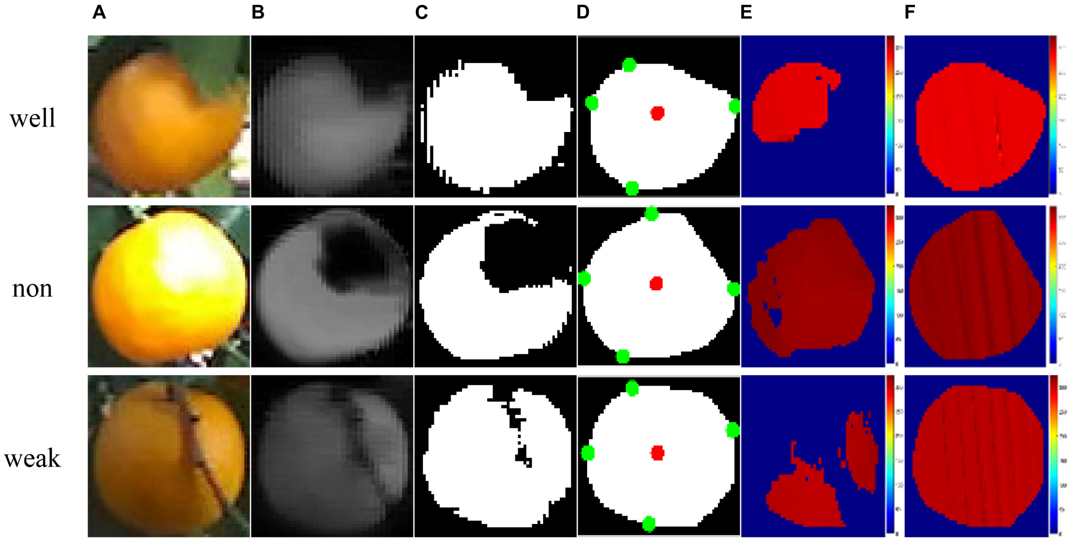

FIGURE 10
Results of samples under different illumination conditions: **(A)** RGB image, **(B)** Cr-Cb chromatic mapping, **(C)** Otsu segmentation, **(D)** morphological operations, where the red point is the center point and the green point is the maximum and minimum point of the citrus fruit, **(E)** color map of the original depth map, and **(F)** color map on depth map restored by the kriging method.

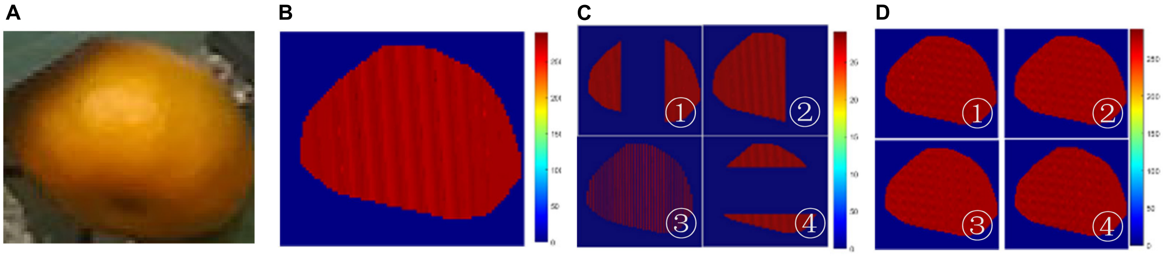

FIGURE 11
Experiment results using the Kriging method: **(A)** color map of depth values, **(B)** RGB image of a citrus fruit, **(C)** color map by setting depth values zero at random pixels, and **(D)** color map of restoration by kriging.

immature citrus and yellow insect-attracting boards, can lead to false detection of the citrus target. It is likely that immature green citrus has similar texture and shape properties as mature citrus, and the yellow insect-attracting board has similar color characteristics as citrus. On the other hand, citrus that is occluded by leaves, branches, or other backgrounds objects may be misclassified as background, i.e., missed detection. For such citrus fruits, it is likely that only a few features can be extracted from the image, resulting in a hard negative sample that is difficult to distinguish from the background.

Figure 9A shows that our proposed loss function achieves the best detection performance. Specifically, the penalty for false detection is enhanced by the penalty function f_P during the training process, and citrus targets are displaced from the background. As a result, the probability of missed detection is reduced substantially, and the detection performance of citrus is thus improved. Note that $Loss_P$ uses a penalty function similar to f_P and also achieves better performance than that of $Loss_F$ and $Loss_B$. Indeed, it was developed based on $Loss_F$. However, $Loss_F$ cannot push the object further from the background, which may not be an effective improvement on our dataset.

Performance evaluation of citrus region extraction and depth value restoration

Figure 10 illustrates the results of citrus region extraction and depth map restoration under variable illumination conditions. Under the well illumination conditions, the citrus occluded by leaves is shown in the first row of Figure 10A. The results of Cr-Cb chromatic mapping and Otsu thresholding are presented in Figures 10B–C. Image noise, holes, and weakly

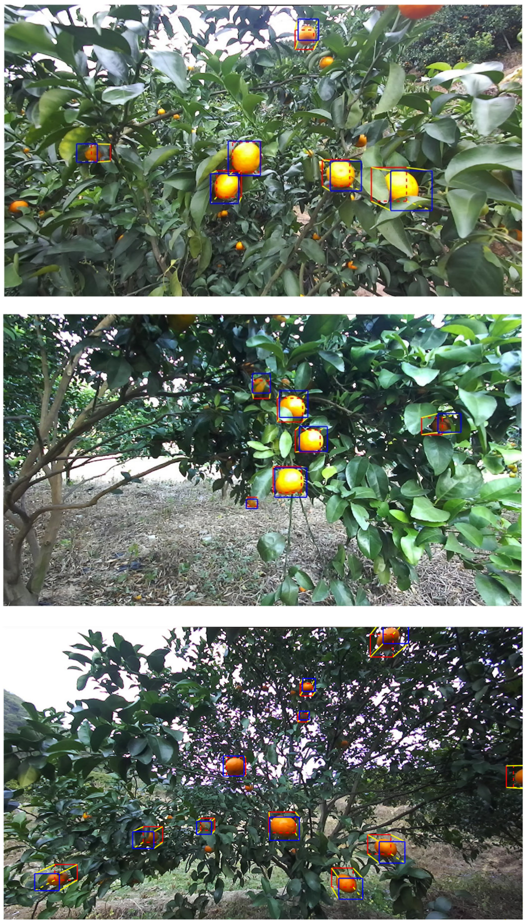

FIGURE 12
Examples of 3D bounding boxes for citrus fruits.

connected regions can exist in the binary image obtained via Otsu thresholding. The citrus region is likely blurred, mainly due to the far distance from the camera. The result of morphological processing is shown in Figure 10D. The image noise was completely removed, and contour smoothing was achieved, such that the majority of the citrus region occluded by the leaves was filled perfectly. As shown in Figure 10E, the depth map of the extracted citrus fruit region after the convex hull operation is incomplete, i.e., the area of missing values covers approximately large than half of the area of the citrus fruit region, which may be caused by camera performance limitations. As shown in Figure 10F, the missing depth values are restored using the kriging method, thereby estimating the complete depth values of the fruit region.

The results of citrus fruit extraction and depth map restoration under the non-uniform illumination conditions are presented in the second row of Figure 10. The shape of the extracted citrus region is obviously incomplete, which may result from overexposure to the citrus surface. As shown in Figure 10D, the incomplete part was restored by morphological operations. Subsequently, the missing depth values in the citrus region (Figure 10E) were recovered, as shown in Figure 10F. Similarly, the results under weak illumination conditions are illustrated in the third row of Figure 10. The citrus fruit region occluded by branches is extracted almost completely, as shown in Figure 10D. Due to the lack of light and other factors, the depth map of the extracted citrus region is sparse, as shown in Figure 10E. After using the kriging method, the missing depth values are effectively restored, as shown in Figure 10F.

To evaluate the accuracy of the kriging method to recover depth values on the occluded citrus region, an experiment was conducted by simulating the restoration using the incomplete depth map. Figure 11 shows the results of using the kriging method on an extracted citrus region. Figure 11A is the complete depth map of Figure 11B. Figure 11C shows that the incomplete depth map was generated by setting the corresponding depth values to zero with four schemes. About 50% of the pixels are set as missing values. Specifically, the incomplete depth maps ① and ④ were created by setting the pixels of the central part to zero in the vertical and horizontal directions. The incomplete depth map ② was created by setting the pixels of the right part to zero, and ③ was created by setting the interleaving pixels to zero. As shown in Figure 11D, the missing values are recovered using the kriging method, such that the depth map of the fruit region is completely restored.

Compared with the original depth map of Figure 11B, the average restoration error of depth map ①, ②, ③, and ④ is 2.29, 2.15, 2.08, and 2.31 mm, respectively, such that the average of the all the restoration errors is 2.21 mm. The minimum error was performed in the depth map ③, indicating that the estimate of missing depth value is recovered with high accuracy when the depth values are only missing randomly in the depth map. On the other hand, the maximum error was performed in the depth map ① and ④, indicating that the restoration error is large when the missing depth values are in the most discontinuous part of the depth map. In total, the mean relative error is 1.36%, indicating that the kriging method effectively restored the depth map with high accuracy.

Performance evaluation of citrus 3D localization

Citrus diameter d_x, d_y, and d_z, coordinates of citrus $Q_0(X_q, Y_q, Z_q)$, the distance between the citrus and camera d, the 3D coordinates of the citrus 3D bounding box (X_i, Y_i, Z_i), and its 2D coordinates (u_i, v_i) are calculated using Algorithm 1. Specifically, to obtain the regression model for d_z, as mentioned in Eqn. (12), a total of 137 citrus samples were collected in the orchard. The diameter d_x, d_y, and d_z of each fruit were measured

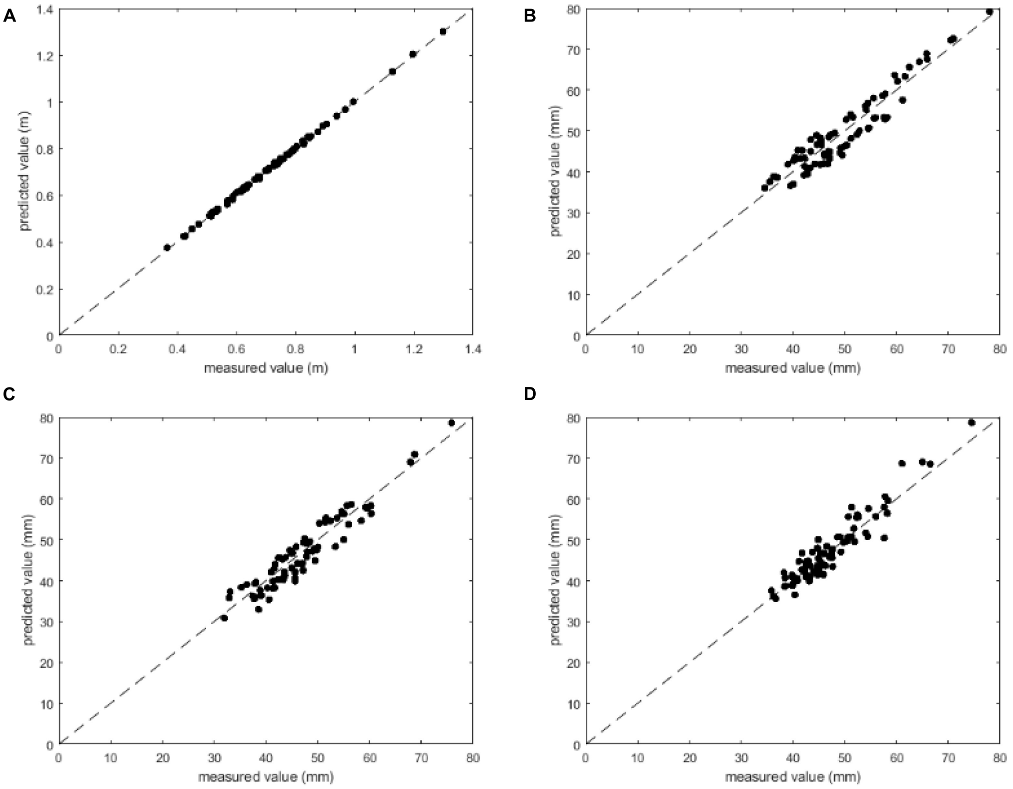

FIGURE 13
Comparison between the measured values and predicted values: **(A)** *d*, **(B)** d_x, **(C)** d_y, and **(D)** d_z.

by a Vernier caliper (Pro skit, PD-151). The quadratic polynomial function fitted for d_z is determined as follows:

$$\hat{d}_z = 16.0728 + 0.0028d_x^2 + 0.0018d_y^2 + 0.0264d_x + 0.4133d_y, \tag{15}$$

where the root mean square error (RMSE) is 4.51 mm and the coefficient of determination $R^2 = 0.940$, indicating a good model for estimating d_z .

Figure 12 shows the result of 3D bounding boxes predicted for each citrus fruit. The boxes are drawn by connecting the adjacent vertices (u_i, v_i), for $i = 1, 2, ..., 8$, with a straight line. The front face of the 3D bounding box was drawn by the blue rectangle, the back face of the 3D bounding box was drawn by the red rectangle, and the side face of the 3D bounding box was drawn by the yellow line. The citrus fruits near the center of the image are correctly detected with the 3D bounding boxes. Moreover, the four edge lines (yellow lines) of the 3D bounding box disappear in the center of the image, which is consistent with the principle of parallel perspective (Cai et al., 2021). Thus, our proposed method achieves accurate localization results.

To evaluate the localization accuracy of citrus 3D localization, 22 images of citrus fruits were considered. The distance between the citrus and the camera *d* was measured by a laser rangefinder (UNI-T, UT392B), and citrus diameters d_x, d_y, and d_z were measured with a Vernier caliper (Pro skit, PD-151). A scatter plot of the measured values and the values predicted by our method is presented in Figure 13. Our method obtains good accuracy for predicting *d*, d_x, d_y, and d_z: the closer the measured values and the predicted value are to the 45-degree line, the higher the accuracy. Figure 13A shows the best prediction and fewer errors between the measured value and predicted values for *d*, where all the plotted points lie almost on the 45-degree line. Furthermore, Figures 13B–D shows that the predicted values of d_x, d_y, and d_z are generally close to the 45-degree line, indicating that our proposed method is able to achieve accurate localization results.

Overall, the average error of distance *d* between the citrus and camera is 3.98 mm, which is better than the 15 mm achieved in Wang et al. (2016). The average errors of citrus diameters d_x, d_y, and d_z were 2.75, 2.52, and 2.11 mm, respectively, which is almost the same precision as (Yang et al., 2020) and better than the 10 mm achieved in (Nguyen et al., 2016) and the 4.9 mm achieved in (Wang et al., 2017).

Our method can accurately locate citrus under variable illumination and different occlusion conditions in natural orchards. The distance *d* can be used to determine the extension length of the robot hand, and the coordinates of citrus $Q_0(X_q, Y_q, Z_q)$ can be used to manipulate the robot hand's the series of joints or articulations. The diameter d_x, d_y, d_z and the 3D bounding box (X_i, Y_i, Z_i) can be used to finetune the posture of grasping structures.

Conclusion

This paper aims to address the problem of the lower detection rate for mature citrus under variable illumination and occlusion conditions. We proposed a novel method to detect and localize citrus fruits in natural orchards using binocular cameras and deep learning. The main conclusions are as follows:

1. A new loss function $Loss_{PB}$ for YOLO v5s is proposed to calculate the loss value for class probability and objectness score, with a penalty function f_p developed to account for the disparity between citrus and background. As a result, the citrus detection performance of our loss function is improved by pushing the citrus further from the background in the training process, even under variable illumination and different occlusion conditions. The recall values of the three groups of illumination conditions were 99.55%, 98.47%, and 98.48%, the precision values were 95.79%, 96.13%, and 96.64%, respectively, and the F_1-scores were close to 0.98. The average detection time was 78.97 ms per image. Compared with the original YOLO v5s, the performance improvement was 2-9% on average.
2. Based on the detected 2D bounding box for a citrus, the potential fruit region of mature citrus was segmented completely using Cr-Cb chromatic mapping, Otsu thresholding and morphology processing. In particular, the difference in color intensity between citrus targets and background objects is enhanced using Cr-Cb chromatic mapping, which helps to extract the complete shape of citrus fruit using Otsu thresholding and morphology processing.
3. To recover the missing depth values in the citrus region under different occlusion states, the kriging method was applied based on the spatial proximity among neighboring points. The experimental results show that the average error of the restored depth values was 2.02 mm and the relative error was 1.26%, indicating that the method can accurately restore the depth map of citrus fruit.
4. Based on the ellipsoid characteristic of citrus fruit, the 3D localization information of citrus is accurately determined using the camera imaging model and a restored depth map. The experimental results show that the average error of the distance d between the citrus fruit and the camera was 3.98 mm, and the average errors of the citrus diameter d_x, d_y and d_z were 2.75, 2.52, and 2.11 mm, respectively, which is better than the results achieved in other research.

Our method can provide 3D citrus position data under variable illumination and different occlusion conditions in natural orchards. Future work will focus on few-shot learning and reduce the number of citrus fruits in the training dataset to improve citrus detection and localization.

Data availability statement

The original contributions presented in this study are publicly available. This data can be found here: https://github.com/AshesBen/citrus-detection-localization.

Author contributions

All authors contributed to the method and result of the study, dataset generation, model training and testing, analysis of results, and the drafting, revising, and approving of the contents of the manuscript.

Funding

We acknowledged support from the Natural Science Foundation of Guangdong Province, China (Grant Nos. 2020B1515120070 and 2021A1515010824), the Planned Science and Technology Project of Guangdong Province, China (Grant Nos. 2019B020216001, 2019A050510045, and 2021A0505030075), the National Natural Science Foundation of China (Grant No. 32071895), the Key Project of Universities in Guangdong Province, China (Grant No. 2020ZDZX1061), the Innovation Team Project of Universities in Guangdong Province, China (Grant No. 2021KCXTD010), the Planned Science and Technology Project of Guangzhou, China (Grant Nos. 202002020063, 202007040007, and 202103000028), and the Rural Revitalization Strategy Project of Guangdong Province, China (Grant No. 2019KJ138).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Buda, M., Maki, A., and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural netsworks. *Neural Netw.* 106, 249–259. doi: 10.1016/j.neunet.2018.07.011

Cai, W., Liu, D., Ning, X., Wang, C., and Xie, G. (2021). Voxel-based three-view hybrid parallel network for 3D object classification. *Displays* 69:102076. doi: 10.1016/j.displa.2021.102076

Chen, M., Tang, Y., Zou, X., Huang, Z., Zhou, H., and Chen, S. (2021). 3D global mapping of large-scale unstructured orchard integrating eye-in-hand stereo vision and SLAM. *Comput. Electron. Agric.* 187:106237. doi: 10.1016/j.compag.2021.106237

Chu, P., Li, Z., Lammers, K., Lu, R., and Liu, X. (2021). Deep learning-based apple detection using a suppression mask R-CNN. *Pattern Recognit. Lett.* 147, 206–211. doi: 10.1016/j.patrec.2021.04.022

Ge, Y., Xiong, Y., and From, P. J. (2020). Symmetry-based 3D shape completion for fruit localisation for harvesting robots. *Biosyst. Eng.* 197, 188–202. doi: 10.1016/j.biosystemseng.2020.07.003

Gongal, A., Amatya, S., Karkee, M., Zhang, Q., and Lewis, K. (2015). Sensors and systems for fruit detection and localization: a review. *Comput. Electron. Agric.* 116, 8–19. doi: 10.1016/j.compag.2015.05.021

He, Z., Xiong, J., Chen, S., Li, Z., Chen, S., Zhong, Z., et al. (2020). A method of green citrus detection based on a deep bounding box regression forest. *Biosyst. Eng.* 193, 206–215. doi: 10.1016/j.biosystemseng.2020.03.001

Huang, M., Lu, Q., Chen, W., Qiao, J., and Chen, X. (2019). Design, analysis, and testing of a novel compliant underactuated gripper. *Rev. Sci. Instrum.* 90:045122. doi: 10.1063/1.5088439

Jiang, Z., Zhao, L., Li, S., and Jia, Y. (2020). Real-time object detection method based on improved YOLOv4-tiny. *ArXiv [preprint]*

Jocher, G., and Stoken, A. (2021). *ultralytics/yolov5: v5.0.*

Kang, H., and Chen, C. (2020). Fast implementation of real-time fruit detection in apple orchards using deep learning. *Comput. Electron. Agric.* 168:105108. doi: 10.1016/j.compag.2019.105108

Liang, C., Xiong, J., Zheng, Z., Zhong, Z., Li, Z., Chen, S., et al. (2020). A visual detection method for nighttime litchi fruits and fruiting stems. *Comput. Electron. Agric.* 169:105192. doi: 10.1016/j.compag.2019.105192

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2020). Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 318–327. doi: 10.1109/TPAMI.2018.2858826

Liu, W., Chen, X., Yang, J., and Wu, Q. (2017). Robust color guided depth map restoration. *IEEE Trans. Image Process.* 26, 315–327. doi: 10.1109/TIP.2016.2612826

Nguyen, T. T., Vandevoorde, K., Wouters, N., Kayacan, E., De Baerdemaeker, J. G., and Saeys, W. (2016). Detection of red and bicoloured apples on tree with an RGB-D camera. *Biosyst. Eng.* 146, 33–44. doi: 10.1016/j.biosystemseng.2016.01.007

Noorizadeh, S., Golmohammadi, M., Bagheri, A., and Bertaccini, A. (2022). Citrus industry: phytoplasma-associated diseases and related challenges for Asia, America and Africa. *Crop Prot.* 151:105822. doi: 10.1016/j.cropro.2021.105822

Onishi, Y., Yoshida, T., Kurita, H., Fukao, T., Arihara, H., and Iwai, A. (2019). An automated fruit harvesting robot by using deep learning. *Robomech J.* 6:13. doi: 10.1186/s40648-019-0141-2

Rahman, S., Khan, S. H., and Barnes, N. (2020). Polarity loss for zero-shot object detection. *ArXiv [preprint]*

Tang, Y., Dananjayan, S., Hou, C., Guo, Q., Luo, S., and He, Y. (2021). A survey on the 5G network and its impact on agriculture: challenges and opportunities. *Comput. Electron. Agric.* 180:105895. doi: 10.1016/j.compag.2020.105895

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

Wan, S., and Goudos, S. (2020). Faster R-CNN for multi-class fruit detection using a robotic vision system. *Comput. Netw.* 168:107036. doi: 10.1016/j.comnet.2019.107036

Wang, C., Wang, Y., Liu, S., Lin, G., He, P., Zhang, Z., et al. (2022). Study on pear flowers detection performance of YOLO-PEFL model trained with synthetic target images. *Front. Plant Sci.* 13:911473. doi: 10.3389/fpls.2022.911473

Wang, C., Zou, X., Tang, Y., Luo, L., and Feng, W. (2016). Localisation of litchi in an unstructured environment using binocular stereo vision. *Biosyst. Eng.* 145, 39–51. doi: 10.1016/j.biosystemseng.2016.02.004

Wang, D., and He, D. (2021). Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. *Biosyst. Eng.* 210, 271–281. doi: 10.1016/j.biosystemseng.2021.08.015

Wang, J., Chen, Y., Gao, M., and Dong, Z. (2021). Improved YOLOv5 network for real-time multi-scale traffic sign detection. *ArXiv [preprint]*

Wang, Z., Walsh, K. B., and Verma, B. (2017). On-tree mango fruit size estimation using RGB-D Images. *Sensors* 17:2738. doi: 10.3390/s17122738

Xiong, J., Liu, Z., Chen, S., Liu, B., Zheng, Z., Zhong, Z., et al. (2020). Visual detection of green mangoes by an unmanned aerial vehicle in orchards based on a deep learning method. *Biosyst. Eng.* 194, 261–272. doi: 10.1016/j.biosystemseng.2020.04.006

Xu, D., Anguelov, D., and Jain, A. (2018). "PointFusion deep sensor fusion for 3D bounding box estimation," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE). doi: 10.1109/CVPR.2018.00033

Yan, B., Fan, P., Lei, X., Liu, Z., and Yang, F. (2021). A real-time apple targets detection method for picking robot based on improved YOLOv5. *Remote Sens.* 13:1619. doi: 10.3390/rs13091619

Yang, C. H., Xiong, L. Y., Wang, Z., Wang, Y., Shi, G., Kuremot, T., et al. (2020). Integrated detection of citrus fruits and branches using a convolutional neural network. *Comput. Electron. Agric.* 174:105469. doi: 10.1016/j.compag.2020.105469

Zheng, Z., Xiong, J., Lin, H., Han, Y., Sun, B., Xie, Z., et al. (2021). A method of green citrus detection in natural environments using a deep convolutional neural network. *Front. Plant Sci.* 12:705737. doi: 10.3389/fpls.2021.705737

Zhuang, J., Hou, C., Tang, Y., He, Y., Guo, Q., Zhong, Z., et al. (2019). Computer vision-based localisation of picking points for automatic litchi harvesting applications towards natural scenarios. *Biosyst. Eng.* 187, 1–20. doi: 10.1016/j.biosystemseng.2019.08.016

Zhuang, J. J., Luo, S. M., Hou, C. J., Tang, Y., He, Y., and Xue, X. Y. (2018). Detection of orchard citrus fruits using a monocular machine vision-based method for automatic fruit picking applications. *Comput. Electron. Agric.* 152, 64–73. doi: 10.1016/j.compag.2018.07.004