A Pear Tree Trunk Recognition and Positioning Method Based on Improved YOLO_v5s

Yifan Hou , Lijian Yao , Zidong Yang , Gaozhong Liu , Rong Ma *

Posted Date: 27 May 2025

doi: 10.20944/preprints202505.2012.v1

Keywords: YOLOv5s; target detection; attention mechanism

 doi:10.20944/preprints202505.2012.v1

Article

A Pear Tree Trunk Recognition and Positioning Method Based on Improved YOLO_v5s

Yifan Hou [1], Lijian Yao [1,2,3], Zidong Yang [1,2,3], Gaozhong Liu[1], and Rong Ma [1,2,3,*]

[1] College of Optical, Mechanical and Electrical Engineering, Zhejiang A&F University, Hangzhou 311300, China
[2] National Engineering Technology Research Center of State Forestry and Grassland Administration on Forestry and Grassland Machinery for Hilly and Mountainous Areas, Hangzhou 311300, China
[3] Key Laboratory of Agricultural Equipment for Hilly and Mountainous Areas in South-Eastern China (Co-Construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Hangzhou 311300, China
* Correspondence: 20200001@zafu.edu.cn

Abstract: In order to achieve the recognition and detection of pear tree trunks in natural environment and solve the problem of low recognition rate caused by the complex recognition environment, this study uses binocular camera to construct a visual localisation system for pear tree trunks, and proposes a pear tree trunk recognition and localisation model based on the improved YOLOv5s model. The model replaces C3 with C3TR module and embeds the CA attention mechanism, replaces the standard convolution in the Neck part with GSConv, and replaces it with the bidirectional feature pyramid network BiFPN structure; combined with the acquired depth information of the pear tree trunk, the 3D spatial coordinate information of the pear tree trunk is obtained by using the stereo imaging principle of the binocular camera. In order to verify the effectiveness of the proposed method, the effectiveness of YOLOv5s-pear is compared and analysed with the other three models, YOLOv4, YOLOv4-Tiny and YOLOv5s, in terms of the detection results, and the experimental results show that the accuracy of YOLOv5s-pear is improved compared with YOLOv5s, YOLOv4-Tiny and YOLOv4 by 2.1%, 6.2% and 7.7%; the positioning test shows that the error rate between the shooting distance and the actual distance is 1.48%. The improved YOLOv5s model, YOLOv5s-pear, improves the model's performance in dealing with target detection in multi-scale and complex environments, and can realise the rapid identification and precise positioning of pear tree trunks, which can provide a reference for the research and development of autonomous navigation devices.

Keywords: YOLOv5s; target detection; attention mechanism

1. Introduction

China is one of the world's largest fruit producers, and the fruit industry has become an important pillar of China's rural economic development and a major source of income for orchard farmers[1].Orchard weeding is a key part of agricultural production in orchards, especially the raw grass method can effectively control orchard weeds and improve soil quality[2].In traditional agriculture, agricultural operations are mainly done by hand. Farmers have high labour intensity and low operational efficiency, and the development of intelligent orchard mowers is an inevitable trend for future development[3].Mainstream navigation technologies include satellite-based positioning navigation, LIDAR navigation and vision-based navigation[4].The unique environments of garden fruit growing are, firstly, the dense foliage shading and the complex background of the operational field of view, and secondly, the terrain, mostly hilly and mountainous[5].Satellite navigation struggles to consistently obtain reliable positional information of machinery in this environment. The construction of LiDAR navigation maps demands high computational power, and the extraction of navigational features is challenging with a lot of redundant information, leading to high equipment costs and limited use in orchards [6]. With the advancement of deep learning technology in the field

 doi:10.20944/preprints202505.2012.v1

of image recognition and the low-cost nature of visual navigation, visual navigation has become a hot research topic in agricultural robotics [7]. Visual navigation technology, known for its wide range of detection information, completeness of information acquisition, and low cost, is widely applied in local path planning for autonomous driving. This technology reliably and stably identifies navigation paths through image processing methods, thereby guiding the machinery [8].

To promote the intelligence of the garden fruit industry, it is necessary to build a trunk recognition system for weeding equipment to locate the position of the mower during weeding operations. This paper focuses on the research of pear tree trunk recognition during mower operations. In recent years, many countries have researched visual navigation robots and achieved significant results. However, existing visual navigation robots have low speed and accuracy in target recognition, making the development of high-precision and fast target detection algorithms particularly important [9]. With the continuous development of artificial intelligence technology, deep learning-based target detection methods have gradually gained widespread application, even surpassing traditional image recognition methods [10]. Deep learning target detection algorithms can automatically select and extract features, effectively improving the quality and efficiency of feature extraction [11]. For example, Peng et al. [12] used the Otsu algorithm to segment jujube trees from the background in a dwarf dense planting jujube orchard and selected the intersection of the trunk and the ground. The weeds between the jujube tree rows significantly affected the trunk segmentation, leading to inaccurate selection of the trunk and ground intersection and causing deviations in the final navigation. Wang Yi et al. [13] used deep learning methods to recognize citrus tree trunks under different environments and lighting conditions, achieving an overall recognition rate of 95.37%. Zhang et al. [14] detected apple tree branches using RCNN and fitting methods, with average recall and accuracy rates of 91.5% and 85.5%, respectively.

From the above literature, it can be seen that, apart from pear orchards, intelligent mowers also have applications and needs for specific position recognition in dense jujube orchards, citrus orchards, and apple orchards. However, there is less research on the recognition and positioning of pear tree trunks, and there is significant room for improvement in the recognition rate and positioning accuracy of existing achievements. This paper proposes a pear tree trunk position recognition algorithm based on improved YOLO v5s, named YOLO v5s-pear, considering the complex natural environment in actual detection processes. The algorithm incorporates a CA attention mechanism, replaces the C3TR module, changes the convolutional module to GSConv, and integrates four improved strategies for feature extraction network changes to achieve pear tree trunk recognition. Based on the YOLO v5s-pear recognition algorithm, a stereo vision system was built, completing coordinate conversion and obtaining the three-dimensional coordinates of pear tree trunks. The results show that the improved method proposed in this paper can effectively enhance the efficiency and accuracy of pear tree trunk recognition and can obtain pear tree trunk coordinates within the application's allowable error margin.

2. Materials and Methods

2.1. Pear Tree Trunk Image Collection and Dataset Creation

From June to November 2024, the pear tree trunk dataset was collected at the pear tree plantation of Lin'an Garden Co., Ltd. in Hangzhou, Zhejiang Province. The pear orchard is designed for mechanized planting, with appropriate horizontal and vertical spacing. The row spacing of pear trees in the orchard is 3 meters, and the plant spacing ranges from 1.4 to 1.8 meters. The binocular camera was mounted at a height of 1 meter. A binocular camera (model PXYZ-S-AR135-030T160 from Pixel XYZ) was used, with a baseline set at 140 mm and a ranging distance of 0.8 to 9 meters. The camera operated in global shutter exposure mode and moved along a straight line to capture images of pear tree trunks in a complex environment with a wide field of view. The image resolution was set to 640*480 pixels.

Under these conditions, images were captured under varying light intensities and weed densities. After screening and comparison, 1,000 images were selected to form the pear tree trunk

 doi:10.20944/preprints202505.2012.v1

dataset for the orchard. As shown in Table 1, the dataset includes 200 images under strong lighting (50,000-80,000 lux), 600 images under moderate lighting (30,000-60,000 lux), and 200 images under low lighting (8,000-30,000 lux). Figure 1 displays the data collection site, where pear trees are neatly arranged, and some trunks are partially obscured.

Table 1. Statistical table of the data set.

light intensity	realm	quantities	resolution
Higher light	60000-90000lx	200	640*480
moderate light	30000-60000lx	600	640*480
Less light	8000-30000lx	200	640*480

Figure 1. Dataset image acquisition site.

To enhance the diversity of the data, improve the generalization ability of the model, better extract the features of pear tree trunks, and avoid overfitting or underfitting during the training process, the collected dataset images were augmented using methods such as mirroring, rotation, vertical flipping, affine transformation, brightness variation, and noise perturbation. Among these, mirroring, rotation, and affine transformation can simulate the sensitivity of target positions in images obtained from different orientations of a lawn mower[15]. The brightness variation method can simulate the impact of different lighting conditions, and the noise perturbation method is used to simulate varying imaging qualities. After data augmentation, the dataset was expanded to 6,000 images. The processed images are shown in Figure 2.

(a) original (b) mirroring (c) revolve

 doi:10.20944/preprints202505.2012.v1

(d) brightness variation (e) noise disturbance (f) affine transformation

Figure 2. Schematic diagram of data enhancement.

2.2. Detection and Localisation Methods for Pear Tree Trunks

According to the actual growth and distribution characteristics of pear tree trunks in the orchard, this paper selects YOLOv5s as the base network for pear tree trunk detection to be improved, and the pear tree trunk detection accuracy is further improved by optimising and improving the model structure. In addition, the pear tree trunk detection results are used as inputs to locate the spatial coordinates of pear tree trunks.

Combining the internal and external parameters calibrated by the binocular camera and the transformation relationship between the pixel coordinates and the world coordinate system, the spatial coordinates of the pear tree trunk in the orchard environment can be obtained. Compared with other network models in the YOLO series, the detection accuracy and detection speed of YOLOv5s are better, so in this paper, we choose YOLOv5s as the base network for pear tree trunk detection to be improved, and the structure of the improved YOLOv5s network model (i.e., YOLOv5s-pear) is shown in Figure 3. Firstly, the C3 module in the penultimate layer of Backbone, the backbone feature extraction network, is replaced with the C3TR module to enhance the feature extraction capability of the backbone network; The CA attention mechanism is embedded to enhance the feature fusion and adaptation capability of the backbone network; The standard convolution in the Neck part is replaced with GSConv, and the standard convolution inside the C3 module is also replaced with GSConv and named as C3GS, which reduces the number of model parameters and improves the detection speed without affecting the model detection accuracy; The feature extraction network in the Neck part is replaced with a bidirectional feature pyramid network BiFPN structure to improve the performance of the model in handling target detection in multi-scale and complex environments. Then, the information about the positioning of the pear tree trunk is obtained by taking the pear tree trunk detection results as spatial coordinates. Combined with the internal and external parameters of the binocular camera calibration, the actual distance between the binocular camera and the pear tree trunk is calculated. Next, using the transformation relationship between the pixel coordinate system and the world coordinate system, the pixel coordinates of the spatial positioning reference point of the pear tree trunk in the image are converted to the actual spatial coordinates of the pear tree trunk in the orchard environment.

 doi:10.20944/preprints202505.2012.v1

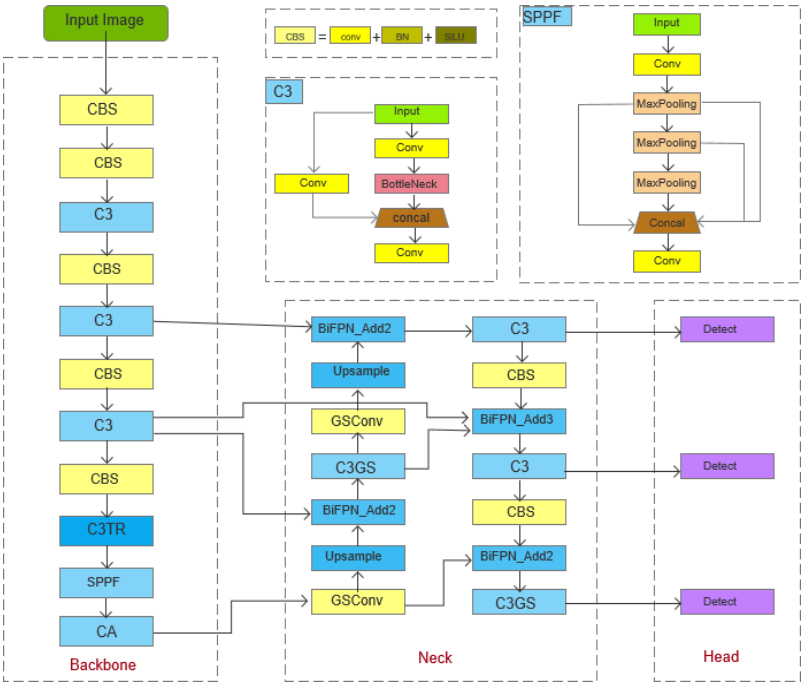


Figure 3. Improved YOLOv5s network structure diagram.

2.2.1. Pear Tree Trunk Detection in Orchard Environment Based on Improved YOLOv5s

Compared with other network models in the YOLO series, YOLOv5s has significantly improved detection accuracy and detection speed, and YOLOv5s can be adapted to the needs of object detection in different environments. Its model structure includes Input, Backbone, Neck, and Head. The backbone block includes CBS, C3TR and CA modules, which are responsible for extracting the feature information of foreground objects in the image. In order to enhance the feature extraction of pear tree trunks in a wide range of complex environments, and to reduce the impact of branch and leaf occlusion, exposure interference, etc. on the detection accuracy of pear tree trunks in the image, targeted improvements were made to YOLOv5s. The images were taken backlit under the influence of the open air environment. The results show that under the influence of strong exposure, the detected features of the pear tree trunk cannot be clearly characterised in the image, thus seriously affecting the detection accuracy of the pear tree trunk image by YOLOv5s. Therefore, on the basis of the original YOLOv5s model structure, according to the improvement method shown in Figure 4, the C3 module in the penultimate layer of the backbone unit is replaced by the C3TR module, and the CA (Coordinate Attention) attention mechanism is added at the end; the standard convolution in the Neck part is replaced by GSConv, and the standard convolution in the C3 module is also replaced by GSConv and named C3GS; the feature extraction network in the Neck part is replaced by a bidirectional feature pyramid network BiFPN structure, which enhances the feature extraction of pear tree trunks in the complex environment of the orchard, and reduces the influence of occlusion and exposure interference on the detection accuracy of pear tree trunks in the image (the improved YOLOv5s model is named YOLOv5s-pear).

 doi:10.20944/preprints202505.2012.v1

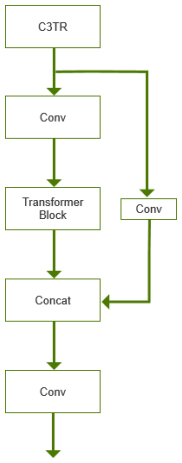


Figure 4. Structure of C3TR module.

In this paper, the original convolutional structure is replaced with the Transformer module, whose Transformer structure module is shown in Figure 5, and richer image information extraction is achieved by replacing the Bottleneck module in the C3 module. In this paper, the Transformer encoder module is added to the last C3 module in the trunk part of Yolov5s, which can improve the feature extraction ability of the model for the pear tree trunk region in complex environments on the basis of almost no increase in computational resources.

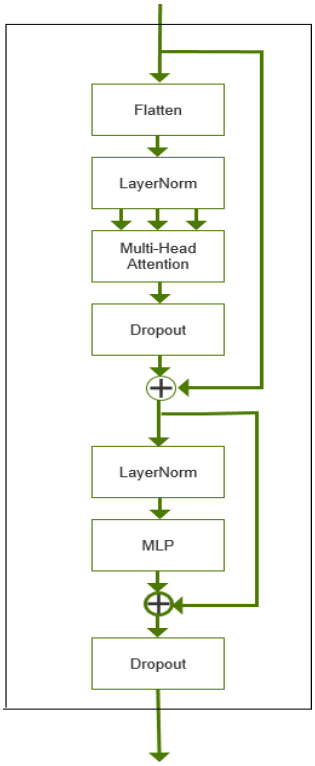


Figure 5. Structure of Transformer encoder module.

The CA attention mechanism is a hybrid domain attention mechanism [16].Adding the CA module to the last layer of the backbone network of Yolov5s not only captures cross-channel and location-sensitive information, which helps the model to locate and identify objects more accurately; compared with other attention mechanisms, CA is relatively more lightweight, can be easily inserted into the target detection model of Yolov5s, and does not significantly increase the computational overhead.The CA structure is shown in Figure 6.

 doi:10.20944/preprints202505.2012.v1

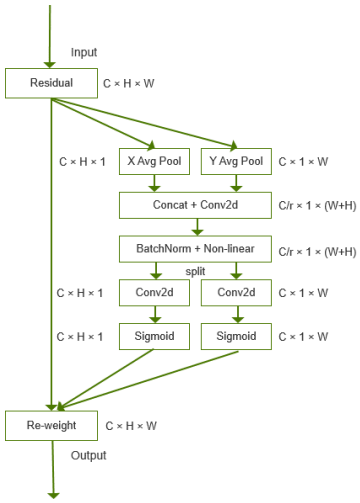


Figure 6. Structure of CA attention mechanism.

The implementation process of the CA attention mechanism can be divided into two steps: in the first step, given an input layer X, an average pooling operation is performed horizontally and vertically using pooling kernels of size (H,1) and (1, W), respectively. For the cth channel with height H, and the cth channel with width W. In the second step, the feature maps in the width and height directions are spliced together and then passed through a 1 × 1 convolutional layer and a Sigmoid activation function to obtain a feature map f of shape 1 (W + H) × C/r, where C is the number of channels of the input feature map, and r is a scaling factor used to reduce the dimensionality of the feature map. The feature map f is then partitioned according to the original height and width and passed through a 1×1 convolutional layer and a Sigmoid activation function, respectively, to obtain the attention weights in the height and width directions Finally, the attention weights in the height and width directions are multiplied by the original feature map, respectively, to obtain the final output feature map.

GSConv combines the advantages of traditional convolution and depth-separable convolution. Through channel reordering operations, GSConv enables the exchange of information between different convolutional groupings, maintaining the superiority of traditional convolution in feature extraction and fusion, while successfully achieving the goal of lightweighting [17]. The use of GSConv in Neck replaces the traditional convolutional layer, due to the fact that the Neck feature map has the largest channel size and the smallest width and height dimensions, the serial processing of the feature map using GSConv reduces the repetitive information and does not require compression, which achieves an increase in the efficiency of the convolutional layer with little or no compromise in the quality of the features [18]. In this paper, the standard convolution module in the neck part is replaced with GSConv, and the network structure of GSConv is shown in Figure 7.

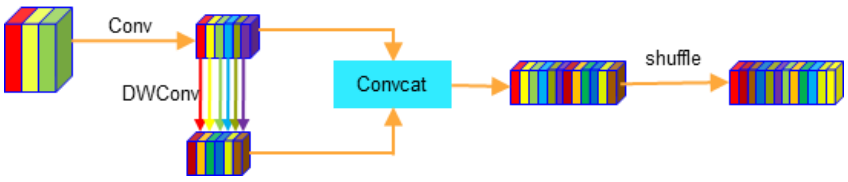


Figure 7. Structure of GSConv network.

GSConv implementation steps, firstly, perform standard convolution on the input feature map and halve the number of channels; secondly, perform deep convolution on the output of the standard convolution and the other half of the channels; then, stitch the results of the two convolutions in the channel dimension; finally, perform a shuffle operation on the spliced feature map, so that the information from the standard convolution and the deep convolution can be exchanged in the channel dimension and fusion. In order to further improve the computational cost-effectiveness of

 doi:10.20944/preprints202505.2012.v1

the network, this paper also replaces the standard convolution module in the C3 module of the neck section with GSConv and names it C3GS, and the C3GS network structure is shown in Figure 8.

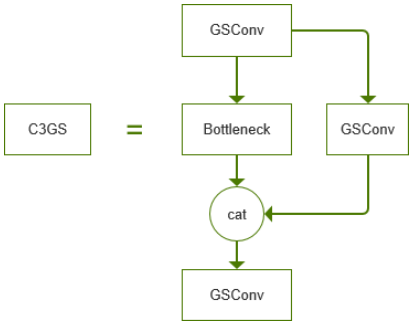


Figure 8. Structure of C3GS network.

Considering that input features of different scales contribute differently to the final output, this paper introduces a bidirectional feature pyramid network (BiFPN) to improve the feature extraction network in order to balance the contribution of each feature layer, and its network structure is shown in Figure 9. The main optimisation consists of three aspects: first, simplifying the network structure without compromising performance by removing nodes with only one input edge and no feature fusion; second, adding additional edges between the original input and output nodes in the same layer to fuse more features without significantly increasing the computational cost; and finally, treating bi-directional paths (top-down and bottom-up) as one feature network layer and repeated multiple times to achieve higher level feature fusion. Compared to PANet, BiFPN uses fast normalised fusion to balance the contributions of different input features by introducing learnable weights. This process normalises the weights to between [0,1], which improves the computational speed and efficiency.

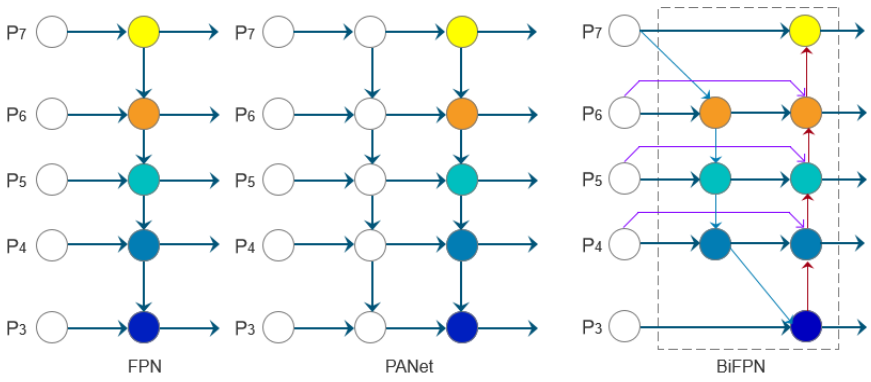


Figure 9. Architecture of FPN, PANet, BiFPN networks.

The pear tree trunk dataset is manually labelled one by one using the graphical image annotation tool Labelimg, and saved as an XML file in accordance with the PASCAL VOC dataset format, where each XML file contains information about the coordinates of the trunk pixel positions for network training. In order to achieve accurate pear tree trunk detection, the labelling rules are as follows, the minimum outer rectangular box of the trunk is in accordance with the actual size of the trunk, to improve the accuracy of pear tree trunk positioning. The complete trunk of the pear tree is labelled in the rectangular box, ensuring that the connection between the roots of the trunk and the soil is close to the midpoint of the bottom edge of the labelling box. The results of the labelling example are shown in Figure 10.

 doi:10.20944/preprints202505.2012.v1

Figure 10. Example of pear tree trunk labelling.

The software environment for the deep learning experiment in this paper is a framework based on Unbuntu20.04, python3.9, pytorch2.0.0, and CUDA11.7, and the hardware environment is configured as follows: the CPU is an Inter Core 12400, and the graphics card is a GeForce RTX3060 graphics card with 12GB of video memory. The training configuration includes a sample batch size set to 16, a stochastic gradient descent (SGD) optimiser with an initial learning rate of 0.01, a minimum learning rate of 0.0001, a momentum of 0.957, and a weight-decay of 0.0005. The batch size is set to a header of 16 during the iterative training process. from Figure 11, we can see that in the early stage of training, the batch size is set to 16. it can be seen that at the beginning of training, the loss shows a large decrease with the increase of the number of iterations. When the number of generations is increased to 300, the loss is basically stable. The loss value starts to stabilise when 800 rounds are trained, and the total loss value of YOLOv5s converges at 0.03, and the total loss value of YOLOv5s-pear is 0.02.

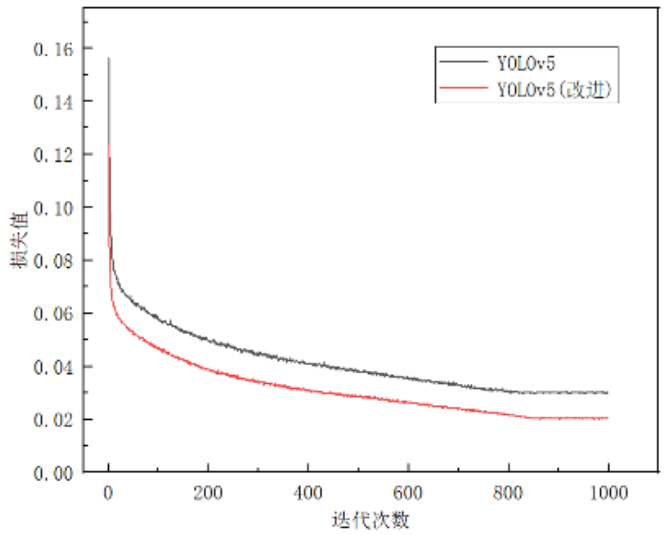


Figure 11. Variation curve of training loss value.

After the training of the YOLOv5s-pear model is completed, the performance of the YOLOv5s-pear model is verified using the validation set of images:The results are shown in Figure 12.

 doi:10.20944/preprints202505.2012.v1

Figure 12. Pear tree trunk detection based on YOLOv5s-pear.

2.2.2. 3D Localisation of Pear Tree Trunks in an Orchard Environment

To achieve spatial localisation of pear tree trunks in complex background environments based on pear tree trunk target detection, it is necessary to find the coordinate information of pear tree trunks, determine the distance between pear tree trunks and the optical centre of the camera (depth information), and use the transformation relationship between the world coordinate system and the pixel coordinate system to convert the coordinates of the pear tree trunks in the pixel coordinate system to the coordinates in the world coordinate system. The 3D spatial coordinate information of the pear tree trunk is then obtained by combining the acquired depth information of the pear tree trunk with the stereo imaging principle of the binocular camera.

In order to obtain the distance (depth information) between the pear tree trunk and the optical centre of the camera, it is necessary to use the binocular camera to achieve effective acquisition of the depth information of the pear tree trunk, following the principle of stereo imaging of the binocular camera and using the calculation rule of similar triangles, as shown in Figure 13.

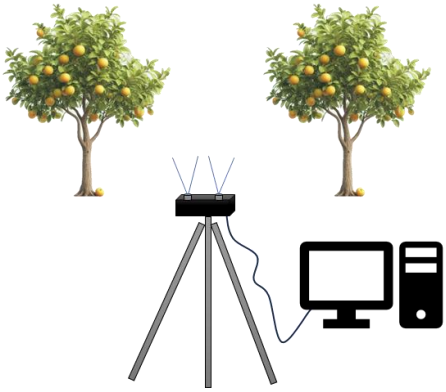


Figure 13. Positioning principle for photographing a pear tree using a binocular camera.

In this paper, the binocular camera is calibrated in the Camera Calibration Toolbox of Matlab R2021b, which has higher accuracy and stronger robustness compared to manual calibration. In the experiments of this paper, due to the measurement distance is far, in order to ensure the accuracy of the positioning of the pear tree trunk, used is a larger calibration board, selected 8x12 grid, a single checkerboard grid side length of 40mm, the overall is about the size of A3. An example of the calibration is shown in Figure 14.

 doi:10.20944/preprints202505.2012.v1

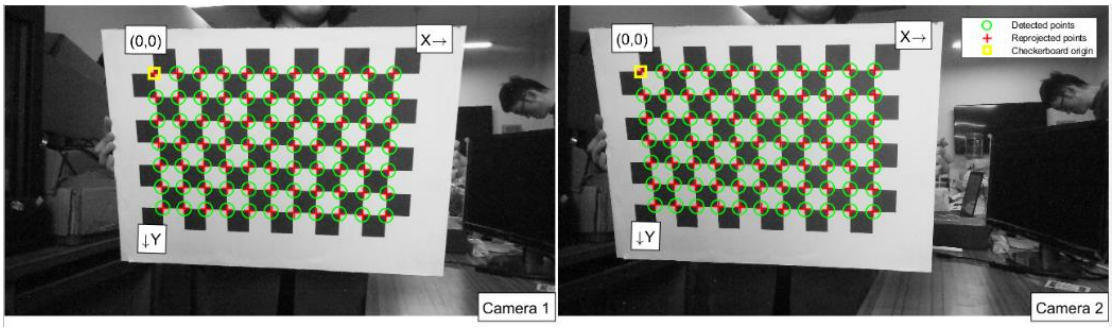


Figure 14. Binocular camera corner point detection results.

The binocular camera calibration results are shown in Table 2:

Table 2. Binocular camera calibration parameters.

Calibration results	left camera	right camera
inner parameter matrix M_1	$\begin{bmatrix} 925.1938 & 0.0000 & 650.1896 \\ 0.0000 & 924.2407 & 363.9331 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}$	$\begin{bmatrix} 923.4747 & 0.0000 & 652.1444 \\ 0.0000 & 922.1971 & 358.3224 \\ 0.0000 & 0.0000 & 1.0000 \end{bmatrix}$
Vector of distortion coefficients	[0.0514 -0.048 0.0000 0.0000]	[0.0770 -0.0967 0.0000 0.0000]
External parameters M_2	rotation matrix R	$\begin{bmatrix} 0.9998 & -0.0045 & -0.0200 \\ 0.0045 & 1.0000 & 0.0013 \\ 0.0200 & -0.0014 & 0.9998 \end{bmatrix}$
	translation vector T	[-129.0403 -0.3687 3.1518]

After obtaining the internal reference and distortion parameters of the binocular camera, in order to obtain the three-dimensional spatial information of the pear tree trunk, the conversion of the pixel coordinate system, the image coordinate system, the camera coordinate system and the world coordinate system was carried out through the pinhole imaging principle of the camera and the method of matrix operation: the schematic diagram of the spatial coordinate system conversion is shown in Figure 15.

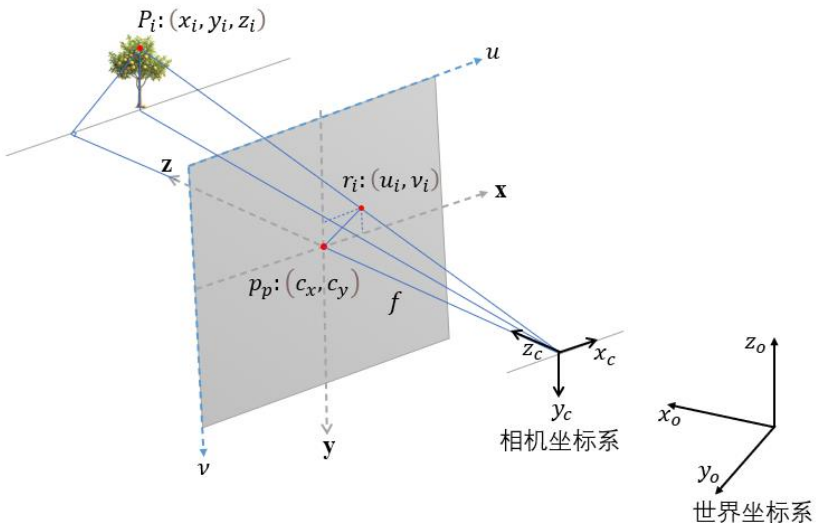


Figure 15. Schematic diagram of coordinate system coordinate conversion.

 doi:10.20944/preprints202505.2012.v1

In this paper, we choose the midpoint of the lower border of the rectangular box of the pear tree trunk output from target recognition as the positioning point of the pear tree trunk, and the coordinates of the upper left corner of the rectangular box are , The coordinates of the lower right corner are , The coordinates of the midpoint of the lower border of the rectangular box are calculated as follows:

$$\begin{cases} x_n = \dfrac{x_l + x_r}{2} \\ y_n = y_r \end{cases} \quad (1)$$

Extracted pear tree trunk loci are indicated by blue dots as shown in Figure 16.

Figure 16. Pear tree trunk localisation point extraction.

The positioning of the pear tree trunk is that the binocular camera determines the position of the pear tree trunk by improving the pixel coordinates of the midpoints of the lower border of the YOLOv5s model detection frame, and obtains the parallax information of the same pear tree trunk after a successful matching by the SGBM algorithm, and calculates the position of the pear tree trunk in the camera coordinate system by the principle of triangulation of the binocular camera.

3. Results

3.1. Validation of the Accuracy of Pear Tree Trunk Detection in Orchard Environments

3.1.1. Performance Comparison and Validation of Different Detection Models

Images from the test set are selected to test the performance training YOLOv5s-pear pear tree trunk detection model, to verify the effectiveness of YOLOv5s-pear in detecting pear tree trunks in the orchard background environment. The detection results are shown in Figure 17. Figure 17 shows that YOLOv5s-pear can accurately detect most of the pear tree trunks in the image.

 doi:10.20944/preprints202505.2012.v1

Figure 17. Results of YOLOv5s-pear detection of pear tree trunks.

In addition, in order to fully evaluate the detection performance of the YOLOv5s-pear pear tree trunk detection model, comparative experiments were conducted on YOLOv5s-pear, YOLOv4, YOLOv4-Tiny, and YOLOv5s in the same configuration environment using the loss value, the precision P recall R and the precision mean AP, as well as the processing frame rate per second, FPS, to assess the detection performance of the model. The formulae for each evaluation metric are shown below:

$$P = \frac{TP}{TP + FP} \tag{2}$$

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

$$AP = \int_0^1 P(R) dR \tag{4}$$

Where, in equations (2) and (3), TP is the number of pear tree trunks (positive samples) that were correctly detected, FP is the number of pear tree trunks (negative samples) that were mistakenly detected as pear tree trunks (negative samples), TN is the number of non-pear tree trunks (negative samples) that were correctly detected (generally not of concern), and FN is the number of non-pear tree trunks (positive samples) that were mistakenly detected as non-pear tree trunks (positive samples).

As shown in Table 3, YOLOv5s-pear processed the pear tree trunk image with P-value, R-value, and AP-value of 97.3%, 96.1%, and 97.2%, respectively, with a detection speed of 14.3 frames/s and a model size of 11.6M. Among them, the P-value of YOLOv5s-pear is 7.7%, 6.2% and 2.1% higher than that of YOLOv4, YOLOv4-Tiny and YOLOv5s, respectively; The R-value is 10.4%, 3.9%, and 2.6% higher than YOLOv4, YOLOv4-Tiny, and YOLOv5s, respectively; and the AP-value of YOLOv5s-pear is 10.8%, 8%, and 2.1% higher than YOLOv4, YOLOv4-Tiny, and YOLOv5s, respectively; The detection speed is improved by 9.4 %, 1.5 % and 4 % over YOLOv4, YOLOv4-Tiny and YOLOv5s, respectively; YOLOv5s-pear is 62.8 and 8.1 smaller than YOLOv4, YOLOv4-Tiny and and only 2.1 larger than the original YOLOv5s. The results show that YOLOv5s-pear is basically able to realise the fast detection of pear tree trunks in the background environment of the orchard under the premise of ensuring higher target object detection accuracy. In addition, our proposed YOLOv5s-pear model is smaller in size, has fewer parameters, and is faster in reasoning. Moreover, due to its small size and limited resource consumption, it is easier to deploy and integrate in mobile devices, embedded systems, or low-power environments, which enables it to operate efficiently on resource-limited devices, making it ideal for applying the model in agricultural robot automation. As the detection

 doi:10.20944/preprints202505.2012.v1

model required for the vision cell of the orchard mowing robot, YOLOv5s-pear can effectively reduce the amount of computation required for image processing and improve the efficiency of object detection.

Table 3. Comparison of detection results of multiple detection algorithms.

mould	P%	R%	AP%	Average time spent on computer testing /ms	model parameter /M
YOLOv4	89.6	85.7	86.4	23.7	78.4
YOLOv4-Tiny	91.1	92.2	89.2	15.8	23.7
YOLOv5s	95.2	93.5	95.1	18.3	13.5
YOLOv5s-pear	**97.3**	**96.1**	**97.2**	**14.3**	**15.6**

3.1.2. Detection Performance of YOLOv5s-Pear on Pear Tree Trunk Images Under Different Acquisition Conditions

In order to further verify the detection effect of YOLOv5s-pear on pear tree trunks, this study uses the YOLOv5s-pear detection model to process pear tree trunks collected under different weather conditions and at different time points, and the images are all randomly selected from the test set of the experimental sample library, which verifies the adaptability of YOLOv5s-pear to different light intensities. The detection results of pear tree trunks with different light intensities are shown in Figure 18: the detection results of pear tree trunks under weak light conditions are shown in Figure 18(a), the detection results of pear tree trunks under moderate light intensity conditions are shown in Figure 18(b), and the detection results of pear tree trunks under strong light conditions are shown in Figure 18(c).

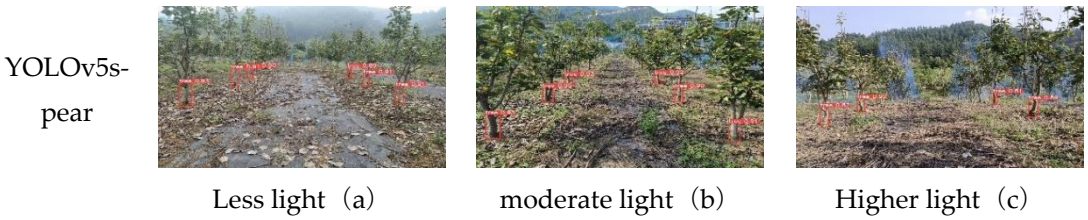


Figure 18. Detection results of YOLOv5s-pear under different acquisition transitions.

As shown in Table 4, the statistical results of the two models for different light intensities are presented. YOLOv5s-pear tested the images under three conditions of strong, moderate, and weak light intensity. The P-value, R-value, and AP-value of the pear tree trunk image acquired under strong light were 95.7%, 94.4%, and 96.4%, respectively, which were 1.1%, 2.9%, and 1.3% higher than those of YOLOv5s. The P-value, R-value and AP-value of the pear tree trunk images collected in moderate light were 98.2%, 97.2% and 97.7%, which were 3.1%, 4.4% and 1.5% higher than those of YOLOv5s, respectively; and the P-value, R-value and AP-value of the pear tree trunk images collected in low light were 93.5%, 95.1% and 96.5%, which were 2.3%, 5.4% and 5.2% higher than those of YOLOv7, respectively. The results show that YOLOv5s-pear can effectively mitigate the effects of strong light irradiation and low light irradiation on detection accuracy and model performance. For object detection under unstable light intensity or strong light environment, YOLOv5s-pear has better performance, so the model is more suitable as an object detection model in the construction of the vision cell of orchard mower robot.

Table 4. Detection results for different light intensities.

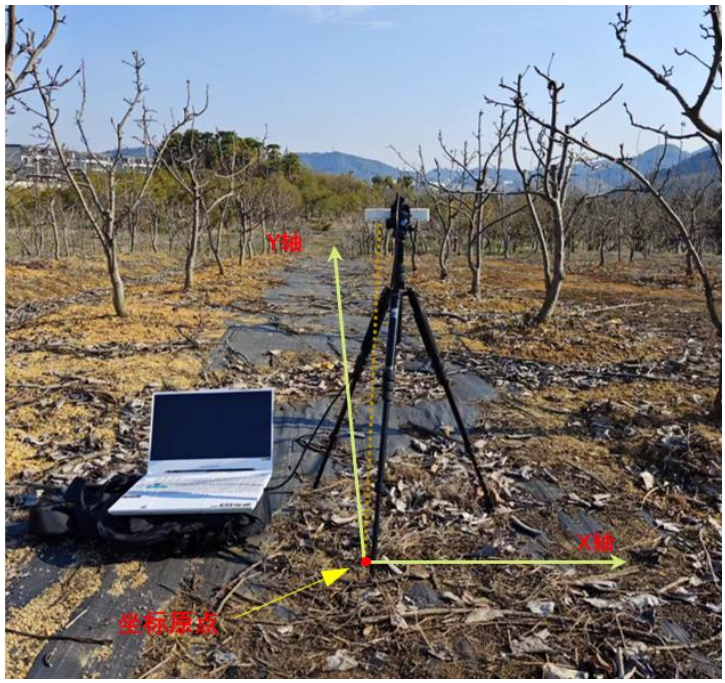
 doi:10.20944/preprints202505.2012.v1

Image acquisition conditions	Detection performance					
light intensity	P/%		R/%		AP/%	
	YOLOv5s	YOLOv5s-pear	YOLOv5s	YOLOv5s-pear	YOLOv5s	YOLOv5s-pear
Higher light	94.6	**95.7**	91.5	**94.4**	95.1	**96.4**
moderate light	95.1	**98.2**	92.8	**97.2**	96.2	**97.7**
Less light	91.2	**93.5**	89.7	**95.1**	91.3	**96.5**

3.2. Verification of Pear Tree Trunk Positioning Accuracy in Orchard Environment

3.2.1. Object Positioning Accuracy and System Performance Verification of Binocular Camera Positioning System

According to the spatial positioning principle and positioning method of binocular camera to locate the spatial coordinates of pear tree trunks described in section 2.2.2, the binocular camera establishes the camera coordinate system with the centre of light of the left eye camera as the origin for visual localisation, and the coordinates under the camera coordinate system of the pear tree trunks obtained by binocular stereo matching are (x_i, y_i, z_i), x_i, z_i represent the estimates of the lateral distance of the pear tree trunk locus from the origin of the camera coordinate system and the longitudinal distance of the tree trunk locus from the origin of the camera coordinate system, respectively. y_i represents the vertical distance between the positioning point of the pear tree trunk in the camera coordinate system and the origin of the camera coordinate system. Because this paper only two-dimensional spatial positioning of the orchard pear tree, it does not consider the distance between the positioning point of the pear tree and the vertical direction of the camera coordinate system, and the intersection point of the binocular camera coordinate system under the current location of the binocular camera axes perpendicular to the ground as the origin of the two-dimensional map of the orchard as shown in Figure 19, to establish the two-dimensional coordinate system of the orchard, and the length of unit m. The two-dimensional spatial positioning of the pear tree in this paper is based on a two-dimensional spatial positioning of the orchard.



 doi:10.20944/preprints202505.2012.v1

Figure 19. Establishment of the orchard world coordinate system.

In this paper, the pear tree positioning test is carried out in the pear tree plantation of Hangzhou Lin'an Garden Co., Ltd. in Zhejiang Province, the binocular camera is fixed on a tripod and placed between the rows of pear trees, and the binocular camera is used to take binocular pictures of two groups of orchard rows with different angles, i.e., one group of cameras shoots perpendicularly to the centre line of the road and the pear trunks are positioned by using the pear trunk binocular visual positioning algorithm combined with the YOLOv5s designed in this paper. Positioning, the two groups were selected the same six pear tree positioning points, to obtain its two-dimensional coordinates under the orchard coordinate system, and then compare it with the actual two-dimensional coordinates of the pear tree positioning points measured using a laser range finder. The results are shown in Table 5.

Table 5. Error statistics for 2D positioning of pear tree trunks during filming.

groups	Matching Distance /m	practical distance /m	inaccuracy /%	Parallel x-coordinate /m	Measured x-coordinate /m	inaccuracy /%	Parallel z-coordinate /m	Measured z-coordinate /m	inaccuracy /%
1	4.33	4.38	1.14	-1.74	-1.76	1.14	3.96	4.01	1.25
2	5.76	5.85	1.53	-1.82	-1.85	1.62	5.43	5.51	1.45
3	7.29	7.17	1.67	-2.05	-2.02	1.49	6.93	6.81	1.76
4	4.22	4.27	1.17	0.91	0.92	1.09	4.12	4.17	1.20
5	5.62	5.71	1.58	0.62	0.63	1.59	5.55	5.64	1.59
6	6.82	6.70	1.79	0.41	0.40	2.50	6.74	6.63	1.65
inaccuracy			1.48			1.57			1.48

The statistical results show that the average error rate between the matched distance and the actual distance of parallel shooting is 1.48%, the average error of the matched x-coordinate value is 1.57%, and the average error of the matched z-coordinate value is 1.48%. The experimental results well verify the matching effect and can meet the needs of positioning. The distance between 3 and 4 metres has a small error, the error rate is 1.1%, and the error at about 7 metres becomes larger, the error rate is 1.7%. The above error can be the positioning requirement of orchard mower. The positioning speed of this research method is fast, meets the requirements of positioning accuracy, and provides a reference for the research on the intelligentisation of autonomous walking equipment for lawn mowers.

4. Discussion

Analysis of the problem of pear tree trunk detection in an orchard environment. During the detection of pear tree trunks in an orchard environment, misdetection or omission may occur, as shown in Figure 20, where the yellow rectangular box in the figure indicates that a pear tree trunk was omitted from detection, and the possible reasons for the omission are as follows.

 doi:10.20944/preprints202505.2012.v1

Figure 20. Leakage of pear tree trunks.

The light in the natural environment is greatly affected by weather factors, when shooting the pear tree trunk on a sunny day, the strong light will lead to overexposure of some areas of the picture, overexposed areas of pear tree trunks will change in colour and texture characteristics, resulting in the loss of features in some areas, and it is difficult for the detection model to extract the complete features of the pear tree trunk, which leads to leakage (circled by the yellow square box in Figure 20).

Analysis of pear tree trunk positioning problems in the orchard environment. In the process of pear tree trunk localisation in the orchard environment, there are also individual pear tree trunks with large errors in localisation or not localised, due to the fact that pear tree trunk localisation is built on the basis of successful detection of the target trunks, and due to the lack of detection of some of the trunks, this type of fruit cannot be localised further successfully.

5. Conclusions

In this study, the YOLOv5s-pear object detection model combined with a binocular camera is used to achieve the detection and spatial localisation of pear tree trunks in an orchard environment. The module of the YOLOv5s backbone network is replaced by the C3TR module to enhance the feature extraction ability of the backbone network; the CA attention mechanism is added to enhance the feature fusion and adaptation ability of the backbone network; and the standard convolution is replaced by the GSConv to reduce the number of model parameters without affecting the accuracy of the model detection and to improve the detection speed. The bidirectional feature pyramid network BiFPN structure improves the performance of the model to deal with target detection in multi-scale and complex environments. Through the detected pear tree trunks, the pixel coordinates of the pear tree trunks are combined with the depth information to obtain the 3D spatial coordinates of the pear tree trunks in the spatial growth environment, and the images of the pear tree trunks are obtained by using the stereoscopic imaging principle of the binocular camera, which achieves effective detection and accurate spatial localisation of the pear tree trunks in the orchard environment.

In the subsequent experiments, the four models, YOLOv4, YOLOv4-Tiny and YOLOv5s, are compared and analysed to verify the effectiveness of the four models for pear tree trunk detection in the complex environment of the orchard.YOLOv5s-pear processes the pear tree trunk image with P-value, R-value and AP-value of 97.3%, 96.1% and 97.2%, respectively, and the detection speed is 14.3 frames/s, and the model size is 11.6 M. The performance is significantly better than that of YOLOv4, YOLOv4-Tiny, and YOLOv5s. The results show that YOLOv5s-pear is basically able to achieve fast detection of pear tree trunks in the background environment of the orchard under the premise of guaranteeing a higher detection accuracy of target objects. Through the comparative analysis of the two models, YOLOv5s and YOLOv5s-pear, the detection results of the two models on the field-of-

 doi:10.20944/preprints202505.2012.v1

view images of pear tree trunks obtained under different acquisition conditions are verified. The results show that the P value, R value and AP value of the pear tree trunk image acquired under strong light are 95.7%, 94.4% and 96.4%, which are 1.1%, 2.9% and 1.3% higher than that of YOLOv5s; the P value, R value and AP value of the pear tree trunk image acquired under moderate light are 98.2%, 97.2% and 97.7%, which are 3.1%, 4.4% and 1.5 The P-value, R-value and AP-value of the pear tree trunk images collected under weak illumination were 93.5%, 95.1% and 96.5%, which were 2.3%, 5.4% and 5.2% higher than those of YOLOv7, and significantly higher than those of YOLOv5s, which proves that YOLOv5s-pear has excellent performance in detecting objects under unstable illumination. detection with excellent performance. In this study, the real 3D spatial coordinates of the target pear tree trunk are obtained through the spatial coordinate positioning test built, and the corresponding spatial coordinates can be obtained by combining with the binocular camera positioning system. The test results show that the device is able to lock onto the target object and achieve efficient localisation in the 3D spatial localisation test. This study has advantages and contributions, and the proposed information perception technique can be migrated to help in other autonomous navigation operations. The average error rate between the matched distance and the actual distance for parallel shooting is 1.48%, the average error of the matched x-coordinate value is 1.57%, and the average error of the matched z-coordinate value is 1.48%. The experimental results well verify the matching effect and can meet the needs of positioning. The positioning speed of this research method is fast, meets the needs of positioning accuracy, and provides a reference for the research on the intelligentisation of autonomous walking equipment for lawn mowers. In the follow-up study, we will further optimise the YOLOv5s-pear detection model to improve the positioning accuracy of the binocular camera positioning system.

Author Contributions: All authors have made equal contributions to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yu J; Jiang X. Discussion of several ways of weeding in orchards and their effectiveness. Shandong Mechanization of Agriculture,2021,01,41.
2. Ma J; Chen X; Jiang R. Review of Researches on Key Technologies for Autonomous Navigation of Intelligent Lawn Mowers. Mechanical & Electrical Engineering Technology.2024,07,41-45.
3. Yang T; Zhao W; Chen B. Development Status and Prospect of Orchard Lawn Mower in China, Agricultural Engineering .2022,01,5-14.
4. Zhou Z; Hu J; Zhang C .Design of Obstacle Avoidance Lawn Mower Control System Based on Image and Lidar. Journal of Agricultural Mechanization Research.2022,09,80-84.
5. Lu Shi; Hongjie Liu; Liu W; Research on Obstacle Avoidance Method of Orchard Mower Based on Lidar. Journal of Agricultural Mechanization Research.2023,02,62-66.
6. Liu L, Wang X, Liu H, et al. A Full-Coverage Path Planning Method for an Orchard Mower Based on the Dung Beetle Optimization Algorithm[J].Agriculture,2024,14(6),865.
7. Li Y, Wang X, Liu D.3D Autonomous Navigation Line Extraction for Field Roads Based on Binocular Vision[J].Journal of Sensors,2019,2019(1),6832109.
8. Fei K, Mai C, Jiang R, et al. Research on a Low-Cost High-Precision Positioning System for Orchard Mowers[J].Agriculture,2024,14(6),813-813.

 doi:10.20944/preprints202505.2012.v1

9. Nizam M S H, Nurul L S, Mohd A F, et al. Vision Based Row Guidance Approach for Navigation of Agricultural Mobile Robots In Orchards[J].Modern Applied Science,2024,18(1),60-67.
10. Chen P, Fei Z, G.V S.GNSS-Free End-Of-Row Detection and Headland Maneuvering for Orchard Navigation Using a Depth Camera[J].Machines,2023,11(1),84-84.
11. Peng S; Li J; Design and realization of visual navigation path extraction software in jujube garden. Jiangsu Agricultural Sciences,2018,10,213-217.
12. Yan C; Liu Y; Wang Y; Research and Experiment on Recognition and Location System for Citrus Picking Robot in Natural Environment. Transactions of the Chinese Society for Agricultural Machinery.2019,12,14-22+72.
13. Zhang M, Zhang Y, Zhou M, et al. Application of Lightweight Convolutional Neural Network for Damage Detection of Conveyor Belt[J].Applied Sciences,2021,11(16),7282-7282.
14. Wang H. Detection of Personal Protective Equipment (PPE) using an Anchor Free-Convolutional Neural Network[J].International Journal of Advanced Computer Science and Applications (IJACSA),2024,15(2),366-374.
15. GuanJ, Liu J, Feng P, et al. Multiscale deep neural network with two-stage loss for SAR target recognition with small training set[J]. IEEE geoscience and remote sensing letters, 2021, 19: 1-5.
16. Zhu X, Liu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on transformer predicti on head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF internat ional conference on computer vision. 2021, 2778-2788.
17. Zhou F, Zhao H, Nie Z. Safety helmet detection based on YOLOv5[C]//2021 IEEE Internationa l conference on power electronics, computer applications (ICPECA). IEEE, 2021: 6-11.
18. Nguyen D T, Nguyen T N, Kim H, et al. A High-Throughput and Power-Efficient FPGA Imple mentation of YOLO CNN for Object Detection[J].IEEE Transactions on Very Large Scale Integrati on (VLSI) Systems, 2019:1-13.