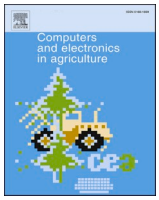
Computers and Electronics in Agriculture 217 (2024) 108621

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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



Improved binocular localization of kiwifruit in orchard based on fruit and calyx detection using YOLOv5x for robotic picking

Changqing Gao [a], Hanhui Jiang [a], Xiaojuan Liu [a], Haihong Li [a], Zhenchao Wu [a], Xiaoming Sun [a], Leilei He [a], Wulan Mao [a,e], Yaqoob Majeed [f], Rui Li [a,*], Longsheng Fu [a,b,c,d,*]

[a] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi, 712100, China
[b] Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi 712100, China
[c] Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Service, Yangling, Shaanxi 712100, China
[d] Northwest A&F University Shenzhen Research Institute, Shenzhen, Guangdong 518000, China
[e] Institute of Agricultural Mechanization, Xinjiang Academy of Agricultural Sciences, Urumqi 830000, China
[f] Faculty of Agricultural Engineering and Technology, University of Agriculture, Faisalabad 38000, Pakistan

ARTICLE INFO

Keywords:
Fruit detection
Calyx localization
Binocular stereo vision
YOLOv5x
Robotic harvesting

ABSTRACT

Localization is the first critical step for picking robots to successfully grasp fruit. However, classical binocular localization methods adopted global matching for kiwifruit, which may result in a large amount of mismatching feature points in complex orchard and thus cause low localization accuracy. Therefore, an improved binocular localization method of calyxes based on deep learning was proposed to accurately detect and locate kiwifruit for robotic harvesting. Calyxes in the binocular images and kiwifruit in the left images of the binocular images were detected using You Only Look Once version 5 xlarge (YOLOv5x). The detected calyxes were matched in the binocular images using kiwifruit and calyx pairing and kiwifruit matching. The matched calyxes in the binocular images were used to locate calyxes using three localization methods. Specifically, three binocular localization methods, i.e., calyx localization (CL), fruit localization (FL), and depth information from depth map (DIDM), were compared to find the optimal one. Ground truth three-dimensional coordinates of calyxes was measured by laser rangefinder and coordinate paper on a self-designed experimental platform. Results showed that YOLOv5x achieved an average precision (*AP*) of 99.5 % on kiwifruit detection and a mean *AP* of 93.5 % on kiwifruit and calyx detection with a detection speed of 108 ms per image. Average deviation of X-axis, Y-axis, and Z-axis obtained by the CL method were 7.9 mm, 6.4 mm, and 4.8 mm, respectively. Compared with the FL and DIDM methods, localization error rate of the proposed CL method was reduced by 55.1 % and 53.8 %, respectively. It indicates that the proposed CL method is promising for robotic harvesting.

1. Introduction

Global kiwifruit planting area exceeded 270 thousand ha and production reached 4.4 million tons in 2020 (UN Food & Agriculture Organization, 2022). Such large kiwifruit production requires labor force while kiwifruit is mainly hand-picked, which is labor-intensive (Fu et al., 2020a). The labor force involved in fruit picking accounts for over 25 % of annual production costs, which is a strong desire of automation of kiwifruit harvesting (Suo et al., 2021).

The first and foremost task of picking robots is localization, which guides harvesting robot to perform subsequent actions. Classical localization methods include structured light (SL), Time of Flight (ToF) and monocular movement (Robin and Lacroix, 2016; Liu et al., 2023a), all of which yield depth information. The SL projects a known pattern onto a scene and analyzes its deformation to calculate depth, while the ToF relies on measuring the time it takes for light to travel to an object and return to determine distance for depth mapping. Depth information principles of the SL and the ToF make them sensitive to light variations (Fu et al., 2020b; Gené-Mola et al., 2020; Hu et al., 2019; Li et al., 2023a; Mehranian et al., 2016; Neupane et al., 2021). The monocular movement estimates depth information by analyzing changes in visual information over time, such as displacement and rotation of the camera, which requires accurate control of camera position (Mehta et al., 2014). Besides, the monocular movement is not suitable for scenes with a large

* Corresponding authors.
E-mail addresses: ruili1216@nwafu.edu.cn (R. Li), fulsh@nwafu.edu.cn (L. Fu).

https://doi.org/10.1016/j.compag.2024.108621
Received 1 August 2023; Received in revised form 19 October 2023; Accepted 4 January 2024
Available online 22 January 2024

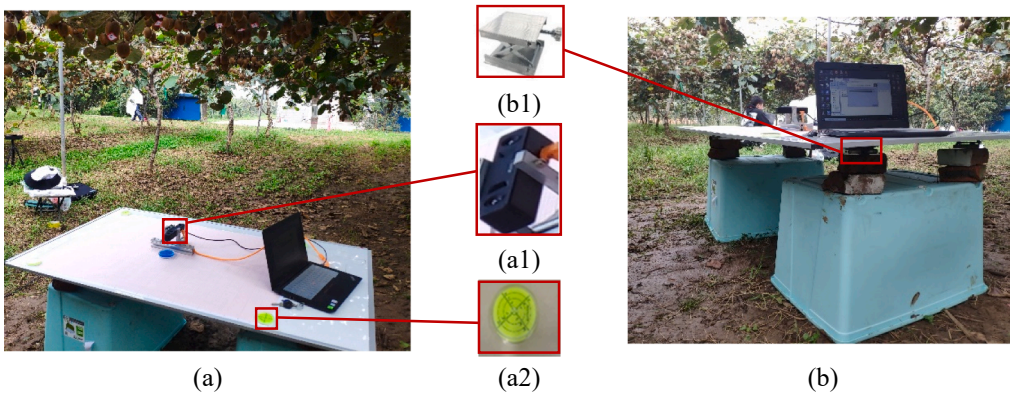


Fig. 1. Experiment platform of this work. (a) Image acquisition platform; (a1) Binocular camera; (a2) Level gauge; (b) Measurement platform of calyx three-dimensional coordinates; (b1) Micro height adjusting holder.

number of fruits in long-range perspective (Gongal et al., 2015; Mehta et al., 2017). In short, the above localization methods are prone to complex environment. A binocular vision system is based on the principle of similar triangles, which could help localize objects in complex environment with high accuracy and robustness. Therefore, a binocular vision system based on optical geometry is needed to ensure stability in agricultural picking tasks.

Binocular localization method offers exceptional robustness and precision for harvesting fruit in complex environment, featuring a straightforward design and remarkable efficiency (Ding et al., 2021; Li et al., 2023b; Liu et al., 2023a; Xiao et al., 2023). Hsieh et al. (2021) applied ZED mini binocular vision camera to locate beef tomato in greenhouse, which attained average localization error of 0.5 cm. Wang et al. (2017) used two charge-coupled device (CCD) color cameras to develop a binocular system for locating litchi, which acquired matching success rate of 82.6 %. Chen et al. (2021) employed ZED2 binocular camera to locate citrus in orchard, which got average localization error of 12.3 mm. Liu et al. (2023b) utilized two MV-CA060-10GC GigE industrial cameras to detect and locate pineapple fruit, which achieved an average absolute error of 24.4 mm in orchard. In general, the binocular localization method of fruits has stronger robustness than classical localization methods including structured light (SL), Time of Flight (ToF) and monocular movement, which is more suitable to complex environment.

Matching strategies become a key factor influencing binocularly localized fruits. Classical binocular localization adopted global matching method but there are large number of mismatching feature points for fruit matching (Zhang et al., 2021; Li et al., 2023b). Additional strategies for matching fruits based on feature points from environments have also been proposed, but this is only applicable in conditional environments (Wang et al., 2019; Tang et al., 2023). Once global matching is applied to small and dense fruit images in a wide and complex environment, it was prone to large amount of wrong match owing to similar color and texture of kiwifruit in binocular images (Liu et al., 2023b; Tang et al., 2020; Williams et al., 2019; Song, 2021). Contrarily to the classical stereo matching methods based on global matching using a large amount of feature points, deep learning-based kiwifruit and calyx detection and kiwifruit matching could utilize a few feature points limited in detection rectangles to increase matching accuracy of kiwifruit. Additionally, kiwifruit is normally adjacent or overlapped in canopy, where calyx of each kiwifruit is visible and independent (Song et al., 2021). As a result, it is desirable to locate midpoint of calyx of kiwifruit as target of end-effector instead of kiwifruit (Fu et al., 2019). Therefore, calyxes were matched using deep learning-based kiwifruit and calyx detection and

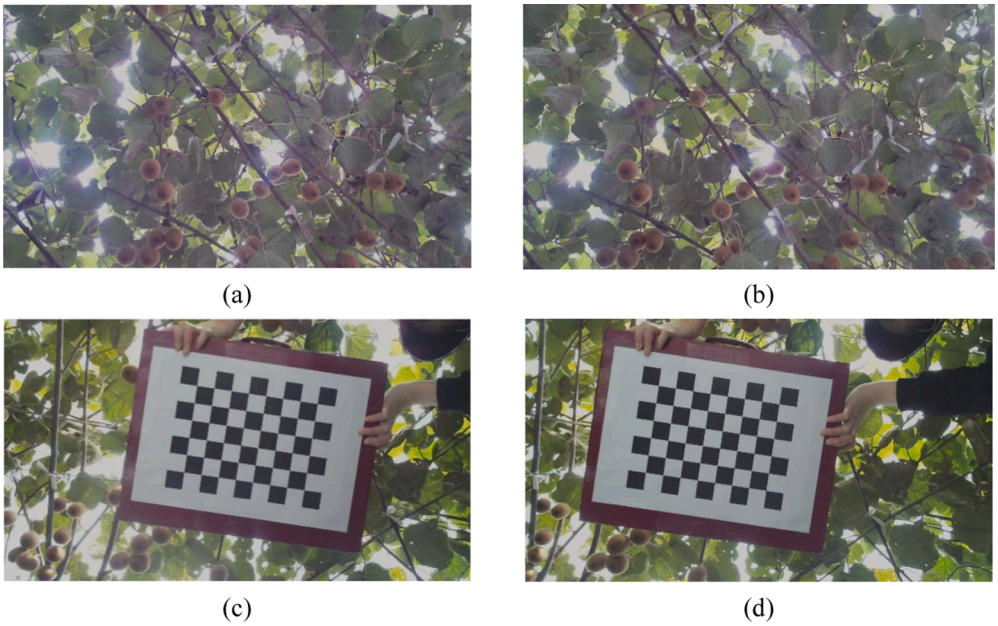


Fig. 2. Examples of kiwifruit images and calibration board images captured by a binocular camera MYNT EYE D1000-50. (a) Image taken by the left lens (i.e., left image); (b) Image taken by the right lens (i.e., right image); (c) Calibration board image taken by the left lens; (d) Calibration board image taken by the right lens.

Table 1
Camera internal reference calibration results.

Calibration parameters	Main calibration values of left lens	Main calibration values of right lens
Main point (u_0, v_0)	(621.2903, 341.0630)	(629.0143, 341.9893)
U-axis scale factor f_x	1037.0	1035.8
V-axis scale factor f_y	1030.8	1028.6
Distortion parameters k_c	(0.0410, −0.3724, 1.2469, −0.0108, 0.0019)	(0.0821, −0.3955, −0.8229, −0.0111, 0.0037)

Table 2
Camera external reference calibration results.

Calibration parameters	Calibration values
Baseline	120.00 mm
Rotation vector (R)	$\begin{pmatrix} 1.0000 & -0.0029 & -4.3146e-4 \\ 0.0029 & 1.0000 & -3.0873e-4 \\ -4.3237e-4 & 3.0745e-4 & 1.0000 \end{pmatrix}$
Translation vector (T)	$\begin{pmatrix} -118.3872 \\ -0.0850 \\ -0.9161 \end{pmatrix}$

kiwifruit matching, which is promising for kiwifruit localization.

In this paper, an improved binocular localization method of calyxes based on You Only Look Once version 5 xlarge (YOLOv5x) was proposed to detect and locate kiwifruit for robotic harvesting. The proposed method extracts kiwifruit and calyx using detection rectangles generated by YOLOv5x model. Then calyx matching based on kiwifruit and calyx pairing and kiwifruit matching was innovatively proposed to match calyxes between binocular images. Feature points of the matched calyxes in the binocular images were used to calculate depth values. The feature points that integrated with depth values are transformed into three-dimensional coordinates of calyxes to locate kiwifruit.

2. Materials and methods

2.1. Image acquisition and coordinates measuring

Kiwifruit images were captured in October 2020 and 2021 using an experimental platform during the harvesting season of Hayward variety at Meixian Kiwifruit Experimental Station of Northwest A&F University, Shaanxi, China, as shown in Fig. 1. A common binocular camera (MYNT EYE D1000-50/Color, Slightech, WuXi, China) shown in Fig. 1(a1) was mounted on the experimental platform, where the kiwifruit images were collected in a vertical upward manner about one meter below kiwifruit canopy, as shown in Fig. 1(a). The images were taken during different times of the day without artificial lighting and saved in 'PNG' format with a resolution of 1280 × 720 pixels. A total of 1266 pairs of kiwifruit binocular images were collected using the binocular camera, out of which 1250 pairs were used for training the network and 16 pairs were tested for kiwifruit calyx localization. Whereas, the binocular images used for calyx localization (16 pairs) were collected from 16 subregions according to the size and structure of kiwifruit orchard to make captured images to be representative.

Ground truth of three-dimensional coordinates of calyxes was measured by a laser rangefinder (VCHON H-40, JinYun, China) on the experimental platform. Firstly, five level gauge and four micro height adjusting holders, as shown in Fig. 1(a2) and Fig. 1(b1), were applied to adjust the experiment platform to be horizontal, respectively. Secondly, three-dimensional coordinates of the camera's left lens were recorded as coordinate system origin. Thirdly, ten fruits in field view of cameras were randomly selected and the midpoints of their calyxes were located using the laser rangefinder. Fourthly, three-dimensional coordinates of kiwifruit were recorded in combination with the coordinate paper, as shown in Fig. 1. Finally, the above steps were repeated by shifting camera position.

Meanwhile, 20 pairs of calibration board images were taken with the binocular camera. These images were employed to calibrate the camera's internal references, and thus to enable spatial coordinate transformation. The camera was calibrated using the calibration method proposed by Zhang (1999). Examples of kiwifruit images and calibration board images captured by the binocular camera are shown in Fig. 2. Images from the left lens of binocular camera will be represented as the left images and images from the right lens will be represented as the right images. Different environments were simulated by changing shooting areas and angles to make collected data representative during different natural light.

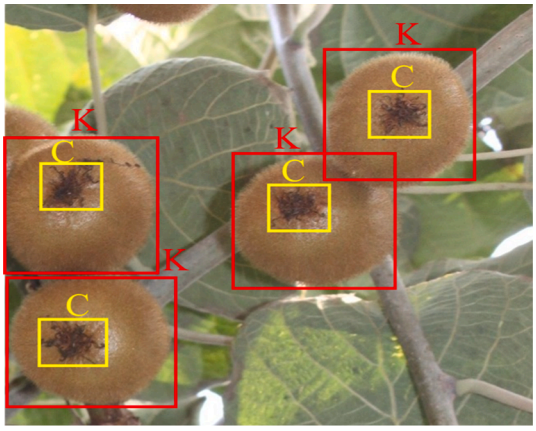


Fig. 3. Labeling examples of kiwifruit and calyx. Kiwifruit is labeled as "K" using a red rectangle, and calyx is labeled as "C" using a yellow rectangle. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In this work, both the internal and external parameters of the camera are accurately calculated by the calibration method. The coordinate of a main point is (u_0, v_0). U-axis scale factor and V-axis scale factor are f_x and f_y, respectively. And array kc is a distortion parameter of the camera, and five parameters in the array correspond to the internal parameters are listed in Table 1. The Rotation vector (R) and Translation vector (T) are the external parameters of the camera, which are the relative rotation and translation of the camera as shown in Table 2. And baseline of the camera is 120 mm.

2.2. Image dataset and annotation

A total of 1250 pairs of images (i.e., 2500 images) were utilized for YOLOv5 network training and testing. With LabelImg1.8 (https://github.com/tzutalin/labelImg), kiwifruit and calyx in the images were manually annotated as rectangles with labels "K" and "C", respectively, which were saved as annotation files in corresponding 'xml' files. Labeling examples of kiwifruit and calyx are shown in Fig. 3. Annotated dataset was randomly divided into training set (1750 images), validation set (500 images), and testing set (250 images) at a ratio of 7:2:1 to ensure reliability of later evaluation indicators.

2.3. Data augmentation

A small number of the images are prone to overfitting, resulting in an unsatisfactory detection accuracy. To improve overall learning procedure and performance, data augmentation techniques were used to artificially enlarge the number of images in the training set. Data augmentation methods involve image rotation in 90°, 180°, 270° and image mirroring in horizontal and vertical axis, thereby increasing the number of images in the training set from 1750 to 10500. The original and augmented images datasets are available on https://github.com/

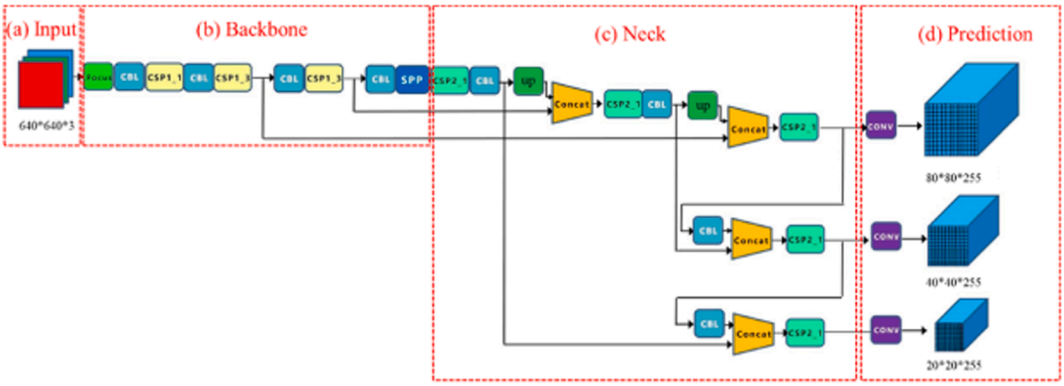


Fig. 4. Architecture of YOLOv5x. It contains four parts, i.e., (a) Input, (b) Backbone: CSPNet, (c) Neck: FPN + PANet, and (d) Prediction.

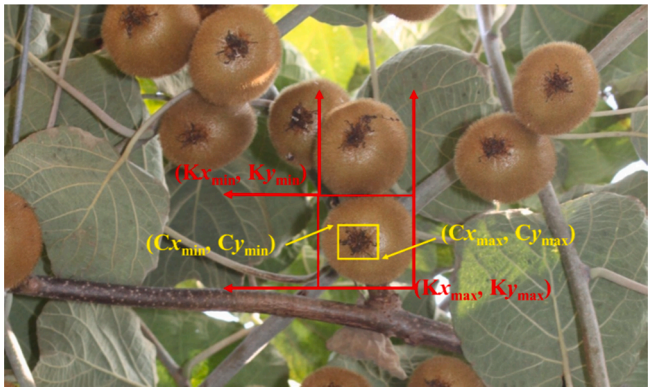


Fig. 5. Meaning of parameters in the rectangle. "K" and "C" represent kiwifruit and calyx, respectively. Then, x_{min} and y_{min} represent distances from the upper left corner of a detection rectangle to left edge and the top edge of the image, respectively. Besides, x_{max} and y_{max} represent distances from the bottom right corner of a detection rectangle to left edge and top edge of the image, respectively.

fu3lab/Kiwifruit_improved-binocular-localization_images.

2.4. Network architecture and training

Training architecture for kiwifruit and calyx detection based on YOLOv5x contains input, backbone, neck and prediction (Subedi et al., 2023), as shown in Fig. 4. The backbone mainly adopts focus structure, cross stage partial (CSP) structure and spatial pyramid pooling (SPP) structure (Bist et al., 2023). Neck section consists of FPN (Feature Pyramid Networks) and PAN (Path Aggregation Network). FPN adopts a top-down feature pyramid to transfer and connect high-level semantics information by means of up-sampling, thereby gaining a predicted feature map (Arifando et al., 2023). Meanwhile, PAN applies a bottom-up feature pyramid to transmit lower-level locating features through down sampling to enhance location information (José et al., 2023). In addition, the prediction of YOLOv5x generates feature maps of three different scales, enabling multi-scale predictions and enhancing the ability to predict small, medium and large targets (Nepal and Eslamiat, 2022).

In this study, the training platform was a desktop computer equipped with AMD R7-5800X (3.80 GHz) quad-core CPU, GeForce GTX 3080 Ti GPU (10240 CUDA cores), and 12 GB of memory, running on a Windows 10 64 bits system. Software tools included CUDA 10.1, CUDNN 11.0, and OpenCV 4.1. Experiments of this work were implemented in PyTorch framework. The network input size was 640 × 640 pixels, with a batch size of 16 and 350 epochs. Stochastic gradient descent was applied for training with a momentum of 0.937 and a weight decay of 0.0005. An initial value of 0.001 was set as learning rate of the network.

2.5. Pairing and matching of kiwifruit and calyx

After detecting kiwifruit and calyx, kiwifruit needs to be paired with corresponding calyx before matching calyx in the left image. Detection results of kiwifruit and calyx were described as detection rectangles using their coordinate values of two corner points, as shown in Fig. 5. And coordinate values of two corner points were defined as (x_{min}, y_{min}) and (x_{max}, y_{max}), respectively. According to size constraints of kiwifruit and calyx detection rectangles ($Cx_{min} \geq Kx_{min}$, $Cy_{min} \geq Ky_{min}$, $Cx_{max} \leq Kx_{max}$, and $Cy_{max} \leq Ky_{max}$), kiwifruit was successfully paired with corresponding calyx in the left image.

The whole process of calyx matching was based on pairing of kiwifruit and calyx, and template matching of kiwifruit, as shown in Fig. 6. Kiwifruit and calyx were paired in the left image. Then, paired kiwifruit in the left image was matched with that in the right image using template matching. Furthermore, the matched kiwifruit in the right image was paired with calyx detected by YOLOv5x in the right image thus to make calyxes matched in the binocular images.

A correlation coefficient template match (TM_CCOEFF) was employed to search for a similar patch between the matched detection rectangles from template images and target images. Specifically, patches that were divided from the matched detection rectangles were used to calculate the similarity, described as similarity matrixes, where the location of the extreme largest value indicates the most similar patch. The similarity (i.e., $R(x, y)$) of the TM_CCOEFF method is calculated using Eq. (1). $T'(x', y')$ represents pixel value in the matched detection rectangle from template image, as shown in Eq. (2). $I'(x', y')$ represents pixel value in matched detection rectangle from target image, as shown in Eq. (3).

$$R(x,y) = \sum_{x',y'} (T'(x',y') \times I'(x+x', y+y')) \tag{1}$$

$$T'(x',y') = T(x',y') - \frac{\sum T(x',y')}{w \times h} \tag{2}$$

$$I'(x',y') = I(x',y') - \frac{\sum I(x',y')}{w \times h} \tag{3}$$

2.6. Binocular localization of calyxes

Calyx is more visible and has less occlusion on the canopy image than kiwifruit and is an ideal option for a picking position. Thus, the calyx needs to be located and its three-dimensional coordinates are required based on binocular localization principle. The principle was applied to the matched calyxes to observe a feature point together in space on the left and right lens of MYNT EYE D1000-50, as shown in Fig. 7. O_l and O_r indicate the origins of left and right lenses coordinate systems,

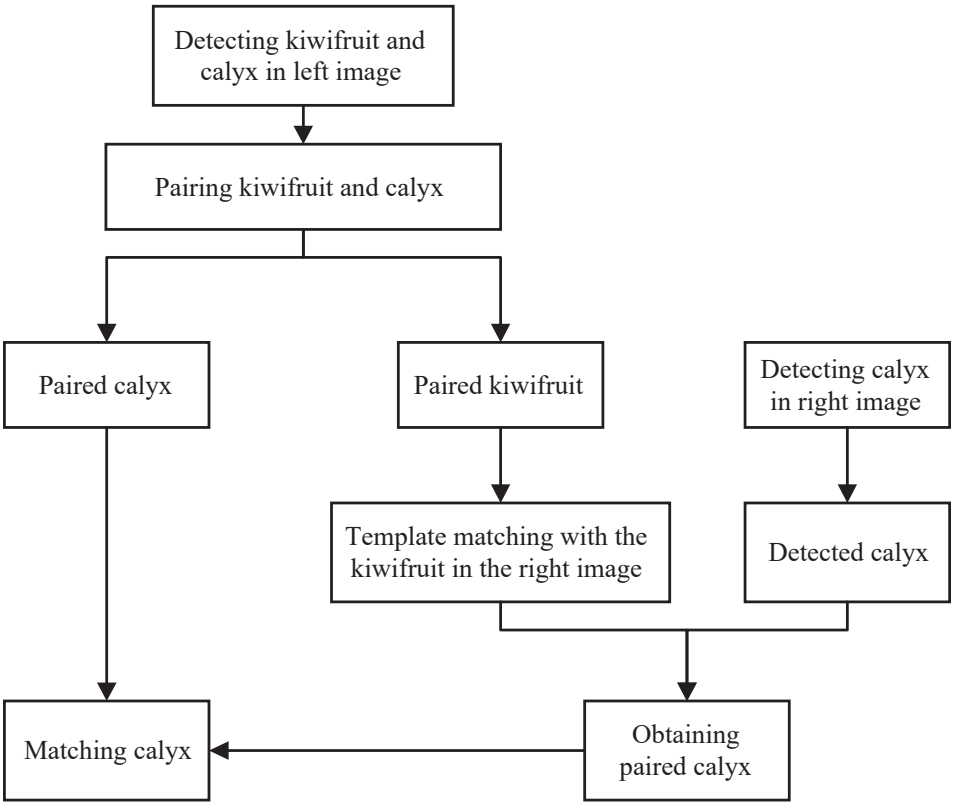


Fig. 6. Flowchart of calyx matching.

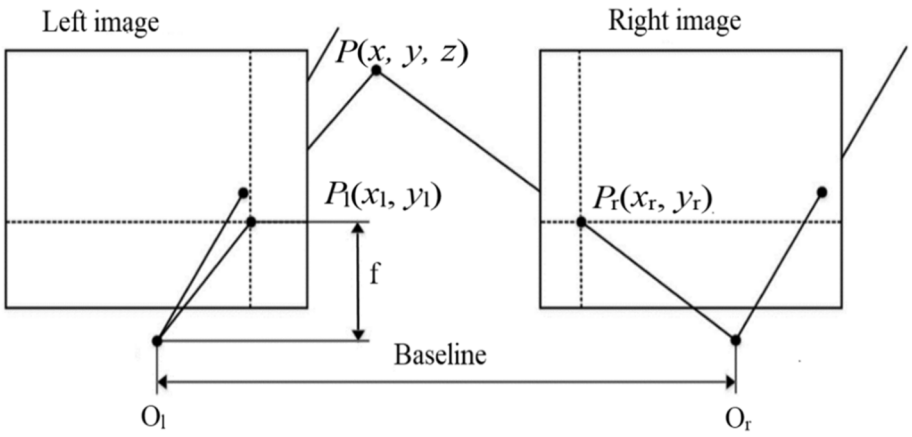


Fig. 7. Principle of binocular localization. Black intersection of line between P and origin O_l and imaging plane of the left image is P_l, while P_r is the intersection of P with origin O_r and the imaging plane of the right image.

respectively.

The feature point and depth value $P(x, y, z)$ were simultaneously detected by two lenses, and its three-dimensional coordinates contain X, Y, and Z. The projection of $P(x, y, z)$ on the left lens is $P_l(x_l, y_l)$ and on the right lens is $P_r(x_r, y_r)$. Eqs. (4)–(6) are specific expressions about the principle of similar triangles in plane geometry, where baseline represents the distance between O_l and O_r in Eq. (4), and f represents focal length of the binocular camera in Eq. (6). Overall, it can be seen from Eqs. (4)–(6) that the critical step for fruit localization is to find corresponding coordinates of the points in binocular images.

$$x = \frac{\text{Baseline}}{x_r - x_l} \times x_l \tag{4}$$

$$y = \frac{\text{Baseline}}{x_r - x_l} \times y_l \tag{5}$$

$$z = \frac{\text{Baseline}}{x_r - x_l} \times \text{f} \tag{6}$$

Three methods, i.e., calyx localization (CL) method, fruit localization (FL) method, and depth information from depth map (DIDM), were compared to obtain the most accurate three-dimensional coordinates of calyxes for kiwifruit robot harvesting. The difference between the CL method and the FL is that the CL utilized midpoints of calyx detection rectangles as feature points while the FL method employed the midpoints from connecting regions of calyxes as feature points after adaptive threshold binarization to detection rectangles of kiwifruit and calyx.

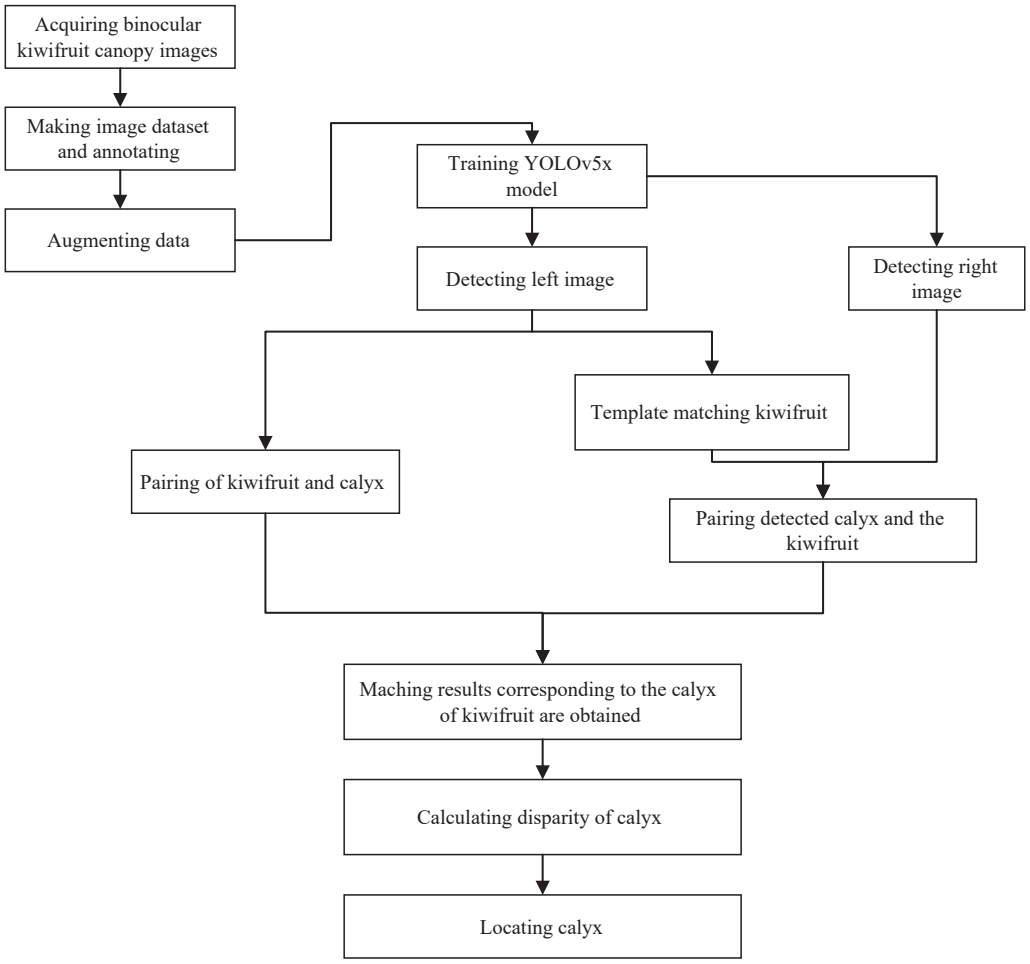


Fig. 8. Flowchart of the calyx (CL) method.

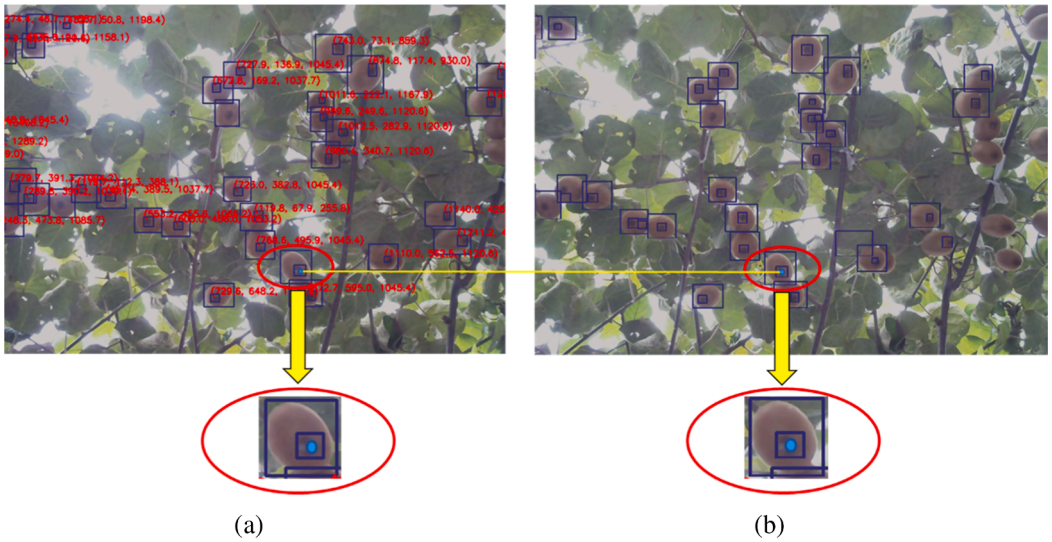


Fig. 9. Examples of feature points selected for localization using the CL method. Red ellipse in binocular images are manually drawn to magnify manually drawn blue feature points while a yellow line represents disparity in binocular images. (a) Left image; (b) Right image. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Both of them are based on the principle of binocular localization described in Section 2.6 to calculate depth while depth information from DIDM was obtained using Semi-Global Block Matching (SGBM) method (Hirschmüller, 2005). For DIDM, midpoint coordinates from detection rectangles of calyxes in left images were calculated. Then, the midpoint coordinates were utilized to extract depth values from a depth map, which was generated by the binocular camera through the SGBM. The midpoint coordinates were transformed using Eq. (4) and Eq. (5) and

Fig. 10. Examples of visual localization results using the FL method. In kiwifruit detection rectangles, black connecting region surrounded by white connecting region represents the calyx. Red numbers represent the three-dimensional coordinates of calyxes while connecting region represents the kiwifruit and the rest are background. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

then integrated with extracted depth values into three-dimensional coordinates of calyxes.

For the CL method, calyxes were firstly matched in binocular images based on pairing of kiwifruit and calyx and template matching of kiwifruit, which is same as in Section 2.5. Moreover, the midpoints from detection rectangles of the matched calyxes were selected as feature points and thus utilized to obtain three-dimensional coordinates using Eqs. (4)–(6). Fig. 8 depicts the flowchart of the CL method. Examples of feature points selected for localization in the CL method are shown in Fig. 9.

Contrarily to the CL method, the FL method binarized pixels in kiwifruit detection rectangles based on an adaptive threshold, as shown in Fig. 10. For a binarized kiwifruit detection rectangle, a black connecting region surrounded by a white connecting region represents calyx while the white connecting region represents the kiwifruit. Then midpoints of the calyx were used as feature points and thus utilized to obtain three-dimensional coordinates using Eqs. (4)–(6). Meanwhile, for those binarized kiwifruit detection rectangles that does not have the black connecting region surrounded by white connecting region partly or completely, midpoints of kiwifruit detection rectangles is utilized for localization.

2.7. Performance evaluation

Performance of kiwifruit and calyx detection was evaluated by precision (*P*), recall (*R*), average precision (*AP*), mean average precision (*mAP*) and average detection speed. *R* was a measure value of how many truly relevant detection results returned, while *P* was a measure value of detection results relevancy. *P* and *R* are defined as shown in Eq. (7) and Eq. (8), respectively. If a labeled kiwifruit or calyx was the same as detected kiwifruit or calyx, it was true positive (*TP*). If a kiwifruit or calyx was labeled but detected differently, it was false negative (*FN*). If a kiwifruit or calyx did not exist but was detected, it was false positive (*FP*).

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

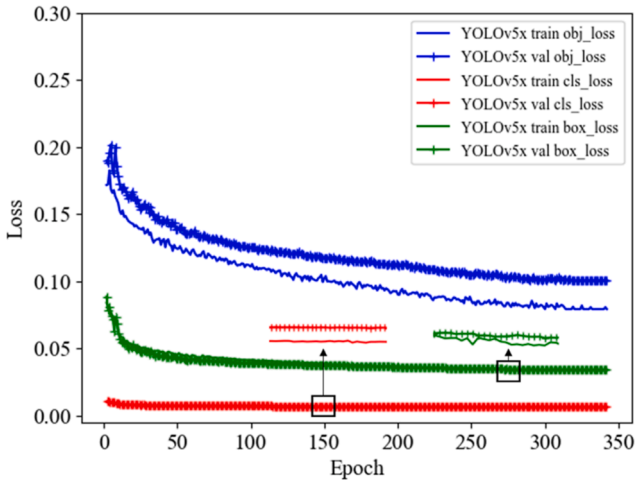


Fig. 11. Loss curves of YOLOv5x. Blue, red, and green curves represent obj-loss, cls-loss, and box-loss curves of YOLOv5x, respectively. Black rectangles represent local enlargement curves. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The integration of *AP* based on accuracy is defined as shown in Eq. (9). The *AP* is defined as the area under P-R curve, which aims to evaluate the performance of model in detecting each class. Meanwhile, the *mAP* is defined as mean *AP* of the two classes (i.e., kiwifruit and calyx) in Eq. (10). The value *k* indicates the two classes of objects.

$$AP = \int_0^1 P_{(R)} dR \tag{9}$$

$$mAP = \frac{1}{k} \sum_{i=1}^{k} AP_i \tag{10}$$

$$SD = \sqrt{\sum_{i=1}^{n} (D_i - D')^2 / n} \tag{11}$$

Localization accuracy was evaluated by calculating the standard

Table 3
Results of kiwifruit and calyx detection by YOLOv4, YOLOv5s and YOLOv5x.

Model	*AP* (%) Kiwifruit	Calyx	*mAP* (%)	Detection speed (ms / image)
YOLOv4	99.2	79.8	89.5	21.5
YOLOv5s	99.3	84.5	91.9	20.8
YOLOv5x	99.5	87.4	93.5	108.0

deviation (*SD*), as shown in Eq. (11). Where D_i indicates the value of depth obtained by localizing each feature point, $\bar{D}$ refers to the average deviation of kiwifruit, and n represents the number of kiwifruit. Besides, detection speed was calculated to evaluate the performance of YOLOv5x model.

3. Results and discussion

3.1. Training evaluation

As mentioned before, a two-class training dataset of kiwifruit binocular images was applied to train YOLOv5x model under PyTorch framework. There are six loss curves of YOLOv5x, as shown in Fig. 11, where different colors and line types represent different losses in training (train) or validation (val) set. Fig. 11 illustrates results of loss curves of the training set and validation set for 350 epochs. The values of classification loss (cls_loss), bounding box loss (box_loss) and object loss (obj_loss) decreased as the number of epochs increased, but generally stabilized when the number of epochs reached 350, and gradually approached the lowest value of 0.012, 0.043, and 0.15, respectively. Training results demonstrated that YOLOv5x adopted in this work efficiently learned the features with good convergence ability, and the model trained can be used for detecting kiwifruit and calyx.

3.2. Kiwifruit and calyx detection

YOLOv5x have shown promising results on kiwifruit and calyx detection. As can be seen from Table 3, YOLOv5x achieved a highest *AP* of 87.4 % compared with YOLOv4 (79.8 %) and YOLOv5s (84.5 %) on calyx detection. All three models reached acceptable *AP* of over 99 % on kiwifruit detection. Notably, there was no false positive on kiwifruit detection, as shown in Fig. 12. Deep learning has been widely employed to detect kiwifruit. Liu et al. (2020) adopted the Image-Fusion method with VGG16, which reached the highest *AP* of 90.7 % on kiwifruit detection. Zhou et al. (2020) applied MobileNetV2 to detect the images of kiwifruit, which obtained the *AP* of 90.8 %. Williams et al. (2020) reached the *AP* of 94.0 % on kiwifruit detection and 91.0 % on calyx detection. Whereas, the *mAP* of 93.5 % on kiwifruit and calyx detection was obtained by YOLOv5x. Compared to YOLOv4 and YOLOv5s, YOLOv5x seems the most suitable on kiwifruit and calyx detection.

YOLOv5x showed an acceptable accuracy in detecting kiwifruit and calyx, although the difference of detection accuracy between kiwifruit and calyx exceeds 10 %. *P-R* curves by YOLOv5x model on the testing dataset are shown in Fig. 13. As can be seen from it, the *AP* on kiwifruit detection is higher than that of calyx detection, which may be due to calyx takes smaller proportion in the image than kiwifruit (examples shown in Fig. 12). Besides, it makes calyx to take less pixels thus to lower detection accuracy of calyx than before when the images were down sampling from 1280 × 720 to 640 × 640 (network input size) pixels.

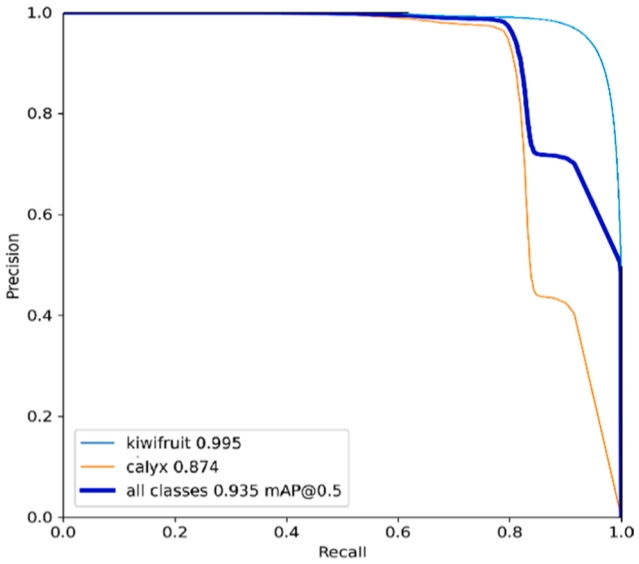


Fig. 13. *P-R* curves of YOLOv5x model on the testing dataset.

Fig. 12. Examples of kiwifruit and calyx detection by YOLOv5x. "K" and "C" represent kiwifruit and calyx, respectively.

Fig. 14. Examples of kiwifruit paired with calyx in the left image. Paired kiwifruit and calyx are marked as the same colors in the detection rectangles. The red circle represents pairing failed calyxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Detection speed

Average detection speed of YOLOv5x on detecting kiwifruit in an image with 1280 × 720 pixels was 108 ms, which can meet requirement for real-time kiwifruit picking. YOLOv4 and YOLOv5s were employed to detect kiwifruit in an image of 1280 × 720 pixels resolution, which took an average detection speed of 21.5 ms and 20.8 ms. This is due to the fact that YOLOv5x has more convolutional layers, which costs more computation time. Picking efficiency is not affected when the detection speed is faster than picking speed because picking turn goes on round by round. It is acceptable for all three models in terms of detection speed, as they are shorter than picking speed of current kiwifruit robots, which take approximately 2.16 s to pick a kiwifruit using a single-armed gripper (Williams et al., 2020). Therefore, it is undeniable that the detection speed of YOLOv5x can meet requirement for real-time kiwifruit detection. Considering accuracy and speed of all the three models, YOLOv5x was chosen to detect kiwifruit and calyxes in our following study.

It should be noted that traditional image processing method for kiwifruit and calyx detection in small area are accurate but time-consuming. Fu et al. (2017) calculated the V (Value) component of the HSV (Hue, Saturation, Value) color model for detecting the calyx from kiwifruit, which achieved correct detection rate of 94.3 % and cost 1.16 s to detect an image with 640 × 360 pixels, which is much longer than the detection of a 1280 × 720 pixel image using any of the three deep learning models.

3.4. Pairing and matching of kiwifruit and calyx

The pairing of kiwifruit and calyx lays foundation of calyx matching.

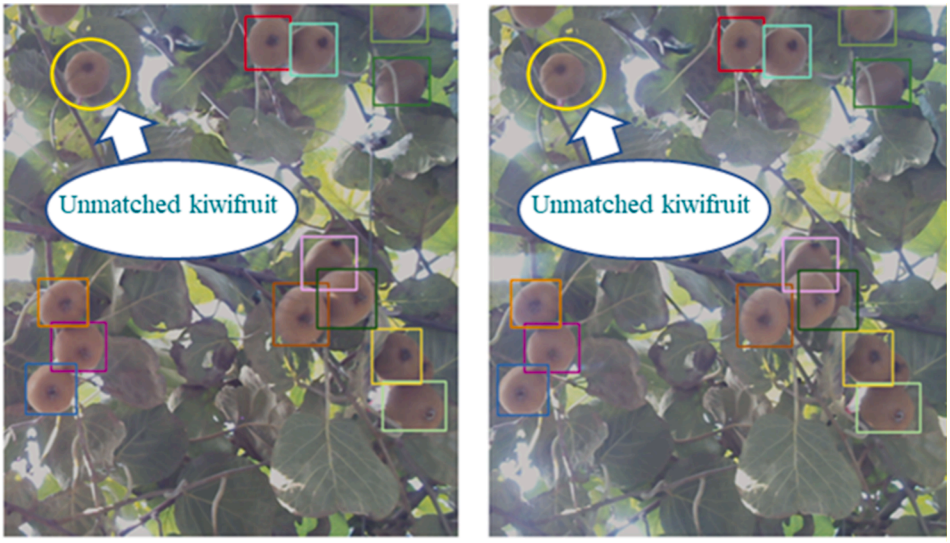


Fig. 15. Examples of kiwifruit detection and matching results in binocular images. (a) Kiwifruit detection results in the left image; (b) Kiwifruit matching results in the right image. The rectangles of matched kiwifruit in the right image are given the same colors as the left image. The yellow circles emphasize unmatched kiwifruit. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

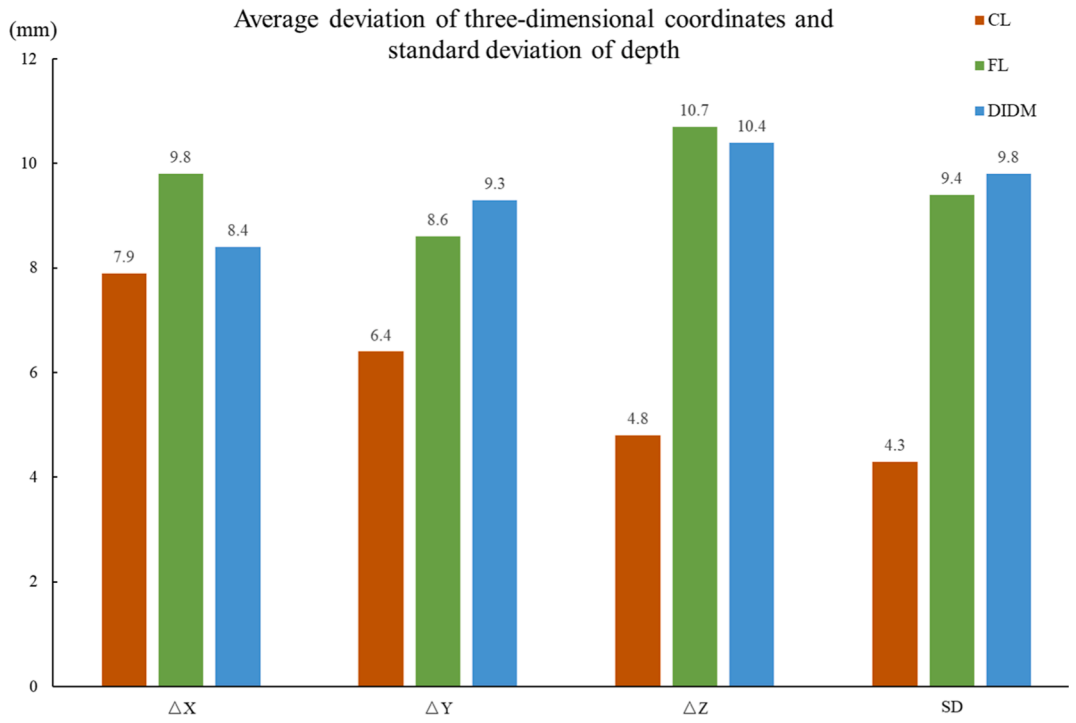


Fig. 16. Average deviation of three-dimensional coordinates and standard deviation of depth obtained by three different methods. △X, △Y, and △Z represent the average deviation of X-axis, Y-axis, and Z-axis, respectively. *SD* refers to the standard deviation of depth obtained by three different methods based on binocular localization principle.

As shown in Fig. 14, paired kiwifruit and calyx were marked using the detection rectangles with the same color. It should be mentioned that the average success pairing rate of kiwifruit and calyx is 86.9 %. Failed pairing was due to non-detection of the calyx and was illustrated using a manually drawn red circle. Thus, pairing success rate of the calyx was supported based on kiwifruit detection (*AP* of 99.5 %) and calyx detection (*AP* of 87.4 %).

Moreover, matching accuracy was promising for calyx localization. Despite the fact that the matching accuracy of the kiwifruit was 86.6 %, it is still higher than the results presented in previously similar studies. Scafe (2012) adopted a theoretical fruit removal process to match kiwifruit, which obtained matching accuracy of 83.56 %. And Wang et al. (2017) used circle Hough transform (CHT) to match litchi fruit, which obtained matching accuracy of 80.04 %. Matching accuracy was connected with detection results and thus pairing results. The detection rectangles of kiwifruit in the right image in Fig. 15(b) were given same color shown in the left image if kiwifruit were matched correctly. A failed matching example was circled using yellow ellipse in Fig. 15 and showed that kiwifruits were not matched if a kiwifruit was not detected and paired with its calyx.

3.5. Accuracy of binocular localization

The CL method has highest localization accuracy than the two other methods, which is acceptable for kiwifruit harvesting. The average deviation of three-dimensional coordinates and standard deviation of depth were calculated using three different methods, as shown in Fig. 16. The average deviation of X-axis, Y-axis, and Z-axis using the CL method were 7.9 mm, 6.4 mm, and 4.8 mm, respectively, which can meet the design requirement of 25 mm tolerance for kiwifruit harvesting robot arm (Mu et al., 2020). Compared with the results of the FL and DIDM methods, the localization error rate of the CL method was reduced by 55.1 % and 53.8 %, respectively.

The larger localization error of the FL method and the DIDM method compared to CL may be due to the fact that the DIDM method employed the SGBM based on global matching while the FL method was based on the traditional image processing algorithms. Owing to the similar color and texture of fruit, global matching result in a large amount of mismatching feature points in the complex orchard and thus cause low localization accuracy (Tang et al., 2020). The traditional image processing algorithms capture whole calyxes hardly although it was applied in detection rectangles. Therefore, it is practical to adopt the CL method to locate calyxes as picking positions.

Table 4
results from previous studies on kiwifruit localization.

	localization methods	objects	environments	localization error (mm)
Scarfe (2012)	classical binocular	kiwifruit	orchard	16.0
Liu (2020)	structured light	kiwifruit	indoor	3.6
Song (2021)	classical binocular	kiwifruit	orchard	10.4
Our method	improved binocular	kiwifruit and calyx	orchard	4.8

3.6. Results from other studies on kiwifruit localization

In our work, calyx localization was proposed for fruit picking while other methods focused on locating kiwifruit itself. As shown in Table 4., Scarfe (2012) achieved absolute distance error of 16 mm on kiwifruit localization based on stereo matching. Liu, 2020 obtained the smallest average error of 3.6 mm of kiwifruit localization, which was implemented indoors and may not be suitable in orchard. Song, 2021 acquired depth values from midpoints of kiwifruit detection rectangles, which achieved average error of 10.4 mm. Considering the findings from previous research on kiwifruit localization, our study achieved an average deviation of 4.8 mm in calyx localization using the CL method, which is likely to be the best outdoor localization error reported so far

for kiwifruit robotic picking. In light of these outcomes, it is reasonable to conclude that the improved binocular localization method of calyxes based on deep learning proposed in this paper is promising for locating and picking fruits.

4. Conclusions

This study proposed an improved binocular localization method of calyxes based on deep learning named CL to accurately detect and locate kiwifruit for robotic harvesting. The CL method achieved the best localization performance compared with FL method and DIDM method. Specifically, YOLOv5x achieved the desired results of kiwifruit and calyx detection. A few feature points were selected from detection rectangles generated by YOLOv5x and improve matching accuracy of kiwifruit in binocular images. In terms of calyx matching, it was found that the kiwifruit and calyx pairing facilitated the calyx matching, where the calyx matching effectively avoided feature points mismatching in global matching method. The CL method based on YOLOv5x could reflect actual localization of kiwifruit, which is conducive to the operation of picking robots. Furthermore, matching calyxes directly is unacceptable because there are a lot of missing calyxes when detecting calyxes. Further improvement on calyx detection is needed such as enhancing model performance and utilizing the state-of-the-art network due to detection accuracy of calyx is lower than that of kiwifruit.

CRediT authorship contribution statement

Changqing Gao: Data curation, Investigation, Methodology, Writing – original draft. **Hanhui Jiang:** Data curation, Investigation, Methodology, Writing – original draft. **Xiaojuan Liu:** Data curation, Investigation, Methodology, Writing – original draft. **Haihong Li:** Methodology, Writing – review & editing. **Zhenchao Wu:** Methodology, Supervision, Writing – review & editing. **Xiaoming Sun:** Conceptualization, Methodology, Writing – review & editing. **Leilei He:** Methodology, Supervision, Writing – review & editing. **Wulan Mao:** Investigation, Methodology, Writing – review & editing. **Yaqoob Majeed:** Investigation, Methodology, Writing – review & editing. **Rui Li:** Conceptualization, Methodology, Writing – review & editing. **Longsheng Fu:** Conceptualization, Data curation, Methodology, Supervision, Writing – review & editing.

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 Natural Science Foundation of China (32371999); Key Research and Development Program of Shaanxi (Program No. 2023JBGS-21); National Foreign Expert Project, Ministry of Science and Technology, China (DL2022172003L, QN2022172006L).

References

Arifando, R., Eto, S., Wada, C., 2023. Improved YOLOv5-based lightweight object detection algorithm for people with visual impairment to detect buses. Appl. Sci. 13, 5802. https://doi.org/10.3390/app13095802.

Bist, R.B., Yang, X., Subedi, S., Chai, L., 2023. Mislaying behavior detection in cage-free hens with deep learning technologies. Poult. Sci. 102, 102729 https://doi.org/10.1016/j.psj.2023.102729.

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

Ding, J., Yan, Z., We, X., 2021. High-accuracy recognition and localization of moving targets in an indoor environment using binocular stereo vision. ISPRS Int. J. Geo-Inf. 10 (4), 234. https://doi.org/10.3390/ijgi10040234.

Fu, L., Sun, S., Manuel, V., Li, S., Li, R., Cui, Y., 2017. Kiwifruit recognition method at night based on fruit calyx image. Trans. Chin. Soc. Agric. Eng. 33 (2), 199–204. https://doi.org/10.11975/j.issn.1002-6819.2017.02.027.

Fu, L., Tola, E., Al-Mallahi, A., Li, R., Cui, Y., 2019. A novel image processing algorithm to separate linearly clustered kiwifruits. Biosyst. Eng. 183, 184–195. https://doi.org/10.1016/j.biosystemseng.2019.04.024.

Fu, L., Feng, Y., Wu, J., Liu, Z., Gao, F., Majeed, Y., Al-Mallahi, A., Zhang, Q., Li, R., Cui, Y., 2020a. Fast and accurate detection of kiwifruit in orchard using improved YOLOv3-tiny model. Precis. Agric. 22 (3), 754–776. https://doi.org/10.1007/s11119-020-09754-y.

Fu, L., Gao, F., Wu, J., Li, R., Karkee, M., Zhang, Q., 2020b. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. Comput. Electron. Agric. 177, 105687 https://doi.org/10.1016/j.compag.2020.105687.

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J., Morros, J., Ruiz-Hidalgo, J., Vilaplana, V., Gregorio, E., 2020. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. Comput. Electron. Agric. 169, 105165 https://doi.org/10.1016/j.compag.2019.105165.

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

Hirschmüller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutua information. In: 2005 IEEE Comp. Society Conference on Comp. Vision and Pattern Recognition (CVPR). 807–814. https://doi.org/10.1109/CVPR.2005.56.

Hsieh, K., Huang, B., Hsiao, K., Tuan, Y., Shih, F., Hsieh, L., Chen, S., Yang, I., 2021. Fruit maturity and location identification of beef tomato using R-CNN and binocular imaging technology. J. Food Meas. Charact. 15, 5170–5180. https://doi.org/10.1007/s11694-021-01074-7.

Hu, X., Bowen, N., Chai, J., 2019. Research on the location of citrus picking point based on structured light camera. In: 2019 IEEE 4th Int. Conference on Image, Vision and Computing (ICIVC). 567–571. https://doi.org/10.1109/ICIVC47709.2019.8980938.

José, B., Frizzo, S., Singh, G., Zanetti, R., 2023. Hybrid-YOLO for classification of insulators defects in transmission lines based on UAV. Int. J. Electr. Power Energy Syst. 148, 108982 https://doi.org/10.1016/j.ijepes.2023.108982.

Li, T., Fang, W., Zhao, G., Gao, F., Wu, Z., Li, R., 2023b. An improved binocular localization method for apple based on fruit detection using deep learning. Inf. Process. Agric. 10 (2), 276–287. https://doi.org/10.1016/j.inpa.2021.12.003.

Li, Q., Sun, X., Jiang, H., Wu, A., Fu, L.R., 2023a. Design and test of intelligent spraying unmanned vehicle for greenhouse tomato based on YOLOv4-tiny. J. Intell. Agric. Mech. 4 (2), 44–52. https://doi.org/10.12398/j.issn.2096-7217.2023.02.005.

Liu, Z., 2020. Kiwifruit detection and localization methods based on multi-source information fusion. Master Thesis, Northwest A&F University, Shaanxi, China. 10.27409/d.cnki.gxbnu.2020.000944.

Liu, T., Kang, H., Chen, C., 2023a. ORB-Livox: A real-time dynamic system for fruit detection and localization. Comput. Electron. Agric. 209, 107834 https://doi.org/10.1016/j.compag.2023.107834.

Liu, T., Nie, X., Wu, J., Zhang, D., Liu, W., Cheng, Y., Zheng, Y., Qiu, J., Qi, L., 2023b. Pineapple (Ananas comosus) fruit detection and localization in natural environment based on binocular stereo vision and improved YOLOv3 model. Precis. Agric. 24, 139–160. https://doi.org/10.1007/s11119-022-09935-x.

Liu, Z., Wu, J., Fu, L., Majeed, Y., Feng, Y., Li, R., Cui, Y., 2020. Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion. IEEE Access 8, 2327–2336. https://doi.org/10.1109/ACCESS.2019.2962513.

Mehranian, A., Kotasidis, F., Zaidi, H., 2016. Accelerated time-of-flight (ToF) PET image reconstruction using ToF bin subsetization and ToF weighting matrix pre-computation. Phys. Med. Biol. 61 (3), 1309–1331. https://doi.org/10.1088/0031-9155/61/3/1309.

Mehta, S., Burks, T., 2014. Vision-based control of robotic manipulator for citrus harvesting. Comput. Electron. Agric. 102, 146–158. https://doi.org/10.1016/j.compag.2014.01.003.

Mehta, S., Ton, C., Asundi, S., Burks, T., 2017. Multiple camera fruit localization using a particle filter. Comput. Electron. Agric. 142, 139–154. https://doi.org/10.1016/j.compag.2017.08.007.

Mu, L., Cui, G., Liu, Y., Cui, Y., Fu, L., Gejima, Y., 2020. Design and simulation of an integrated end-effector for picking kiwifruit by robot. Inf. Process. Agric. 7 (1), 58–71. https://doi.org/10.1016/j.inpa.2019.05.004.

Nepal, U., Eslamiat, H., 2022. Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. Sensors 22 (2), 464. https://doi.org/10.3390/s22020464.

Neupane, C., Koirala, A., Wang, Z., Walsh, K., 2021. Evaluation of depth cameras for use in fruit localization and sizing: Finding a successor to Kinect v2. Agronomy 11, 1780. https://doi.org/10.3390/agronomy11091780.

Robin, C., Lacroix, S., 2016. Multi-robot target detection and tracking: taxonomy and survey. Auton. Robots 40, 729–760. https://doi.org/10.1007/s10514-015-9491-7.

Scarfe, A., 2012. Development of an autonomous kiwifruit harvester. Massey University, Manawatu, New Zealand. Doctor Thesis,.

Song, Z., Zhou, Z., Wang, W., Gao, F., Fu, L., Li, R., Cui, Y., 2021. Canopy segmentation and wire reconstruction for kiwifruit robotic harvesting. Comput. Electron. Agric. 181, 105933 https://doi.org/10.1016/j.compag.2020.105933.

Subedi, S., Bist, R., Yang, X., Chai, L., 2023. Tracking pecking behaviors and damages of cage-free laying hens with machine vision technologies. Comput. Electron. Agric. 204, 107545 https://doi.org/10.1016/j.compag.2022.107545.

Song, Z., 2021. Kiwifruit canopy image segmentation and multi-classes fruit localization methods based on deep learning. Master Thesis, Northwest A&F University, Shaanxi, China. 10.27409/d.cnki.gxbnu.2021.000573.

Suo, R., Gao, F., Zhou, Z., Fu, L., Song, Z., Dhupia, J., Li, R., Cui, Y., 2021. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking. Comput. Electron. Agric. 182, 106052 https://doi.org/10.1016/j.compag.2021.106052.

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.

UN Food & Agriculture Organization Production/Yield quantities of kiwi fruit in World https://www.fao.org/faostat/zh/#data/QCL/visualize 2022 Retrieved 2022–08-12, from.

Wang, C., Tang, Y., Zou, X., Luo, L., Chen, X., 2017. Recognition and matching of clustered mature litchi fruits using binocular charge-coupled device (CCD) color cameras. Sensors 17 (11), 2564. https://doi.org/10.3390/s17112564.

Wang, C., Luo, T., Zhao, L., Tang, Y., Zou, X., 2019. Window zooming-based localization algorithm of fruit and vegetable for harvesting robot. IEEE Access 7, 103639–103649. https://doi.org/10.1109/ACCESS.2019.2925812.

Williams, H., Jones, M., Nejati, M., Seabright, M., Bell, J., Penhall, N., Barnett, J., Duke, M., Scarfe, A., Ahn, H., Lim, J., MacDonald, B., 2019. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms. Biosyst. Eng. 181, 140–156. https://doi.org/10.1016/j.biosystemseng.2019.03.007.

Williams, H., Ting, C., Nejati, M., Jones, M., Penhall, N., Lim, J., Seabright, M., Bell, J., Ahn, H., Scarfe, A., Duke, M., MacDonald, B., 2020. Improvements to and large-scale evaluation of a robotic kiwifruit harvester. J. Field Robot. 37 (2), 187–201. https://doi.org/10.1002/rob.21890.

Xiao, Z., Luo, L., Chen, M., Wang, J., Lu, Q.L.S., 2023. Detection of grapes in orchard environment based on improved YOLO-v4. J. Intell. Agric. Mech. 4 (2), 35–43. https://doi.org/10.12398/j.issn.2096-7217.2023.02.004.

Zhang, Z., 1999. Flexible camera calibration by viewing a plane from unknown orientations. In: Proceedings of the Seventh IEEE Int. Conference on Comp. Vision (ICCV) 666–673. https://doi.org/10.1109/ICCV.1999.791289.

Zhang, J., Zhang, Y., Wang, C., Yu, H., Qin, C., 2021. Binocular stereo matching algorithm based on MST cost aggregation. Math. Biosci. Eng. 18 (4), 3215–3226. https://doi.org/10.3934/mbe.2021160.

Zhou, Z., Song, Z., Fu, L., Gao, F., Li, R., Cui, Y., 2020. Real-time kiwifruit detection in orchard using deep learning on Android™ smartphones for yield estimation. Comput. Electron. Agric. 179, 105856 https://doi.org/10.1016/j.compag.2020.105856.