electronics

MDPI

Article

A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5

Jia Yao [1,2], Jiaming Qi [1,2], Jie Zhang [1,2,*], Hongmin Shao [1,2], Jia Yang [3] and Xin Li [1,2]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an 625000, China; 201902257@stu.sicau.edu.cn (J.Y.); qijiaming@stu.sicau.edu.cn (J.Q.); shaohongmin@stu.sicau.edu.cn (H.S.); 201902150@stu.sicau.edu.cn (X.L.)
[2] Sichuan Key Laboratory of Agricultural Information Engineering, Ya'an 625000, China
[3] School of Computing, National University of Singapore, Singapore 119077, Singapore; yangjia9809@gmail.com
* Correspondence: 12340@sicau.edu.cn; Tel.: +86-135-1834-0890

Abstract: Defect detection is the most important step in the postpartum reprocessing of kiwifruit. However, there are some small defects difficult to detect. The accuracy and speed of existing detection algorithms are difficult to meet the requirements of real-time detection. For solving these problems, we developed a defect detection model based on YOLOv5, which is able to detect defects accurately and at a fast speed. The main contributions of this research are as follows: (1) a small object detection layer is added to improve the model's ability to detect small defects; (2) we pay attention to the importance of different channels by embedding SELayer; (3) the loss function CIoU is introduced to make the regression more accurate; (4) under the prerequisite of no increase in training cost, we train our model based on transfer learning and use the CosineAnnealing algorithm to improve the effect. The results of the experiment show that the overall performance of the improved network YOLOv5-Ours is better than the original and mainstream detection algorithms. The mAP@0.5 of YOLOv5-Ours has reached 94.7%, which was an improvement of nearly 9%, compared to the original algorithm. Our model only takes 0.1 s to detect a single image, which proves the effectiveness of the model. Therefore, YOLOv5-Ours can well meet the requirements of real-time detection and provides a robust strategy for the kiwi flaw detection system.

Keywords: deep learning; real-time detection; fruit defect detection; YOLOv5; loss function of bounding-box regression

Citation: Yao, J.; Qi, J.; Zhang, J.; Shao, H.; Yang, J.; Li, X. A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5. *Electronics* **2021**, *10*, 1711. https://doi.org/10.3390/electronics10141711

Academic Editor: Byung Cheol Song

Received: 15 June 2021
Accepted: 15 July 2021
Published: 17 July 2021

1. Introduction

China is a giant producer of kiwi, whose output ranks first in the world [1]. Defect detection plays a significant role in the postpartum reprocessing of kiwifruit. Through defect detection, we can grade and price different kiwifruit based on their quality, which helps to change the phenomenon that the price of kiwifruit was difficult to increase in the past [2]. It also guarantees food safety. However, detection technology is very traditional and outdated. Most manufacturers and workers mainly rely on manual detecting, which wastes too much labor and has poor efficiency [3].

In recent years, computer-vision-based object detection technology has gradually become matured [4,5]. Shah et al. use Faster RCNN to identify plants and weeds [6]. Zeze et al. use CNN to realize the recognition of apples [7]. Computer vision has the obvious advantages of high accuracy and fast speed [8]. Defect detection based on computer vision is an automatic and nondestructive fruit detection method [9]. It overwhelms manual detection on precision and efficiency; hence, it will bring the inevitable trend of application in fruits in the future [10].

In current fruit defect detection algorithms, it is difficult to balance speed and accuracy simultaneously. Dong et al. [11] used computer vision technology to detect the surface

defects of Korla fragrant pears. Under the condition of guaranteeing accuracy, it still takes 2.5 s to detect a single image. Wang et al. [12] conducted rapid detection of pomegranate leaf diseases, but the accuracy was only 87%. Xing et al. [13] used the BP neural network in mango quality inspection to increase the speed as much as possible while ensuring accuracy. The final speed also took 0.8 s.

The development of deep learning algorithms in recent years has led to major breakthroughs in the field of computer vision. In terms of target recognition, deep learning algorithms represented by convolutional neural networks (CNNs) have improved the accuracy and detection speed, compared with traditional methods [14]. At present, target recognition algorithms are mainly divided into two types: one is a two-stage algorithm based on the detection frame and classifier, such as the R-CNN [15] series algorithm, which is of higher accuracy, but its deeper network structure also leads to a slower speed, failing to meet real-time the requirements of the target recognition detection. The other is a regression-based first-order algorithm, such as SDD [16], YOLO [17] series algorithms, etc., with faster inference speed and stronger practicability, which can meet real-time object recognition and detection.

This paper takes kiwifruit defect as the research object, collects four types of common flaw photos to make a kiwi flaw dataset, and uses the characteristics of high detection speed and high accuracy of the YOLOv5 [18] algorithm in the field of image detection. We ameliorated the problem and compared the improved model with the original one. The use of the CosineAnnealing [19] decay method in the training process can improve the model effect without increasing the cost of training. The result proves that the improved model leads to significant progress, which proves the effectiveness of the improved model.

2. Related Work

2.1. YOLO Algorithm

The main current object recognition algorithms include the R-CNN series and the YOLO series. The R-CNN series is superior in target detection requiring higher accuracy, but its detection speed is lower than that of the YOLO series. In practical scenarios, it cannot meet the real-time performance of object detection. In this context, the YOLO series of algorithms use the idea of regression to make it easier to learn the generalized characteristics of the target and solve the speed problem. The YOLO series of algorithms use a one-stage neural network to complete detection object positioning and classification directly [20,21].

YOLO views image detection as a regression problem with a simple pipeline and fast speed. It can process streaming video in real-time with a delay of fewer than 25 s. During the training process, YOLO can look over the entire image with more attention on global information in target detection. The core idea of YOLO is to use the entire picture as the input of the network, and directly return to the position of the bounding box and the category to which the bounding box belongs at the output. In YOLO, each bounding box is predicted by the characteristics of the entire image, and each bounding box contains five predictions and confidences, which are relative to the grid unit in the center of the bounding box of the boundary. The basic frame of YOLO is as follows: w and h are the predicted width and height of the entire image (relative to the entire image). The YOLO is mainly composed of three main components:

- Backbone: A convolutional neural network that aggregates and forms image features on different types of image granularity;
- Neck: A series of network layers that mix and combine image features and pass the image features to the prediction layer;
- Head: It can predict image features, generate bounding boxes, and predict categories. The confidence indicates the accuracy of classification under the specific condition.

YOLOv2 [22] uses a new training algorithm. YOLOv2 uses the k-means clustering method to cluster the bounding boxes in the training set. As the main purpose of setting, the a priori box is to make the IOU between the prediction box and the ground truth better,

the IOU value between the box and the cluster center box is used as the distance indicator in the cluster analysis. Compared with YOLOv1, it significantly improves the accuracy and the recall rate. YOLOv3 [23] uses a better basic classification network-class ResNet [24] and classifier Darknet-53. At the same time, the FPN [25]-like network structure is used to realize multiscale prediction. The detection accuracy and speed are greatly improved, and the false background detection rate is effectively reduced. YOLOv4 [26] retains the head part of YOLOv3, changes the backbone network to CSPDarknet53, and uses the idea of SPP [27] (spatial pyramid pooling) to expand the receptive field, with PANet [28] as the neck part. The structure of CSPNet [29] can achieve richer gradient combination information and reduce the amount of calculation. The PANet structure fully integrates the different feature layers, which can effectively improve the feature extraction ability of defects.

YOLOv5 continues to use the three main components of the YOLO series. The network structure is shown in Figure 1.

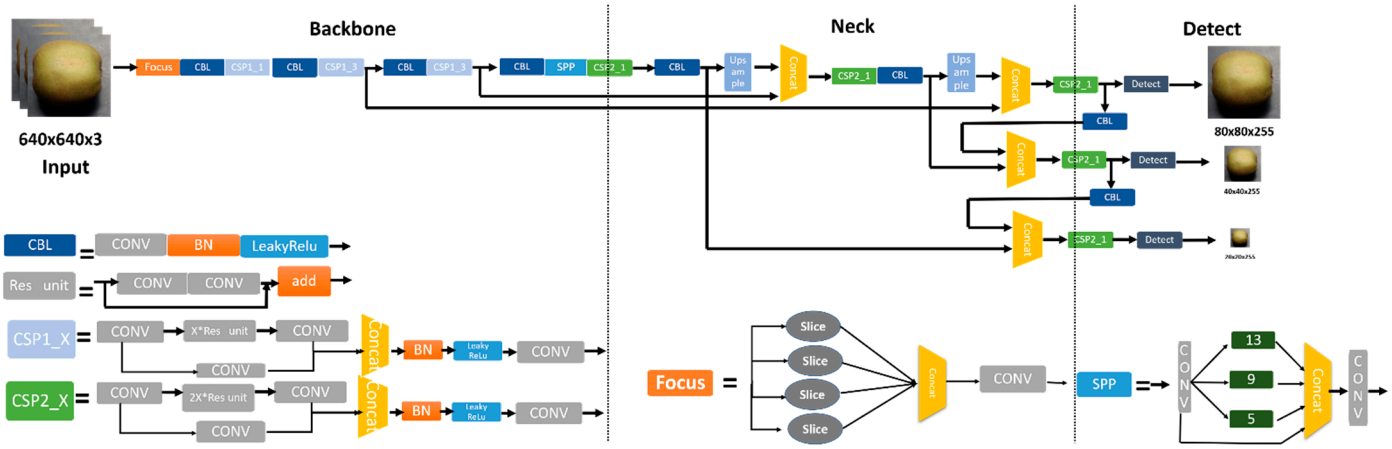


Figure 1. YOLOv5 network structure diagram.

2.1.1. Input

The input end of YOLOv5 uses the same mosaic data enhancement method as YOLOv4, which performs better in small target detection. YOLOv5 adds the function of adaptive anchor frame calculation. During each training, the value of the optimal anchor frame in different training sets is calculated adaptively.

2.1.2. Backbone

YOLOv5 adds the Focus structure to realize the slicing operation. Taking the structure of Yolov5 s as an example, the original 640 × 640 × 3 image is input into the Focus structure, and the slicing operation is used first to form a 320 × 320 × 12 feature map, and then after a convolution operation of 32 convolution kernels, it finally constructs a feature map of 320 × 320 × 32.

2.1.3. Neck

Yolov5 uses the FPN-PAN structure, CSP2 structure designed by CSPNet, and PANET as Neck to aggregate features. The neck is mainly used to generate feature pyramids, enhance the model's detection of objects of different scales, and realize the recognition of the same object of different sizes and scales. The feature extractor of the network uses a new FPN structure, which enhances the bottom-up path and improves the propagation of low-level features.

2.2. Algorithm Optimization

2.2.1. Small Target Recognition Layer

As the kiwifruit has small-size flaws and few pixel features, the inspection model is required to have a strong inevitable ability for small defects. In the original YOLOv5 model, the feature map of the last layer of the convolutional network structure is too small to meet the requirements of the subsequent detection and regression. To solve this problem, we add a small target detection layer and continue to process the feature map for expansion. The main purpose of upsampling is to enlarge the original image so that it can be displayed on a higher resolution display device. The zoom operation of the image cannot bring more information about the image; hence, the quality of the image will inevitably be affected. However, there are indeed some zooming methods that can increase the information of the image such that the quality of the zoomed thick image exceeds the quality of the original image. Upsampling adopts the interpolation method, that is, on the basis of the original image pixels, a suitable interpolation algorithm is used to insert new elements between pixels, as shown in the following Figure 2. At the same time, the acquired feature map and the feature map of the second layer in the backbone network are Concat Fusion in order to obtain a larger feature map for small target detection.

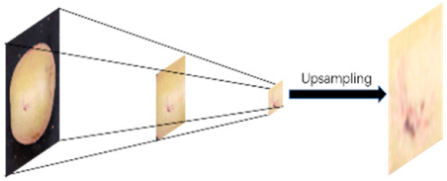


Figure 2. Upsampling process.

2.2.2. SELayer

In order to obtain more detailed information about the target that needs attention and suppress other useless information from different channels, we introduce the Attention network, SElayer [30]. SENet is a network structure proposed by Jie et al., which mainly focuses on the feature fusion among channels of the convolution operation in the backbone network. The main innovation of this network is that the model can automatically learn the importance of different channel features by focusing on the relationship between channels. The SE module mainly includes operations through compression (Squeeze) and excitation (Excitation). The Squeeze operation takes global average pooling to encode the entire spatial feature on a channel as a local feature. The calculation method is as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

In this formula, the second two-dimensional matrix in the three-dimensional matrix after convolution represents the result of the Squeeze operation, and the subscript represents the number of channels.

After the Squeeze operation obtains the channel information, it uses two fully connected layers to form a gate mechanism and activates it with Sigmod. The calculation method is as follows:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

where δ is the ReLu activation function, σ is the Sigmoid function, and W_1 and W_2 is the weight of the two fully connected layers used for dimensionality reduction and dimension upgrade, which, respectively, equals to $\frac{C}{r} \times C$ and $C \times \frac{C}{r}$, r is the scaling parameters to limit model complexity and improve model capabilities. s represents the weight set of the feature maps obtained through the fully connected layer and the nonlinear layer. Finally,

the weight of the output is assigned to the original feature. The calculation formula is as follows:

$$\widetilde{x}_c = s_c \times u_c \tag{3}$$

In the formula, $\widetilde{x}_c$ is a feature map of a featured channel of $\widetilde{X}$, s_c is a weight, and u_c is a two-dimensional matrix. After modification, the network structure is shown in Figure 3:

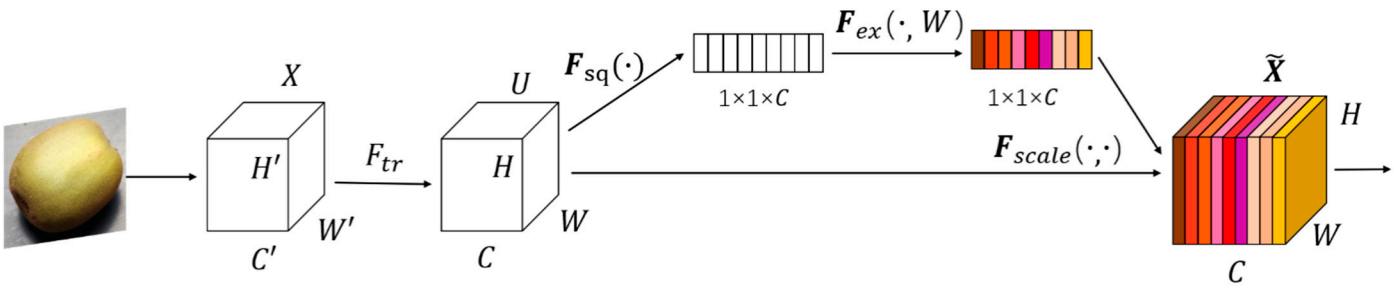


Figure 3. The structure of SELayer.

This article considers embedding SELayer in the backbone. The improved network structure is shown below in Figure 4.

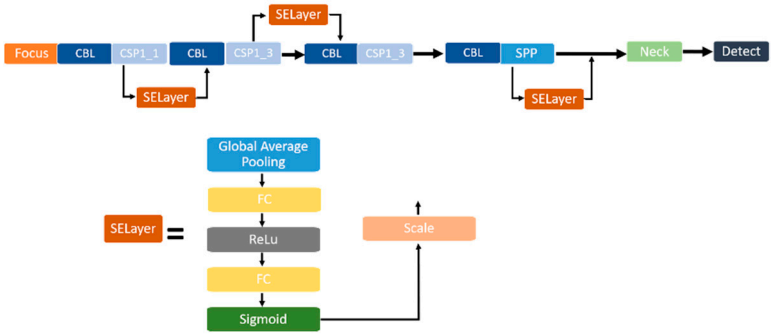


Figure 4. The structure of the improved network.

2.2.3. Boundary Loss Function

IoU [31] is the intersection over union, a common indicator in target detection, whose main function is to determine the positive sample and the negative sample and to evaluate the distance between the output box and the correct label. IoU is scale invariant, which means that it is not sensitive to scale. Therefore, in the regression task, IoU is the most direct indicator for judging output madness and correct labeling. However, there is a problem with the definition of IoU itself. IoU is 0 if the two boxes do not intersect. At the same time, due to the 0 loss, there is no gradient back; hence, learning and training cannot be performed. To solve these problems, Rezatofighi et al. proposed the idea of GIoU [32] and directly set IoU as the return loss. Since IoU is a ratio concept, it is not sensitive to the scale of the target object. However, the BBox regression loss (MSE loss, l1-smooth loss, etc.) optimization and IoU optimization in the detection task is not completely equivalent, the Ln norm is also sensitive to the scale of the object, and IoU cannot optimize the part that does not overlap directly. The principle of GIoU is as follows:

$$L_{GIoU} = 1 - \mathrm{IoU} + \frac{\left|C - B \cup B^{gt}\right|}{|C|} \tag{4}$$

However, there are still some problems with the GIoU such as the unstable target frame regression and the easy divergence during training. Some frames of the target detection without overlapping GIoU regression strategies may degenerate into IoU regression strategies. In order to directly minimize the normalized distance between the anchor box

and the target box to achieve a faster convergence rate and make the regression more accurate and faster when it overlaps or even contains the target box, Zheng et al. put forward the idea of DIoU and CIoU [33]. The principle is as follows:

$$L_{DIoU} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} \tag{5}$$

where b and b^{gt} represent the center points of the prediction box and the real box, respectively, ρ represents the Euclidean distance between the two center points, and c represents the diagonal distance of the smallest closed area that can contain the prediction box and the real box at the same time.

Comparatively, DIoU is more in line with the target frame regression mechanism than GIou. For the situation that contains two frames in the horizontal and vertical directions, the DIoU loss can make the regression very fast, while the GIoU loss almost degenerates into the IoU loss. DIoU can also replace the common IoU evaluation strategy and apply it to NMS, making the results obtained by NMS more reasonable and effective.

The DIoU calculation does not take the aspect ratio into consideration but only considers the overlapping area of the bounding box and the center point distance of b and b^{gt}. However, the consistency of the ratio of w and h between the anchor box, and the target box is also of high significance. Based on this, the author proposes complete loU loss.

The penalty term of CIoU is based on the penalty term of DIoU plus an impact factor α, ν, which takes into account the aspect ratio of the predicted frame to fit the target frame. The specific principle is as follows:

$$L_{CIoU} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\nu \tag{6}$$

As shown in Figure 5, the upper left block represents the target frame, the lower right block represents the prediction frame, and the dashed block represents the smallest bounding rectangle, and c and d, respectively, represent the diagonal distance of the smallest enclosing rectangle and the Euclidean distance between the center points of the two boxes.

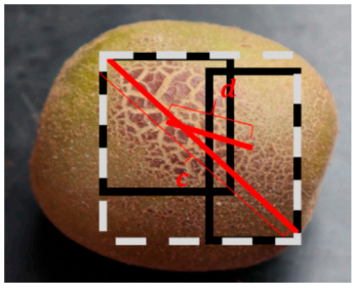


Figure 5. The normalized distance between the prediction frame and target frame.

The expressions of the weight function and the parameters for measuring the consistency of the aspect ratio are shown in Equations (7) and (8).

$$a = \frac{v}{1 - \mathrm{IoU} + v} \tag{7}$$

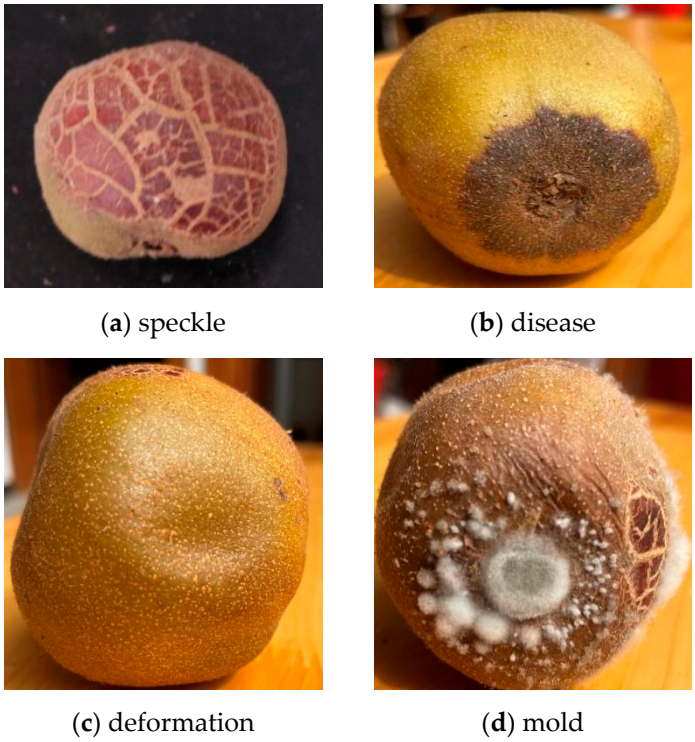
$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w^{p}}{h^{p}}\right)^2 \tag{8}$$

Among them, w^{gt} and h^{gt} represent the width and height of the target frame, and w^p and h^p represent the width and height of the prediction frame, respectively.

2.3. Experimental Setup

2.3.1. Dataset Production and Preprocessing

From September 2019 to December 2019, three different types of kiwifruit were randomly collected at Ya'an Hongming Farm. Different types of kiwifruit vary in size and shape. To improve the effectiveness of training and increase the diversity of samples, the collected image data were screened before training. The image preprocessing software was Labelimg, which is a software used to annotate image labels. Finally, 1600 images were obtained and stored in JPG format with a resolution of 6000 px × 4000 px. In the next step, 1000 pictures were randomly selected as the training set. The dataset was enhanced by adaptive contrast, rotation, translation, cropping, and other methods, and the dataset was expanded to 2000. The dataset was divided into 4 categories, which are disease, mold, speckle, and deformation. Then, 300 pictures were randomly selected as the verification set, and 2200 pictures were annotated as kiwifruit. There were 300 unlabeled kiwifruit images left as the test set. The dataset is shown in Figure 6.



(**a**) speckle (**b**) disease

(**c**) deformation (**d**) mold

Figure 6. Kiwi defect sample image.

2.3.2. Migration Network Initialization

Transfer learning is a common machine learning method, whose key is to transfer the knowledge that has been trained in a certain field to another new field. As for this paper, it concerns the completion of the model pretraining. The results will be migrated to the YOLO v5 network of kiwi flaw detection to help the training of the detection model. To initialize the model parameters of a small training set, a pretrained network model is selected with a good learning ability to complete. Since the kiwi flawed image samples in this paper are limited and few, migration learning is also chosen to initialize the parameters of the YOLO v5 network, which can ensure the successful migration of the learned knowledge and the capability to make the new network capable to learn quickly. In this way, the overfitting caused by insufficient kiwi samples can be improved to a certain degree. At the same time, the generalization ability of kiwi flaw detection can also be improved correspondingly so that the recognition model can be facilitated. Even under complex natural conditions, the model has a good recognition ability to perform migration learning. We need to understand the datasets, because there are many datasets in the field of image deep learning, and they

have their characteristics. This paper selects one of the most common and widely used datasets—ImageNet. This dataset shows outstanding performance in image classification, detection, positioning, and other fields.

2.3.3. CosineAnnealing

The CosineAnnealing is different from the traditional method. The learning rate will decrease rapidly with the increase of epoch, and the model will find the local optimal point and save the current model. After that, the learning rate will abruptly increase to a larger value, escape from the current local optimal point, find a new local optimal point, and then repeat this process to adjust the learning rate according to the cycle until the training is completed. As shown in Equation (6), l_{min} represents the minimum learning rate, l_{init} represents the initial learning rate, T_{max} represents a quarter of the change period of the learning rate, and l_{new} represents the new learning rate obtained.

$$l_{new} = l_{\min} + (l_{init} - l_{\min}) \times (1 + \cos(\frac{epoch}{T_{\max}}\pi)) \tag{9}$$

In this training, T_{max} is set to 5, l_{min} is 0.00001, and the learning variability curve of the first 100 epochs is shown in Figure 7.

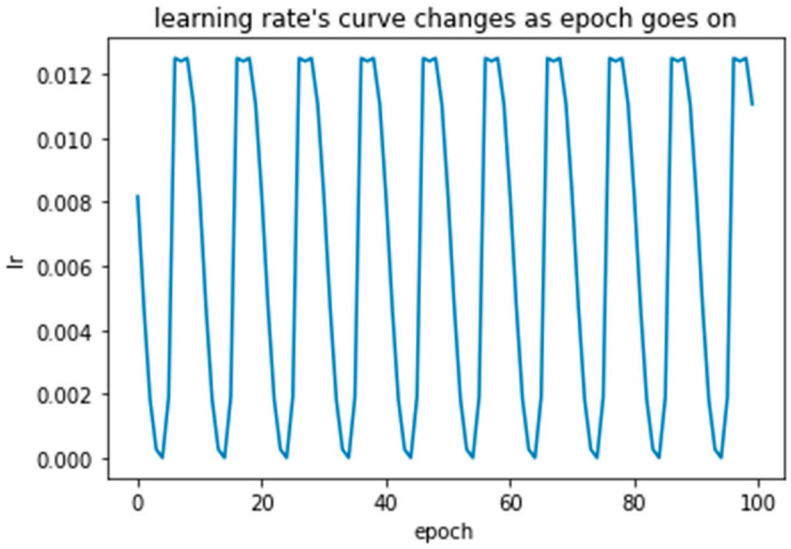


Figure 7. The learning rate using the cosine annealing decay method varies with the number of epochs.

2.3.4. Experimental platform

The training of the model was completed based on the Windows 10 operating system and the Pytorch framework. The CPU model of the test equipment is Intel®Core™ i9_10900K CPU@3.70 GHz, the GPU model is GeForce RTX 5000 16 G, and the software environment is CUDA 10.1, CUDNN 7.6, Python3.7.

The original YOLOv5 and the improved YOLOv5 were trained separately. The parameters were set as follows: the maximum number of iterations was 1000, the momentum was 0.95, the CosineAnnealing of base learning rate was 0.01.

2.3.5. Model Evaluation Indicators

This paper introduces precision (P), which is precision rate, recall rate (R), and mean average precision (mAP) to evaluate the performance of the kiwi flaw detection model. The expressions of P and R are as follows:

$$\mathrm{P} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \tag{10}$$

$$R = \frac{TP}{(TP + FN)} \tag{11}$$

Among them, true positives (TP), false positives (FP), and false negatives (FN), respectively, represent positive samples with correct classification, negative samples with incorrect classification, and positive samples with incorrect classification.

AP is the average accuracy rate, which is the integral of the P index to the R index, that is, the area under the P–R curve; mAP is the average accuracy of the mean, which means that the AP value of each category is summed, and then divided by all categories, i.e., the average value. They are defined as follows:

$$AP = \int_0^1 P(R) dR \tag{12}$$

$$mAP = \frac{1}{|Q_R|} \sum_{q=Q_R} AP(q) \tag{13}$$

where Q_R is the number of categories.

3. Results

3.1. Experimental Results

In order to judge the quality of the detection model accurately, the evaluation in this paper is based on the loss function curve (Loss) and average accuracy value (mAP).

During the network training process, the loss function can intuitively reflect whether the network model can converge stably as the number of iterations increases. The specific loss function of the model is shown in Figure 8 below.

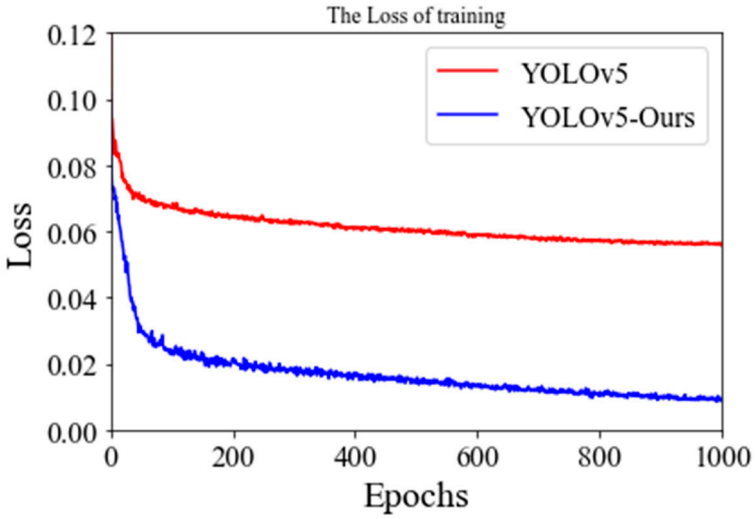


Figure 8. The training of loss.

From the figure, it is found that as the number of iterations gradually increases, the improved YOLOv5 algorithm curve gradually converges, and the loss value becomes smaller and smaller. When the model is iterated 600 times, the loss value is basically stable and has dropped to near 0, and the network basically converges. Compared with the original YOLOv5, the regression is faster and more accurate, which proves the validity and effectiveness of the model.

The mAP is used to measure the quality of the defect detection model. The higher the value is, the higher the average detection accuracy and the better the performance will be.

Figure 9 shows that after about 200 iterations of the YOLOv5-Ours model, the mAP reaches about 94%, and has gradually stabilized, reaching a maximum of 98%, indicating

that the improved YOLOv5 model has an average accuracy rate for defect detection. The overall model performance has met and even exceeded expectations.

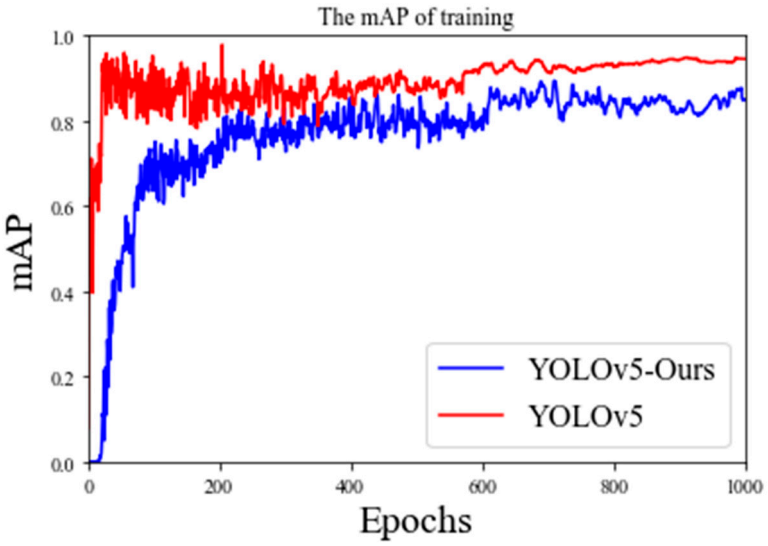
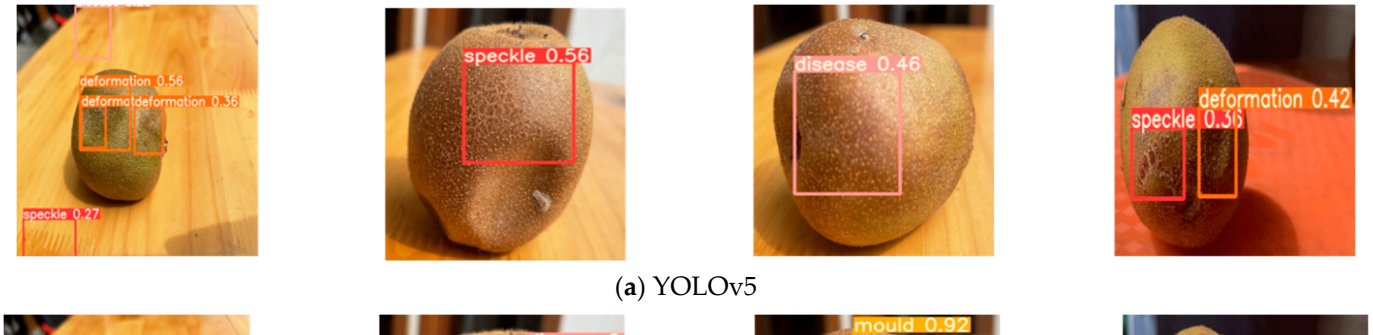


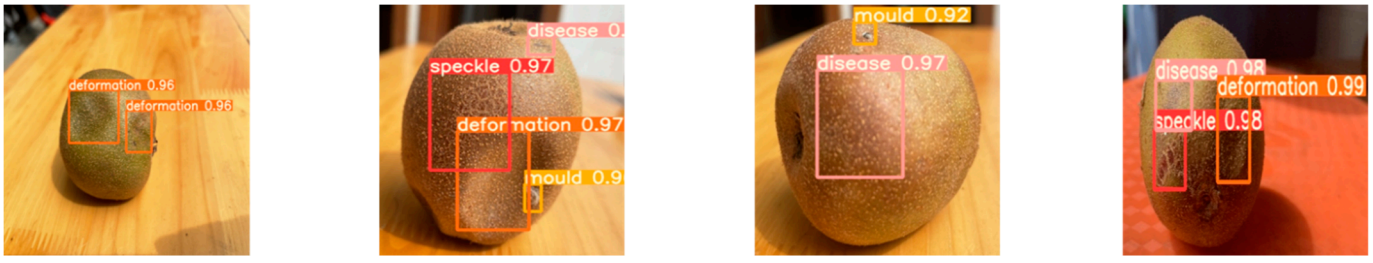
Figure 9. The training of mAP.

3.2. Analysis

The following Figure 10 shows the improved YOLOv5 network and the YOLOv5-Ours network in the kiwifruit dataset part of the detection results, respectively, for different defect categories and defect sizes.



(**a**) YOLOv5



(**b**) YOLOv5

Figure 10. Comparison of detection algorithm before and after improvement.

As the results show, our improved YOLOv5 can accurately detect defects in complex environments, such as tiny defects, and the return positioning frame is more accurate. Embedding SELayer discards unimportant features, significantly improves the robustness of the model, and proves the effectiveness of the network.

Under the condition that the IoU threshold is 50%, the mAP@0.5 of the original YOLOv5 is 85%, and the mAP@0.5 of the improved YOLOv5 is 94.7%.

Table 1 below shows the accuracy comparison between the original model and the improved one.

Table 1. Comparison between the original model and the improved model.

Category	Original YOLOv5 Model	YOLOv5-Ours
speckle	0.95	0.96
disease	0.71	0.93
deformation	0.74	0.92
mould	0.84	0.95
Total mAP	0.85	0.947

According to Table 1, the improved model has improved mAP by nearly 8%. Through testing, it is found that despite the increased complexity of the model, the improved network still only takes 0.1 s to detect a single image, which is in line with real-time detection.

It can be inferred from Table 2 that, compared with mainstream detection algorithms, our network has a higher mAP. Although Fast R-CNN performs well on mAP, it takes 0.79 s to detect a single image, which cannot meet the requirements of real-time detection.

Table 2. Comparison between mainstream detection algorithms.

Category	YOLOv5-Ours	SSD300	YOLOv3	FAST R-CNN
mAP	0.947	0.855	0.821	0.939

4. Discussion

This paper explores an automatic detection method for kiwifruit defects in real time. To meet the needs of farmers to understand the states of kiwifruit at any time and in real time, we use the YOLOv5 model for deeper research. By adding a small target detection layer, the ability to detect small defects is improved. The layer was embedded to enhance useful features and suppress less important features. The CIoU was used as the loss function to make the regression more stable. The feasibility of this method is as follows:

- In terms of processing accuracy, the dataset of this study is manually captured images; hence, the background information is relatively simple. In slightly complex background conditions, the accuracy may be reduced. However, this research is based on unnatural or industrial scenes. Thus, there will be no complex background in practical application.
- In terms of processing speed, in order to meet the real-time needs of farmers, it is necessary to process the images collected by the camera. The initial consideration is using an object detection model to replace models such as, for instance, segmentation or semantic segmentation (the latter two are relatively slow in processing speed). To detect models in multiple objects, the YOLOv5 model for processing is considered, which is a useful model in an advanced single-stage method in the field of object detection. Compared with the two-step method, the former has a higher processing speed based on the same hardware environment. Compared with other one-stage methods (such as YOLOv2), the related reasons have been described in Section 2.1. The optimized YOLOv5 network structure is complex. Compared with the YOLOv5-Ours, the detection speed is reduced, but a single image only takes 0.1 s, which can meet the above requirements.
- In terms of model generalization ability, YOLOv5 uses a mosaic data enhancement strategy to improve the model's generalization ability and robustness.

Based on the above discussion, we believe that the method we proposed is an effective exploration and can promote the development of postproduction reprocessing of crops.

5. Conclusions and Future Work

In this research, Deep learning technology was applied to kiwi flaw detection. Based on YOLOv5, a high-precision kiwi flaw detection method was proposed. First, a kiwifruit dataset containing four types of defects was collected. As far as we know, this is the first kiwifruit defect dataset in the world and even the first agricultural product postproduction defect dataset. At the same time, this is the first time that the YOLOv5 network has been applied to crops. Then, through the improvement of YOLOv5, a small target detection layer was added to the backbone network, and SELayer was embedded to improve the feature extraction ability of the model. In addition, we modified the DIoU loss function to the CIoU loss function to improve the accurate positioning ability of the model prediction frame and enhance the model convergence effect. Compared with the original YOLOv5 model, mAP@0.5 increases 9%. It can detect a single image in only 0.1 s (base on GPU 1050Ti) and has better robustness to the environment, which proves the effectiveness of the model and provides farmers with more efficient and intelligent postproduction reprocessing strategies.

This paper mainly researches and develops kiwifruit defects under the requirement of real-time detection. However, fast detection still needs specific hardware configuration. In the future, we will continue to optimize YOLOv5-Ours and use pruning technology to optimize the model. At the same time, we will continue to increase the research on more kiwifruit varieties and increase the scope of application.

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

Funding: This research was funded by the Sichuan Provincial Federation of Social Sciences, Youth Project, Grant Number SC19C032, and funded by Sichuan Agricultural University Scientific Research Interest Program, Grant Number 2021663.

Acknowledgments: We would like to thank Ya'an Hongming Farm for its help in collecting the dataset, and Jiaoyang Jiang, Yuan Ou, for providing English language support. We also thank Hongming Shao, Jingyu Pu, and Ying Xiang for their advice on the dataset.

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

References

1. Food Industry Network. China's kiwifruit production ranks first in the world. *Food Saf. Guide* **2018**, *33*, 6.
2. Fayuan, W.; Wenkai, W. *Introduction to Frontier Knowledge and Skills of Modern Agricultural Economic Development*; Hubei Science and Technology Press: Wuhan, China, 2010.
3. Li, Q. *Research on Non-Destructive Testing and Automatic Grading of Kiwifruit Based on Computer Vision*; Anhui Agricultural University: Hefei, China, 2020.
4. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
6. Shah, T.M.; Nasika, D.P.B.; Otterpohl, R. Plant and Weed Identifier Robot as an Agroecological Tool Using Artificial Neural Networks for Image Identification. *Agriculture* **2021**, *11*, 222. [CrossRef]
7. Zeze, F.; Qian, L.; Jiewei, C.; Xiaofeng, Y.; Haifang, L. Apple tree fruit detection and grading based on color and fruit diameter characteristics. *Comput. Eng. Sci.* **2020**, *42*, 82–90.
8. Pan, Y.; Wei, J.; Zeng, L. Farmland Bird Target Detection Algorithm Based on YOLOv3. Available online: http://kns.cnki.net/kcms/detail/31.1690.TN.20210409.0942.050.html (accessed on 16 July 2021).
9. Qingzhong, L.; Maohua, W. Development and prospect of real-time fruit grading technology based on computer vision. *Trans. Chin. Soc. Agric. Mach.* **1999**, *6*, 1–7.
10. Xu, T. *Research on Classification and Recognition of Fruit Surface Grade Based on Machine Vision*; Chongqing Jiaotong University: Chongqing, China, 2018.

11. Jianwei, D.; Yuanyuan, L.; Fei, C.; Tongxuan, W.; Shengsheng, D.; Yankun, P. Surface Defect Detection of Korla Fragrant Pear Based on Multispectral Image. *J. Agric. Mech. Res.* **2021**, *43*, 41–46.
12. Yanni, W.; Li, H. Detection method of pomegranate leaf diseases based on multi-class SVM. *Comput. Meas. Control.* **2020**, *28*, 197–201.
13. Huajian, X. Research on the Application of Computer Vision in Mango Quality Detection. *J. Agric. Mech. Res.* **2019**, *1*, 190–193.
14. Du, Z.; Fang, S.; Zhe, L.; Zheng, J. *Tomato Leaf Disease Detection Based on Deep Feature Fusion of Convolutional Neural Network*; China Sciencepaper: Beijing, China, 2020.
15. Liu, X. *Research on Tomato Diseased Leaf Recognition Based on Mask R-CNN and Its Application in Smart Agriculture System*; Xidian University: Xian, China, 2020.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision—ECCV 2016 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *IEEE* **2016**, *1*, 779–788.
18. Shao, H.; Pu, J.; Mu, J. Pig-Posture Recognition Based on Computer Vision: Dataset and Exploration. *Animals* **2021**, *11*, 1295. [CrossRef] [PubMed]
19. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the ICLR 2017 (5th International Conference on Learning Representations), Toulon, France, 24–26 April 2017.
20. Ruan, J. *Design and Implementation of Target Detection Algorithm Based on YOLO*; Beijing University of Posts and Telecommunications: Beijing, China, 2019.
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; 25, pp. 1097–1105.
22. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition IEEE, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
23. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. In Proceedings of the CVPR 2018: IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–22 June 2018.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. Available online: https://arxiv.org/abs/1512.03385 (accessed on 16 July 2021).
25. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
26. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. Available online: https://arxiv.org/abs/2004.10934 (accessed on 16 July 2021).
27. Luvizon, D.; Tabia, H.; Picard, D. SSP-Net: Scalable Sequential Pyramid Networks for Real-Time 3D Human Pose Regression. Available online: https://arxiv.org/abs/2009.01998 (accessed on 16 July 2021).
28. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, Salt Lake City, UT, USA, 18–23 June 2018.
29. Wang, C.Y.; Liao, H.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh I, H. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, Seattle, WA, USA, 14–19 June 2020.
30. Jie, H.; Li, S.; Gang, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 7132–7141.
31. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of Localization Confidence for Accurate Object Detection. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018.
32. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, Long Beach, CA, USA, 15–20 June 2019.
33. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. Available online: https://arxiv.org/abs/2005.03572 (accessed on 16 July 2021).