Enhanced precision in greenhouse tomato recognition and localization: A study leveraging advances in Yolov5 and binocular vision technologies

Shuangyou Wang[1], Zhanying Shao[2,*], Yongjian Zhang[3]

[1]School of Software, Handan University, Handan, China; [2]Hebei Construction Material Vocational and Technical College, Qinhuangdao, China; [3]School of Information and Electrical Engineering, Hebei University of Engineering, Handan, China

***Corresponding Author:** Zhanying Shao, Hebei Construction Material Vocational and Technical College, Qinhuangdao, China. Email: shaokunpeng8@163.com

Received: 8 February 2024; Accepted: 16 August 2024; Published: 29 August 2024

RESEARCH ARTICLE

Abstract

The core challenge in realizing automatic tomato harvesting in greenhouse environments lies in the precise identification and localization of the fruits. This paper introduces a comprehensive approach based on an improved YOLOv5 detection algorithm and optimized binocular stereo vision technology. Firstly, by introducing the C3-Transformer Encoder (CTM) structure and Bidirectional Feature Pyramid Network (Bi-FPN), this study enhanced the model's ability to recognize tomatoes, especially under complex backgrounds and occlusion conditions. After field testing, the mAP50 accuracy reached 97.1%, an increase of 1.2 percentage points, enhancing detection precision. In addition, the ZED binocular camera was used, and the census stereo matching algorithm was optimized, significantly reducing disparity errors, thereby improving the accuracy of depth information. This allows the model to accurately calculate the three-dimensional spatial position of tomatoes obscured by branches and leaves, greatly improving the efficiency of the harvesting robot. Through field debugging verification with the harvesting robot, the method proposed in this study has shown high accuracy and reliability in the recognition and localization of tomatoes in complex greenhouse environments.

Keywords: Improved YOLOv5; Tomato Recognition and Detection; Binocular Camera; Stereo Matching

Introduction

With modernization of agriculture, intelligent and automated picking robots are playing an increasingly important role in agricultural production. As tomatoes are a common crop in greenhouses, manual picking requires significant labor and high costs. Therefore, research on tomato picking robots has broad application prospects and practical significance. However, the harvesting of tomatoes mainly relies on manual labor, which is labor-intensive and costly, creating a pressing need for agricultural robots capable of automated picking. Currently, the technical bottleneck limiting harvesting robots primarily revolves around the recognition and localization of targets. Especially, in greenhouses, the growth posture of tomato fruits varies, with fruits overlapping, and severe occlusions caused by leaves, branches, and peduncles, combined with complex lighting environments and backgrounds. These factors pose significant challenges for the recognition and localization capabilities of harvesting robots. Therefore, the rapid and precise recognition and localization of tomatoes in complex greenhouse environments is a critical issue that needs to be addressed in the development of tomato-harvesting robots.

Related Research

Fruit and vegetable recognition and detection technology

Target detection methods based on machine vision have been widely applied to harvesting robots worldwide. In recent years, scholars both domestically and internationally have conducted research on the recognition and detection of fruits and vegetables.

For instance, a computer vision system is designed that enhances image contrast using the R component of RGB images and employs the Sobel operator for edge detection (Arianti *et al*., 2023; Benavides *et al*., 2020; Naik and Thimmaiah, 2021; Ratha *et al*., 2023; Rajpoot *et al*., 2022; Selçuk and Tütüncü, 2023; Trinh and Nguyen, 2023; Wang *et al*., 2023; Zhang, 2023). Based on grayscale or intensity segmentation, it eliminates the effects of image noise and shadows, applies dilated morphology, and finally achieves the detection of tomato fruits through segmentation by size, as shown in Figure 1. The detection rate for cluster tomatoes is 87.5%, and for beef tomatoes, it is 80.8%.

Stereo cameras were selected to collect images in greenhouses, establish a dataset of red and green tomatoes, and propose SSD and YOLO deep learning models (Magalhaes *et al*., 2021). Magalhaes *et al*. (2021) compared five deep learning models: SSD MobileNet v2, SSD Inception v2, SSD ResNet 50, SSD ResNet 101, and YOLOv4 Tiny. The results showed that SSD MobileNet v2 performed the best, with a mean average precision (mAP) of 51.46%, an F1 score of 66.15, and an inference time of 16.44 ms. However, improvements are needed to address the issue of occlusion by branches and leaves. They used Intel D435 cameras to collect images in greenhouses, and the Mask-RCNN method was applied to detect greenhouse tomatoes (Afonso Manya *et al*., 2022). They conducted a comparative analysis of the detection capabilities of ResNext 101, ResNet 50, and ResNet 101 network architectures, and the results showed that the ResNext 101 architecture achieved a prediction accuracy of 88% and a recall rate of 91%. They employed the deep learning Mask-RCNN algorithm with a backbone network of Resnet50, combined with a Feature Pyramid Network (FPN) architecture to extract features, and used a Region Proposal Network to complete the selection of candidate boxes, achieving automatic detection of strawberries (Yu *et al*., 2019). Their experimental report stated that the average detection accuracy was 95.78%, and the recall rate was 95.41% under different lighting conditions and fruit occlusion. An improved tomato detection model based on YOLOv3 (YOLO-Tomato) was proposed, which included a dense architecture to enhance feature fusion, making the feature learning more compact and accurate (Liu *et al*., 2020). They improved the traditional rectangular bounding box (R-Bbox) to a circular bounding box (C-Bbox), and the YOLO-Tomato model achieved a correct recognition rate of 94.58%. It was proposed as a model based on YOLOv4, which integrates a new backbone network, R-CSPDarknet53, constructed by fusing residual neural networks and enhancing feature information reuse and multiscale fusion by replacing the original SPP network's max pooling with depth-wise convolution (Zheng *et al*., 2022). The improved model achieved tomato detection accuracy and recall rates of 88% and 89%, respectively, with an average detection accuracy of 94.44%. It was aimed to meet the real-time requirements of automatic tomato harvesting, improve the YOLOv3 algorithm, and propose a real-time tomato ripeness detection model, SE-YOLO-MobileNetV1 (Su *et al*., 2022). They used depth-wise separable convolutions to increase computational speed, and applied data augmentation, K-means clustering algorithm, and SE attention mechanism module to improve accuracy. The accuracy rates for different categories of tomatoes were above 71.80%.

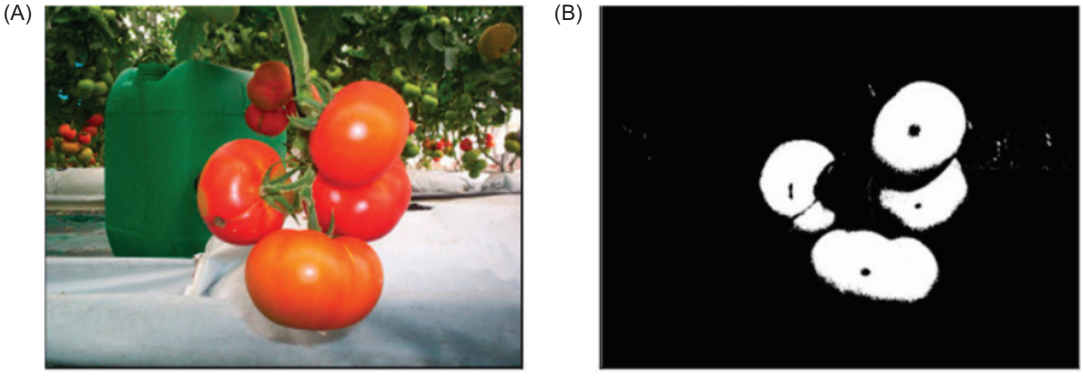

Figure 1. Images from the University of Almeria in Spain. (A) Original image. (B) Detection image.

In summary, although significant improvements have been made in the recognition and detection of tomatoes by domestic and international research, the studies mainly focus on image analysis and lack sufficient sample data of tomatoes in real-field environments. Especially, in the context of on-site robotic harvesting operations, the precision, real-time performance, and robustness of tomato recognition and detection still cannot meet the needs of harvesting robots. Further strengthening and improvements are needed to better enhance the environmental perception capabilities of harvesting robots.

Fruit and vegetable localization technology

The accuracy and speed of tomato localization determine the precision and efficiency of harvesting robots. In recent years, binocular stereo vision has become one of the important research hotspots in the field of computer vision. The main process of binocular localization includes binocular image acquisition, binocular calibration, stereo matching, and depth calculation, among which stereo matching is key to the distance measurement of binocular localization. The method of stereo matching determines the accuracy of localization. Scholars at home and abroad have conducted research on stereo matching methods. A stereo matching algorithm was proposed based on minimum spanning tree (MST) cost aggregation, through an improved root to leaf (L2R) matching cost aggregation algorithm (Zhang *et al*., 2021). By combining stereo matching technology with parallel computing technology, experimental results show that the improved stereo matching algorithm has high accuracy and matching speed for binocular vision. It was proposed as a real-time distance measurement method based on parallel binocular vision (Xu *et al*., 2017). In Xu *et al*. (2017) study, a single parameter division model was used. Then, the parameters of the binocular camera were calibrated using a mode with three parameters {C, dcu, dcv}. When the distance was between 0.4 m and 1.1 m, the error of the measured distance was less than 5 cm, and the average time for distance measurement was 30 ms. When the measured distance exceeded 1.1 m, the error of distance measurement increased. It was proposed as a stereo matching algorithm based on census transform and texture filtering (Hou *et al*., 2022). In Hou *et al.* (2022) work, the weighted census transform circular template was used for matching cost to reflect the impact of neighboring pixels distance to the target pixel on computation. Experimental results show that their method can reduce the mismatch rate of images, obtain disparity maps with less noise, and achieve better pattern texture matching effects.

It was used as a calibration method based on point distance constraints and image space error, obtaining initial values and refining by minimizing reprojection error through the Levenberg–Marquardt method (Zhang *et al*., 2022). Within a rotation angle range of −45° to 45°, the measurement error was less than ±0.032°, and within a displacement range of 0–39 mm, the error was less than ±0.047 mm. When measuring lengths of 300*225 mm, the error was less than ±0.039 mm. An end-to-end stereo matching algorithm was proposed for a "miniaturized" convolutional neural network (CNN) (Liu *et al*., 2020). The loss function errors on the KITTI 2012 and KITTI 2015 datasets were reduced to 2.62% and 3.26%, respectively. This algorithm can obtain dense disparity maps with higher accuracy and efficiency. An improved method of histogram equalization, novel feature extraction, spatial gradient model, and matching cost was proposed, and it was tested on the Middlebury dataset under different lighting and exposure values. This method reduced the average percentage of bad pixels to 3.35 and decreased the relative mean square error (RMSE) to 30.08 (Qazi Mazhar *et al*., 2022). A stereo matching algorithm based on an improved adaptive support window was proposed (Qi and Liu, 2022). In their research, a cross base arm was obtained according to the preset arm length and color threshold. Then, adaptive areas for the vertical and horizontal arms were constructed separately. Finally, the union of the two adaptive areas was used as the final support window for census. This algorithm can improve the matching accuracy in weak texture areas and discontinuous disparity areas. It was proposed as an algorithm to assess the ripeness of truss tomatoes and a comprehensive stem localization method based on experimental errors of various methods (Miao *et al*., 2023). Tests were conducted both indoors and outdoors. The results show that the proposed algorithm has high accuracy under different lighting conditions, with an average deviation of 2 mm. It can guide robots to effectively harvest truss tomatoes, with an average operating time of 9 seconds per cluster.

From the above literature analysis, it is clear that binocular vision technology is already being applied to agricultural fruit and vegetable harvesting. However, due to the lack of research on stereo matching of tomatoes in actual greenhouse harvesting scenes, existing research results cannot be applied. Therefore, there is an urgent need for research on stereo matching algorithms for tomatoes in the field harvesting environment to obtain more accurate three-dimensional spatial information of tomatoes.

Image Data

Image dataset

The tomato image data were sourced from the greenhouse tomato cultivation bases of the national

Figure 2. Tomato image data under natural conditions.

agricultural park. A binocular camera was selected for image collection. The shooting distance for the tomato images ranged from 300 mm to 1,100 mm, with both left and right images having a resolution of 640*480. All images were captured under natural greenhouse conditions, including variations in tomato size, overlapping tomatoes, occlusion by branches and leaves, and different light intensities. The images making up the tomato dataset, as shown in Figure 2, feature black edges due to the calibration correction of the binocular camera.

The learning and generalization abilities of the training models for deep learning neural networks come from the training dataset (Mirhaji *et al*., 2021). Therefore, besides richness, the dataset also needs to be complete. Image flipping and rotation can enhance the detection capability and stability of the network model, while brightness balance can prevent performance deviations due to differences in sensors and changes in environmental lighting (Tian *et al*., 2019). Ultimately, a total of 2,000 images were obtained for the dataset, with the training, validation, and test sets divided in an 8:1:1 ratio.

Image data annotation

Training deep learning models cannot be separated from the annotation of image data. Accurate marking is crucial for precise localization during the tomato detection process. The network can only accept the accuracy provided by the labeled training set, necessitating that each tomato data in the image data be marked in the same manner. This ensures that the trained model is more accurate. A zoomed-in tomato, with T, R, D, and L representing the tangent points at the four edges of the tomato, is shown on the left side of Figure 3. The method of annotation

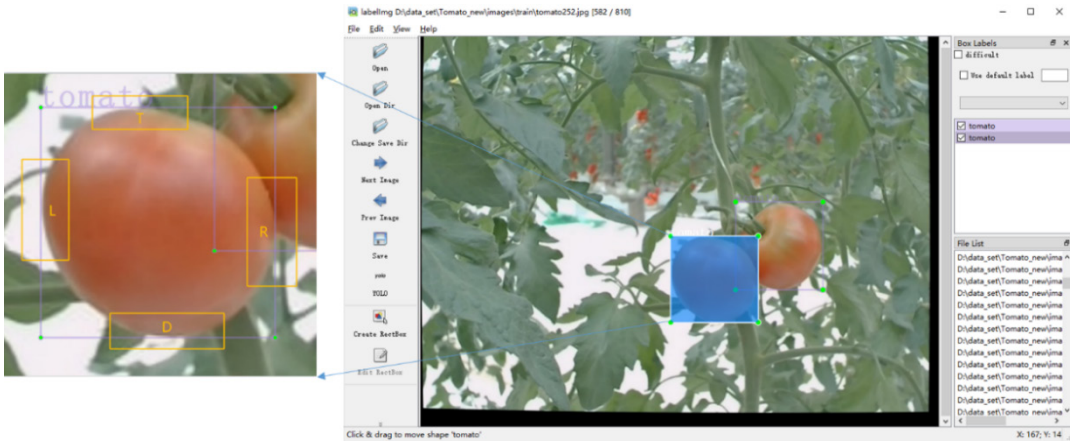

Figure 3. Dataset annotation.

uniformly follows the direction of the tangent line to the circular edge of the tomato to mark the rectangular box, thereby obtaining more precise and reliable data.

Improved YOLOv5 Detection Algorithm

Model structure

In 2015, it was first introduced as a one-stage deep learning CNN model for object detection named YOLO (You Only Look Once) (Redmon and Farhadi, 2017 and 2018). It directly addresses the object detection problem as a regression problem to be solved.

In 2020, YOLOv5 was released, achieving good performance in terms of accuracy and speed. It mainly consists of three parts: the backbone network (Darknet53), the neck of the feature pyramid (PANet), and the prediction head. To design a tomato recognition and detection model with higher accuracy suitable for complex greenhouse environments, the YOLOv5s network model structure was improved, as shown in Figure 4.

CTM structure

Inspired by the work of Dosovitskiy *et al*. (2020) at Google, a transformer encoder module was introduced and combined with the C3 module to design a new structure called the CTM structure, as shown in Figure 5. The CTM structure replaces the C3 module after the SPP block in the original YOLOv5 network. The transformer module includes multihead attention and multilayer perceptron (MLP) as two sublayers, using residual connections. To improve model generalization and reduce computational cost, there is a dropout layer after each sublayer. Different feature vectors containing spatial positional information in the transformer are transmitted to the dropout layer through the attention module of the feedforward neural network. The dropout layer ignores elements with low weight and then transfers effective elements to the fully connected layer, preventing overfitting and enhancing generalization ability. Compared to before replacement, the CTM module can capture more local features and global information, increasing the ability to focus on current pixels and acquire context semantics.

Bi-FPN

The original PANet structure was replaced with a Bi-FPN structure, adding bidirectional weighted fusion. In most networks, when different image features are fused, there is no distinction made. They are simply stacked or added together. However, since different features have different resolutions, their contributions to the output features after fusion also vary. Bi-FPN introduces weighted feature fusion to learn the importance of different input features, while repeatedly applying bidirectional (top-down and bottom-up) features and adding lateral connections between input and output features of the same scale (Tan *et al*., 2020). This reduces the loss of feature information and achieves multiscale and cross-scale optimization, as shown in Figure 6.

The computation expression for Level-4 layer is:

$$P_4^{td} = Conv\left(\frac{\omega_1 \cdot P_4^{in} + \omega_2 \cdot Resize\left(P_5^{in}\right)}{\omega_1 + \omega_2 + \in}\right) \quad (1)$$

$$P_4^{out} = Conv\left(\frac{\omega_1' \cdot P_4^{in} + \omega_2' \cdot P_4^{td} + \omega_3' \cdot Resize\left(P_3^{out}\right)}{\omega_1' + \omega_2' + \omega_3' + \in}\right) \quad (2)$$

Here, the resize operation usually refers to down-sampling or up-sampling, and w is the parameter to distinguish the importance of different features during the fusion process.

Experiments and analysis

Experimental platform

The overall experimental setup is as follows: Intel i5-9400F CPU, Nvidia GeForce RTX 1660 graphics card, Windows 10 operating system, 16 GB RAM, with the PyTorch framework used for the experiments. The resolution of the experimental images was 640*640.

Result analysis

The training loss curves for the training and validation sets of the improved model are shown in Figure 7. Both losses begin to decrease and converge quickly. After 150 iterations, the decline tends to slow, and by 300 iterations, it stabilizes, indicating the model's stability.

The training set mAP is shown in Figure 8. After 100 iterations, the curve reaches near its highest value and tends to stabilize, with the process being relatively smooth. The training process is stable, without any overfitting.

Figure 9 presents the PR (precision–recall) curves of the model before and after improvement. The mAP of the YOLOv5 (baseline) model algorithm was 95.9%, and the improved YOLOv5 model algorithm achieved a higher mAP of 97.1% accuracy, an increase of 1.2 percentage points over the baseline. This enhances the precision of tomato recognition and detection, providing higher accuracy for tomato harvesting.

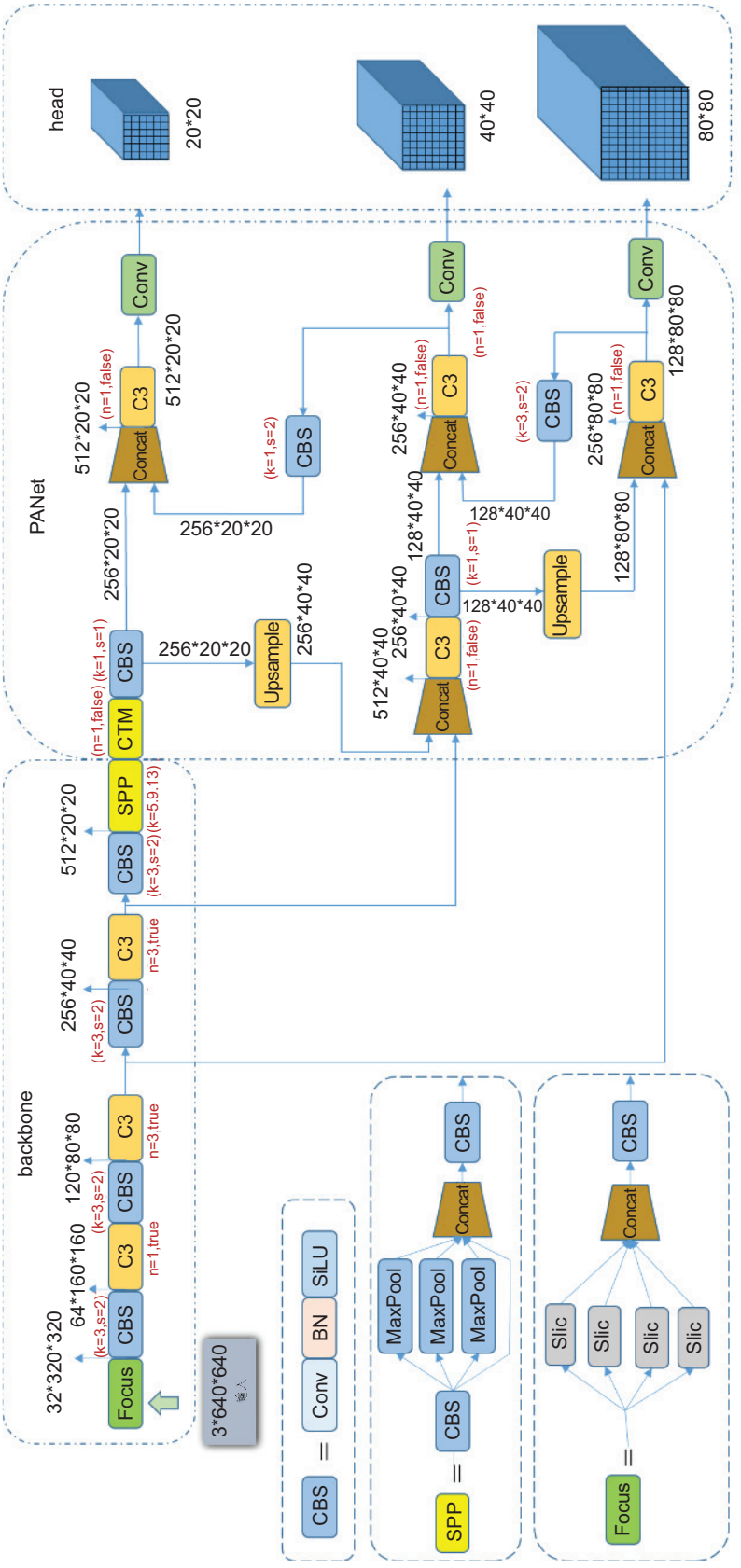

Figure 4. Improved YOLOv5 network model structure.

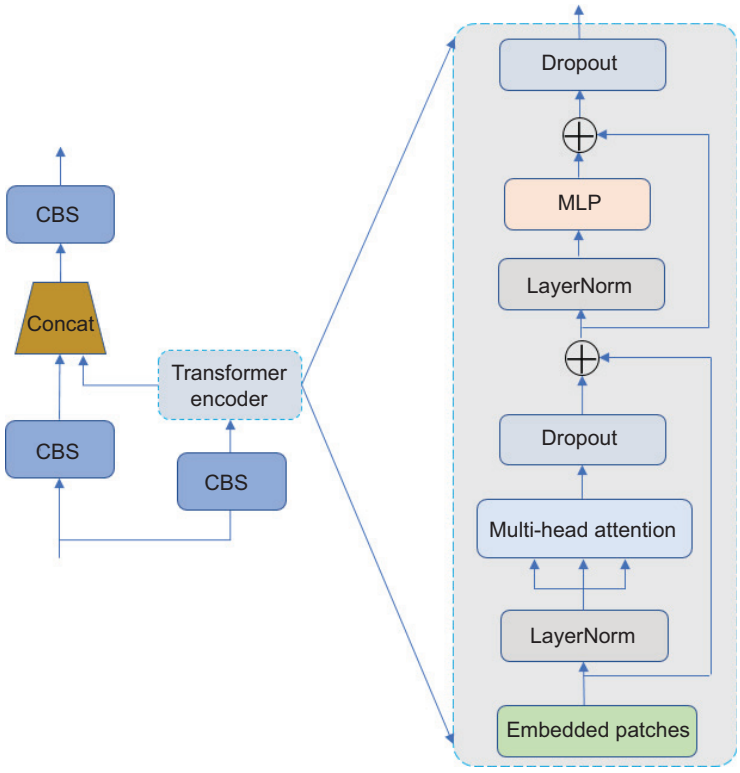

Figure 5. CTM structure.

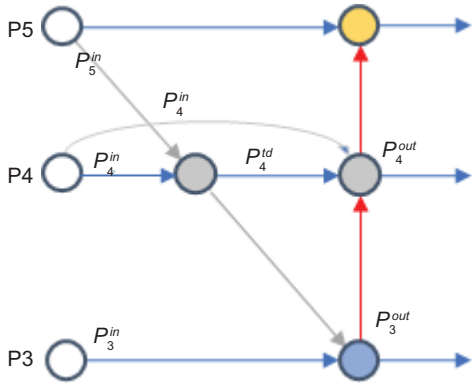

Figure 6. Improved Bi-FPN structure.

The image results of recognition and detection are shown in Figure 10. In the natural growing environment, the model can effectively recognize tomatoes even when the overlapping parts reach 80%. Furthermore, it can accurately detect fruits when the degree of foliage occlusion is at the level of half a leaf and double half leaf. The experimental results show that the improved model has good generalization ability.

To showcase the performance of the improved model, Table 1 presents a comparison of parameters using common model methods. The improved YOLOv5 model has the highest mAP50, outperforming Faster-RCNN, YOLOv3, and YOLOv5 by 30.9%, 9.9%, and 1.2%, respectively. The improved model has an inference time of 55 ms per image, which is 17 ms more than the fastest original YOLOv5 model (38 ms). In terms of model size, the improved YOLOv5 model is 26 MB, which is 12 MB larger than the smallest original YOLOv5 model and 62 MB smaller than the largest Faster-RCNN model (98 MB). The current time and size already meet the needs of the tomato-picking robot.

Tomato Localization Technology

Principle of binocular ranging

The principle of measurement with binocular stereo vision is similar to human eyes, using cameras and a computer to simulate the perception and understanding of objects by human eyes and brain. When the left and right cameras of a binocular camera shoot the same object, two different images can be obtained, and there exists a disparity between their positions in the images, which is known as disparity (Hartley, 2003). Accurate calculation of disparity is key to obtaining depth information, and its principle is shown in Figure 11. Where Ol and Or are the optical centers of the left and right cameras, respectively, $P(x,y,z)$ is a point in space, and and are

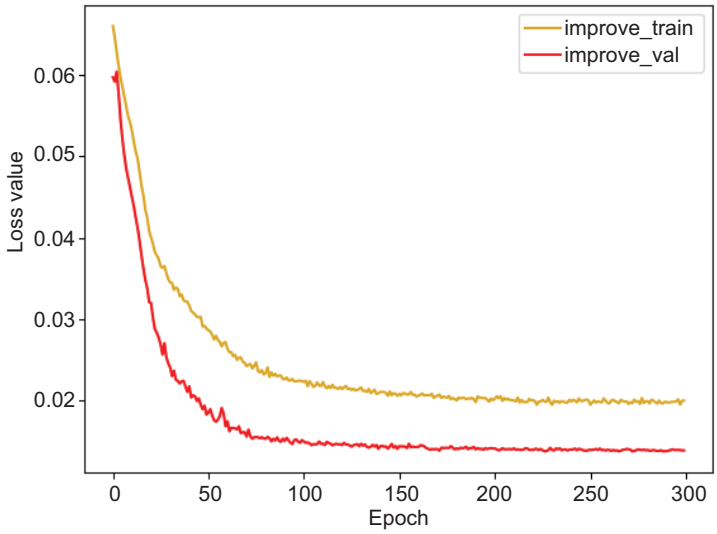

Figure 7. Loss curve of the improved model.

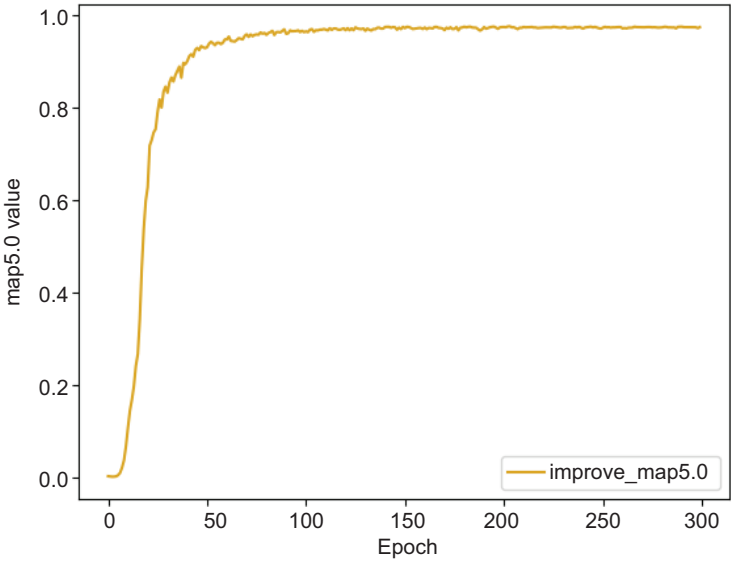

Figure 8. mAP curve of the improved model.

the coordinates of the projection points of P on the left and right cameras, respectively.

The projection of the binocular vision structure on the XOZ plane is shown in Figure 12. The specifications and parameters of the two cameras selected are the same. b represents the baseline of the camera, which is the distance between the two cameras; f is the focal length of the camera; and z is the perpendicular distance from point P to the baseline, also representing the depth information. Using the left camera coordinate system as the world coordinate system, P(x,y,z) represents the three-dimensional spatial coordinates in the binocular camera coordinate system. $P_l(x_{lp},y_{lp})$ and $P_l(x_{rp},y_{rp})$ are the projection coordinates on the left and right cameras, respectively.

In this study, the camera coordinate system of the left camera was taken as the world coordinate system, meaning the origin of the world coordinate system coincides with the origin of the left camera. From this, the transformation relationship between world coordinates and pixel coordinates can be simplified as Formula (3):

$$K\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix}, \quad (3)$$

where K is the scaling factor; R and T represent the position and orientation in the real world; f_x, f_y, f_z, u_0, v_0 are obtained from the intrinsic and extrinsic parameters of

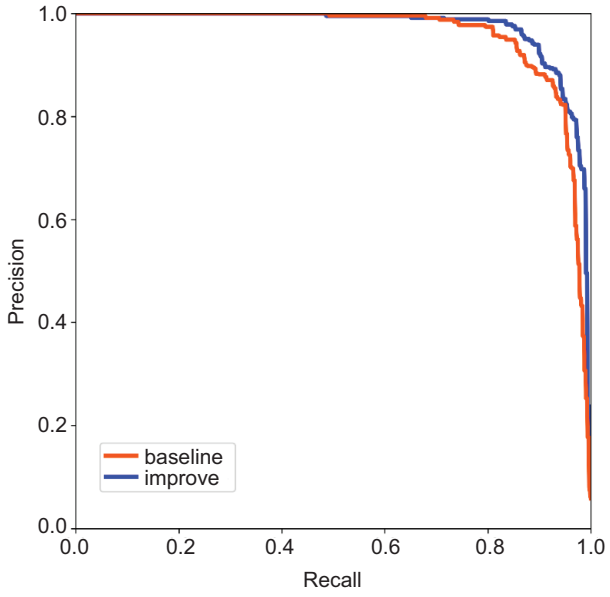

Figure 9. PR Curves of the model before and after improvement.

the binocular's left and right cameras, where the intrinsic and extrinsic parameters of the left and right cameras can be obtained through calibration and correction using Zhang's (2000) calibration method referred in literature. Finally, the three-dimensional coordinates P (x, y, z) in the world coordinate system can be solved.

Stereo matching algorithm

Due to the different environments and lighting changes in different greenhouses, the census algorithm showed better performance under varying lighting conditions (Müller *et al.*, 2011). To better adapt to different environments and improve accuracy, the following improvements were made to the census algorithm.

Improvement of the central pixel

The traditional census algorithm relies too heavily on the central pixel during the transformation process. Once the

Figure 10. Detection of tomato images.

Table 1. Different model performance indicators.

Detection method	P (%)	R (%)	mAP50 (%)	F1 (%)	Single image inference time (ms)	Model size
Faster-RCNN	82.9	58.4	66.2	69	426	98 M
YOLOv3	93.1	67.3	87.2	78	189	78 M
YOLOv5	94.5	85.6	95.9	90	38	14 M
Improve	93.9	90.5	97.1	92	55	26 M

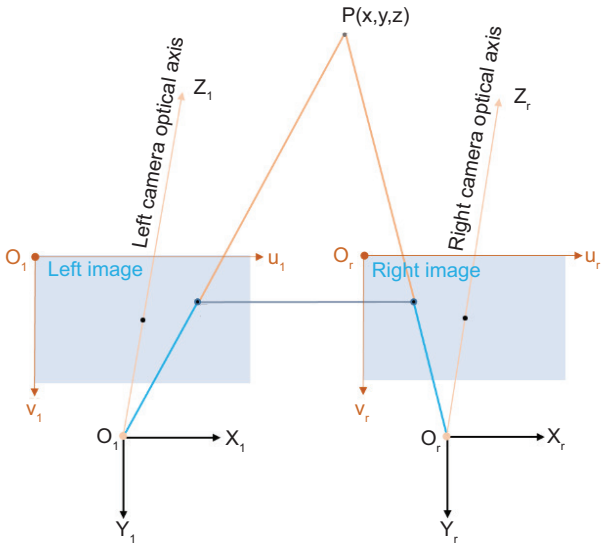

Figure 11. Structure of the binocular vision model.

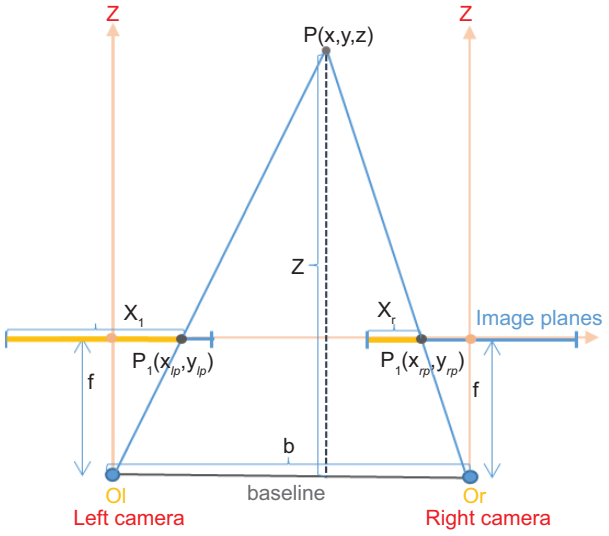

Figure 12. Projection of binocular vision on the XOZ plane.

central pixel is disturbed by noise, the resulting binary code will be affected (Chen Lv *et al.*, 2021), thereby affecting the matching accuracy. Therefore, the average value of the grayscale values of the neighboring pixels within the window is used as the grayscale value of the central pixel, with the improved formula as follows:

$$B(i_k, j_k) = \begin{cases} 0, C(i_k, j_k) < \overline{C}(i, j) \\ 1, C(i_k, j_k) \geq \overline{C}(i, j) \end{cases}, \quad (4)$$

where $\overline{C}(i, j)$ is the average value of the grayscale values of the neighboring pixels within the window; $B(i_k,j_k)$ is the transformation code of the neighboring pixels within the window; $C(i_k,j_k)$ is the grayscale value of the neighboring pixels within the window.

Improvement of the matching range

The section "Improved YOLOv5 Detection Algorithm" has already detected the tomatoes, providing the position information of the tomatoes in the image in the form of rectangular boxes (center coordinates and width and height). As shown in Figure 13, it is only necessary to match within the width range of the rectangular box in the tomato image. The improved formula is as follows:

$$C_n(i, j, d) = min(Hamming_n(L_{str_n(i,j)} \oplus R_{str_n(i,j-d)})) \quad (5)$$

where $C_n(i,j,d)$ is the minimum matching cost for the n-th tomato at disparity d, namely, the minimum Hamming distance; *Hamming* is the Hamming distance corresponding to the n-th tomato in the image; $L_{str(i,j)n}$ is the code at the center coordinate point of the n-th tomato in the left image; $R_{str_n(i,j-d)}$ is the transformed code at the matching point of the corresponding n-th tomato in the right image within disparity d.

In the left image, the detected central pixel point C serves as the target matching point, and it is necessary to match the corresponding point C′ in the right image. From this, it is apparent that the matching range is located between C1 and C2. As can be seen in Figure 13, the number of pixel points matched for the tomato is reduced to the width w pixels recognized and detected in the right image, and the matching process data are presented in Table 2.

Figure 14 shows the Hamming distance for the central pixel point of a tomato in the left image within the matching range in the right image. From Figure 14 and Table 2, it is observed that there are many matching points with a minimum Hamming distance value of 0. This makes it difficult to determine the true matching pixel point. Therefore, further improvements are needed to ensure that the matching point is unique.

Improvement of constraint conditions

As illustrated in Figure 13, the points matched for the tomato pixel coordinates in the left image are also near the center coordinates in the right image. When there are multiple pixels with a Hamming distance of 0 or the minimum value within the matching range of the tomato width in the right image, it is not possible to determine the specific matching point, leading to a large matching precision error. Based on the principle that the matching point is near the detected central pixel point in the

Figure 13. Tomato process matching.

Table 2. Match the procedure data.

Tomato Number	Left Image Center Pixel Coordinates L(i,j)	Right Image Center Pixel Coordinates R(x,y)	Width Detected in Right Image w	Matching Pixel Range R(i,y-w/2)~ R(i,y+w/2)	Corresponding Minimum Hamming Distance	Matched Pixel Point
1	(265,345)	(263,279)	83	(265,238)~(265,320)	0	(265,269) (265,270) (265,271) (265,273) (265,274) (265,275) (265,314)

right image, when the Hamming distance is minimal and the matched pixel point is closer to the detected central pixel point of the right image (i.e., the Euclidean distance between the matched pixel point and the detected central pixel point is minimal), the similarity of the matched point is higher, and the precision is higher. The improved formula is as follows:

$$C_n = short\_\rho(C_n(i,j),R_n(x,y)) \quad (6)$$

where C_n is the minimum Euclidean distance between the matched point of the n-th tomato and the center coordinates in the right image; (i,j) are the detected tomato center coordinates, with i and j being the row and column pixels of the image, respectively; $C_n(i,j)$ is the pixel point in the right image that matches the n-th tomato with a Hamming value of 0 or the minimum; $R_n(i,j)$ is the center pixel point of the rectangle detected for the n-th tomato in the right image; $short\_\rho()$ calculates the minimum Euclidean distance between the center pixel point in the right image and the matched pixel point.

Therefore, the pixel point in closest to is the matching point, and its corresponding disparity d is the disparity value for the center pixel point of the n-th tomato detected in the left image.

Localization of tomatoes occluded by branches and leaves

In the greenhouse field environment, it is inevitable that branches and leaves will occlude tomatoes, as shown in Figure 15. When the detected central pixel point is located on branches or leaves, the depth information calculated after stereo matching is the depth of the branches or leaves rather than the depth of the tomato, leading to errors that result in failure to harvest the tomato.

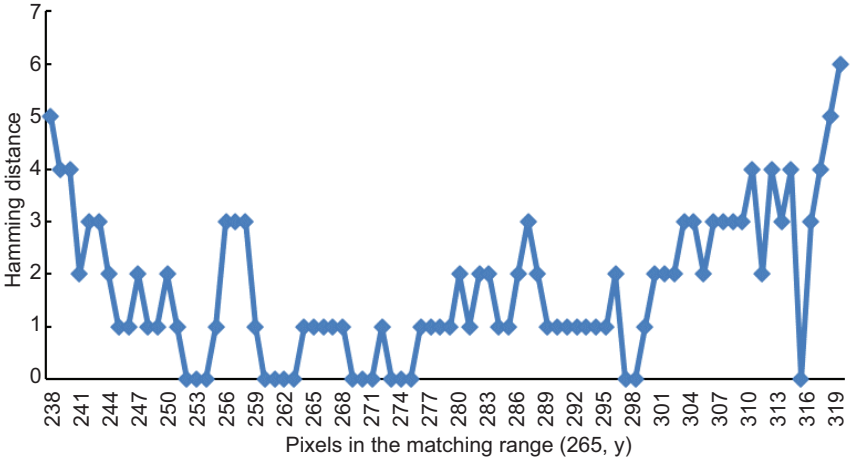

Figure 14. Hamming distance of matches pixels.

In Figure 15, the detected central pixel points of the tomatoes, Al and Bl, are located on a branch and a leaf, respectively. Only by ensuring that the pixel points fall on the tomatoes within the detected region and calculating their depth information can the data be considered reliable. Therefore, it is necessary to segment the occluded area of the image to isolate the tomato region. The centroid of the largest connected area within this segmented region can then be taken as the tomato's pixel point for matching. In image segmentation algorithms, the Otsu threshold segmentation method (Guo and Fei, 2010) has been widely used in fruit and vegetable image recognition due to its speed and ease of implementation. The result of segmenting the images in Figure 16 using the Otsu threshold method is shown in Figure 16. From the left images in Figure 16, it is possible to see the segmented tomatoes' connected regions. The centroids of the largest connected areas are identified as Al′ and Bl′, which are then matched in the right images as Ar′ and Br′ using the improved census matching algorithm. The disparity can then be calculated to obtain the depth information of the tomatoes. Table 3 provides the calculated disparity

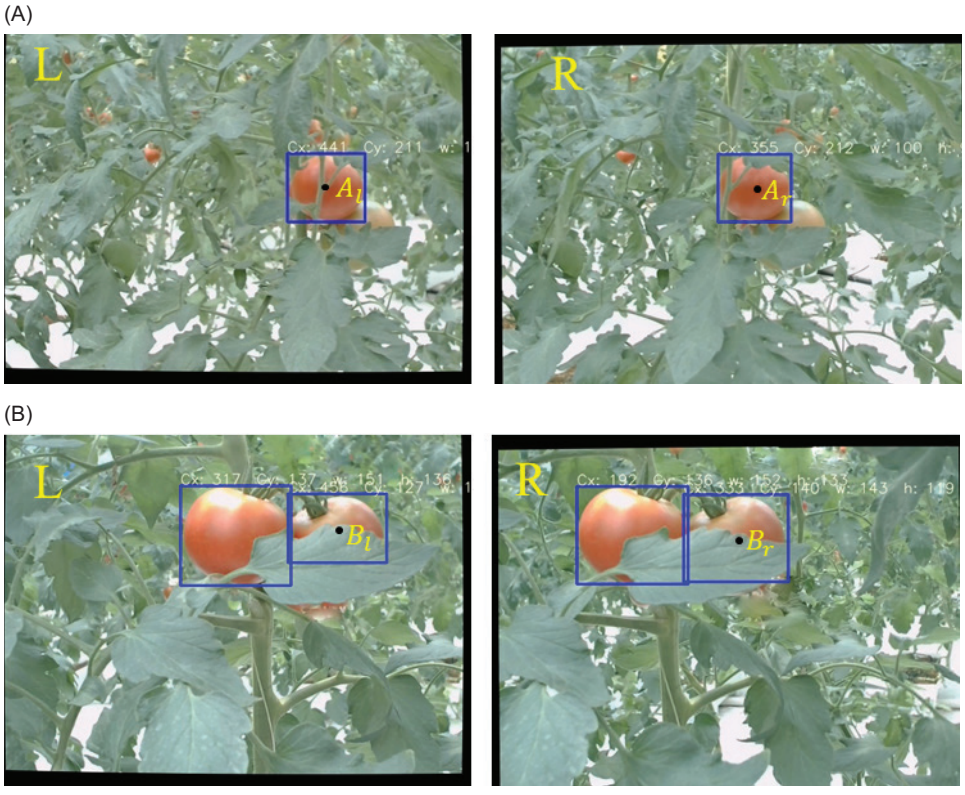

Figure 15. Branches and leaves occlusion. (A) Branch occlusion; (B) Leaf occlusion.

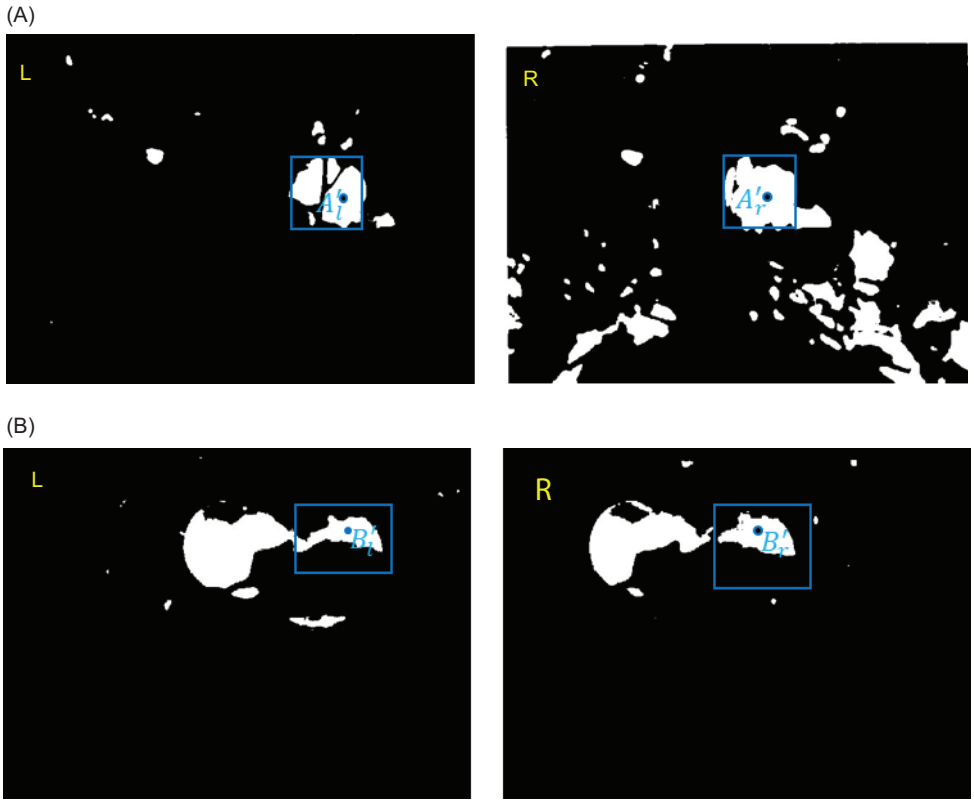

Figure 16. Image segmentation based on branch and leaf occlusion. (A) Branch occlusion image segmentation. (B) Leaf occlusion image segmentation.

Table 3. Calculation results of branch and leaf occlusion.

Number	Left Image A_l' Coordinates (pixels)	Right Image Matched A_r' Coordinates	Tomato Width in Left Image (pixels)	Disparity Value (pixels)	Three-Dimensional Coordinates (mm)
Branch Block	(226,463)	(226,374)	100	89	(91.58, −8.97,473.29)
Leaf Block	(112,474)	(112,355)	140	119	(73.76, −61.31,353.97)

values and three-dimensional coordinates corresponding to these pixel points.

Conclusion

The challenge in robotic harvesting technology lies in the recognition and localization of tomatoes, with their accuracy and real-time performance being critical to the efficiency of the harvesting robot. This research improved the YOLOv5 model through the enhanced YOLOv5 algorithm, incorporating the CTM structure and BiFPN, which brought the tomato's average detection accuracy to 97.1% and an average single image recognition and detection time of 55 ms, addressing the issue of tomato detection in complex environments. Furthermore, by utilizing the principle of binocular distance measurement and improving the census stereo matching algorithm, the three-dimensional coordinates of tomatoes under complex greenhouse conditions were calculated. This information was provided on-site to the tomato-picking robot, achieving precise tomato harvesting and offering a reference for the development of intelligent agricultural equipment.

Acknowledgments

The work is supported by Science and Technology Program of Qinhuangdao (Grant No.: 202301A293) and Science and Technology Program of Handan (Grant No.: 23422012099).

References

Afonso, M., Fonteijn, H., Fiorentin, F.S., Lensink, D., Mooij, M., Faber, N., et al. 2022. Tomato fruit detection and counting in greenhouses using deep learning. Frontiers in Plant Science 11: 571299. https://doi.org/10.3389/fpls.2020.571299

Arianti, N.D., Muslih, M., Irawan, C., Saputra, E., Sariyusda and Bulan, R., 2023. Classification of harvesting age of mango based on NIR spectra using machine learning algorithms. Mathematical Modelling of Engineering Problems 10(1): 204–211. https://doi.org/10.18280/mmep.100123

Benavides, M., Cantón-Garbín, M., Sánchez-Molina, J.A. and Rodríguez, F., 2020. Automatic tomato and peduncle location system based on computer vision for use in robotized harvesting. Applied Sciences 10: 5887. https://doi.org/10.3390/app10175887

Chen Lv, Li J., Q.Q. Kou, H.D. Zhuang, S.F. Tang. 2021. Stereo matching algorithm based on HSV color space and improved census transform. Mathematical Problems in Engineering 1857327. https://doi.org/10.1155/2021/1857327

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. 2020. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Guo, J.C. and Fei, Y.L., 2010. Research on vein image preprocessing based on NiBlack algorithm. The Ninth International Conference on Information and Management Sciences 8: 190–197.

Hartley, R., 2003. Multiple view geometry in computer vision (2nd edition). Cambridge: University Press. pp. 1–532. https://doi.org/10.1017/CBO9780511811685

Hou, Y., Liu, C., An, B. and Liu, Y., 2022. Stereo matching algorithm based on improved census transform and texture filtering. Optik 249: 168186. https://doi.org/10.1016/j.ijleo.2021.168186

Liu, G., Nouaze, J.C., Mbouembe, P.L.T. and Kim, J.H., 2020. YOLO-tomato: a robust algorithm for tomato detection based on YOLOv3. Sensors 20(7): 2145. https://doi.org/10.3390/s20072145

Liu, Y., Lv, B., Wang, Y. and Huang, W., 2020. An end-to-end stereo matching algorithm based on improved convolutional neural network. Mathematical Biosciences and Engineering 17(6): 7787–7803. https://doi.org/10.3934/mbe.2020396

Magalhaes, S.A., Castro, L., Moreira, G., Santos, F.N., Cunha, M., Dias, J., et al. 2021. Evaluating the single-shot multibox detector and YOLO deep learning models for the detection of tomatoes in a greenhouse. Sensors 10: 3569–3593. https://doi.org/10.3390/s21103569

Miao, Z., Xu, X. and Li, N., 2023. Efficient tomato harvesting robot based on image processing and deep learning. Precision Agriculture 25: 254–287. https://doi.org/10.1007/s11119-022-09944-w

Mirhaji H.M., Soleymani M., Asakerehet A., Mehdizadeh S.A., 2021. Fruit detection and load estimation of an orange orchard using the YOLO models through simple approaches in different imaging and illumination conditions. Computers and Electronics in Agriculture 191: 106533. https://doi.org/10.1016/j.compag.2021.106533

Müller, T., Rabe, C., Rannacher, J., Franke, U. and Mester, R., 2011. Illumination-robust dense optical flow using census signatures. The 33rd Joint Pattern Recognition Symposium pp. 236–245. https://doi.org/10.1007/978-3-642-23123-0_24

Naik, A.J. and Thimmaiah, G.M., 2021. Detection and localization of anamoly in videos using fruit fly optimization-based self-organized maps. International Journal of Safety and Security Engineering 11(6): 703–711. https://doi.org/10.18280/ijsse.110611

Qi, J. and Liu, L., 2022. The stereo matching algorithm based on an improved adaptive support window. IET Computer Vision 16(10): 2803–2816. https://doi.org/10.1049/ipr2.12527

Qazi, M.H., Chang H.L., Shanq-Jang, R. and Derlis, G., 2022. An edge-aware based adaptive multi-feature set extraction for stereo matching of binocular images. Journal of Ambient Intelligence and Humanized Computing 13: 1953–1967. https://doi.org/10.1007/s12652-021-02958-8

Rajpoot, V., Dubey, R., Mannepalli, P.K., Kalyani, P., Maheshwari, S., Dixit, A., et al. 2022. Mango plant disease detection system using hybrid BBHE and CNN approach. Traitement du Signal 39(3): 1071–1078. https://doi.org/10.18280/ts.390334

Ratha A.K., Barpanda, N.K., Sethy, P.K. and Behera, S.K., 2023. Papaya fruit maturity estimation using Wavelet and ConvNET. Ingénierie des Systèmes d'Information 28(1): 175–181. https://doi.org/10.18280/isi.280119

Redmon, J. and Farhadi, A., 2017. YOLO9000: better, faster, stronger. Conference on Computer Vision and Pattern Recognition (CVPR) 2017: 690. https://doi.org/10.1109/CVPR.2017.690

Redmon, J. and Farhadi, A., 2018. Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767.

Selçuk, T. and Tütüncü, M.N., 2023. A raspberry pi-guided device using an ensemble convolutional neural network for quantitative evaluation of walnut quality. Traitement du Signal 40(5): 2283–2289. https://doi.org/10.18280/ts.400546

Su, F., Zhao, Y., Wang, G., Liu, P., Yan, Y. and Zu, L., 2022. Tomato maturity classification based on SE-YOLOv3-MobileNetV1 network under nature greenhouse environment. Agronomy-Basel 12(7): 1638. https://doi.org/10.3390/agronomy12071638

Tan, M., Pang, R. and Le, Q., 2020. Efficientdet: scalable and efficient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. 2020. https://doi.org/10.1109/CVPR42600.2020.01079

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E. and Liang, Z., 2019. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Computers and Electronics in Agriculture 157: 417–426. https://doi.org/10.1016/j.compag.2019.01.012

Trinh, T.H. and Nguyen, H.H.C., 2023. Implementation of YOLOv5 for real-time maturity detection and identification of pineapples. Traitement du Signal 40(4): 1445–1455. https://doi.org/10.18280/ts.400413

Wang, S.Y., Gao, G.H. and Shuai, C.Y., 2023. Study on feedback and correction of tomato picking localization information. Traitement du Signal 40(1): 81–90. https://doi.org/10.18280/ts.400107

Xu, H., Liu, X., Zhu, C., Li, S. and Chang, H., 2017. A real-time ranging method based on parallel binocular vision. International Symposium on Computational Intelligence and Design pp.183–187. https://doi.org/10.1109/ISCID.2017.33

Yu, Y., Zhang, K., Yang, L. and Zhang, D., 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Computers and Electronics in Agriculture 163. https://doi.org/10.1016/j.compag.2019.06.001

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

Zhang, Y., 2023. Information acquisition method of tomato plug seedlings based on cycle-consistent adversarial network. Acadlore Transactions on AI and Machine Learning 2(1): 46–54. https://doi.org/10.56578/ataiml020105

Zhang, Z., 2000. A flexible new technique for camera calibration. IEEE transactions on pattern analysis and machine intelligence 22(11): 1330–1334. https://doi.org/10.1109/34.888718

Zhang, Z., Kai, X., Wu, Y., Zhang, S. and Qi, Y., 2022. A simple and precise calibration method for binocular vision. Measurement Science and Technology 33(6): 065016. https://doi.org/10.1088/1361-6501/ac4ce5

Zheng, T., Jiang, M., Li, Y. and Feng, M., 2022. Research on tomato detection in natural environment based on RC-YOLOv4. Computers and Electronics in Agriculture 198: 107029. https://doi.org/10.1016/j.compag.2022.107029