agronomy

MDPI

Article

RGB-D Heterogeneous Image Feature Fusion for YOLOfuse Apple Detection Model

Liqun Liu * and Pengfei Hao

College of Information Science and Technology, Gansu Agricultural University, Lanzhou 730070, China; haopf@st.gsau.edu.cn

* Correspondence: liulq@gsau.edu.cn; Tel.: +86-139-0948-2426

Abstract: Heterogeneous image features are complementary, and feature fusion of heterogeneous images can increase position effectiveness of occluded apple targets. A YOLOfuse apple detection model based on RGB-D heterogeneous image feature fusion is proposed. Combining the CSPDarknet53-Tiny network on the basis of a YOLOv5s backbone network, a two-branch feature extraction network is formed for the extraction task of RGB-D heterogeneous images. The two-branch backbone network is fused to maximize the retention of useful features and reduce the computational effort. A coordinate attention (CA) module is embedded into the backbone network. The Soft-NMS algorithm is introduced, instead of the general NMS algorithm, to reduce the false suppression phenomenon of the algorithm on dense objects and reduce the missed position rate of obscured apples. It indicates that the YOLOfuse model has an AP value of 94.2% and a detection frame rate of 51.761 FPS. Comparing with the YOLOv5 s, m, l, and x4 versions as well as the YOLOv3, YOLOv4, YOLOv4-Tiny, and Faster RCNN on the test set, the results show that the AP value of the proposed model is 0.8, 2.4, 2.5, 2.3, and 2.2 percentage points higher than that of YOLOv5s, YOLOv3, YOLOv4, YOLOv4-Tiny, and Faster RCNN, respectively. Compared with YOLOv5m, YOLOv5l, and YOLOv5x, the speedups of 9.934FPS, 18.45FPS, and 23.159FPS are obtained in the detection frame rate, respectively, and the model are better in both of parameter's number and model size. The YOLOfuse model can effectively fuse RGB-D heterogeneous source image features to efficiently identify apple objects in a natural orchard environment and provide technical support for the vision system of picking robots.

Keywords: object detection; heterogeneous images; YOLOv5s; fruit detection; attention mechanism

Citation: Liu, L.; Hao, P. RGB-D Heterogeneous Image Feature Fusion for YOLOfuse Apple Detection Model. *Agronomy* **2023**, *13*, 3080. https://doi.org/10.3390/agronomy13123080

Academic Editor: Baohua Zhang

Received: 30 September 2023
Revised: 24 November 2023
Accepted: 24 November 2023
Published: 18 December 2023

1. Introduction

Nowadays, intelligent robot picking of fruits has become possible, and intelligent picking robots have become a current research hotspot [1]. In the study of apple picking robots, due to the randomness of an apple location, the apple images collected by the vision system often have a large number of obscured targets, and the reasons for obscuring can be generally attributed to leaf obscuring, branch obscuring, and fruit overlapping obscuring. The defective information of the obscured apples will make it more difficult for the vision system to identify and locate the apple fruits and directly affect the grasping accuracy. Designing a target detection model that can effectively identify the obscured apples has important research value.

1.1. Related Works

The YOLOv5 network architecture has the advantages of high detection accuracy and fast operation [2]. There are four model architectures, including YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, with increasing model volume in sequence. Zhang et al. 2022 embedded the transformer module [3] into the backbone network of YOLOv5 and combined the self-attentive mechanism with a convolutional neural network to improve the detection accuracy of the model for cherry fruit with a mAP of 95.20% [4]. Yan et al.

2022 embedded the squeeze and excitation (SE) channel attention mechanism [5] into backbone network of YOLOv5m to improve the feature map spanning fusion of the input medium-sized target detection layer, and the improved model had an mAP of 80.70 and an average detection time of 25 ms for one image [6]. He et al. 2022 used YOLOv5 for recognition of tomato in the nighttime environment of a heliostat and achieved tomato fruit recognition under dark light features by recomputing the adaptive frame with an improved CIOU objective loss function [7]. Sun et al. 2022 embedded the convolutional block attention module (CBAM) [8] into YOLOv5s and used the phantom structure of GhostNet [9] to reduce the model complexity. The YOLOv5s was improved by adding the transformer structure to the model and applying it to apple fruit disease detection with a mAP of 88.2% and a model parameter count of only 2.06 MB, which is ideal for deployment on mobile devices [10]. Huang et al. 2022 introduced the CBAM attention mechanism and α-IoU loss to YOLOv5 for improvement and applied the improved model to citrus fruit target detection with an AP of 91.3% and a single image detection speed of 16.7 ms [11]. Lyu et al. 2022 adopted a lightweight YOLOv5-CS to recognize green citrus in wild field [12]. Xu et al. 2023 proposed HPL-YOLOv4 to improve the recognition of citrus-fruit and ensured real-time efficiency [13].

YOLOv5s has little volume and quicker detection speeds than the other three architectures but has the lowest detection accuracy. It is more suitable for detecting large target scenarios. The volume of lightweight model determines that YOLOv5s runs with fewer computational and memory resources, making it more suitable for deployment on the embedded devices commonly used by picking robots.

We propose a new YOLOfuse model with YOLOv5s combined with CSPDarknet53-Tiny [14] and a coordinate attention (CA) mechanism [15]. The features of heterogeneous source images separately are extracted by the backbone and then fused by the fusion strategy. Then, the fused features are fed into the neck network for target detection after the CA attention mechanism. Experiments are designed to evaluate the model performance including detection accuracy, detection speed, and model parametric number under the test set.

1.2. Highlights

First, this paper embeds CSPDarknet53-Tiny as a feature extraction network for a depth map on the basis of YOLOv5s, which can extract features of both the depth map and RGB image for model judgment.

Second, we introduce a more lightweight coordinate attention (CA) module, which can ensure the lightweight and improve the accuracy and efficiency.

Third, we design a new feature fusion strategy for feature fusion, which multiplies the sparse features of the depth map by the scaling factor and adds the features of the RGB map. This way ensures the dominance of RGB features and guarantees the accuracy.

2. Materials and Methods

2.1. Materials

2.1.1. Datasets

There are two datasets for the experiments, including one public dataset and one self-built dataset. The cameras in the two datasets use the technology of time of flight (ToF) to capture the depth images. The ToF technology adopts the time difference between the active transmission signal and the received reflection signal to perform centimeter-level precise ranging, providing pixel values as distance information. The proposed YOLOfuse model is suitable for fruit detection of a pair of heterogeneous images with RGB and ToF technology in natural orchard scene.

The public dataset used for the experiments is from the publicly available apple RGB-D image dataset PApple_RGB-D-Size [16] from the Universitat de Lleida (Spain). The dataset contains 4017 pairs of RGB-D images from six Fuji apple trees, including 2407 pairs of mature apple images and 1610 pairs of 75% mature green apple images, containing a total

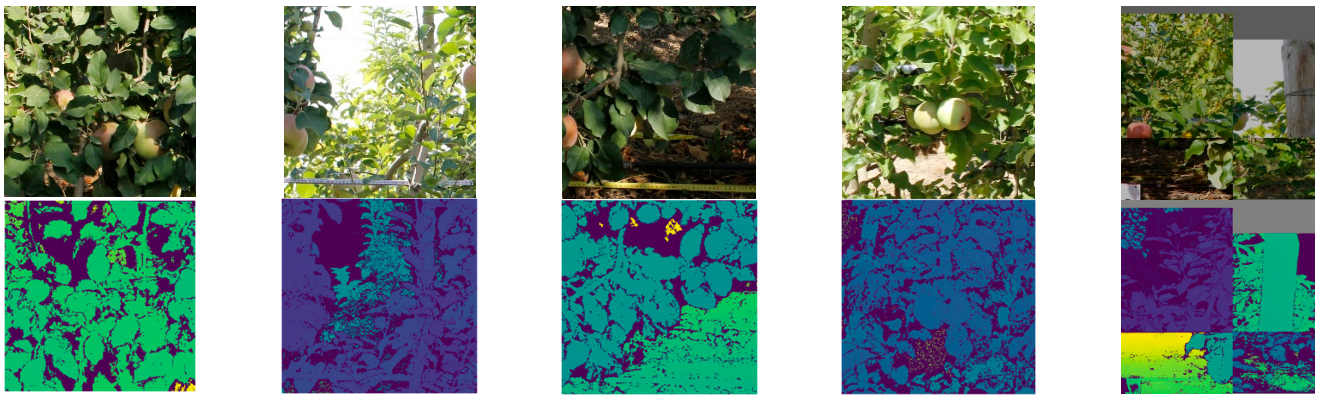
of 16,332 apple objects. In this dataset, there are a large number of apple objects in the state of branch occlusion, inter-fruit overlap occlusion, and edge mutilation. The image acquisition device of this dataset is a Microsoft Kinect v2 depth camera. The capture range of the camera is between 0.3 and 0.8 m. The target imaging distance is about 0.5 m, the images are saved in PNG format, and the image resolution reaches 1024 × 1024.

The self-built dataset RGB-D heterologous images are captured by a depth camera produced by Basler Company in Germany in cooperation with a color camera. The capture range of the camera is between 0.3 and 1.5 m. The target imaging distance is about 0.8 m, the images are saved in TIFF format, and the image resolution reaches 640 × 480. There are four types of camera output, including intensity images, confidence images, depth images, and 3D point cloud images. In an outside scenario, when the camera's deep information lacks performance, we avoid taking photos in direct sunlight or choose to avoid the noon period. The heterologous images were collected from the apple experimental base at Tianshui Fruit Research Institute in Gansu Province, China. Over 2000 RGB-D heterogeneous images were captured at different time periods from 14:00 to 17:00 in August.

2.1.2. Image Enhancement

We use LabelImg image annotation software to target and annotate RGB images in the RGB-D heterogeneous image dataset using XML format for annotation. In order to increase the spatial multi-scale and feature richness of training set images, four image enhancement methods, namely random Affine transformation, random horizontal flip, Mosaic data enhancement, and HSV gamut transformation, are used in series at the input end of the network. The RGB color image and depth image of two datasets are expanded to 10,000 images by using the above four image enhancement methods.

Two datasets are randomly divided into training, validation, and testing sets in an 8:1:1 ratio, respectively. In order to reduce the redundant information of high-resolution images and accelerate the model convergence, the image resolution of two datasets of the input network is uniformly scaled to 480 × 480. Some RGB-D image pairs and the training images of the public dataset after data enhancement are displayed in Figure 1.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 1. Heterologous images of the dataset: (**a**) Severely overlapped; (**b**) Overlapping occlusion; (**c**) Incomplete; (**d**) 75% ripe green apple; (**e**) Train images after data enhancements.

2.2. YOLOfuse Apple Detection Model

2.2.1. YOLOv5s Network Framework

The network framework of YOLOv5s includes backbone, neck network and head network. The first part is used as the feature extractor. It generates feature maps and inputs them into the neck network for subsequent prediction. The first layer of the backbone is the focus module, which splits the input image into four complementary parts by extracting pixels at intervals in the image aspect direction. In this way, the information in the image aspect direction can be concentrated in the channel direction. The loss of direct

downsampling of the image and model computation can be reduced and thus the network convergence can be accelerated as well. The focus module structure is shown in Figure 2.

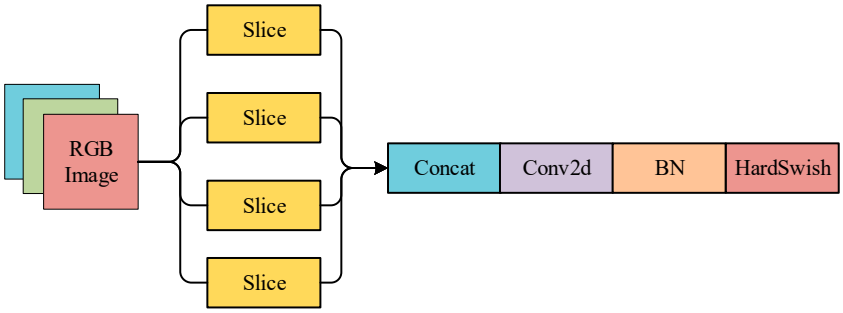


Figure 2. Focus module.

Layer 2 of the backbone is the CBS module, including Conv2d layer, BN layer and SiLU activation function in series for feature map channel number adjustment and downsampling. The details of CBS module is given in Figure 3.

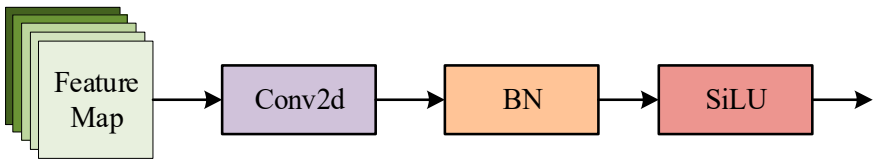


Figure 3. CBS module.

The third layer of the backbone is the CSP1_X module, which is mainly used to strengthen the network's ability to mine deeper features of images. The main component of CSP1_X is the Res_unit residual unit. There are two CBS modules. One is embedded with one convolutional kernel, which size is 1×1. Meanwhile, the other one is embedded with another convolutional kernel, which size is 3×3 connected sequentially. And the output features of this part are summed with the input features of the whole residual unit as the overall output of the Res_unit unit. The CSP1_X module inputs the initial input features into a parallel two-branch network and connects X Res_unit units in series. The Res_unit unit is indicated in Figure 4. The CSP1_X structure is given in Figure 5.

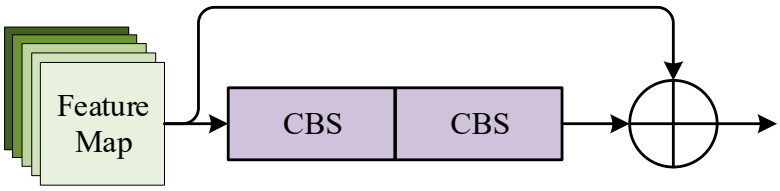


Figure 4. Res_unit module.

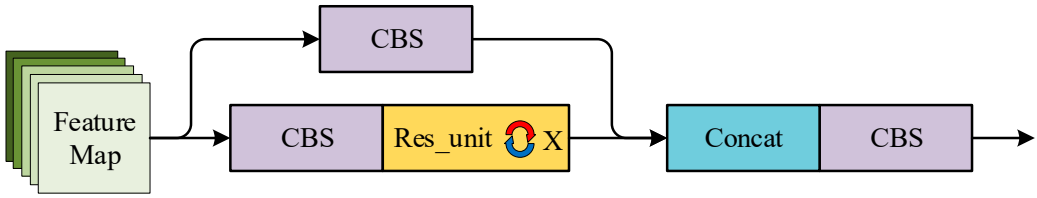


Figure 5. CSP1_X module.

The 9th layer of the backbone is the spatial pyramid pooling (SPP) module [17]. The main purpose of this module is to increase the network receptive field and improve the network's perception ability towards small targets. The SPP module passes the input

features through three pooling windows of size 5 × 5, 9 × 9, and 13 × 13 of a two-dimensional maximum pooling layer Maxpool2d to obtain three groups of features in the same space with input. Finally, the three groups of features are channel-spliced with the input features as the module output, so as to effectively fuse local features and global features. The SPP module is listed in Figure 6.

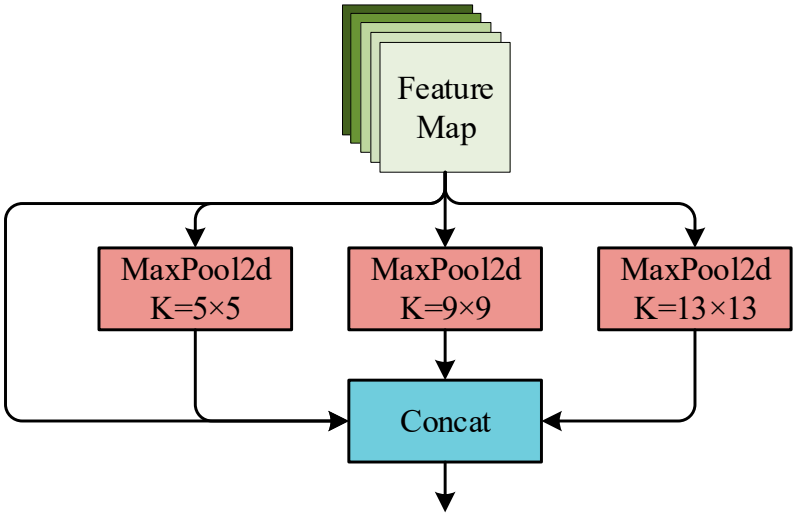


Figure 6. SPP module.

The 10th layer of the backbone is the CSP2_X module, whose main role is to integrate the fused features output from the SPP module and enhances the ability of feature extraction. This module inputs initial features into a parallel two-branch network. And after, one branch is channel halved by the CBS module and the other branch is channel halved by the CBS module in series with X sequential modules. The other branch consists of two CBS modules. One is embedded with one convolutional kernel, which size is 1 × 1. Meanwhile, another one is embedded with another convolutional kernel, which size is 3 × 3. Finally, the outputs of the two branches are Concat spliced in the channel direction and the feature maps are output through the CBS module. The CSP2_X module details are given in Figure 7.

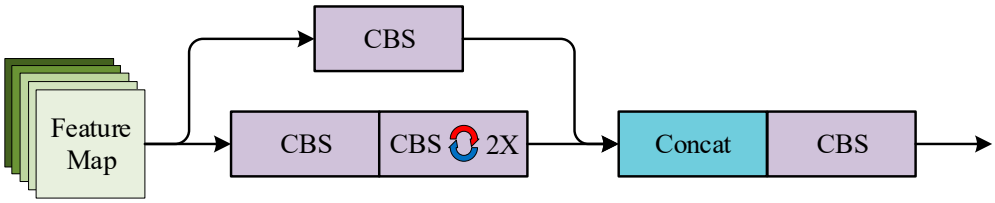


Figure 7. CSP2_X module.

The neck network uses the overall architecture of feature pyramid networks (FPN) [18] combined with path aggregation network (PAN) [19] to fuse features of different scales. The FPN structure can transmit high-dimensional semantic features extracted from deep network to shallow network. These features are more abstract and have a larger receptive field, which is beneficial for the network to classify and judge objects. At the same time, this type of information loses location information, which is not conductive to the network to judge the position of objects. The PAN structure transmits the low-dimensional localization features extracted from the shallow network to the deep network, which is more concrete and has sufficient expression ability for the spatial location of objects. It is beneficial for the network to judge the position of objects.

The Head network consists of 3 prediction modules, which receive the output feature maps of the neck network at 3 different scales. These modules are applied to check different

sizes of targets from three scales. The module mainly consists of a Conv2d layer, and the final output network detection result has a channel number of 18 (the network contains 3 anchor frame outputs and each anchor frame output includes 1 category probability, 1 confidence value, and 4 position coordinates). The YOLOv5s framework is displayed in Figure 8.

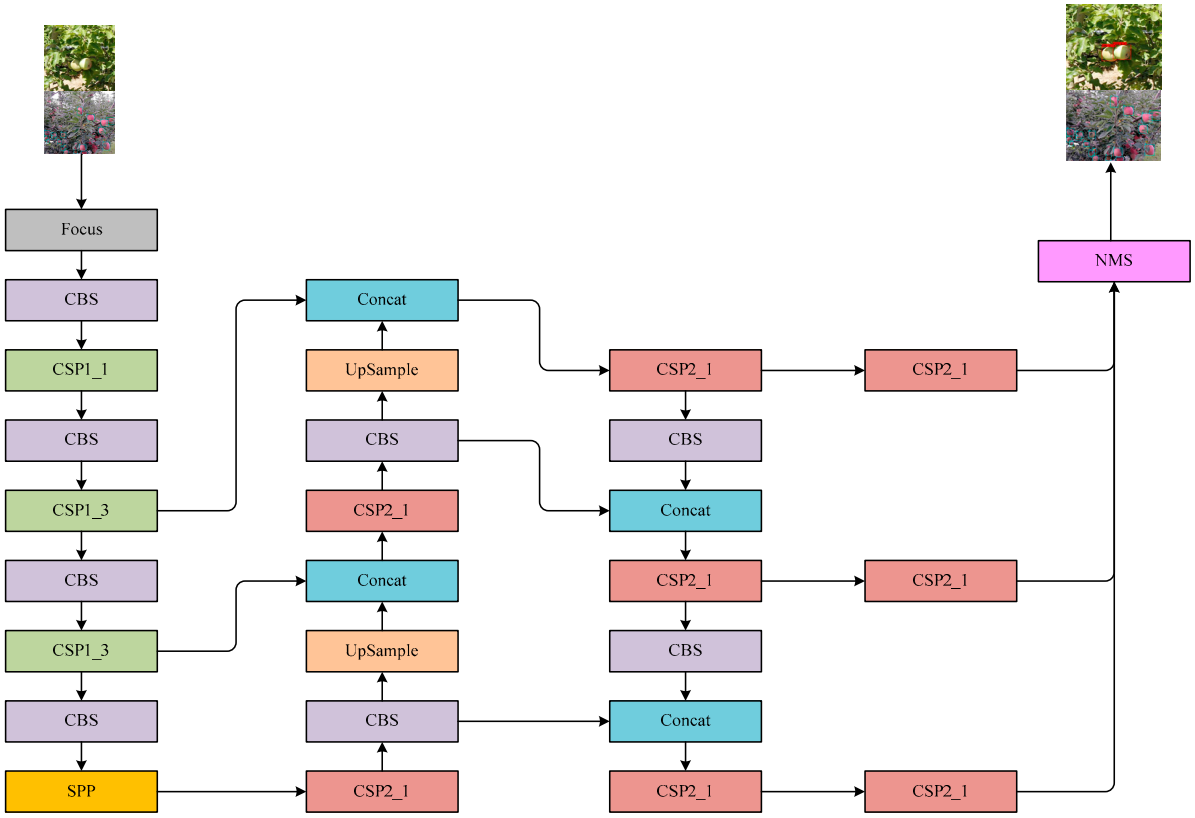


Figure 8. Architecture of original YOLOv5m network.

2.2.2. Embedding the CSPDarknet53-Tiny in Backbone

To effectively extract useful features from depth images and RGB color images in RGB-D images, we combine the CSPDarknet53-Tiny network with the original backbone of YOLOv5s to form a two-branch backbone, and the 2 branches are used for feature extraction of the RGB images and depth images. The main components of the CSPDarknet53-Tiny network are the CBL module and the Resblock_body module. The CBL module consists of the Conv2d layer, the BN layer, and the Leaky ReLU activation function in series. Resblock_body module first uses the CBL module to bisect the output features in the channel direction into two branch networks. Branch 2 network undergoes three rounds of CBL module (with residual connection in the second CBL) and concatenates with branch 1 output in the channel direction through Concat concatenation. Finally, feature downsampling is performed through the Maxpool2d layer with a pooling window of 2×2, as the output of the whole module. The CSPDarknet53-Tiny network consists of two CBL modules, three Resblock_body modules, and one CBL module sequentially connected in series. The Resblock_body module and the CSPDarknet53-Tiny network are shown in Figures 9 and 10, respectively.

The network uses CSPDarknet53-Tiny as the exclusive feature extraction network for depth images and uses a specific feature fusion strategy to fuse the final output features with the input features from the SPP module in YOLOv5s' backbone. The fused features are then further fed into subsequent networks for prediction.

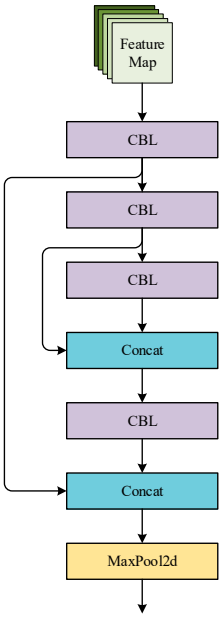


Figure 9. Resblock_body module.

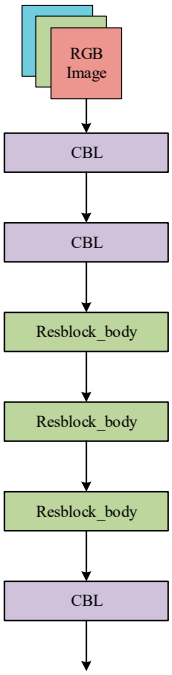


Figure 10. Architecture of CSPDarknet53-Tiny network.

2.2.3. Design Feature Integration Strategy

Feature fusion strategy is designed based on using weighted average. Considering the sparse characteristics of the depth image features, this paper argues that depth features are prone to feature dilution when the proportion of depth features in the final fused features is too high. This causes the neck network and head network to fail to obtain enough useful features and thus causing a decrease in detection accuracy. Therefore, this paper proposes a measure factor α ($\alpha \in [0,1]$) to measure the utilization of depth image features. The feature fusion strategy is shown in Equation (1).

$$F_{fuse} = F_{RGB} + \alpha F_{Depth} \quad (1)$$

2.2.4. Embedded CA Attention Module in the Network

Attention weighting is performed directly on the fused features output in the previous section using the CA attention module. Although channel attention represented by the SE attention module has been shown to be effective in improving model detection, this type of attention ignores the spatial location information of features. It has disadvantage of low efficiency of capture direction-aware and location-sensitive information. The CA attention module adopts the position into channel to effectively utilize the spatial information of the features while removing a large computational overhead. The CA attention module adopts two one-dimensional global averaging pooling layers to aggregate the input features into two independent direction features in H and W directions. Two maps are encoded. The input features are captured along their respective perceptual directions. The CA attention module network structure is given in Figure 11, and YOLOfuse structure is listed in Figure 12.

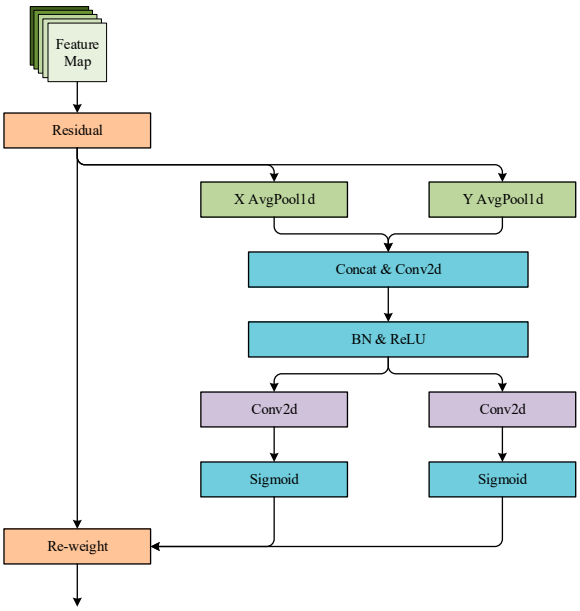


Figure 11. CA attention module.

2.2.5. Soft-NMS Algorithm

The model improvement replaces the non-maximum suppression (NMS) algorithm of YOLOv5s [20] with the flexible non-maximum suppression algorithm (Soft-NMS) [21]. YOLOv5s, as an anchor box based object detection algorithm, usually generates multiple detection boxes for a detected object. The NMS algorithm is applied in YOLOv5s to filter the poorly predicted detection boxes and ensure that the best-detected box is retained for each detected object as much as possible. However, when the overlapping degree of detected objects of the same type is high, the detection boxes of two objects overlap at the same height. The detection boxes of the obscured objects are easily suppressed by the detection boxes of the objects in front of them under the NMS algorithm, resulting in missed detections. The Soft-NMS algorithm can be more effective in solving the above-mentioned phenomenon of missed detection due to excessive suppression. The Soft-NMS algorithm is a flexible improvement in the NMS suppression algorithm. When the detection box suppression condition is established, the Soft-NMS algorithm does not completely suppress the confidence score of the predicted box to 0 but sets a decay function for the confidence score of the predicted box. Through multiple iterations of screening, multiple suppressions are performed as far as possible. The Soft-NMS method is computed in Equation (2).

$$S_i = \begin{cases} S_i & if\ IoU(M, bi) < N_t \\ S_i e^{-\frac{IoU(M,b_i)^2}{\sigma}} & if\ IoU(M, bi) \geq N_t \end{cases} \quad (2)$$

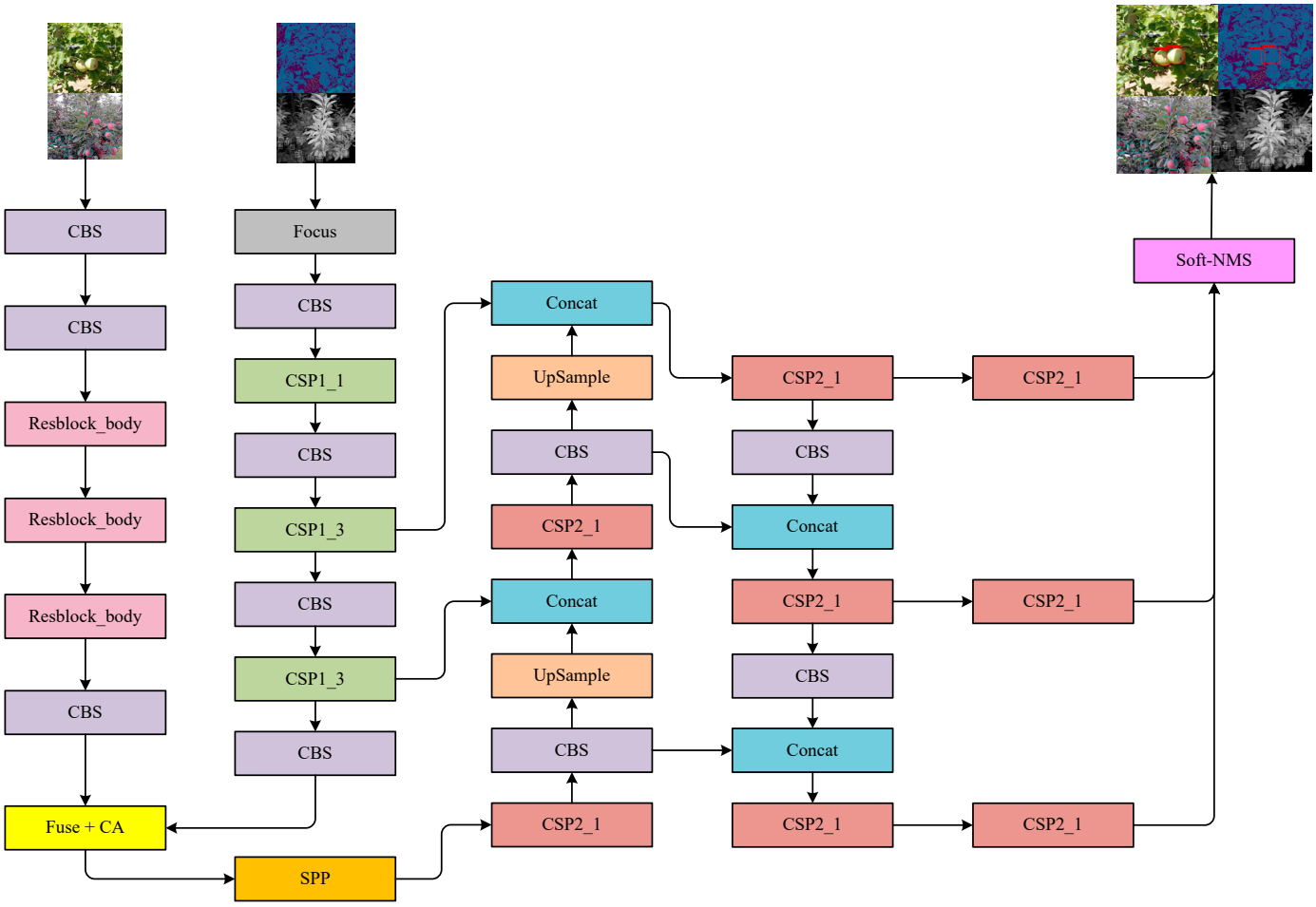


Figure 12. Architecture of YOLOfuse network.

3. Results

3.1. Experiments

The environment configuration used for experiment is listed in Table 1.

Table 1. Environment configuration.

Environment	Configuration
Training and testing platform	ecloud
CPU	Intel(R) Xeon(R) Gold 5118 CPU @ 2.30 GHz
RAM	376 GB
Cloud Storage Space	50 GB
GPU	NVIDIA Tesla V00 16 GB
OS	Ubuntu18.04.03 LST
Virtual Environment	Anaconda
NVIDIA GPU Driver	450.51.05
Programming Languages	Python 3.7.6
Deep Learning Framework	PyTorch 1.8
CUDA Version	11.0
cuDNN Version	7.6.5

The experiments check the detection performance on test set in five aspects, including precision, recall, F1 value, average precision, and frames per second (FPS). They are listed between Equations (3) and (7).

$$P = TP/(TP + FP)) \quad (3)$$

$$R = TP/(TP + FN) \tag{4}$$

$$AP = \int_0^1 P(R)dR \tag{5}$$

$$F1 = 2PR/(P + R) \tag{6}$$

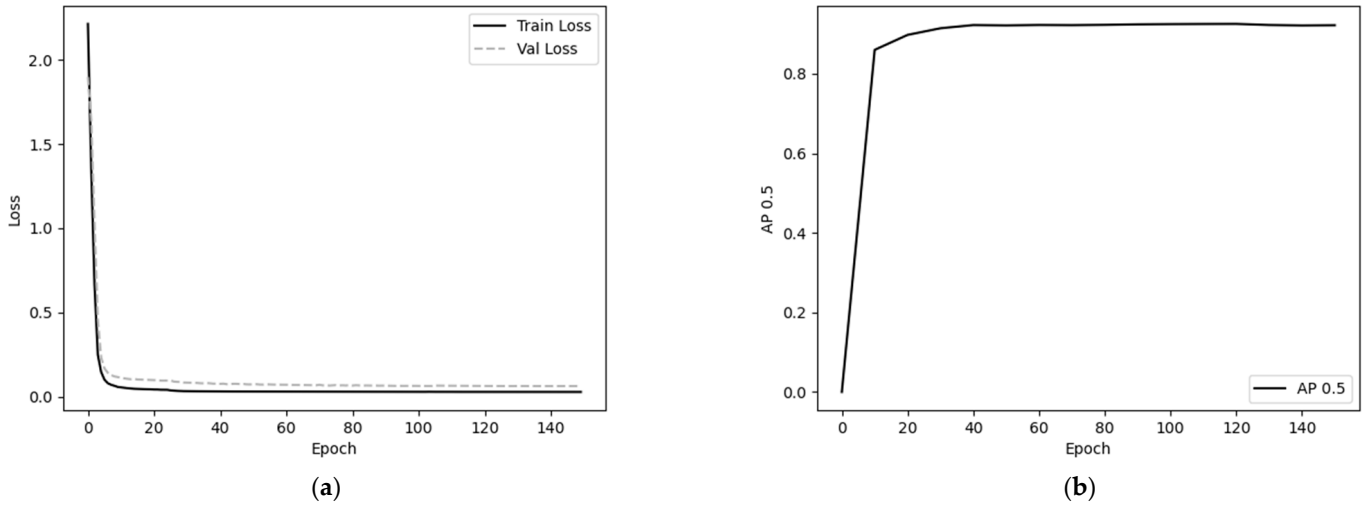
$$FPS = 100/\int_0^{100} \left(T_i^{start} - T_i^{end}\right) \tag{7}$$

3.2. Experimental Results

3.2.1. Training Results

YOLOfuse is trained by migration learning using pre-trained weights loaded into the backbone network before the CA attention mechanism, using the optimization function Adam, and a batch size of eight samples. The number of training rounds (Epochs) is set to 300 rounds. 0.937 is momentum factor. Zero is set as decay rate. The learning rate update strategy of COS is adopted. 0.00001 is initial learning rate, rising rapidly to 0.0003 at the beginning of training and decreasing gradually to the initial learning rate as the number of training rounds increased. Metric factor α is set to 0.2. Every 10 Epochs are saved to the model weights.

The loss curves of the training and validation sets are listed in Figure 13a, and the AP curves of the validation set are displayed in Figure 13b.



(**a**) (**b**)

Figure 13. Loss curve and the curve of each evaluation index under the validation set (**a**) Loss curve; (**b**) AP curve.

3.2.2. Performance Results

To analyze the performance of YOLOfuse, it is compared with the original YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv3 [22], YOLOV4 [23], YOLOV4-Tiny [14], and Faster RCNN [24] on the public test set. Among them, Faster RCNN selects ResNet 50 [25] as the backbone, and the test results using public data are listed in Table 2. Figure 14 displays the detection effect of YOLOfuse on public test set. Figure 15 gives the detection effect of YOLOfuse from self-built test set.

Table 2. Results of each model on public test set.

Models	Precision	Recall	F1	AP	FPS	Param/M	Size (MB)
YOLOv3	94.7%	87.1%	90.8%	91.8%	45.676	61.524	235
YOLOv4	91.8%	90.0%	90.9%	91.7%	36.018	63.938	245
YOLOv5s	95.0%	87.4%	91.0%	93.6%	54.645	7.022	26.9
YOLOv5m	94.7%	90.7%	92.7%	94.4%	41.827	20.871	79.9
YOLOv5l	92.7%	91.5%	92.1%	95.7%	33.311	46.138	176
YOLOv5x	94.0%	90.7%	92.3%	96.1%	28.602	86.217	329
YOLOv4-Tiny	93.1%	85.3%	89.0%	91.9%	154.357	5.874	22.4
Faster RCNN	57.0%	92.7%	70.6%	92.0%	17.599	28.275	109
YOLOfuse	95.4%	89.1%	92.1%	94.2%	51.761	10.701	41.1

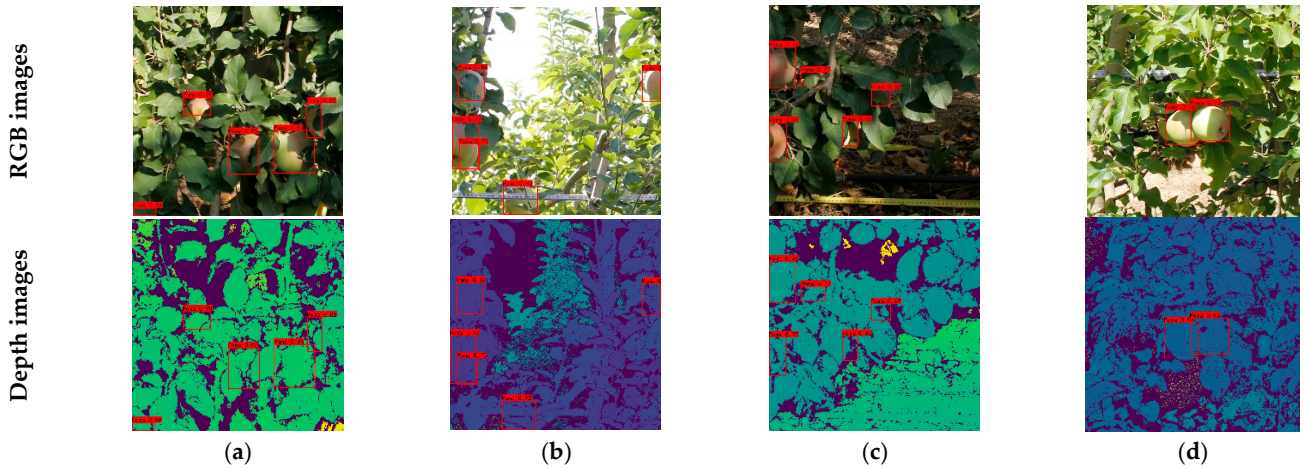


Figure 14. The detection effect of YOLOfuse model on the public test set: (**a**) Severely overlapped; (**b**) Overlapping occlusion; (**c**) Incomplete; (**d**) 75% ripe green apple.

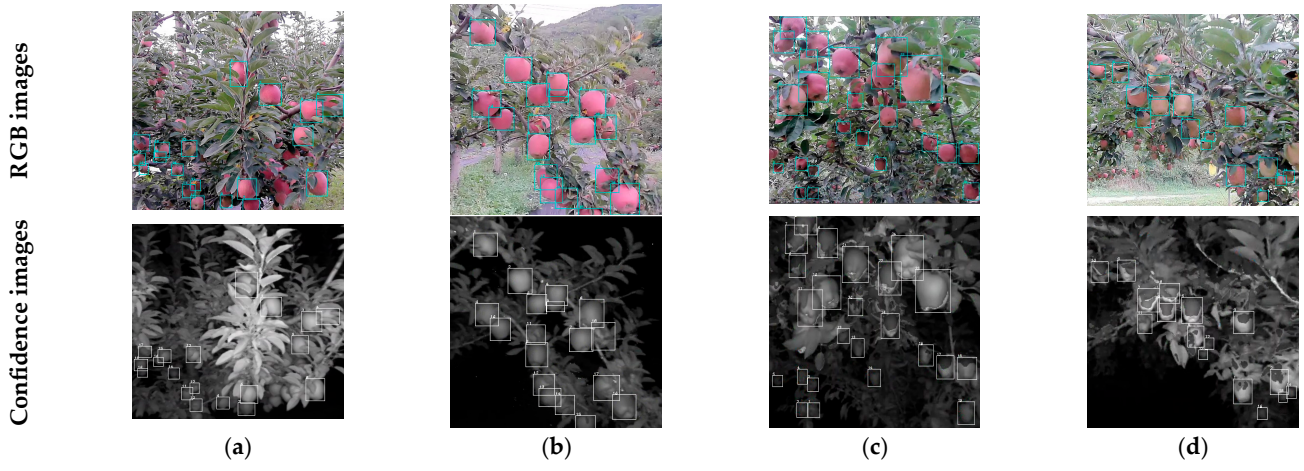


Figure 15. The detection effect of YOLOfuse model on the self-built test set: (**a**) Severely overlapped; (**b**) Overlapping occlusion; (**c**) Incomplete; (**d**) 75% ripe green apple.

3.2.3. Ablation Experiment of Soft-NMS

This experiment verifies the performance difference in YOLOfuse equipped with Soft-NMS compared to the one equipped with the NMS algorithm. It is tested on the public test set for model testing, and the performance metrics are obtained after completing the test as listed in Table 3.

Table 3. Detection results under NMS and Soft-NMS.

Model	Precision	Recall	F1	AP
YOLOfuse with NMS	95.2%	88.3%	91.6%	93.7%
YOLOfuse with Soft-NMS	95.4%	89.1%	92.1%	94.2%

3.2.4. Ablation Experiment of Attention

The experiments verify the performance differences between YOLOfuse equipped with the CA attention mechanism and other attention mechanisms. Considering the plug-and-play and easy replacement of the attention mechanism module, this section's experiments use SE and CBAM, to replace CA. A total of three models are obtained and trained in the network separately, and then, the models are tested on public test set. The performance metrics are obtained from Table 4. Figure 16 gives detection results of embedding three attention modules of some public test set images.

Table 4. Results of the models equipped with different attention modules.

Models	Precision	Recall	F1	AP	FPS (V100)	Param/M	Size (MB)
YOLOfuse SE	94.9%	88.2%	91.4%	93.6%	52.482	10.685	41.0
YOLOfuse CBAM	95.6%	88.6%	92.0%	94.3%	48.273	10.783	41.1
YOLOfuse CA	95.4%	89.1%	92.1%	94.2%	51.761	10.701	41.1

YOLOfuse CA | YOLOfuse SE | YOLOfuse CBAM

Figure 16. Detection effect of this model embedded with different attention modules.

4. Discussion

4.1. Results Discussion

In the training experiment, in Figure 13a, and from the loss curves, we can see that the loss of YOLOfuse are in a fast-decreasing state in the first 25 rounds, and the model loss decreases slowly and finally stabilizes after 25 rounds. In Figure 13b and from the AP curve, the AP value of YOLOfuse in the validation set increases rapidly in the first 40 rounds of training and then stabilizes, and the weight file from the 120 rounds with the highest AP value in the validation set is selected as the final model for this training, and the AP is 94.5% at this time.

In the performance test and from Table 2, we can see that the AP value of YOLOfuse on the public test set reaches 94.2%, the detection frame rate reaches 51.761 FPS, the parameter's number is only 10.701 M. Meanwhile, the model size is 41.1 MB. YOLOfuse has the advantages of high detection accuracy, fast detection speed, a small number of model parameters, and being a lightweight model when compared with the three models YOLOv3, YOLOv4, and Faster RCNN. Taking YOLOv5s as an example, the AP value of YOLOfuse is improved by 0.6%, the detection frame rate is decreased by only 2.884 FPS, the parameter's number is increased by 3.679 M, and the model size is increased by 14.2 M, so that the detection accuracy is higher at a smaller cost. Taking YOLOv5m, YOLOv5l, and YOLOv5x models as examples, the AP value of YOLOfuse decreases by 0.2, 1.5, and

1.9 percentage points, but the detection frame rate gains by 9.934 FPS, 18.45 FPS, and 23.159 FPS, respectively, and it has a clear advantage in both parameter's number and model size. Although the AP value has decreased, the model gains in the detection frame rate and model size. Taking YOLOv4-Tiny model as an example, YOLOfuse improves the detection accuracy by 2.3%, but the detection frame rate decreases by 102.596 FPS.

The model YOLOfuse has better detection results and detection accuracy than other models except YOLOv5m, YOLOv5l, and YOLOv5x in this apple detection task. To analyze the reasons, we believe that depth images can provide distance features that RGB color images do not have, and depth images added to the network greatly enrich the overall richness of useful features and thus, help the model make more correct judgments. The model YOLOfuse occupies less memory space than other models, other than YOLOv5s and YOLOv4-Tiny and has the obvious advantage of light weight, which is more suitable for the embedded devices of picking robots. This paper concludes that YOLOfuse, as a dual-input model, has a certain degree of improvement of model parameter's number compared with YOLOv5s, and preprocessing of depth images consumes computational resources. Therefore, the speed of the original YOLOv5s model decreases to a certain extent, while YOLOv4-Tiny, as a single-input, purely lightweight network, only performs target detection at two feature scales, and YOLOfuse is not comparable to YOLOv4-Tiny in field of detection speed.

In the ablation experiment of Soft-NMS, from Table 3, it indicates that YOLOfuse with the Soft-NMS algorithm has improved 0.2% in detection accuracy, 0.8% in recall, 0.5% in F1 value, and 0.5% in AP value compared with that of the NMS algorithm. It displays that the Soft-NMS algorithm has a practical optimization effect for the apple fruit detection task and can effectively avoid the oversuppression of the dominant target on the invisible target detection frame in the fruit overlap state.

In the ablation experiment of attention mechanisms, from Table 4, it displays that in terms of detection accuracy, the model YOLOfuse has a 0.6% improvement in the AP value with the CA attention module compared to that with the SE attention module and a 0.1% decrease in AP value compared to that with the CBAM attention module. In terms of model size, all three model files are about 41 MB, with no significant volume differences. In terms of parameter's number, the number of all three models is about 10.7 M, with no significant differences. In terms of detection speed, the detection frame rate of YOLOfuse with CA module decreased by 0.721 FPS compared with that of the SE attention module and increased by 3.488 FPS compared with that of the CBAM attention module. We believe that the spatial attention mechanism of the CBAM directly applies attention to overall scale of features. It is a more direct attention weighting process than the CA attention mechanism for spatial feature points but requires a lot of computational resources, resulting in the higher detection accuracy and relatively lower detection speed of YOLOfuse with CBAM module. Considering both the reverse sides of detection speed and detection accuracy, this paper concludes that YOLOfuse equipped with the CA attention mechanism is more suitable for the apple fruit target detection task.

4.2. Discussion Summary

(1) By analyzing data of the public test set, YOLOfuse has relatively higher detection accuracy and relatively lower detection speed, indicating that using RGB-D heterogeneous images as network input can provide richer features to the network than the model with RGB images alone as input. This facilitates the model to make more accurate predictions, but it also increases the computational burden of the whole vision detection system and produces negative effects in terms of model detection speed.

(2) The ablation experiments equipped with both Soft-NMS and NMS demonstrat that the application of Soft-NMS is able to bring a 0.5% improvement in the AP value for the model on the apple fruit detection task. It is verified that the Soft-NMS algorithm has a positive impact on the detection task of occluding overlapping targets.

(3) Ablation experiments equipped with a total of three attention modules, CA, SE, and CBAM prove that each type of attention module can bring some accuracy improvements for the apple fruit detection task, and the reasons for the performance of each model equipped with each of the three modules are analyzed. Considering both detection accuracy and detection speed, we conclude that the CA attention module is more suitable for the apple fruit detection task as it balances the two most important factors of detection accuracy and detection speed for picking robots.

5. Conclusions

A YOLOfuse apple detection model for feature fusion of RGB-D heterogeneous images is proposed, which performs feature fusion on RGB-D heterogeneous images. The fused features are input into the neck network through the CA attention mechanism for object detection. The results indicate that the YOLOfuse model can effectively fuse RGB-D heterogeneous image features and efficiently identify apple targets in natural orchard environments.

Compared with the YOLOv5m, YOLOv5l, and YOLOv5x models, the YOLOfuse model still has the defect of low detection accuracy. Further improvements of the network structure and reduction in parameter quantity are needed to improve the average accuracy. Next, we will research on the YOLO network series and explore the embedded real-time characteristics of YOLOv4-Tiny to achieve efficient real-time detection.

Author Contributions: L.L. and P.H. collected data and analyzed the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Gansu Provincial University Teacher Innovation Fund Project [grant number 2023A-051]; and Young Supervisor Fund of Gansu Agricultural University [grant number GAU-QDFC-2020-08]; and Gansu Science and Technology Plan [grant number 20JR5RA032].

Data Availability Statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

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

References

1. Xiaoyang, L.I.U.; Dean, Z.H.A.O.; Weikuan, J.I.A.; Chengzhi, R.U.A.N.; Wei, J.I. Fruits Segmentation Method Based on Superpixel Features for Apple Harvesting Robot. *Trans. Chin. Soc. Agric. Mach.* **2019**, *50*, 15–23.
2. Ultralytics. YOLOv5[R/OL]. Available online: https://github.com/ultralytics/YOLOv5 (accessed on 1 March 2022).
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
4. Zhang, Z.Y.; Luo, M.Y.; Guo, S.X.; Li, S.; Zhang, Y. Cherry Fruit Detection Method in Natural Scene Base on Improved YOLO v5. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 232–240.
5. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
6. Bin, Y.A.N.; Pan, F.A.N.; Meirong, W.A.N.G.; Shuaiqi, S.H.I.; Xiaoyan, L.E.I.; Fuzeng, Y.A.N.G. Real-time Apple Picking Pattern Recognition for Picking Robot Based on Improved YOLOv5m. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 28–38+59.
7. He, B.; Zhang, Y.B.; Gong, J.L.; Fu, G.; Zhao, Y.; Wu, R. Fast Recognition of Tomato Fruit in Greenhouse at Night Based on Improved YOLO v5. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 201–208.
8. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
9. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
10. Sun, F.G.; Wang, Y.L.; Lan, P.; Zhang, X.D.; Chen, X.D.; Wang, Z.J. Identification of apple fruit diseases using improved YOLOv5s and transfer learning. *Trans. Chin. Soc. Agric. Mach.* **2022**, *38*, 171–179.
11. Huang, T.M.; Huang, H.Q.; Li, Z.; Shilei, L.; Xiuyun, X.; Qiufang, D.; Wei, W. Citrus fruit recognition method based on the improved model of YOLOv5. *J. Huazhong Agric. Univ.* **2022**, *41*, 170–177. [CrossRef]
12. Lyu, S.; Li, R.; Zhao, Y.; Li, Z.; Fan, R.; Liu, S. Green citrus detection and counting in orchards based on YOLOv5-CS and AI edge system. *Sensors* **2022**, *22*, 576. [CrossRef]

13. Xu, L.; Wang, Y.; Shi, X.; Tang, Z.; Chen, X.; Wang, Y.; Zou, Z.; Huang, P.; Liu, B.; Yang, N.; et al. Real-time and accurate detection of citrus in complex scenes based on HPL-YOLOv4. *Comput. Electron. Agric.* **2023**, *205*, 107590. [CrossRef]
14. Jiang, Z.; Zhao, L.; Li, S.; Jia, Y. Real-time object detection method based on improved YOLOv4-tiny. *arXiv* **2020**, arXiv:2011.04244.
15. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–25 June 2021; pp. 13713–13722.
16. Ferrer-Ferrer, M.; Ruiz-Hidalgo, J.; Gregorio, E.; Vilaplana, V.; Morros, J.R.; Gené-Mola, J. Simultaneous Fruit Detection and Size Estimation Using Multitask Deep Neural Networks[EB/OL]. Available online: https://www.grap.udl.cat/en/publications/papple_rgb-d-size-dataset (accessed on 14 August 2022).
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
18. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
19. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
20. Hosang, J.; Benenson, R.; Schiele, B. Learning non-maximum suppression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 15–17 June 2017; pp. 4507–4515.
21. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS--improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
22. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
23. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
24. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–10 December 2015; pp. 1440–1448.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.