Neural Computing and Applications (2023) 35:13895–13906
https://doi.org/10.1007/s00521-021-06029-z

SPECIAL ISSUE ON IOT-BASED HEALTH MONITORING SYSTEM

A detection algorithm for cherry fruits based on the improved YOLO-v4 model

Rongli Gai[1] · Na Chen[1] · Hai Yuan[1]

Received: 29 December 2020 / Accepted: 7 April 2021 / Published online: 26 May 2021

Abstract
"Digital" agriculture is rapidly affecting the value of agricultural output. Robotic picking of the ripe agricultural product enables accurate and rapid picking, making agricultural harvesting intelligent. How to increase product output has also become a challenge for digital agriculture. During the cherry growth process, realizing the rapid and accurate detection of cherry fruits is the key to the development of cherry fruits in digital agriculture. Due to the inaccurate detection of cherry fruits, environmental problems such as shading have become the biggest challenge for cherry fruit detection. This paper proposes an improved YOLO-V4 deep learning algorithm to detect cherry fruits. This model is suitable for cherry fruits with a small volume. It is proposed to increase the network based on the YOLO-V4 backbone network CSPDarknet53 network, combined with DenseNet The density between layers, the a priori box in the YOLO-V4 model, is changed to a circular marker box that fits the shape of the cherry fruit. Based on the improved YOLO-V4 model, the feature extraction is enhanced, the network structure is deepened, and the detection speed is improved. To verify the effectiveness of this method, different deep learning algorithms of YOLO-V3, YOLO-V3-dense and YOLO-V4 are compared. The results show that the mAP (average accuracy) value obtained by using the improved YOLO-V4 model (YOLO-V4-dense) network in this paper is 0.15 higher than that of yolov4. In actual orchard applications, cherries with different ripeness of cherries in the same area can be detected, and the fruits with larger ripeness differences can be artificially intervened, and finally, the yield of cherry fruits can be increased.

Keywords Digital · Agriculture · Improved YOLO-V4 · Target detection · Cherry fruits

1 Introduction

The development of artificial intelligence has brought great convenience to our day-to-day life. Agricultural robot technology is also being developed rapidly and artificial intelligence offers extensive opportunities in the accurate detection of fruits [1].

Currently, the main fruit harvesting method is based on manual picking which is costly and labor intensive. Due to the emergence of smart agriculture, the robots can detect the position of the fruit through computer vision and then accurately pick the fruit at the corresponding position. This reduces the damage to the non-fruit area caused by manual picking, and at the same time reduces the labor costs in fruit harvesting technology. In the process of robotic fruit picking, recognition of the fruit is the most important task in visual control. Only if the fruits are recognized with a certain level of accuracy, then the robot picking can overperform the manual picking [2]. Therefore, it is essential for the improvement of detection accuracy.

Computer vision is a technology that has been progressed as well as objective detection technology. The development of multi-camera combined imaging systems also enables computer vision technology to meet the targeted accuracy and quality requirements [3]. A significant development was made in this research area during recent years. Fangfang Gao et al. proposed an apple detection

✉ Rongli Gai
gairli@sict.ac.cn

Na Chen
13546172387@163.com

Hai Yuan
675718504@qq.com

[1] School of Information Engineering, Dalian University, Dalian, China

system and show that this system enables robotic apple picking performance [4]. Longshen Fu et al. used RGB-D sensors to detect and locate fruits to improve the positioning accuracy of robot picking [5]. Further Linker et al. used color and texture information to classify green apples. By comparing the detection circle with the heuristic model, it is concluded that the color texture affects the result [6].

With the development of machine learning, deep learning technology has been widely used in agriculture. The application of deep learning technology has also improved the key technology of automatic fruit harvesting and has greatly improved fruit recognition. To find a new cherry grading method to replace the traditional method, Mohammad Momeny et al. used an improved neural network (CNN) algorithm to detect the appearance of the cherries to facilitate the classification of cherries and improve their exportability [7].

Furthermore, Fangfang Gao et al. used Faster R-CNN to detect apples, and the accuracy of non-occluded, leaf-occluded, branch-/wire-occluded, and fruit-occluded fruit reached 0.879 on average [4]. It is reported in [4] that it only takes 0.241 s to process each image. Due to the complexity of the orchard environment, the corresponding deep neural network has a relatively high accuracy for fruit recognition. Rui Shi et al. used the YOLO-V3-tiny network to detect mangoes and adopted a generalized attribute pruning detection network method, to achieve an accurate real-time detection [8]. Cherry coffee beans are classified using a classic computer vision system, and the classification effect is measured by calculating accuracy, recall rate and other indicators. Finally, the rate of correct classification accuracy of the improved computer vision system reaches 0.669 [9]. For complex background information, in [10] a new semantic pool is used to evaluate the relevance of each image to the event of interest. The Zero-Shot Object Detection (ZSD) semantic preservation graph propagation model based on the graph scrolling machine network can effectively use the semantic description and structural knowledge shown in the previous category graphs to enhance the generalization ability of the learned projection function [11]. Due to the blurred image acquisition, the use of the image multi-feature fusion paradigm method can solve the complexity of the image [12]

For larger fruits (such as apples), the accuracy of robot picking continues to improve due to improved accuracy. Yaqoob Majeed et al. used simple RGB and foreground RGB to conduct comparative experiments through grid division of the apple trees. The results show that the use of foreground RGB grid line division of the apple trees can improve the accuracy of boundary recognition [13]. Zhihao Liu et al. proposed the use of RGB-D sensors to combine RGB and NIR images with the roll machine neural network for fruit detection, and then applied their system to kiwi fruit detection to improve detection speed and accuracy [14]. It is further shown in [15] that using a deep learning model based on the convolutional neural network and short-term memory of the scroller enables quick selection of the targets.

The Faster R-CNN can also generate a model for fruit detection, and the average accuracy results show that using this method can make the accuracy of fruit recognition better [16]. However, Faster R-CNN is not fast enough. The detection idea of YOLO (You Only Look Once) [17] is different from the idea of the R-CNN series, which solves the target detection as a regression task. The most used YOLO models are YOLO-V3 [18] and Yolo-V4. For object detection, the application of the Yolo model in fruit recognition is relatively small. The improved YOLO-V3 is used to detect tomatoes. The algorithm first improves a dense network structure in YOLO-V3, and secondly improves non-maximum suppression. As a result, the new a priori box is more suitable for tomato detection. The experiment is compared with the YOLO-V3 algorithm. The improved YOLO-V3 is better for tomato detection [19]. Anna Kuznetsova et al. passed the test of robot detection time, using pre-processing and post-processing to enable the YOLO-V3 algorithm to be used in apple detection, which provides great convenience for robot harvesting. Experiments also show that the algorithm can reduce the average detection time of detection and the error rate is low [20]. According to the emergence of the YOLO-V4 model, the recognition of the occluded objects is more accurate [21]. Dihua Wu et al. used YOLO-V4 to detect apple blossoms, simplified the apple detection model, and trimmed the channels to ensure detection efficiency. Compared with Faster R-CNN, YOLO-V2, YOLO-V3 and other algorithms, it has been shown that the YOLO-V4 model performs better and can achieve real-time and accurate detection of apple flowers [22].

Recognition of small fruits is often more difficult. For a long time, the cherry planting industry in our country was solely based on experience and there was a lack of scientific studies and accurate data analysis methods. In the process of cherry identification, issues such as the growth density of the cherries, the existence of more shaded parts, and the small volume of the fruit result in relatively low recognition accuracy. To address this issue, an improved YOLO-V4 model is proposed in this paper to detect the cherry fruits and to provide technical support for high-quality cherry production. The rest of this article is organized as follows. Sect. 2 introduces the theoretical background of the detection method and we introduce the cherry detection method, followed by presenting the experimental results of the proposed method in Sect. 3. The paper is ended by providing conclusions in Sect. 4.

2 Materials

In this study, sweet cherries from Dalian, China were tested. All images were manually annotated with circular box annotations and the corresponding JSON annotation files are generated. YOLO-V4′s backbone network (i.e., CSP-darknet53) is replaced with DenseNet and compares with models such as YOLO-V4 through evaluation indicators including accuracy (P), recall (R), and average accuracy (mAP).

2.1 Image acquisition

Dalian is one of the first regions in China to produce sweet cherries and is in a leading position in the country in terms of industrial-scale production and cultivation techniques. The cherry cultivation area in Dalian has reached 315,500 mu, with a total output of 234,000 tons, accounting for about 24% of the total domestic output. The achievements of cherry planting have made big cherry planting a pillar industry of Dalian's planting industry.

In our experiments, a 3000 × 4000-pixel Sony DSC-HX400 camera and a 40 million pixels Huawei mobile phone are used to capture cherry images. The cherry data set for image collection is from 2016 to 2019, and the location was the Sweet Cherry Park in Dalian of China. Most of the data were collected on a sunny day, and the collection time was between 9–12 am and 1–4 pm. All pictures were taken under natural light, including lighting changes, occlusion and overlap. A total of 400 cherry pictures were collected and divided into training and test sets. Among them, 50 pictures are pictures without cherries. The cherry photographs are all divided into three stages: immature, semi-ripe, and mature, and they are taken from a perspective and a close-up perspective. Changes in the angle of the sun result in changes in the image. Different types of cherry images are used to improve the robustness of the network.

2.2 Fruit classes

The growth environment of cherries and their volume challenge automatic recognition. There are also occluded cherries in the marking process, which are divided into fully occluded, semi-occluded, and un-occluded cherry images. The cherry fruits may also overlap each other. The shooting angle is divided into distant view and close view cherry. As the cherries are growing, the cherries on a branch may be in different maturity stages. Among a bunch of cherries, there are immature, semi-ripe and fully mature cherries, which are marked as immature cherries (cherry_2), semi-ripe cherries (cherry_1), and mature cherries (cherry) during the marking process. Cherry classification photographs are shown in Fig. 1. The image labeling software used in this study is "labelme." The annotation file is saved in "JSON" format.

2.3 Image augmentation

The higher the image quality, the lower the loss of the important content [22]. Here we use OFST to process the images in a lossy manner [23]. The recognition of cherries in this paper is based on the improved YOLO-V4 [21]. In the YOLO-V4 model, Mosaic is used for image enhancement. Mosaic data enhancement refers to the CutMix data enhancement method which utilizes several pictures for stitching. In our approach, Mosaic uses four pictures for stitching to enrich the background of the detected object. The data of four pictures can be directly calculated during the BN calculation. As shown in Fig. 2.

The data enhancement process is as the following.

- Read four pictures at a time,
- Flip, zoom, and change the color gamut of the four pictures, respectively
- Make the combination of pictures and the combination of anchors.

3 Methodologies

3.1 YOLO-V4 model

The YOLO-V4 model is an optimized model based on YOLO-V3. Compared with the YOLO-V3 network structure, the DarketNet53 network in YOLO-V3 is changed to CSP DarketNet53, and the CSP DarketNet53 network is used as the backbone network BackBone. The value generated by the last residual network structure in CSPDarketNet53. Using BoF and Mish for CSPDarknet53 classifier training can improve the accuracy of the classifier and the detector using the classifier pre-training weight. CSPDarknet53 is more suitable for detectors [21]. We use the resblock_body module in the DarketNet53 network which is composed of a stack of downsampling and multiple residual networks. The YOLO-V4 model is divided into three parts: cspdarketnet53, Neck and Head. The cspdarketnet53 structure is the backbone network of the model. The Neck structure is composed of SPP and PAN, and the Head is the prediction part of the model.

In YOLO-V4, the modified part is as follows:

- Modify the activation function of DarknetConv2D from LeaktReLU to Mish, and the winding machine is

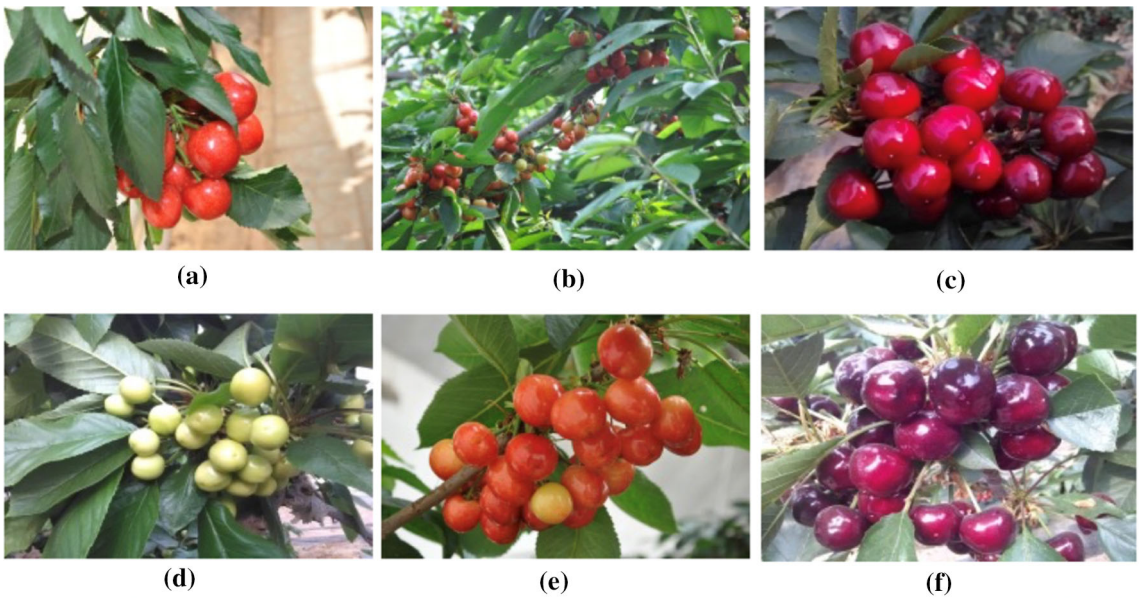


Fig. 1 Classification of pictures of the cherry fruits: **a** cherry occlusion image, **b** cherry perspective image, **c** cherry close-up image, **d** immature cherry image, **e** semi-ripe cherry image, **f** ripe cherry image

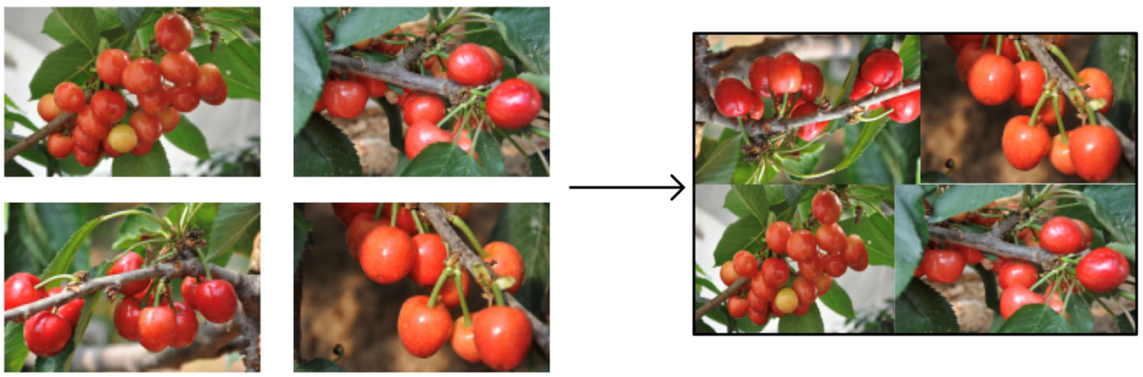


Fig. 2 Image enhancement: select four cherry pictures, flip, zoom, and color gamut changes

changed from DarknetConv2D_BN_Leaky to DarknetConv2D_BN_Mish.

$$\text{Mish} = x \times \tanh(\ln(1 + e^x)) \tag{1}$$

- The resblock_body structure is modified to split the residual network, one part stacks the residual network, and the other part is like a residual edge. After a small amount of processing, it goes directly to the end. This part bypasses many residual structures, collectively referred to as CSPnet structure.
- Adopt SPP structure and PANet structure.

The YOLO network transforms the detection problem into a regression problem and generates boundary coordinates and probabilities for each class. Based on the artificially marked area if the center of the detected object falls within the grid, the network then performs target detection, which is trained by YOLO-V4 loss functions including the bounding box location loss (L_{CIoU}), confidence loss ($L_{\text{confidence}}$) and classification loss (L_{class}) [21].

$$\text{Loss} = L_{\text{CIoU}} + L_{\text{confidence}} + L_{\text{class}} \tag{2}$$

$$\text{CIOU} = \text{IOU} - \frac{\rho^2(b, b^{\text{gt}})}{c^2} - \alpha v \tag{3}$$

where $\rho^2(B, b^{\text{gt}})$ represents the Euclidean distance between the center points of the prediction frame and the real frame, c denotes the diagonal distance of the minimum required area that can contain both the prediction frame and the real frame:

$$\alpha = \frac{v}{1 - \text{IOU} + v} \tag{4}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{\omega}{h}\right)^2 \tag{5}$$

$$L_{\mathrm{CIOU}} = 1 - \mathrm{IOU} + \frac{\rho^2(b, b^{\mathrm{gt}})}{c^2} + \alpha v \tag{6}$$

and IOU is a standard that defines the accuracy of target detection and indicates the intersection ratio of the predicted bounding box and the ground truth bounding box, and d represents the distance between the center points of the two boxes in Fig. 3.

YOLO-V4 can better detect objects that are difficult to identify in the picture and therefore is suitable for cherry fruit detection.

3.2 Dense model

The ResNet [25] makes the neural network deeper and achieves higher accuracy. The core of the ResNet is to help the backpropagation of the gradient by establishing a short-circuit connection between the front and rear layers, thereby training a deep CNN network.

The DenseNet [26] model is based on the same premise as the ResNet network structure, where the link between the front and back layers is established. The difference is that in DenseNet the dense link between all the front layers and the back layer is established, and the feature is reused at the same time. The DenseNet model is composed of DenseBlock and the intermediate interval module Transition Layer. DenseBlock is a unique module in the structure. In the same DenseBlock, the width and height of the feature layer are not changed, but the number of channels is changed accordingly. Transition Layer is a module that connects different DenseBlocks and combines the characteristics of DenseBlock to reduce the width and height of the previous DenseBlock. After the modules are stacked, the features are continuously stacked. This makes the connection between the layers closer. The DenseNet network can also reduce gradient hours, enhance feature propagation, promote feature reuse, and greatly reduce the number of parameters [27]. The DenseNet network structure is illustrated in Fig. 4. Conv refers to convolutional neural network and ReLU refers to Rectified Linear Unit.

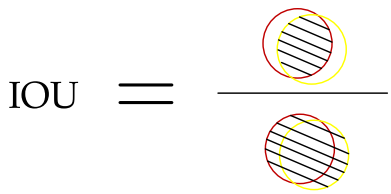


Fig. 3 The model of the IOU [19]

3.3 Improved YOLO-V4 for cherry detection

YOLO-V4 has strong detection capabilities. Because DenseNet reduces gradient disappearance, strengthens feature transfer and reduces the number of parameters and the computational load. Therefore, DenseNet is used to replace the cspdarketnet53 structure in YOLO-V4, and the loss function is also changed to Leaky ReLU. These enhance the denseness between layers and improve feature propagation and promote feature reuse and fusion. By enhancing the density between the layers, the network structure becomes more complex, which is used in cherry detection.

The network structure of YOLO-V4-DenseNet is shown in Fig. 4. The input picture is a 1280×800 cherry picture.

The improved YOLO-V4 model is shown in Fig. 5. The dense block and transition are defined as stage$_n$. The DBL is a special convolutional block consisting of convolutional network, normalization, and Leaky activation functions. Head$_n$ goes through five DBL. After DBL is performed through combination, feature extraction is realized through convolution operation. MP represents three maximum pooling operations.

The dense block model is shown in Fig. 6 where there are several dense layers in a dense block network, and the dense layer structure is shown in Fig. 7. The model structure of Transition is shown in Fig. 6. The structure of stage$_n$ acts as the backbone of the network, thereby improving the feature extraction capability of the picture. The model training indicators of stage$_n$ are presented in Table 1. This allows the network to process low-precision pictures hence expands the range of feature extraction.

The improved YOLO-V4 model continues to use the neck part of the YOLO-V4 model, which is mainly used to fuse the feature information of the feature maps in different scales. Presenting a pyramid form, the model is stacked to increase the detection speed. The a priori box is circular, which better fits the shape of the cherry fruit, and its radius is the mean square error. In the model, cross-entropy and mean square error are used in the center, confidence, and classification parts.

The evaluation indicators of the detection model are precision (precision), recall (recall) and F_1 score. The discrimination values are also True positive (TP), False positive (FP), True Negatives (TN), and False Negatives (FN). We use F_1 score to combine the precision and recall, where

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

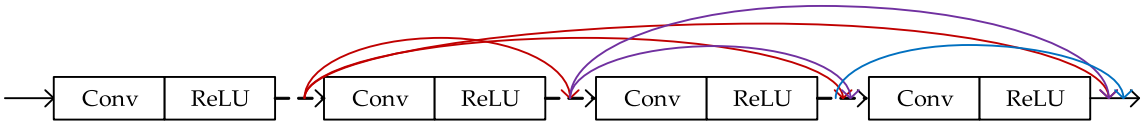


Fig. 4 DenseNet model [26]

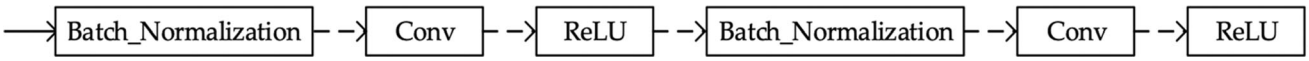


Fig. 5 Dense layer [26]

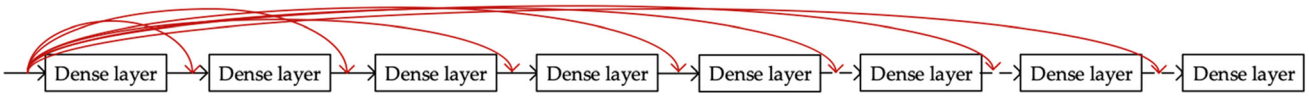


Fig. 6 Dense block. Multiple dense layer composition

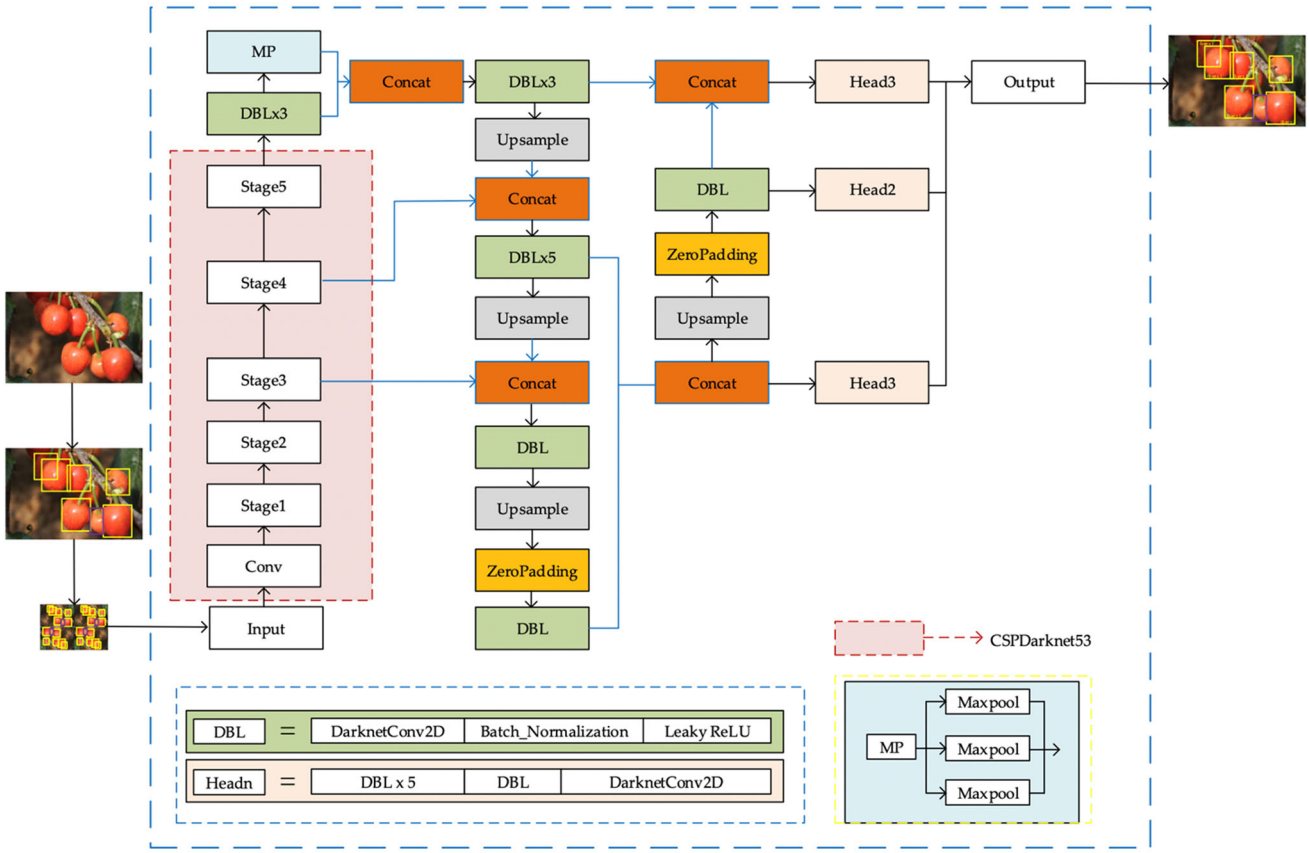


Fig. 7 YOLO-v4-dense structure model

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$F_1\text{score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

The model directly sets the mean square error (MSE) loss function based on the center point coordinates of the prediction box and the real box, and the width and height information [20]. Furthermore, loss is composed of the position loss of the prediction box, the category loss predicted by the prediction box, and the confidence loss of the prediction box:

Table 1 Network model training index

		Number of layers	Growth rate	Input channels	Output channels
Stage 1	Dense block	8	8	24	88
	Transition			88	64
Stage 2	Dense block	8	16	64	192
	Transition			192	128
Stage 3	Dense block	8	32	128	384
	Transition			384	256
Stage 4	Dense block	8	64	256	768
	Transition			768	512
Stage 5	Dense block	8	128	512	1536
	Transition			1536	1024

$$\begin{aligned}\text{loss}_{\text{corrd}} = \lambda_{\text{coord}} \sum_{i=1}^{s^2} \sum_{j=1}^{B} 1_{\text{ij}}^{\text{obj}} \left[(x_i - \hat{x_i})^2 + (y_i - \hat{y_i})^2\right] \\ + \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{\text{ij}}^{\text{obj}} (r_i - \hat{r_i})^2 \end{aligned} \tag{10}$$

where, among them $L_C(x_i, \hat{x_i}) = -x_i \log \hat{x_i} - (1 - x_i)\log(1 - x_i)$, λ_{coord} is the weight of the left error, s^2 is the number of grids of the input image, and B is the number of bounding boxes generated for each grid. With YOLO-V4, $\lambda_{\text{coord}} = 5$,$S = 7$, $B = 9$ are the parameter settings of this article. Note that $1_{\text{ij}}^{\text{obj}} = 1$ is defined as the j frame where the object falls into the grid and $1_{\text{ij}}^{\text{obj}} = 1$ is the center of the frame, $(\hat{x_i}, \hat{y_i}, \hat{r_i})$ are the center coordinates and radius of the predicted bounding circle. Furthermore, (x_i, y_i, r_i) are the center coordinates and radius of the true bounding circle and

$$\begin{aligned}\text{loss}_{\text{IOU}} = \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{\text{ij}}^{\text{obj}} (-c_i \log \hat{c_i}) \\ + \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{\text{ij}}^{\text{noobj}} [-(1 - c_i)\log(1 - \hat{c_i})] \end{aligned} \tag{11}$$

where C_i is the confidence of the true value, $\hat{C_i}$ is the confidence of the predicted value. We also set $\lambda_{\text{noobj}} = 0.5$, which is the corresponding weight of the IOU loss. Furthermore,

$$\begin{aligned}\text{loss}_{\text{classification}} = \sum_{i=0}^{s^2} \sum_{j=1}^{B} 1_{\text{ij}}^{\text{obj}} \sum_{c \in \text{classes}} [-p_i(c) \log \widehat{p_i}(c) \\ - (1 - p_i(c))\log(1 - \widehat{p_i}(c))] \end{aligned} \tag{12}$$

where c is the classification of the detected target, $p_i(c)$ is the true probability of the target, $\widehat{p_i}(c)$ is the predicted classification of the target and $\text{loss}_{\text{classification}}$ is the sum of the error classifications in the classification grid. This indicator is also used in [19, 22].

The YOLO-V4-Densenet model adopted in this paper then predicts that the target detection of cherries could be carried out through classification by proportion.

4 Experiments and discussion

With the development of deep learning, many deep learning algorithms have been applied to target detection tasks. Jing Zhang et al. used deep features and regional convolutional neural networks to establish branch detection methods. They then used these deep learning algorithms to detect apple trees. The convolutional neural network of the algorithm is composed of an improved AlexNet network. Using this network improves the detection accuracy by using pseudo-color and depth images [28]. Shaohua Wan et al. also used the improved Faster R-CNN deep learning framework in the research process of multi-category fruit detection. The results indicated that the algorithm makes the detection speed and short detection time advantages [29]. Yangyu et al. employed mask region convolutional neural network (Mask R-CNN) to detect mature strawberries. This algorithm combines the Resnet50 backbone network with the FPN and achieves highly accurate detection results for overlapping fruits [30]. Zhi-Feng Xu et al. utilized YOLO-V3-tiny for real-time detection of tomato, enhanced the tomato data, improved the model to detect tomatoes in complex environments, improved the detection speed, and applied it to complex robotic picking [31]. These deep learning target detection algorithms can detect objects in different environments and states therefore enable robots to pick them and select appropriate algorithms to improve the accuracy and speed of detection.

To identify cherries quickly and accurately, we need to analyze cherry detection performance. This paper mainly studies different target detection algorithms such as YOLO-V3, YOLO-V3-Dense, and YOLO-V4 to train cherry fruit detection models and compares the models with the same data set. Through experiments, we know that

the accuracy and training speed of YOLO-V4 are better than YOLO-V3 and YOLO-V3-Dense, and YOLO-V4 detects more cherries than YOLO-V3 and YOLO-V3-Dense models. This article mainly uses a model that combines YOLO-V4 and DenseNet network models to identify cherry fruits. DenseNet is closer to the backbone network model of YOLO-V4, which makes the network level deeper. The results show that the combination of YOLO-V4 and dense achieves accurate and rapid detection of cherry fruits.

This article uses the YOLO-V4-Dense model. The processor intel core(TM) i7-7700, the main frequency is 3.60 GHz, the memory is 4 GB, and the GPU is NVIDIA TesIa V100 to train the detection model. We also use Python3.7.6, the compilation script is pycharm, pytroch = 1.2.0, torchvision = 0.4.0, and the PyTorch framework that supports Cuda (v10.0) has been used for training. To reduce the requirements of hardware equipment, we set the epoch value to 8, the learning rate to 0.001, the momentum value to 0.9, the weight attenuation to 0.00005, and the training period to 150. During the training, a 10% training set is randomly selected to adjust the training parameters. The number is 30,000, the test set processes an image in an average of 1 s, and the cherry image is adjusted to 1280 × 800 pixels.

4.1 The impact of data set size

The amount of training data affects cherry model training. Here we consider three data sets for training to evaluate the impact of the size of the dataset on the model. In the mature stage, semi-mature stage, an immature stage of the data set, 10, 50, 100, 200 and 400 pictures are randomly selected for training. The results of different data sets are presented in Table 2. As it is seen, the larger the number of train sets, the higher the accuracy of the model detection results. The F_1 scores of the trained with different size of datasets are shown in Fig. 8.

4.2 Long and short view, no cherry fruit influence on fruit

In the datasets, there are distant and close cherries shots. In distant pictures, it is relatively difficult to recognize cherries in the process of acquiring pictures of cherries. Here, we use 50 cherry vision images and 50 cherry fruits before growth for model detection. The results show that in the context of real applications, due to the limitation of the size of cherries, the visibility of cherries is reduced when shooting in the distant view thus detection of cherries becomes more difficult. Using this model, only the cherry fruits with obvious characteristics can be detected in distant pictures.

Table 2 F_1 scores of models trained with different numbers of images

Number of images	10	50	100	200	400
F_1 score	0.457	0.615	0.727	0.851	0.947

The experimental results show that the YOLO-V4-dense model improves the detection accuracy of cherry fruits. Using 50 pictures of the external environment of the cherry orchard and cherry flowers for testing, the results showed that the model is unable to identify photographs without cherry fruits. It shows that the model can eliminate other interferences and better detect cherry fruits.

4.3 The effect of occlusion and overlap

In the growth environment of cherry fruits, there are problems such as overlap between fruits and occlusion of leaves and branches which make fruit detection difficult. From the IOU and F_1 score values of the YOLO-V4-dense network shown in Table 4, it can be seen that when the fruit is occluded, YOLO-V4-Dense can detect most of the occluded immature, semi-ripe, and mature cherry fruits. During the detection process, some fruits are blocked by leaves that cannot be detected, and some of the blocked branches and leaves are also mistakenly detected as cherry fruits. To a certain extent however it seems that the detection of occlusion and overlapping cherries is optimized. It is further seen that YOLO-V4 is better than YOLO-V3 in occlusion detection. The generated PR curve by model training is presented in Fig. 9. By comparing with the mAP value of YOLO-V4, the mAP value of YOLO-V4-dense increased by 0.15. The higher the mAP value is, the better the detection result of convolutional neural network is [32].

4.4 Comparison of different models in detecting ripe cherries

To verify the superiority of the proposed model this article uses the three growth stages of immature, semi-ripe and mature cherries as the training sets. The data sets are trained in YOLO-V3, YOLO-V3-Dense and YOLO-V4, respectively. The results show that during the training process, YOLO-V3-dense has a better convergence effect than that of YOLO-V33, and the loss reduction is close to 1. This suggests that YOLO-v3-dense improves performance. In terms of speed, YOLO-V3-dense is relatively slow due to its more complex model than YOLO-V3. YOLO-V4 can also detect objects more widely and the detection types are more abundant than that of YOLO-V3.

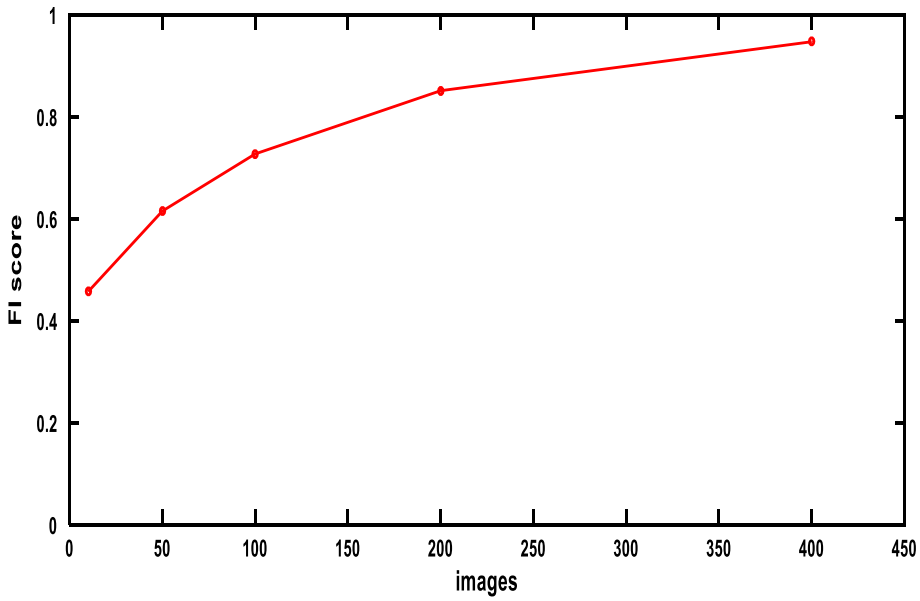


Fig. 8 The F_1 scores of the trained with different size of datasets

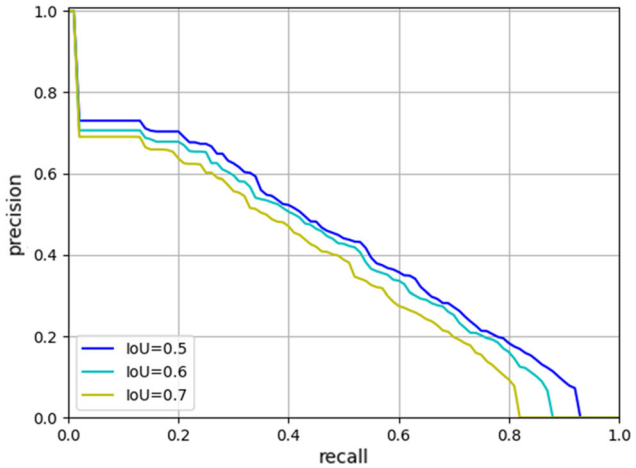


Fig. 9 PR curve for cherry detection

Fig. 10 Improved YOLO-V4 to detect cherry images

YOLO-V4 can detect objects that are missed by YOLO-V3 hence improve the detection performance. The YOLO-V4-Dense used in this paper is the best amongst the examined techniques for cherry fruit recognition. The F_1 scores, IOU, and average detection time of different models are shown in Table 3. The improved YOLO-V4 model for cherry fruit detection with and without a round frame is shown in Figs. 10 and 11, respectively.

In the YOLO-V4-Dense network, the loss is reduced by about 0.8 compared to the YOLO-V4 model, and when the loss converges to 39,000 steps, it starts to approach a saturation state. Regarding the detection performance, due to the network complexity and density of the YOLO-V4-dense model, the detection accuracy is higher than other models, and the F_1 scores value of YOLO-V4-dense is 0.947. F_1 score is about 0.15 higher than YOLO-V3, 0.13 higher than YOLO-V3-Dense, and 0.02 higher than YOLO-V4. The results also indicate that the accuracy of the circular bounding box used by YOLO-V4-dense is higher than that of the other three models. The training time of the YOLO-V4-Dense network model is relatively

Table 3 $F1$ scores, IOU, average time(s) of different models

Models	YOLO-V3	YOLO-V3-dense	YOLO-V4	YOLO-V4-dense
F_1 scores	0.801	0.816	0.935	0.947
IOU	0.879	0.896	0.844	0.856
Average time (s)	0.299	0.304	0.414	0.467

Fig. 11 Improved YOLO-V4 to detect cherry images with round frame

high compared to other models, but the accuracy and confidence values are high. The results show that the YOLO-V4-Dense network model can improve the accuracy of cherry detection. The YOLO-V3 and YOLO-V4 models detect cherry fruits as shown in Fig. 12.

4.5 Data category impact

This paper selects three growing stages of cherry fruits for testing including ripe cherries, semi-ripe cherries and immature cherries. By testing different types of cherries, it is better to observe the growth of cherries. If the same cluster is detected in the case of uneven growth during the fruit growth process, adjustments can be made urgently to ensure the high-quality production of cherry fruits. The YOLO-V4-Dense model is used to detect the IOU and F_1 scores of different levels of ripeness of cherries. The F_1 score and IOU of the model training are shown in Table 4.

The IOU of the immature cherries (Cherry_2), semi-mature cherries and mature cherries tested by the improved YOLO-V4 model are 0.889, 0.895 and 0.905, respectively. F_1 scores are 0.949, 0.949 and 0.958. It can be seen from the table that the detection accuracy of ripe cherries is higher than that of those in other stages, with an average of 0.08 higher. And it can be seen from the experimental results that the model can successfully identify cherries. In the real training process, the unlabeled area defaults to the background when manually labeled. During the training process, three categories are used for output, including ripe, semi-ripe, and immature cherries. In a cherry picture, the positions of cherries at different growth periods can be then detected. Experiments show that the YOLO-V4-Dense network overperforms the other methods for cherry fruit detection.

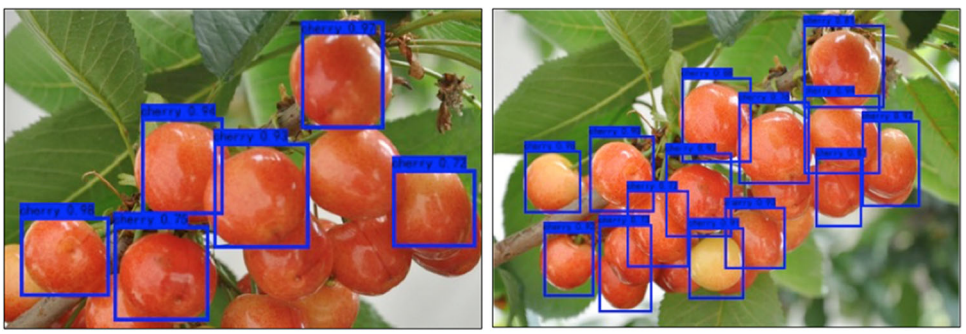
Table 4 $F1$ scores of models trained with different numbers of images

Growth period	IOU	F_1 scores
Cherry_2	0.889	0.949
Cherry_1	0.895	0.952
Cherry	0.905	0.958

5 Conclusion

This paper proposes a detection method for cherry fruits based on an improved YOLO-V4 model. To improve the detection accuracy of cherry fruits, we substitute the backbone network of the YOLO-V4 model with the DenseNet network model. We also introduce a new data set for our experiments. The proposed model, YOLO-V4-dense. Accurately detects ripe, semi-ripe, and unripe fruits. Ripe cherry fruits. Comparisons with YOLO-V3, YOLO-V3-dense, and YOLO-V4 models, confirm that the proposed YOLO-V4-Dense model can improve the detection accuracy by training based on the data set, and its mAP is increased by 0.15 compared with the YOLO-V4 model. The model takes 0.467 for a cherry picture with a resolution of 1280 × 800 pixels. The F_1 scores of YOLO-V4-Dense, IOU is 0.947 and 0.856. The model takes 0.467 for a cherry picture with a resolution of 1280 × 800 pixels.

Furthermore, the real-time detection efficiency of the proposed model is better than the existing models where



(a) (b)

Fig.12 YOLO-V3 and YOLO-V4 to detect cherry images

the cherry fruits in the image are overlapped and/or occluded.

Detecting the ripeness of cherry fruits enables robotic intelligent picking, helps the robot to plan the picking path and results in efficient and intelligent cherry-picking hence greatly reducing required labor costs. Through manual intervention at different stages, the efficiency of cherry-picking can be further increased and contribute to the digitalization and informatization of agriculture.

Our future work will continue to explore new techniques to increase the output of cherry-picking given the challenges caused by cherry flowers, leaves, and buds, and the complex environment.

Author's contribution All authors contributed equally.

Funding The research was funded by Dalian Science and Technology Bureau in Grant No. 2020JJ26SN058.

Declarations

Conflict of interest None.

Ethics approval Confirm.

Consent to participate Confirm.

Consent for publication Confirm.

Project Research on Feature Extraction and Modeling of Growth Dynamics of Sweet Cherries in solar greenhouse based on intelligence of IOT.

References

1. Zhao Y, Gong L, Huang Y, Liu C (2016) A review of key techniques of vision-based control for harvesting robot. Comput Electron Agric 127:311–323. https://doi.org/10.1016/j.compag.2016.06.022
2. Sparrow R, Howard M (2020) Robots in agriculture: prospects, impacts, ethics, and policy. Precis Agric. https://doi.org/10.1007/s11119-020-09757-9
3. Li JB, Huang WQ, Zhao CJ (2014) Machine vision technology for detecting the external defects of fruits — a review. J Photogr Sci 63(5):241–251. https://doi.org/10.1179/1743131X14Y.0000000088
4. Gao F, FuZhang LX, Majeed Y, Li R, Karkee M, Zhang Q (2020) Multi-class fruit-on-plant detection for apple in SNAP system using faster R-CNN. Comput Electron Agric 176:105634. https://doi.org/10.1016/j.compag.2020.105634
5. Fu L, Gao F, Wu J, Li R, Karkee M, Zhang Q (2020) Application of consumer RGB-D cameras for fruit detection and localization in field: a critical review. Comput Electron Agric 177:105687. https://doi.org/10.1016/j.compag.2020.105687
6. Linker R, Cohen O, Naor A (2012) Determination of the number of green apples in rgb images recorded in orchards. Comput Electron Agric 81:45–57. https://doi.org/10.1016/j.compag.2011.11.007
7. Momeny M, Jahanbakhshi A, Jafarnezhad K, Zhang YD (2020) Accurate classification of cherry fruit using deep cnn based on hybrid pooling approach. Postharvest Biol Technol 166:111204. https://doi.org/10.1016/j.postharvbio.2020.111204
8. Shi R, Li T, Yamaguchi Y (2020) An attribution-based pruning method for real-time mango detection with YOLO network. Comput Electron Agric 169:105214. https://doi.org/10.1016/j.compag.2020.105214
9. Rodríguez JP, Corrales DC, Aubertot JN, Corrales JC (2020) A computer vision system for automatic cherry beans detection on coffee trees. Pattern Recogn Lett 136:142–153. https://doi.org/10.1016/j.patrec.2020.05.034
10. Chang X, Yu YL, Yang Y, Xing EP (2016) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Mach Intell 39:1617–1632. https://doi.org/10.1109/TPAMI.2016.2608901
11. Yan C, Zheng Q, Chang X, Luo M, Yeh CH, Hauptman AG (2020) Semantics-preserving graph propagation for zero-shot object detection. IEEE Trans Image Process 29:8163–8176. https://doi.org/10.1109/TIP.2020.3011807
12. Wang H, Li Z, Li Y, Gupta BB, Choi C (2018) Visual saliency guided complex image retrieval. Pattern Recogn Lett 130:64–72. https://doi.org/10.1016/j.patrec.2018.08.010
13. Majeed Y, Zhang J, Zhang X, Fu L, Karkee M, Zhang Q, Whiting MD (2020) Deep learning based segmentation for automated training of apple trees on trellis wires. Comput Electron Agric 170:105277. https://doi.org/10.1016/j.compag.2020.105277
14. Liu Z, Wu J, Fu L, Majeed Y, Cui Y (2020) Improved kiwifruit detection using pre-trained vgg16 with rgb and nir information fusion. IEEE Access 8(1):2327–2336. https://doi.org/10.1109/ACCESS.2019.2962513
15. Sedik A, Hammad M, Abd El-Samie FE et al (2021) Efficient deep learning approach for augmented detection of coronavirus disease. Neural Comput Appli. https://doi.org/10.1007/s00521-020-05410-8
16. Bargoti S Underwood J (2017) Deep fruit detection in orchards.In: IEEE International Conference on Robotics and Automation (ICRA) pp. 3626–3633. https://doi.org/10.1109/ICRA.2017.798 9417
17. RedmonJ Divvala S Girshick R Farhadi A (2016) You only look once: unified, real-time object detection.In: IEEE conference on Computer Vision and Pattern Recognition (CVPR) pp. 779–788. https://doi.org/10.1109/CVPR.2016.91
18. Redmon J , Farhadi A (2018) Yolov3: an incremental improvement. arXiv e-prints.
19. Liu G, Nouaze JC, Touko PL, Kim JH (2020) Yolo-tomato: a robust algorithm for tomato detection based on yolov. Sensors. https://doi.org/10.3390/s20072145
20. Kuznetsova A, Maleva T, Soloviev V (2020) Using yolov3 algorithm with pre- and post-processing for apple detection in fruit-harvesting robot. Agronomy 10(7):1016. https://doi.org/10.3390/agronomy10071016
21. Bochkovskiy A , Wang C Y , Liao H Y M (2020) Yolov4: optimal speed and accuracy of object detection
22. Wu D, Lv S, Jiang M, Song H (2020) Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. Comput Electron Agric 178:105742. https://doi.org/10.1016/j.compag.2020.105742
23. Alsmirat MA, Al-Alem F, Al-Ayyoub M et al (2019) Impact of digital fingerprint image quality on the fingerprint recognition accuracy. Multimed Tools Appl 78:3649–3688. https://doi.org/10.1007/s11042-017-5537-5
24. Yu C, Li J, Li X et al (2018) Four-image encryption scheme based on quaternion Fresnel transform, chaos and

computer generated hologram. Multimed Tools Appl 77(4):4585–4608. https://doi.org/10.1007/s11042-017-4637-6
25. Zhang X, Fu L, Karkee M, Whiting MD, Zhang Q (2019) Canopy segmentation using resnet for mechanical harvesting of apples. IFAC-PapersOnLine 52(30):300–305. https://doi.org/10.1016/j.ifacol.2019.12.550
26. Huang G, Liu Z, Laurens VDM, Weinberger KQ (2016) Densely connected convolutional networks. https://doi.org/10.1109/CVPR.2017.243
27. Tian Y, Yang G, Wang Z, Wang H, Li E, Liang Z (2019) Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Comput Electron Agric 157:417–426. https://doi.org/10.1016/j.compag.2019.01.012
28. Zhang J, He L, Karkee M, Zhang Q, Zhang X, Gao Z (2018) Branch detection for apple trees trained in fruiting wall architecture using depth features and Regions-Convolutional Neural Network (R-CNN). Comput Electron Agric 155:386–393. https://doi.org/10.1016/j.compag.2018.10.029
29. Wan S, Goudos S (2019) Faster R-CNN for multi-class fruit detection using a robotic vision system. Comput Netw 168:107036. https://doi.org/10.1016/j.comnet.2019.107036
30. Yu Y, Zhang K, Yang L, Zhang D (2019) Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Comput Electron Agric 163:104846. https://doi.org/10.1016/j.compag.2019.06.001
31. Xu Z, Jia R, Liu Y, Zhao C, Sun H (2020) Fast method of detecting tomatoes in a complex scene for picking robots. IEEE Access 8:55289–55299. https://doi.org/10.1109/ACCESS.2020.2981823
32. Zhang J, Karkee M, Zhang Q, Zhang X, Yaqoob M, Fu L, Wang S (2020) Multi-class object detection using faster R-CNN and estimation of shaking locations for automated shake-and-catch apple harvesting. Comput Electron Agric 173:105384. https://doi.org/10.1016/j.compag.2020.105384