applied sciences

MDPI

Article

Detection of *Camellia oleifera* Fruit in Complex Scenes by Using YOLOv7 and Data Augmentation

Delin Wu *, Shan Jiang, Enlong Zhao, Yilin Liu, Hongchun Zhu, Weiwei Wang and Rongyan Wang

School of Engineering, Anhui Agricultural University, Hefei 230036, China
* Correspondence: wudelin@ahau.edu.cn; Tel.: +86-158-0560-2399

Abstract: Rapid and accurate detection of *Camellia oleifera* fruit is beneficial to improve the picking efficiency. However, detection faces new challenges because of the complex field environment. A *Camellia oleifera* fruit detection method based on YOLOv7 network and multiple data augmentation was proposed to detect *Camellia oleifera* fruit in complex field scenes. Firstly, the images of *Camellia oleifera* fruit were collected in the field to establish training and test sets. Detection performance was then compared among YOLOv7, YOLOv5s, YOLOv3-spp and Faster R-CNN networks. The YOLOv7 network with the best performance was selected. A DA-YOLOv7 model was established via the YOLOv7 network combined with various data augmentation methods. The DA-YOLOv7 model had the best detection performance and a strong generalisation ability in complex scenes, with mAP, Precision, Recall, F1 score and average detection time of 96.03%, 94.76%, 95.54%, 95.15% and 0.025 s per image, respectively. Therefore, YOLOv7 combined with data augmentation can be used to detect *Camellia oleifera* fruit in complex scenes. This study provides a theoretical reference for the detection and harvesting of crops under complex conditions.

Keywords: object detection; YOLOv7; data augmentation; convolutional neural network; *Camellia oleifera* fruit

Citation: Wu, D.; Jiang, S.; Zhao, E.; Liu, Y.; Zhu, H.; Wang, W.; Wang, R. Detection of *Camellia oleifera* Fruit in Complex Scenes by Using YOLOv7 and Data Augmentation. *Appl. Sci.* **2022**, *12*, 11318. https://doi.org/10.3390/app122211318

Academic Editor: Xiumin Wang

Received: 25 October 2022
Accepted: 5 November 2022
Published: 8 November 2022

1. Introduction

Camellia oleifera fruits can be used to extract tea oil [1]. Tea oil is a high-quality edible oil with clear colour, rich nutrient content and storage resistance, so it has good economic benefits and market prospects [2]. However, manual picking requires high labour costs because of the complex growth environment of *Camellia oleifera*, such as hills and mountains [3]. Therefore, the mechanisation of harvesting *Camellia oleifera* fruits should be improved to increase efficiency, save labour cost and promote the development of the Camellia industry [4]. A key technology of automatic picking is to rapidly detect the location of the fruit in field environments. Therefore, a rapid and accurate method for detection of *Camellia oleifera* fruits in complex field environments should be developed.

Existing methods for crop detection mainly rely on imaging technology [5–7]. Traditional image detection algorithms are mainly based on the colour and shape of the target, which are different from other objects [8]. These complex algorithms with many fixed thresholds have certain application limitations, such as error detection in complex scenes and lack of sufficient robustness [9]. Deep learning algorithms have been widely used in crop detection to effectively extract target features in complex scenes and overcome the limitations of traditional algorithms [10]. Convolutional neural network (CNN), a kind of feedforward neural networks (FNN), involves convolution computation and has a deep structure. As a representative deep learning algorithm, CNN has been widely used in classification [11], localisation [12], detection [13] and segmentation [14] of crops and fruits. A variety of CNN-based detection algorithms, such as YOLO v3 [15], YOLOv5 [16,17] and Faster R-CNN [18], have been used to detect fruit targets. Therefore, in the present study, image detection technology and CNN will be used to detect *Camellia oleifera* fruits.

YOLO is a commonly used single-stage target detection algorithm with the characteristics of fast and high accuracy [19,20]. It exhibits satisfactory performance in detecting small and occluded targets in complex field environments and has better detection speed than other deep learning algorithms [21,22]. YOLOv7 is the latest detector in YOLO series. This network is designed with trainable bag-of-freebies, which enable real-time detectors to greatly improve the accuracy without increasing the inference cost. It also involves extend and compound scaling so the target detector can effectively reduce the number of parameters and calculations, thereby greatly improving the detection speed [23]. At pre-sent, YOLOv7, as a brand-new detector, has not been applied to fruit detection. Therefore, in the present work, YOLOv7 was used to detect *Camellia oleifera* fruits.

Camellia oleifera orchards have a complex environment, where an equisized image of *Camellia oleifera* fruit will be disturbed by sidelight, backlight, slight occlusion and heavy occlusion, which will lead to false detection or missed detection of targets. The training image should include more scenes to extract features and overcome the interference of complex scenes [24]. However, the number of images is limited due to constraints of or-chard and acquisition time when *Camellia oleifera* fruit images are collected in the field. Therefore, existing research on deep learning usually enhances existing data to obtain more training data and achieve better generalisation of the neural network. Traditional data augmentation methods include mirroring, rotating, changing brightness and adding noise [25]. Mosaic is a new data augmentation method for mixing multiple images, and it greatly enriches the background of detected objects [26]. These methods can increase the number of datasets and improve the robustness of the detection models in complex scenes [27]. In this study, the traditional and Mosaic data augmentation methods are combined to develop a detection model for *Camellia oleifera* fruit in complex scenes.

To solve the difficulty of *Camellia oleifera* fruit detection in complex environments, this study proposes a detection method based on imaging technology and YOLOv7 network combined with image augmentation. This study aims to: (1) acquire and pre-process *Camellia oleifera* fruit images in complex conditions to establish detection datasets; (2) develop a YOLOv7 detection model and compare its performance with Faster RCNN, YOLO v3 and YOLOv5s models in complex environment; and (3) build an augmented dataset by combining multiple augmentation methods, compare the performance of YOLOv7 models based on augmented and original datasets and select the optimal model.

2. Materials and Methods

2.1. Acquisition of Camellia oleifera Fruit Images

The fruits of *Camellia oleifera* in standardised planting orchards were used as the research object. Original images of *Camellia oleifera* fruits were collected from orchards in Qinglongwan Ecological Garden, Tongcheng City, Anhui Province and planting bases in Yongzhou City, Hunan Province. In standard planting mode, the row spacing of Camellia oleifera trees was both 2 m. The plant spacing was about 1 m, and the tree height was about 1.8–2.5 m. All images of *Camellia oleifera* fruit were obtained in August 2021. Image acquisition was conducted in the morning, noon and afternoon under sunny and cloudy weather and natural light condition in the field. A total of 100 *Camellia oleifera* fruit trees with good growth were selected by random sampling, and the tree age was 8 to 10 years. All the *Camellia oleifera* trees were not picked in the same year to ensure that the growth form of the fruit was not destroyed. and different angles were selected to capture images at different shooting distances (0.5–1.5 m), with the camera from the ground height of 1–2 m. The acquired images had the following conditions: slight occlusion, heavy occlusion, overlapped, natural light angle, sidelight angle, backlight angle, etc. Examples of acquired images are shown in Figure 1. Slight occlusion is when the part of the fruit occluded by branches and leaves is less than one third of the total area. Heavy occlusion is when the part of fruit occluded by branches and leaves is more than one third and less than two thirds of the total area. Sidelight angle is when the lens direction and the direct sunlight direction is 90° when shooting the images. Backlight angle is when the lens direction

and the direct sunlight direction is 180° when shooting the images. Eight to 12 images of *Camellia oleifera* fruit were taken for each fruit tree. A total of 873 images of *Camellia oleifera* fruit were obtained after removing the blurred or repeated images. A single-lens reflex camera (Canon 200DII, Canon Inc., Tokyo, Japan) in "AUTO" mode with a resolution of 4608 × 3456 pixels was used to acquire the images saved in JPG format.

Figure 1. Examples of *Camellia oleifera* fruit images acquired under different conditions. (**a**) slight occlusion, (**b**) heavy occlusion, (**c**) overlapped *Camellia oleifera* fruit, (**d**) natural light angle, (**e**), sidelight angle and (**f**) backlight angle.

2.2. Image Preprocessing and Dataset Partitioning

Firstly, 200 images (50 of slight occlusion, 50 of heavy occlusion, 50 of sidelight angle and 50 of backlight angle) were randomly selected from 873 images as the test set to evaluate the generalisation of the detection model. The remaining 673 images were randomly divided into a training set (606 images) and a validating set (67 images) with a ratio of 9 to 1. No repeated images among the training, validation and test sets were ensured to prevent overfitting of the model [28].

Image data annotation software 'LabelImg' was used to draw the outer rectangle of the *Camellia oleifera* fruit target in all images of the training set to complete the manual labelling of the fruit. Images were labelled based on the smallest surrounding rectangle of the *Camellia oleifera* fruit to ensure that the rectangle contains the background area as little as possible. Examples of labeled *Camellia oleifera* fruit images are shown in Figure 2. XML format files were generated after the annotations were saved [29].

Figure 2. Examples of labeled *Camellia oleifera* fruit images.

2.3. Data Augmentation

When establishing a deep learning-based object detection model for *Camellia oleifera* fruit, a high-quality dataset with a large amount of image data can improve the quality of model training and prediction accuracy. Therefore, the acquired *Camellia oleifera* fruit images should be augmented [30].

Several image augmentation methods were utilised for the 606 images of the training set to improve the generalisation ability and avoid the overfitting of the detection model. These methods were based on Pycharm software and its related image processing library. The image augmentation methods included horizontal mirroring, vertical mirroring, brightness enhancement and reduction, multi-angle rotation (90°, 180°, 270°), adding Gaussian noise and Mosaic data augmentation. The detailed steps of the image augmentation methods are illustrated as follows.

Multi-angle rotation of an image enables the deep learning model to learn more object features in different positions and directions during training. OpenCV function "cv2.getRotationMatrix2D" and "cv2.warpAffine" based on Python were employed to rotate the original image. Image rotation was conducted by changing the parameter "angle" in the function as 90°, 180° and 270°.

Image mirroring (horizontal and vertical mirroring) can increase the viewing angle of the *Camellia oleifera* fruit. Opencv function 'flip' was used to mirror the original image. The image was divided into left and right parts for symmetrical transformation of the image centred on the vertical axis to achieve horizontal mirroring when the parameter 'dim' was set to 1. The image was divided into upper and lower parts for symmetric transformation of the image centred on the horizontal axis to achieve vertical mirroring.

Image brightness was enhanced and reduced. The complex light conditions of the plantation caused differences in *Camellia oleifera* fruit images, thereby interfering with the detection results. Therefore, the values of the three channels of the pixel points of the original image were multiplied by 0.5 and 1.5 to enhance and reduce the brightness of the image. This method improved the robustness of the model.

Adding Gaussian noise to the image was also conducted. The unclear or blurred images captured by the shaking of the equipment or the branches would affect the accuracy of the detection model. A Gaussian noise with a parameter 'sigma' of 25 was added to the original image to simulate the low-quality image that the model may capture in practical applications.

Mosaic data augmentation was performed referring to CutMix data augmentation method. During training, the input size of the model was assumed as $S \times S$ and a $2S \times 2S$ grey image was marked as a canvas. A point from the rectangle framed by point A (S/2, S/2) and point B (3S/2, 3S/2) was set as the reference point coordinate. Four images were randomly selected and stitched into the image by random scaling, cutting and arrangement. The images and labelled boxes beyond the canvas were ignored. Mosaic data augmentation increased the training data in each BatchSize without increasing the BatchSize to reduce the memory requirements of the model. The mean and variance of each feature layer were calculated during the batch normalisation (BN) operation and were closer to the mean and variance of the entire dataset. Mosaic data augmentation enriched the background of the image, and the image formed by splicing multiple images added numerous small-object *Camellia oleifera* fruit, thereby improving the detection accuracy of the detection model.

The final augmented training set consists of 5854 images, including 606 original images and 5248 enhanced images. The detailed distribution of the augmented training set is shown in Figure 3.

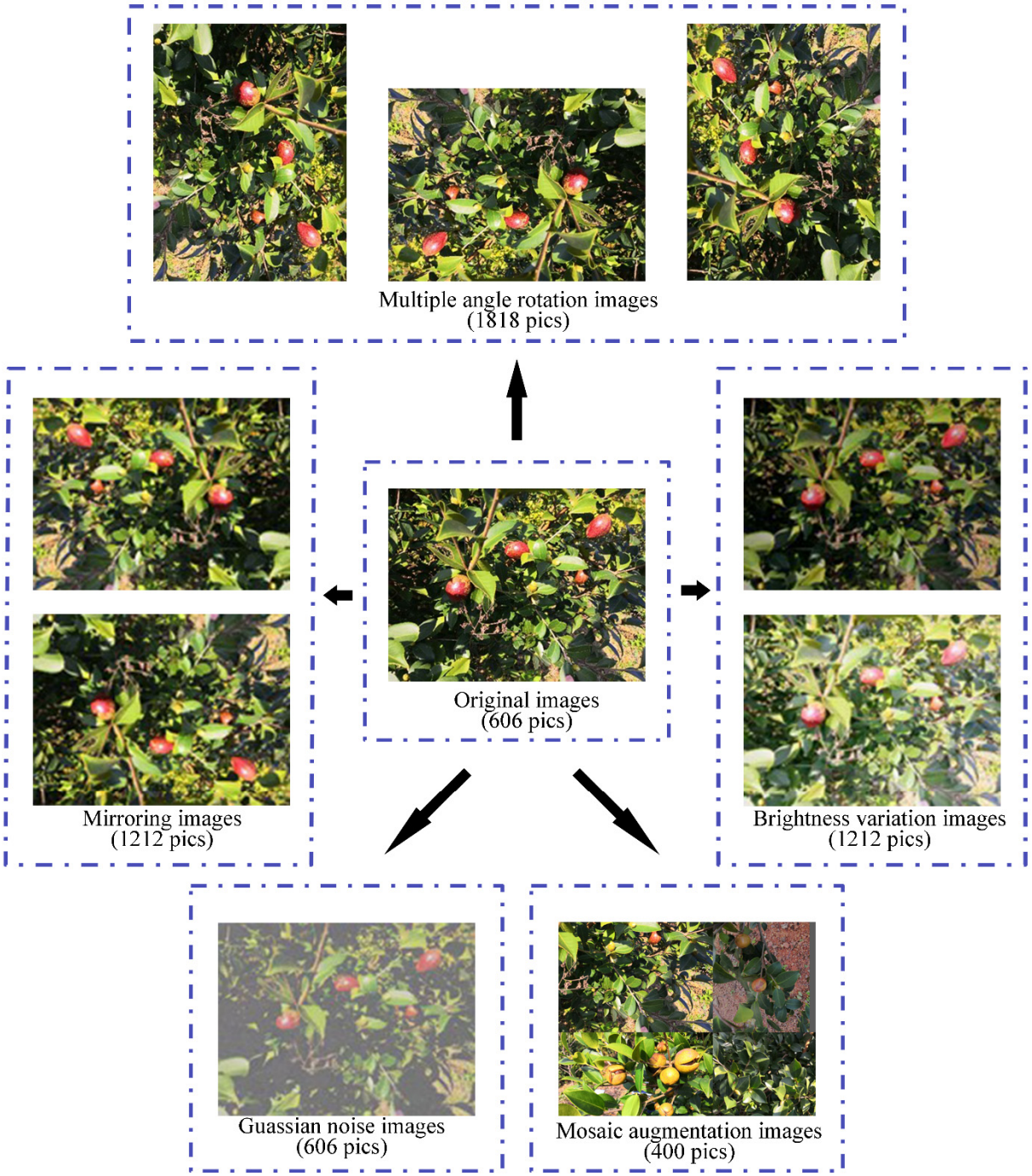

Figure 3. Distribution of the augmented training set.

2.4. YOLO v7 Network Architecture

YOLOv7, a latest detector with YOLO architecture, is an object detection network that has fast detection speed, high precision and easy to train and deploy characteristics. The speed and accuracy of the network is within the range of 5–160 FPS, surpassing currently known object detectors. The network is 120% faster than YOLOv5 in the same volume (FPS). The test results on the MS COCO dataset outperform the YOLOv5 detector [31]. Figure 4 shows the network structure of YOLOv7.

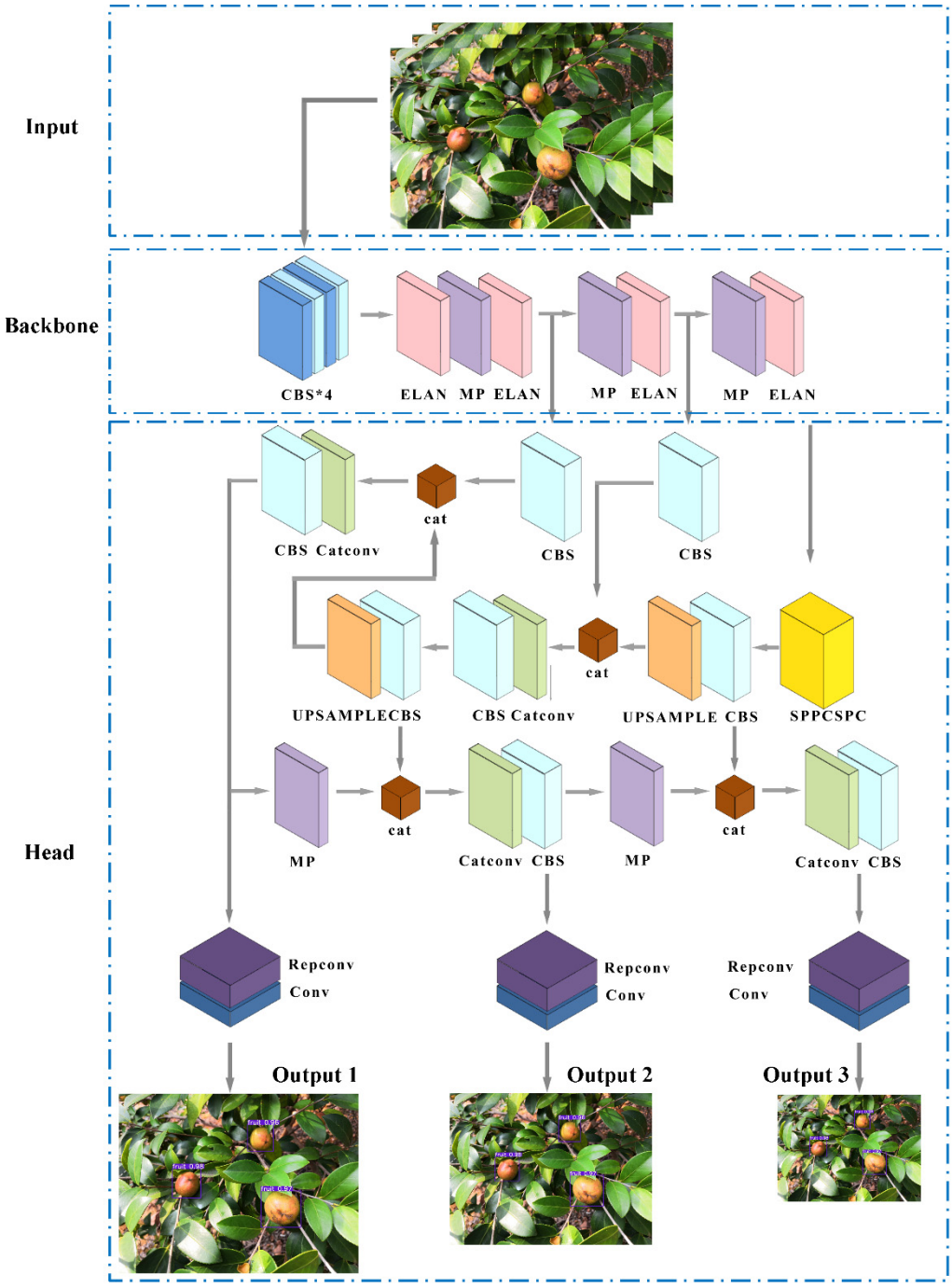

Figure 4. Architecture of YOLOv7 network.

Based on the structure diagram, the YOLOv7 network consists of three parts, namely, input network, backbone network and head network. The YOLOv7 network firstly pre-processed the image, resized it to 640 × 640 × 3 and inputted it into the backbone network. The CBS composite module, efficient layer aggregation networks (ELAN) module and MP module alternately reduced the length and width of the feature map by 1/2, and the number of the output channels was increased to twice the number of input channels. As shown in Figure 5, the CBS composite module performed the convolution + BN + activation function on the input feature map. In YOLOv7, the same as YOLOv5, Silu was used as the activation function. ELAN module was proposed. It used expand, shuffle and merge cardinality to continuously improve the learning ability of the network without destroying the original gradient path, thereby improving the accuracy of the network. The ELAN

structure was composed of different convolutions. The group convolution was used to expand the channel and cardinality of the computational blocks, while ensuring the number of channels in each set of feature maps to be the same as the number of channels in the original architecture. Finally, the number of channels derived from the ELAN module was twice that of the input. The upper branch of the MP module halved the length and width of the feature map by maxpooling operation, and the channels were halved by convolution. The lower branch halved the channels by the first convolution, and the second convolution with kernel size of 3 and stride of 2 halved the length and width of the feature map. The upper and lower branches were combined. Finally, the output feature map with half length and width and equal input and output channels was obtained.

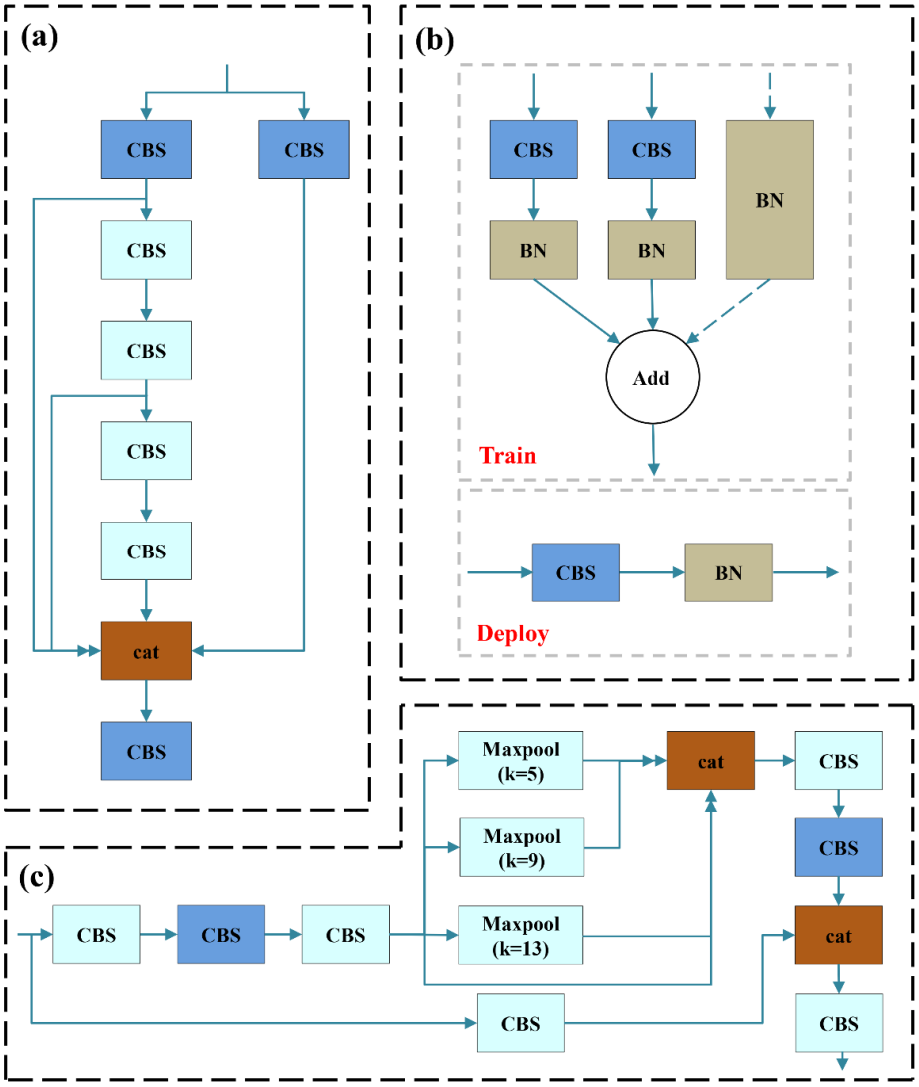

Figure 5. Structure of each module. (**a**) ELAN module; (**b**) Repconv module; (**c**) SPPCSPC module.

Based on the three-layer output in the backbone network, the head network continued to output three layers of feature maps of different sizes. After the Repconv module adjusted the final number of the output channels, three layers of convolution operation of kernel_size = 1 (1 × 1) were used to proceed to objectness, class and bbox prediction tasks for image detection to obtain results. The head network consists of SPPCSPC module, a series of CBS modules, MP module, Catconv module and three subsequent Repconv modules. The SPPCSPC module is similar to the SPPF used by YOLOv5 to increase the receptive field of a network. Firstly, the input feature map with a size of 512 × 20 × 20 was obtained and subjected to three convolution operations. Maxpooling operations with kernel

size of 5, 9 and 13 were performed (for different kernel sizes, padding is adaptive) three times. Finally, the feature map with a size of 512 × 20 × 20 was obtained by combining the results with only 1 × 1 convolution operation data without pooling. The SPPCSPC module can obtain multi-scale object information while keeping the size of the feature maps unchanged. YOLOv7 was used to develop a more standardised model with a re-parameterised structure, namely, Repconv structure [32]. It increased the training time and improved the inference effect [33]. During training, a whole module was split into multiple identical or different module branches and added with 3 × 3 convolution + BN, a 1 × 1 convolution + BN and a BN layer (when the input and output channels were the same) to obtain the training model. During inference, the three parts were re-parameterised, and a 3 × 3 convolution output was used to convert their parameters equivalently to another set of parameters. The multi-branch training model was then transformed into a high-speed single-branch inference model. The final deployed model retained the high accuracy and other excellent properties of the multi-branch model while maintaining high efficiency as well as exhibited good speed and accuracy balance to improve the network performance.

2.5. Training Platform and Parameter Settings

Based on the PyTorch deep learning framework, training and testing were performed on a desktop computer with Windows 10 operating system and Inter Core i7-7800X CPU processor with 32 GB RAM. Considered the needs of the GPU computing power, selected graphics NVIDIA GeForce GTX 3060Ti, video memory 8GB. Python 3.8 was used as the programming language. The software tools included CUDA 11.3, CUDNN 8.2, OpenCV 3.4.5 and Visual Studio 2017.

In this study, YOLOv7 networks trained the *Camellia oleifera* fruit detection model through transfer learning. The training epoch was 300. The batch size of the model training was set to 8. The input size was set to 640 × 640. Regularisation was performed each time through the BN layer to update the model's weight. The momentum factor (momentum) was set to 0.937, and the decay rate (decay) of weight was set to 0.0005. The initial vector was set to 0.01, and the augmentation coefficient of hue (H), saturation (S) and lightness (V) were 0.015, 0.7 and 0.4, respectively. During the training process, Tensorboard visualization tool was used to record data and observe loss, and save the model weight of every epoch.

2.6. Establishment and Evaluation Indicators of Model

2.6.1. Establishment of Model

The establishment of *Camellia oleifera* fruit object detection model was divided into training and testing stages. The YOLOv7 neural network was trained using the training set, and the evaluation indicators were verified on the validation set after model weights were obtained. Finally, the model with the best performance weight was selected as the preliminary model for object detection for *Camellia oleifera* fruit. In the testing phase, the detection model was run on the test set. The prediction results of the models applied to new data were evaluated to ensure the generalisation ability for application to picking machines in the future. The workflow is illustrated in Figure 6. The final output of the neural network is the detection box of the identified *Camellia oleifera* fruit object and the probability (confidence) that the identified object belongs to a specific category.

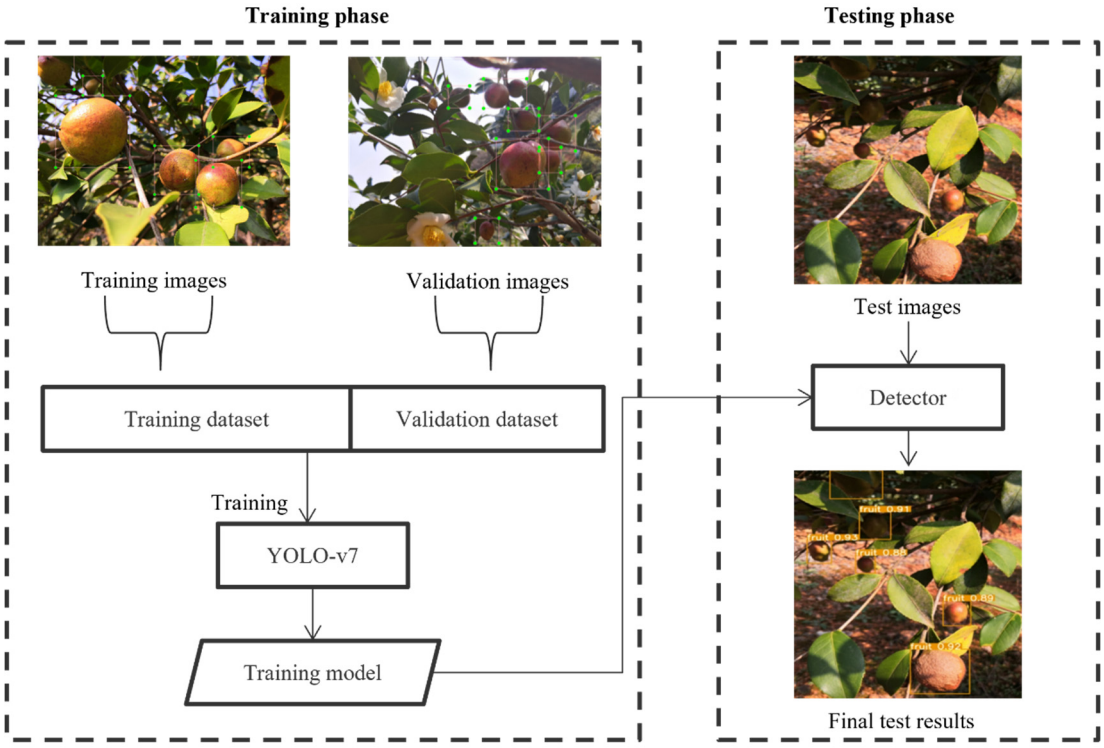

Figure 6. Workflow of the proposed study.

2.6.2. Evaluation Indicators of Model

The function of Complete Intersection over Union (CIoU) loss was used to quantitatively compare the error between the prediction and calibration boxes [34,35]. Figure 7 illustrates the parameters required to calculate CIoU based on the model prediction and calibration boxes, where A is the calibration box, B is the prediction box, l_1 is the distance between the centre points of box A and B, l_2 is the diagonal length of the minimum bounding rectangle of box A and B.

Figure 7. Visualisation of CIoU calculation between model prediction box and the ground truth. Yellow box is the calibration box, blue box is the prediction box.

CIoU was calculated as follows:

$$\text{loss}_{CIoU} = 1 - IOU + \frac{l_1^2}{l_2^2} + \alpha v \tag{1}$$

where v is the similarity of aspect ratio of box A and B and α is the balance factor between the loss caused by v and IoU.

In this paper, Precision, Recall, Mean Average Precision (mAP) and F1 score were used to accurately and objectively evaluate the performance of the model. Precision is the most common evaluation index, and it is the number of right targets divided by the number of detected targets. In general, the higher the Precision is, the better the detection effect will be. Precision is a very intuitive evaluation index, but sometimes high Precision does not represent all. Therefore, mAP, Recall and F1 score were introduced for comprehensive evaluation. Precision, Recall, mAP, and F1 score were calculated as follows:

Precision:

$$\text{P} = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

Recall:

$$\text{R} = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

Average Precision:

$$\text{AP} = \int_0^1 P(r)dr \tag{4}$$

Mean Average Precision:

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} AP_I \tag{5}$$

F1 score:

$$\text{F1} = 2 \times \frac{P \times R}{P + R} \tag{6}$$

where *TP* (True Positive) represents the number of *Camellia oleifera* fruit objects that are correctly detected; *FP* (False Positive) represents the number of other objects detected as *Camellia oleifera* fruit; and *FN* (False Negative) represents the number of *Camellia oleifera* fruit that are undetected/missed.

3. Results and Discussion

3.1. Dataset Training of YOLO v7

The YOLOv7 model for object detection of *Camellia oleifera* fruit was established based on the original dataset and the YOLOv7 network. The fitting curves of training and validation loss for the YOLOv7 model during the process of training are shown in Figure 8, where the horizontal and vertical coordinates represent the number of epochs and the loss value, respectively. The training and validation loss of YOLOv7 decreased rapidly in the first 100 epochs, the value of validation loss is larger than training loss, decreased slowly in the subsequent 100 to 250 epochs and basically stabilised after about 250 epochs. The training and validation loss curves converged, and no overfitting occurred during training. The value of training loss converged to 0.008, and the value of validation loss converged to 0.010. Therefore, this study determined that the model after 300 epochs was the suitable detection model for *Camellia oleifera* fruit.

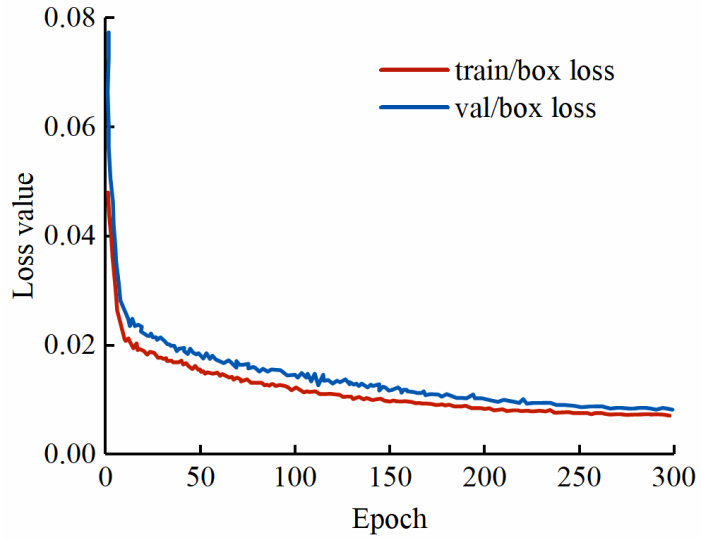

Figure 8. Training and validation loss.

3.2. Comparison of Models

The YOLOv7 model was compared with YOLOv5s, YOLO v3-spp and Faster RCNN networks to verify its accuracy and effectiveness [35]. All these models were established based on the same weights of COCO pre-training and parameters with YOLOv7 model, and the original dataset of *Camellia oleifera* fruit was also used in the models. The comparison of mAP, Precision, Recall, F1 score and detection speed among different models is shown in Table 1. The Precision of YOLOv7 model was 0.4%, 4.51% and 34.86% higher than those of YOLOv5s, YOLO v3-spp and Faster RCNN models, respectively. The mAP of YOLOv7 model increased by 0.98%, 11.44% and 4.24% than those of YOLOv5s, YOLO v3-spp and Faster RCNN models, respectively. Compared with YOLOv5s and YOLO v3-spp, the Recall of YOLOv7 model increased by 3.95% and 5.63%, respectively. The Recall of Faster RCNN model was slightly higher than that of the YOLOv7 model, but the other indicators were significantly lower than those of the YOLOv7 model. The YOLOv7 model had the best F1 score of 93.67%. The detection time of the YOLO models were less than 0.1 s, and that of the YOLOv7 model was only 0.025 s. Faster RCNN is a two-stage target detection model, so the average detection time for a single image is 5.167 s, which is loner than that of the YOLO model. In summary, the YOLOv7 model had higher accuracy and efficiency than YOLOv5s, YOLO v3-spp and Faster RCNN models.

Table 1. Performance comparison among different models.

Target Detection Networks	mAP (%)	Precision (%)	Recall (%)	F1 Score (%)	Average Detection Speed (s/Image)
Faster RCNN	91.50	59.35	93.59	73.00	5.167
YOLO v3-spp	84.30	89.70	87.50	86.90	0.072
YOLOv5s	94.76	93.81	89.18	91.44	0.054
YOLOv7	95.74	94.21	93.13	93.67	0.025

The different models were used to detect the *Camellia oleifera* fruit images on the test set to compare their generalisation ability (Table 2). The YOLOv7 model detected more objects than YOLOv5s, YOLO v3-spp and Faster RCNN models. At the same time, the missed and wrong objects of the YOLOv7 model were both less than YOLOv5s, YOLO v3-spp and Faster RCNN models. Figure 9 shows the detection results of different models in a variety of complex scenes in test set, in which the blue circle is wrong objects and the green circle is missed objects. From the representation images of the four scenes in Figure 8, YOLOv7 can successfully detect all *Camellia oleifera* fruit in sidelight or slightly

occlusion scenes. The YOLOv7 model produced less wrong objects and missed detection than the YOLOv5s, YOLO v3-spp and Faster RCNN models in the scene of backlight or heavy occlusion.

Table 2. Detection result of different models.

Objects Number	Number of Actual Objects	Target Detection Networks			
		Faster RCNN	YOLO v3-spp	YOLO v5	YOLO v7
Number of detected objects	1401	1426	1437	1577	1588
Number of right objects	1401	1256	1081	1308	1327
Number of wrong objects	0	170	356	269	261
Number of missed objects	0	145	320	93	74

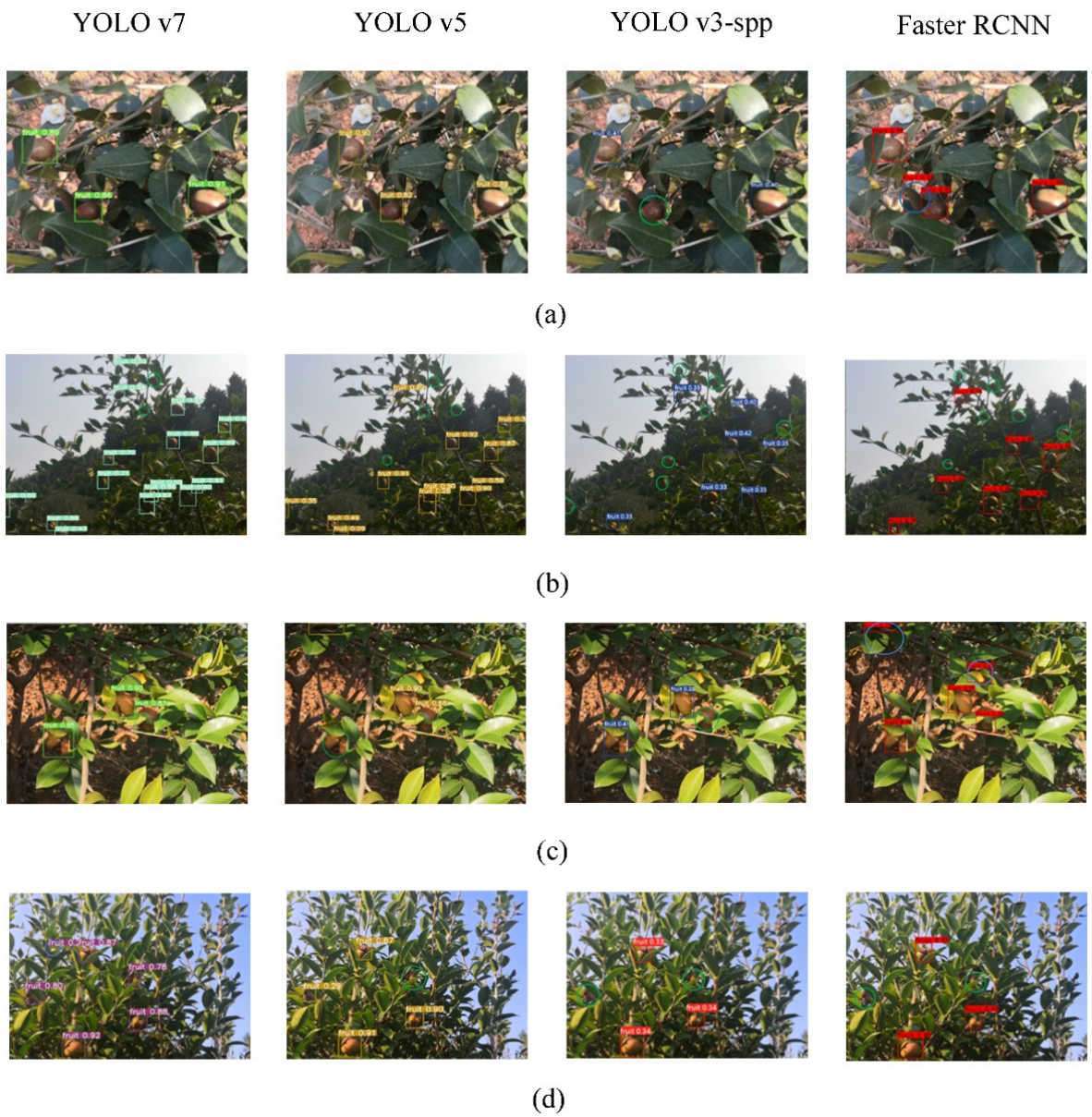

Figure 9. Examples of the detection results of four network models in a variety of complex scenes in test set. (**a**) sidelight angle; (**b**) backlight angle; (**c**) Slightly occlusion; (**d**) Heavy occlusion.

The overall detection result of the YOLOv7 model was better than the current popular *Camellia oleifera* fruit detection models. The YOLOv7 model has high accuracy and real-time detection performance to avoid wrong and missed detection.

3.3. Influence of Data Augmentation

Traditional data augmentation combined with mosaic augmentation were used to augment the dataset and improve the detection ability of the model in complex scenes. The DA-YOLOv7 model was established based on the data augmentation and the YOLOv7 network. The performance indicators of the DA-YOLOv7 model and the YOLOv7 model were compared as shown in the Table 3. It can be seen that the DA-YOLOv7 model had better mAP, Precision, Recall and F1 score than the YOLOv7 model. These indicators were increased by 0.29%, 0.53%, 2.41% and 1.48%, respectively.

Table 3. Performance comparison between DA-YOLOv7 model and YOLOv7 model.

Models	Evaluation Index				Detection Results	Numbers
	mAP (%)	Precision (%)	Recall (%)	F1 Score (%)		
YOLOv7	95.74	94.21	93.13	93.67	Detected objects Right objects Wrong objects Missed objects	1588 1327 261 74
DA-YOLOv7	96.03	94.76	95.54	95.15	Detected objects Right objects Wrong objects Missed objects	1560 1340 220 61

To further compare the generalisation, DA-YOLOv7 and the YOLOv7 models were used to detect the *Camellia oleifera* fruit images in complex scenes on the test set. As shown in Figure 10, those in the blue circle were wrong detection, and those in the green circle were missed objects. Both models detected the sidelight objects. For the scenes with backlight, slightly and heavy occlusion, the YOLOv7 model was difficult to detect the right objects while the DA-YOLOv7 model can well detect them.

In summary, the use of the traditional data augmentation combined with mosaic augmentation can enable the model to learn more features of *Camellia oleifera* fruit in complex field scenes, thereby improving the learning ability of the model and enhancing the generalisation ability of the model. The optimal *Camellia oleifera* fruit detection model is the DA-YOLO v7 model.

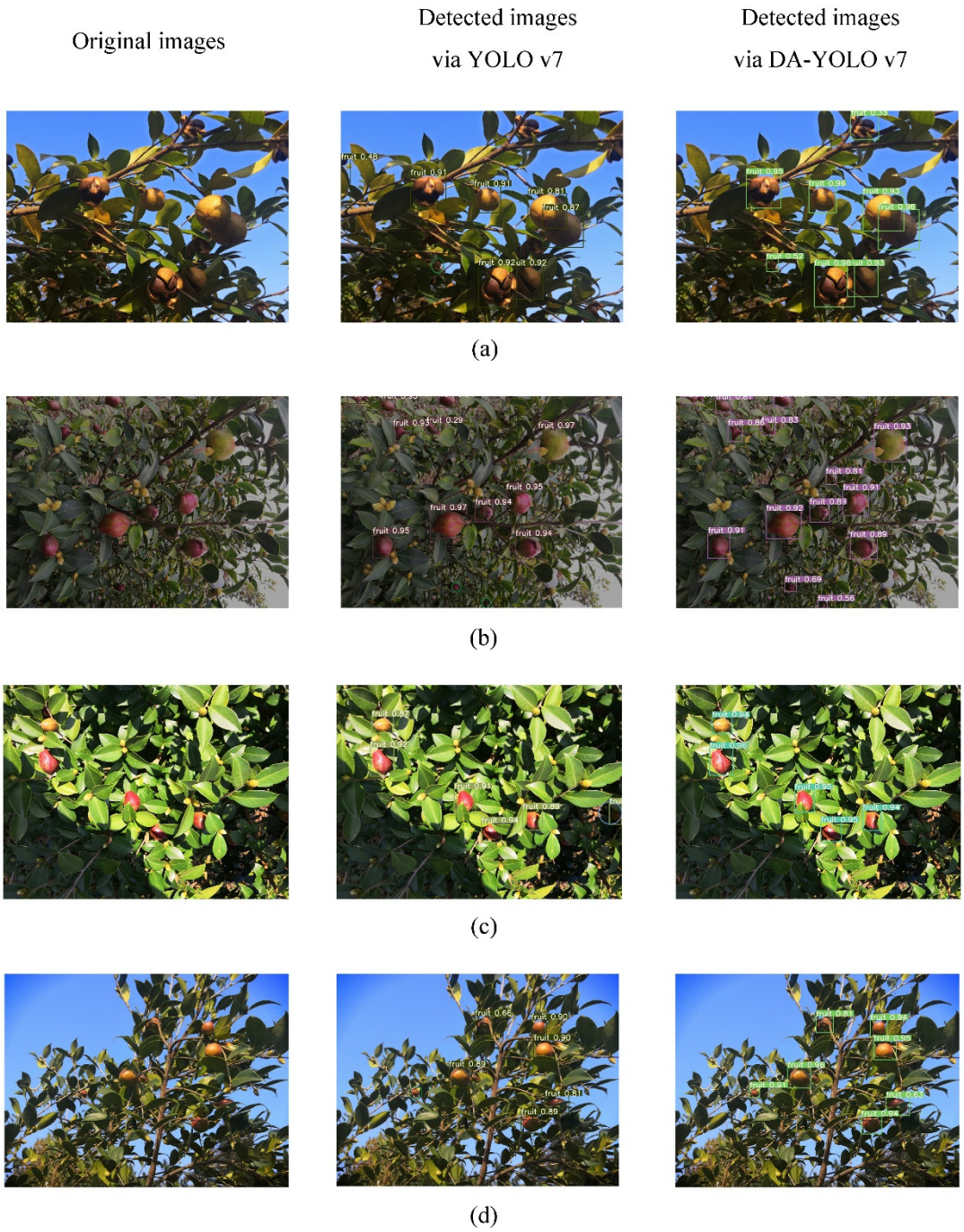

Figure 10. Examples of the detection results between YOLOv7and the DA-YOLOv7models. (**a**) side-light angle; (**b**) backlight angle; (**c**) Slight occlusion; (**d**) Heavy occlusion.

4. Conclusions

A real-time and accurate detection method based on YOLOv7 target detection network and multiple data augmentation was proposed to realize the detection of *Camellia oleifera* fruit in complex scenes of orchard. Firstly, the images of *Camellia oleifera* fruits were collected, and the detection model of *Camellia oleifera* fruits was established by YOLOv7 network, which was compared with YOLOv5s, YOLOv3-spp and Faster R-CNN target detection networks. The results showed that the YOLOv7 model has the best performance with mAP of 95.74%, F1 score of 93.67%, Precision of 94.21%, Recall of 93.13% and the average detection time of 0.025 s. The dataset was further augmented by rotation, mirroring, adding Gaussian noise, increasing or decreasing image brightness and mosaic

augmentation methods, and the DA-YOLOv7 detection model was established by using the augmented dataset and the YOLOv7 network. Data augmentation can effectively improve the detection ability of the model. The optimal *Camellia oleifera* fruit detection model was DA-YOLOv7 model with mAP of 96.03%, Precision of 94.76%, Recall of 95.54% and F1 score of 95.15%. In summary, the YOLOv7 target detection network combined with multiple data augmentation can accurately and quickly detect *Camellia oleifera* fruit in complex scenes. This method has a good application prospect in mechanical harvesting operation. In the future work, we plan to combine the proposed model with the end-effector to realize detection and positioning of fruit, and further adjust the picking angle and the position of the end-effector. At the same time, this study provides a theoretical reference for detection and automatic harvesting of other fruits.

Author Contributions: Conceptualization, D.W. and S.J.; methodology, S.J.; software, S.J. and E.Z.; validation, D.W., E.Z. and Y.L.; formal analysis, Y.L.; investigation, H.Z. and D.W.; resources, W.W.; data curation, D.W., Y.L. and W.W.; writing—original draft preparation, S.J. and Y.L.; writing—review and editing, R.W. and S.J.; visualization, D.W. and S.J.; supervision, D.W. and E.Z.; project administration, S.J., D.W. and E.Z.; funding acquisition, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Anhui Province (NO.2208085ME132) and the National Key Research and Development Program of China (NO.2016YFD0702105).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the editors and all the reviewers who participated in the review.

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

References

1. Wu, D.L.; Fu, L.Q.; Cao, C.M.; Li, C.; Xu, Y.P.; Ding, D. Design and Experiment of Shaking-branch Fruit Picking Machine for *Camellia* Fruit. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 176–182.
2. Wu, D.; Ding, D.; Cui, B.; Jiang, S.; Zhao, E.; Liu, Y.; Cao, C. Design and experiment of vibration plate type camellia fruit picking machine. *Int. J. Agric. Biol. Eng.* **2022**, *15*, 130–138. [CrossRef]
3. Wu, D.L.; Zhao, E.L.; Fang, D.; Jiang, S.; Wu, C.; Wang, W.W.; Wang, R.Y. Determination of Vibration Picking Parameters of *Camellia oleifera* Fruit Based on Acceleration and Strain Response of Branches. *Agriculture* **2022**, *12*, 1222. [CrossRef]
4. Wu, D.L.; Zhao, E.L.; Jiang, S.; Wang, W.W.; Yuan, J.H.; Wang, K. Optimization and Experiment of Canopy Vibration Parameters of *Camellia oleifera* Based on Energy Transfer Characteristics. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 23–33.
5. Liu, J.Z.; Yuan, Y.; Zhou, Y.; Zhu, X.X.; Syed, T.N. Experiments and Analysis of Close-Shot Identification of On-Branch Citrus Fruit with RealSense. *Sensors* **2018**, *18*, 1510. [CrossRef]
6. Zhang, W.; Ma, H.; Li, X.; Liu, X.; Jiao, J.; Zhang, P.; Gu, L.; Wang, Q.; Bao, W.; Cao, S. Imperfect Wheat Grain Recognition Combined with an Attention Mechanism and Residual Network. *Appl. Sci.* **2021**, *11*, 5139.
7. Gill, H.S.; Murugesan, G.; Khehra, B.S.; Sajja, G.S.; Gupta, G.; Bhatt, A. Fruit recognition from images using deep learning applications. *Multimed. Tools Appl.* **2022**, *81*, 33269–33290. [CrossRef]
8. Wei, X.Q.; Jia, K.; Lan, J.H.; Li, Y.W.; Zeng, Y.L.; Wang, C.M. Automatic method of fruit object extraction under complex agricultural background for vision system of fruit picking robot. *Optik* **2014**, *125*, 5684–5689. [CrossRef]
9. Gongal, A.; Amatya, S.; Karkee, M.; Zhang, Q.; Lewis, K. Sensors and systems for fruit detection and localization: A review. *Comput Electron. Agric.* **2015**, *116*, 8–19. [CrossRef]
10. Koirala, A.; Walsh, K.B.; Wang, Z.L.; McCarthy, C. Deep learning—Method overview and review of use for fruit detection and yield estimation. *Comput Electron. Agric.* **2019**, *162*, 219–234.
11. Pu, J.Y.; Yu, C.J.; Chen, X.Y.; Zhang, Y.; Yang, X.; Li, J. Research on Chengdu Ma Goat Recognition Based on Computer Vison. *Animals* **2022**, *12*, 1746. [CrossRef] [PubMed]
12. Yu, Y.; Zhang, K.L.; Yang, L.; Zhang, D.X. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Comput Electron. Agric.* **2019**, *163*, 104846. [CrossRef]
13. Wu, L.; Ma, J.; Zhao, Y.H.; Liu, H. Apple Detection in Complex Scene Using the Improved YOLOv4 Model. *Agronomy* **2021**, *11*, 476.

14. Tian, Y.N.; Yang, G.D.; Wang, Z.; Li, E.; Liang, Z.Z. Instance segmentation of apple flowers using the improved mask R-CNN model. *Biosyst. Eng.* **2020**, *193*, 264–278. [CrossRef]
15. Kuznetsova, A.; Maleva, T.; Soloviev, V. Using YOLOv3 Algorithm with Pre- and Post-Processing for Apple Detection in Fruit-Harvesting Robot. *Agronomy* **2020**, *10*, 1016.
16. Han, W.; Jiang, F.; Zhu, Z.Y. Detection of Cherry Quality Using YOLOV5 Model Based on Flood Filling Algorithm. *Foods* **2022**, *11*, 1127. [CrossRef] [PubMed]
17. Li, S.L.; Zhang, S.J.; Xue, J.X.; Sun, H.X.; Ren, R. A Fast Neural Network Based on Attention Mechanisms for Detecting Field Flat Jujube. *Agriculture* **2022**, *12*, 717. [CrossRef]
18. Halstead, M.; McCool, C.; Denman, S.; Perez, T.; Fookes, C. Fruit Quantity and Ripeness Estimation Using a Robotic Vision System. *IEEE Robot Autom. Let.* **2018**, *3*, 2995–3002. [CrossRef]
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Li, G.; Suo, R.; Zhao, G.A.; Gao, C.Q.; Fu, L.S.; Shi, F.X.; Dhupia, J.; Li, R.; Cui, Y.J. Real-time detection of kiwifruit flower and bud simultaneously in orchard using YOLOv4 for robotic pollination. *Comput. Electron. Agric.* **2022**, *193*, 106641.
21. Yan, B.; Fan, P.; Lei, X.Y.; Liu, Z.J.; Yang, F.Z. A Real-Time Apple Targets Detection Method for Picking Robot Based on Improved YOLOv5. *Remote Sens.* **2021**, *13*, 1619. [CrossRef]
22. Lu, S.Y.; Wang, B.Z.; Wang, H.J.; Chen, L.H.; Ma, L.J.; Zhang, X.Y. A real-time object detection algorithm for video. *Comput Electr. Eng.* **2019**, *77*, 398–408. [CrossRef]
23. Wang, C.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
24. Tian, Y.N.; Yang, G.D.; Wang, Z.; Wang, H.; Li, E.; Liang, Z.Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput Electron. Agric.* **2019**, *157*, 417–426. [CrossRef]
25. Zulkifley, M.A.; Moubark, A.M.; Saputro, A.H.; Abdani, S.R. Automated Apple Recognition System Using Semantic Segmentation Networks with Group and Shuffle Operators. *Agriculture* **2022**, *12*, 756. [CrossRef]
26. Chen, L.Y.; Zheng, M.C.; Duan, S.Q.; Luo, W.L.; Yao, L.G. Underwater Target Recognition Based on Improved YOLOv4 Neural Network. *Electronics* **2021**, *10*, 1634. [CrossRef]
27. Li, R.; Wu, Y.P. Improved YOLOv5 Wheat Ear Detection Algorithm Based on Attention Mechanism. *Electronics* **2022**, *11*, 1673. [CrossRef]
28. Gu, Y.; Wang, S.C.; Yan, Y.; Tang, S.J.; Zhao, S.D. Identification and Analysis of Emergency Behavior of Cage-Reared Laying Ducks Based on YoloV5. *Agriculture* **2022**, *12*, 485. [CrossRef]
29. Jintasuttisak, T.; Edirisinghe, E.; Elbattay, A. Deep neural network based date palm tree detection in drone imagery. *Comput Electron. Agric.* **2022**, *192*, 106560. [CrossRef]
30. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
31. Ahmad, I.; Yang, Y.; Yue, Y.; Ye, C.; Hassan, M.; Cheng, X.; Wu, Y.; Zhang, Y. Deep Learning Based Detector YOLOv5 for Identifying Insect Pests. *Appl. Sci.* **2022**, *12*, 10167.
32. Ding, X.H.; Zhang, X.Y.; Ma, N.N.; Han, J.G.; Ding, G.G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. *arXiv* **2021**, arXiv:2101.03697.
33. Ding, X.H.; Hao, T.X.; Tan, J.C.; Liu, J.; Han, J.G.; Guo, Y.C.; Ding, G.G. ResRep: Lossless CNN Pruning via Decoupling Remembering. *arXiv* **2021**, arXiv:2007.03260.
34. Zheng, Z.H.; Wang, P.; Liu, W.; Li, J.Z.; Ye, R.G.; Ren, D.W. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *arXiv* **2019**, arXiv:1911.08287.
35. Ren, S.Q.; He, K.M.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE T Pattern Anal.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]