electronics

Article

Weed Detection in Potato Fields Based on Improved YOLOv4: Optimal Speed and Accuracy of Weed Detection in Potato Fields

Jiawei Zhao [1], Guangzhao Tian [1,*], Chang Qiu [2], Baoxing Gu [1], Kui Zheng [3] and Qin Liu [4]

[1] College of Engineering, Nanjing Agricultural University, Nanjing 210031, China
[2] College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210031, China
[3] SUNWAY-AI Technology (Changzhou) Co., Ltd., Changzhou 213161, China
[4] School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China
* Correspondence: tgz@njau.edu.cn

Abstract: The key to precise weeding in the field lies in the efficient detection of weeds. There are no studies on weed detection in potato fields. In view of the difficulties brought by the cross-growth of potatoes and weeds to the detection of weeds, the existing detection methods cannot meet the requirements of detection speed and detection accuracy at the same time. This study proposes an improved YOLOv4 model for weed detection in potato fields. The proposed algorithm replaces the backbone network CSPDarknet53 in the YOLOv4 network structure with the lightweight MobileNetV3 network and introduces Depthwise separable convolutions instead of partial traditional convolutions in the Path Aggregation Network (PANet), which reduces the computational cost of the model and speeds up its detection. In order to improve the detection accuracy, the convolutional block attention module (CBAM) is fused into the PANet structure, and the CBAM will process the input feature map with a channel attention mechanism (CAM) and spatial attention mechanism (SAM), respectively, which can enhance the extraction of useful feature information. The K-means++ clustering algorithm is used instead of the K-means clustering algorithm to update the anchor box information of the model so that the anchor boxes are more suitable for the datasets in this study. Various image processing methods such as CLAHE, MSR, SSR, and gamma are used to increase the robustness of the model, which eliminates the problem of overfitting. CIoU is used as the loss function, and the cosine annealing decay method is used to adjust the learning rate to make the model converge faster. Based on the above-improved methods, we propose the MC-YOLOv4 model. The mAP value of the MC-YOLOv4 model in weed detection in the potato field was 98.52%, which was 3.2%, 4.48%, 2.32%, 0.06%, and 19.86% higher than YOLOv4, YOLOv4-tiny, Faster R-CNN, YOLOv5 l, and SSD(MobilenetV2), respectively, and the average detection time of a single image was 12.49ms. The results show that the optimized method proposed in this paper outperforms other commonly used target detection models in terms of model footprint, detection time consumption, and detection accuracy. This paper can provide a feasible real-time weed identification method for the system of precise weeding in potato fields with limited hardware resources. This model also provides a reference for the efficient detection of weeds in other crop fields and provides theoretical and technical support for the automatic control of weeds.

Keywords: weed identification; YOLOv4; attention mechanism; Mobilenetv3; precision agriculture

Citation: Zhao, J.; Tian, G.; Qiu, C.; Gu, B.; Zheng, K.; Liu, Q. Weed Detection in Potato Fields Based on Improved YOLOv4: Optimal Speed and Accuracy of Weed Detection in Potato Fields. *Electronics* **2022**, *11*, 3709. https://doi.org/10.3390/electronics11223709

Academic Editor: George A. Papakostas

Received: 11 October 2022
Accepted: 10 November 2022
Published: 12 November 2022

1. Introduction

The potato is the world's fourth most influential food crop and is loved by people in most countries around the world because of its nutritional value, low-temperature tolerance, and high yield. China has the highest potato cultivation area and potato production in the world, and the country's potato industry has been maintaining a rapid growth trend. During the growth of potatoes, various types of weeds in agricultural fields compete

with potatoes for water, nutrients, light, and space [1,2]. Weeds typically have a well-developed root system, fast growth rate, no natural enemies, high absorption capacity, strong reproduction capacity, and great adaptability, and the growth of these weeds can severely affect the normal growth and development of potatoes [3,4]. Most conventional weed control methods are mechanical and chemical weed controls [5–7]. Mechanical weed controls do not have a generalization, are incomplete weed controls, and have a possibility of crop damage [8]. Chemical weed controls are mostly sprayed over large areas, resulting in large amounts of pesticide waste, and residues from spraying also pose health risks to consumers and can cause damage to the ecological environment [9]. With the spread of smart agriculture, some modern weeding methods, such as laser weeding and precision herbicide spraying, are gradually being applied [10]. The implementation of these new methods is inseparable from the efficient detection of targets [11,12]. In this paper, we propose an efficient and highly accurate weed detection method for potato fields, which also serves as a reference for efficient weed detection in other agricultural crops.

In recent years, the development and application of sensor technology [13], computer technology [14], vision detection technology, and other advanced technologies have considerably accelerated the development of agricultural automation and intelligence [15]. Numerous scholars have investigated methods for crop weed detection, and the methods proposed so far include remote sensing analysis, spectral recognition, and machine vision recognition. Elstone et al. [16] used red, green, and near-infrared reflectance in combination with a size differentiation method for identifying crops and weeds in lettuce fields, but the system relied mainly on the size of the target to be detected to differentiate between weeds and crops and would misidentify rocks, branches, or crops as weeds. Pulido et al. [17] extracted the texture features of weeds using a gray-level co-occurrence matrix, used a principal component analysis to reduce the dimensionality of the features, and finally, used a support vector machine algorithm to complete the classification. While the aforementioned methods achieve the recognition of crops and weeds, they over-rely on manual design and selection of image features, are susceptible to environmental factors such as lighting and have poor stability and low recognition accuracy.

DL techniques are currently developing rapidly [18,19], and numerous detection algorithms based on DL techniques are used for the identification of weeds in crop fields [20–22]. Liu et al. [23] proposed an improved YOLOV4-tiny model fused with multi-scale Retinex with a color restoration (MSRCR) enhancement algorithm. Additionally, the K-means++ clustering algorithm is used for anchor box clustering analysis and channel pruning processing. Zhang et al. [24] fused the SE (squeeze-and-excitation) attention mechanism module with YOLOv5x and proposed SE-YOLOv5x for lettuce and weed detection. Wang et al. [25] constructed a CNN model YOLO_CBAM incorporating YOLOv5s and attention mechanisms for the detection of Solanum rostratum Dunal weeds. They designed a method to slice high-resolution images. This method constructed the dataset by calculating the overlap rate to reduce the possibility of detail loss due to the compression of high-resolution images during training, with a final precision of 92.72%. Dong et al. [26], based on YOLOv4, embedded a SA (shuffle attention) attention mechanism module in its backbone network to enhance the feature extraction ability of the model and introduced a Transformer module to construct the long-range global semantic information of the feature map. The average recognition time of the improved model for a single image was 0.261s, and the average recognition accuracy was 97.49%. The above studies show that although DL can solve the problem of manual feature design in traditional image processing methods, the following problems still exist: (1) although the use of complex network models for weed detection in crops improves recognition accuracy, the recognition speed cannot meet the real-time requirements due to the large size of its network models; (2) the detection speed is improved by simplifying the network models, but frequently, the recognition accuracy cannot meet the requirements [27].

In view of the above issues, in order to obtain a weed detection method in potato fields with both recognition accuracy, and recognition speed, we choose to improve on the

YOLOv4 algorithm in this paper. The original YOLOv4 algorithm has limited detection capability for multiple targets and is prone to miss and make false detections when weeds and crops are dense. Moreover, the network model of the YOLOv4 algorithm is more complicated and less efficient. In this paper, we replaced the backbone network CSPDarknet53 in the YOLOv4 network structure with the lightweight MobileNetV3 network and introduced Depthwise separable convolutions instead of partial traditional convolutions into the PANet structure. In order to improve the detection accuracy, the CBAM was fused into the PANet structure. Various image processing methods, such as CLAHE, MSR, SSR, and gamma, were used to increase the robustness of the model. The K-means++ clustering algorithm was used instead of the K-means clustering algorithm to update the anchor box information of the model. CIoU was used as the loss function, and the cosine annealing decay method was used to adjust the learning rate to make the model converge faster. The above-improved methods were combined to propose the MC-YOLOv4 model for weed identification in potato fields. To the best of our knowledge, this is the first study using the MC-YOLOv4 model for weed-crop classification. The optimized method proposed in this paper outperforms other commonly used target detection models in terms of model footprint, detection time consumption, and detection accuracy and can provide a viable real-time weed identification method for systems with limited hardware resources for precision weeding in the field.

2. Materials and Methods

This section mainly introduces two aspects. First, the collection and processing of the dataset are introduced in detail. Secondly, we detail the specific improvement methods of the model.

2.1. Materials

2.1.1. Data Acquisition

The images used in this study were collected at the Baima test site of the Nanjing Agricultural University in Nanjing, Jiangsu Province, China. The image acquisition device was a 2-megapixel camera module with a resolution of 1920 × 1080, saved in a JPG format. A total of 2000 images were obtained, including the images with only weeds, the images with only potatoes, and the images with both potatoes and weeds. The images were taken in different lighting conditions and at different growth stages, and the shooting height was fixed. Figure 1 shows examples of three types of images. Table 1 shows the number of images per type.

(a) (b) (c)

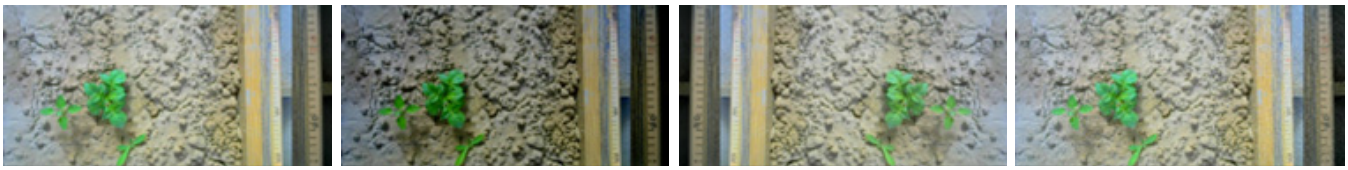
Figure 1. Three types of images. (**a**) The image of only weeds, (**b**) The image of only potatoes, (**c**) The image of both potatoes and weeds.

Table 1. Number of images per type.

Types of Images	Number of Images
The image of only weeds	654
The image of only potatoes	490
The image of both potatoes and weeds	856

2.1.2. Data Preprocessing

To enhance the richness of the experimental dataset, image data enhancement techniques were used to expand the size of the dataset, reduce the dependence of the weed recognition model on certain image attributes, reduce the overfitting of the trained model, and enhance the stability of the model. In this paper, data enhancement was performed by random flip horizontal, brightness variation, the addition of appropriate amounts of Gaussian noise, pretzel noise, and scatter noise in common ways. From the original dataset of 2000 images of potatoes and weeds, 510 suitable images were randomly selected for data enhancement in these ways, resulting in an enhanced dataset of 2550 images. Figure 2 shows examples of images of potatoes and weeds after common enhancement treatments.



(a) (b) (c) (d)

Figure 2. Common image enhancement examples. (**a**) Original image, (**b**) Brightness change processing, (**c**) Flip horizontal, (**d**) Noise processing.

Conventional image enhancement processing usually cannot solve the problems of blurred details and poor contrast in images caused by vibration, light, and weather in real environments. It is necessary to take a deeper enhancement process on the dataset again, which is beneficial to increase the detection accuracy of the subsequent model network. In this paper, CLAHE (contrast limited adaptive histogram equalization), MSR (multi-scale retinex), SSR (singal scale retinex), gamma, laplus, and Log were adopted to enhance the images of 2550 routinely enhanced datasets again. The CLAHE algorithm enhances the image contrast by limiting the height of the local histogram, enhancing the effective signal and suppressing the noisy signal, and increasing the speed by using a linear interpolation algorithm. The SSR removes the low-frequency components from the original image, and naturally, what remains are the high-frequency components in the original image. We know that in the human visual system, it is sensitive to high-frequency information such as edges. Therefore, SSR can better enhance the high-frequency information in the image. MSR is an improvement over SSR as it maintains strong image fidelity and compresses the dynamic range of an image. Under certain circumstances, MSR can achieve color enhancement, color constancy, local dynamic range compression, and global dynamic range compression. Figure 3 shows images of the potatoes and weeds after further enhancement.

The annotation tool Labelimg (the software version number is Labelimg 1.8.6, created by tzutalin, and the creator has released the version of the software in Canada) was used to annotate each of the 2550 images in the annotated format of the Pascal voc dataset: weeds were labeled weed, potatoes were labeled potato, and they were generated in a .xml type annotation file. It was randomly divided into training, validation, and test sets in a 7:2:1 ratio, where the training set consisted of 1785 images, the validation set consisted of 510 images, and the test set consisted of 255 images.

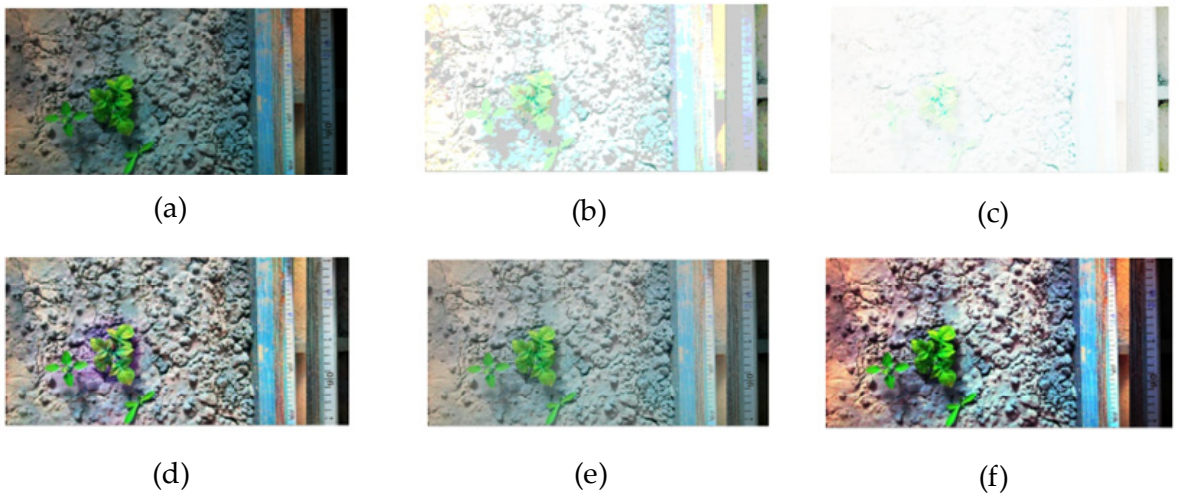


Figure 3. Data enhancement. (**a**) CLAHE enhancement, (**b**) Gamma enhancement, (**c**) MSR enhancement, (**d**) SSR enhancement, (**e**) Log enhancement, (**f**) Laplus enhancement.

2.2. Methods

2.2.1. YOLOv4 Algorithm

At present, there are two main categories of target detection algorithms widely used in DL: the first category is two-stage target detection algorithms, representing algorithms such as the R-CNN [28] series, etc. The two-stage target detection algorithm first forms a series of candidate prediction boxes and then classifies the samples by a CNN, which has high detection accuracy but not fast detection. The second category is the one-stage target detection algorithm represented by the YOLO series, SSD series, Retina-Net, etc. The one-stage target detection algorithm directly regresses the class probability and location coordinate values of the object, which is faster in comparison to the two-stage target detection algorithm, but the detection accuracy is lower [29–31].

Alexey et al. proposed in 2020 the fourth generation of the YOLOv4 object detection algorithm to balance detection accuracy and detection speed. YOLOv4 consists of three main parts: the backbone, neck, and prediction network. The backbone network uses the CSPDarknet53 structure, which is based on the improved backbone network Darknet53 of YOLOv3, with two main improvements. One of them is the introduction of the CSPNet structure in the Resblock body in Darknet53. Second, the activation function of DarknetConv2D is changed from Leaky relu to Mish, and the learning power of the CNN is enhanced by improving the backbone network. These improvements reduce the memory cost required by the model and reduce the computational bottleneck. The neck network of YOLOv4 uses the SPP module and PANet structure. The SPP module processes the feature maps processed by the backbone network using maximum pooling at four different scales (1×1, 5×5, 9×9, 13×13) and then performs a feature stack operation on the feature maps at different scales after processing, which can greatly increase the perceptual field of the convolution kernel and help extract more significant features. The processed feature maps are fed into the PANet network, and the feature extraction is repeated by two upsampling and two downsamplings to complete the feature stack several times. Finally, the extracted three different feature maps are fed into the prediction network YOLO Head for decoding.

2.2.2. Mosaic Data Enhancement Method

The Mosaic data enhancement method is an optimization and extension of the CutMix data enhancement method, which stitches two training images together simultaneously. By contrast, the Mosaic method simultaneously flips, scales, and color gamut's variation in the four training images and arranges them in four orientational positions before stitching

them into a single image, which enriches the target detection background and improves the model efficiency. Figure 4 shows an example of the Mosaic data enhancement method.

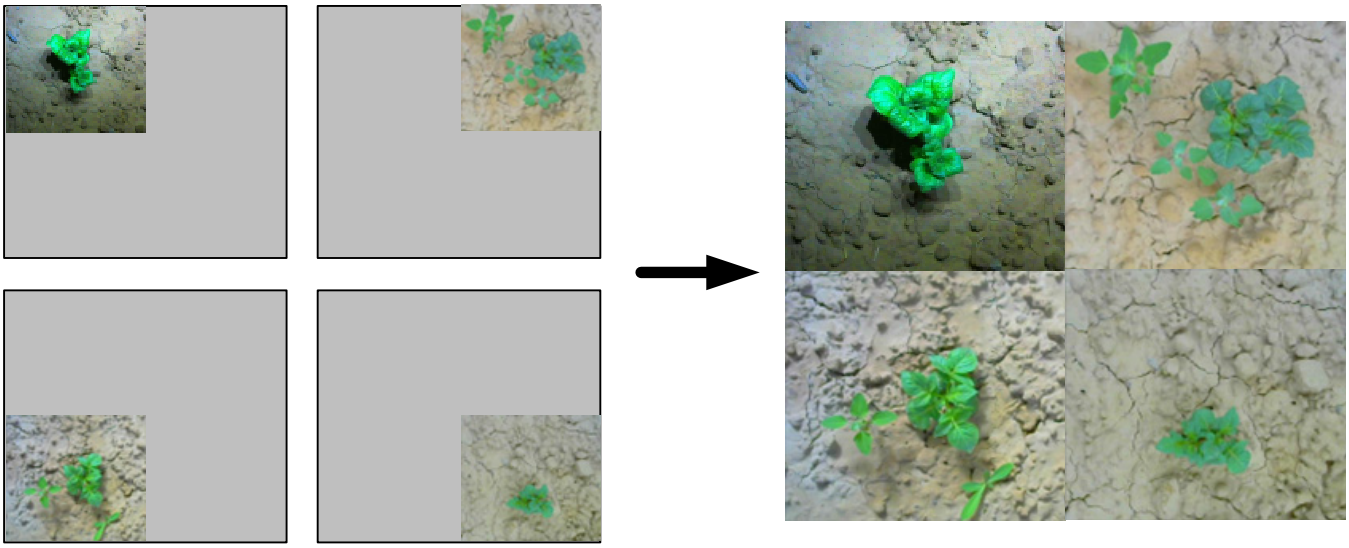


Figure 4. Mosaic data enhancement.

2.2.3. Use of Complete Intersection over Union Loss

The total loss function (L_{Loss}) of the evaluation prediction task is generally composed of three parts: the confidence loss function (L_P), classification loss function (L_C), and regression loss function (L_B). The calculation method is shown in Equation (1).

$$L_{Loss} = L_P + L_C + L_B \tag{1}$$

A good regression localization loss function should consider three geometric parameters: the overlapping area of the prediction box and the ground truth box, the distance between the center point of the prediction box and the ground truth box, and the aspect ratio of the prediction box. The complete intersection over union (CIoU) contains all three geometric parameters, which makes the regression of the target box more stable compared to the conventional loss function intersection over union (IoU) [32] and generalized intersection over union (GIoU) [33]. Compared to the regular loss functions IoU and GIoU, CIoU did not suffer from divergence in training, which makes the regression of the target box more stable. The calculation method of CIoU is shown in Equation (2).

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

The regression loss function L_B is shown in Equation (3).

$$L_B = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{3}$$

where $\rho^2(b, b^{gt})$ represents the Euclidean distance d between the centroids of the prediction box and the ground truth box, IoU represents the ratio of intersection and union of the prediction box and ground truth box. The calculation method of α is shown in Equation (4). The calculation method of v is shown in Equation (5).

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

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

The physical meanings of the above parameters are shown in Figure 5.

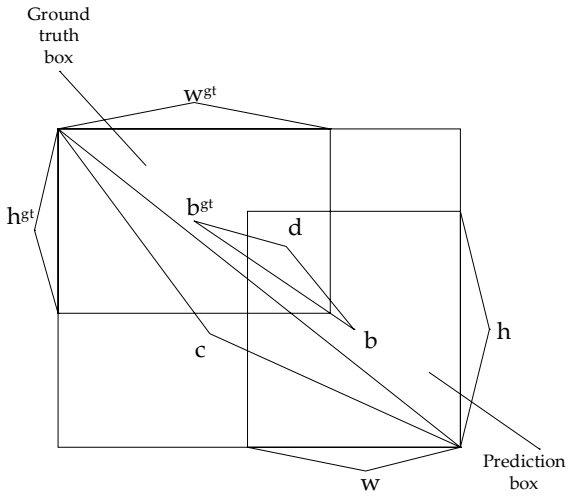


Figure 5. Physical meaning of parameters. w and h represent the width and height of the prediction box; w^{gt} and h^{gt} represent the width and height of the ground truth box; d represents the Euclidean distance between the center of the prediction box and the ground truth box; c represents the diagonal length of the smallest enclosing rectangle of the prediction box and the ground truth box.

2.2.4. Cosine Annealing

To make the training network model converge to the optimal solution as much as possible, the gradient descent algorithm can be used to optimize the objective function. The learning rate should become smaller as the global minimum of the Loss value is gradually approached. The calculation method is shown in Equation (6):

$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \quad (6)$$

where η_t represents the current learning rate; $\eta_{\min}^i$ and $\eta_{\max}^i$ represents the maximum and minimum values of the learning rate, respectively; i is the value of the index run; T_{cur} is the current iteration number; and T_i is the total number of iterations in the current training environment.

In this paper, we adapt the learning rate with a cosine annealed decay method, which is a gradient descent method with a restart. Since the objective optimization function may be multi-peaked, the gradient descent algorithm may fall into a local minimum during training. In this case, the learning rate can be suddenly increased to jump out of the local minimum and find a path to the global minimum.

2.2.5. Replacing the Backbone Network with MobileNetV3

With the hot development of DL, various new CNN models are proposed. The effect of DL network models applied in image processing becomes more and more desirable. However, at the same time, it brings about the increasing size and complexity of neural networks, and the hardware requirements needed to train the network models are gradually increasing. Often, the DL neural network models can only be run on servers with strong computing power. Due to hardware limitations, it is difficult for mobile devices to run complex DL network models. Lightweight network models, such as SqueezeNet, ShuffleNet, NasNet, MnasNet, and MobileNet, have been proposed in the industry. These lightweight models make it possible to run DL network models in mobile terminals and embedded devices. Among them, the MobileNet neural network has a large advantage over lightweight neural networks due to its small size, low computational cost, and high accuracy. Therefore, in this paper, we choose to replace the backbone network CSPDarknet53 in YOLOv4 with the lightweight MobileNetV3 network to reduce the model size.

MobileNetV3 is the latest version of the MobileNet network, which uses a special bneck structure in its network structure. The bneck structure is shown in Figure 6. The bneck structure incorporates the inverted residual with a linear bottleneck of MobileNetV2. It first performs a 1×1 convolution on the input feature map to increase the dimension so that the number of channels is expanded and it has residual edges. The bneck structure also incorporates MobileNetV1's depthwise separable convolutions. It is a key factor in the lightweighting of the MobileNetV1 network by saving 12 times the computational cost compared to regular convolution operations. The bneck structure also introduces a lightweight attention mechanism module to adjust the weights of each channel. This module includes Squeeze and Excitation. Squeeze obtains the global compressed feature vector of the current feature map by global average pooling over the incoming feature map. Excitation obtains the weights of each channel in the feature map by two layers of full connectivity. The weighted feature map is used as the input to the next layer of the network. The bneck structure uses the H-Swish (hard version of swish) activation function instead of the Swish activation function, reducing the required computational effort. The swish activation function is shown in Equation (7).

$$Swish(x) = x \cdot sigmoid(\beta x) \tag{7}$$

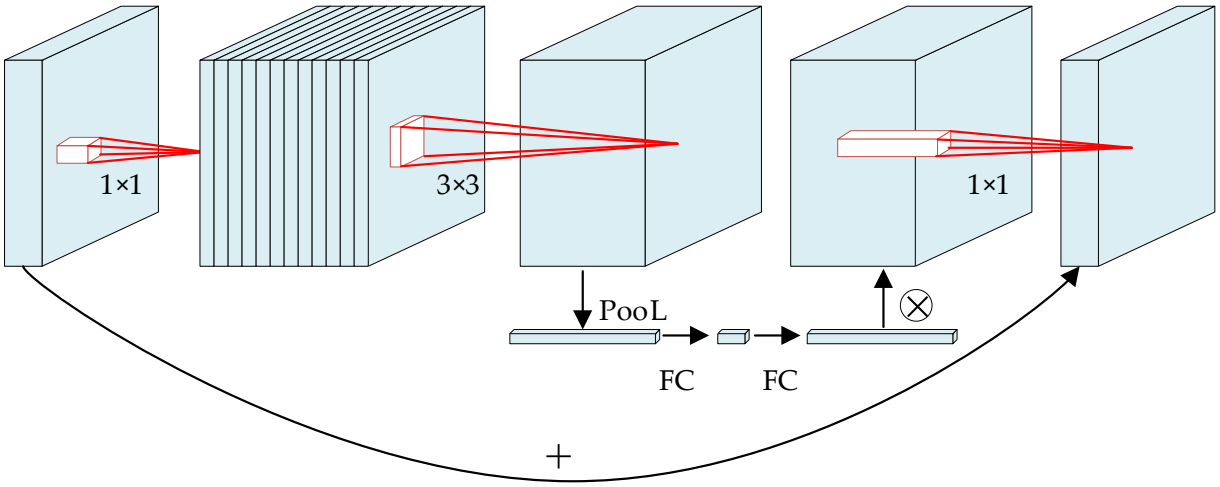


Figure 6. Structure diagram of bneck. PooL represents the global average pooling; FC represents a fully connected network.

The Sigmoid function is shown in Equation (8). β is either a constant or a trainable parameter. The Swish(x) function is a nonlinear interpolation function between the linear function and the ReLU function, while the parameter β controls the degree of interpolation.

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

The H-Swish activation function is shown in Equation (9).

$$\mathrm{H-Swish}(x) = \frac{\mathrm{ReLU6}(x+3)}{6} \tag{9}$$

The ReLU6 function is shown in Equation (10).

$$\mathrm{ReLU6}(x) = \min(6, \max(0, x)) \tag{10}$$

2.2.6. Implementation of CBAM Attention Mechanism

The attention mechanism module is inserted after two upsampling and two downsamplings in the PANet structure to enhance the feature extraction of the input feature map. The role of the attention mechanism module is to adaptively make CNN pay attention to important features. In general, it can be divided into CAM and SAM. The CBAM proposed by Woo et al. [34] is an excellent combination of the CAM module and the SAM module, which can achieve better results, and its structure is shown in Figure 7.

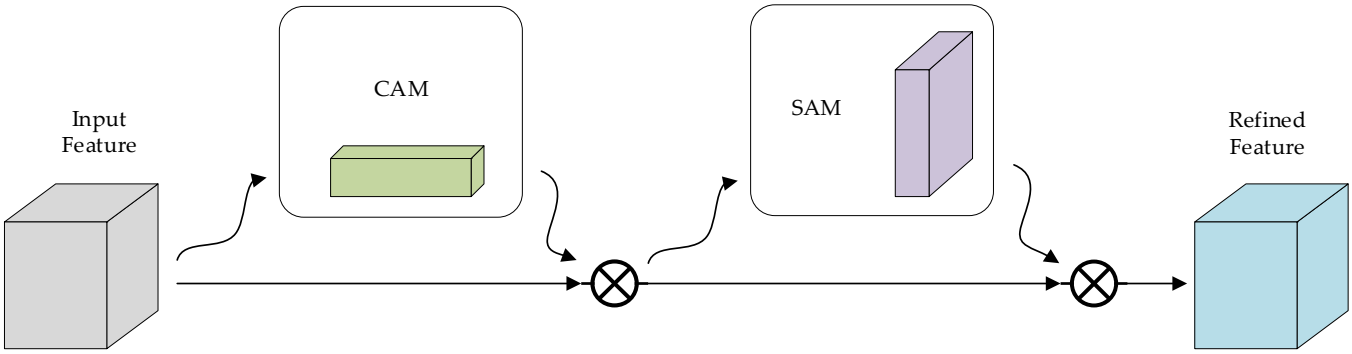


Figure 7. Structure diagram of CBAM. CAM represents channel attention mechanism module; SAM represents spatial attention mechanism module.

The first half of the CBAM structure is the CAM module, whose structure is shown in Figure 8. The CAM structure takes the incoming feature map F (C × H × W) in parallel through global maximum pooling and global average pooling based on the feature map height (H) and feature map width (W) to obtain two C × 1 × 1 feature maps. The obtained feature map is sent to the shared completely connected layer for processing, first compressing the number of channels as C/r (r is the reduction rate) and the activation function as Relu. Then, it expands the number of channels back to C. The two activated results are obtained, and the sum operation and sigmoid operation are performed, at which time the weight M_C of each channel of the input features is obtained. This weight is then multiplied with the corresponding elements of the input features to obtain the input feature map F'(C × H × W) of SAM, and the channel attention operation is completed.

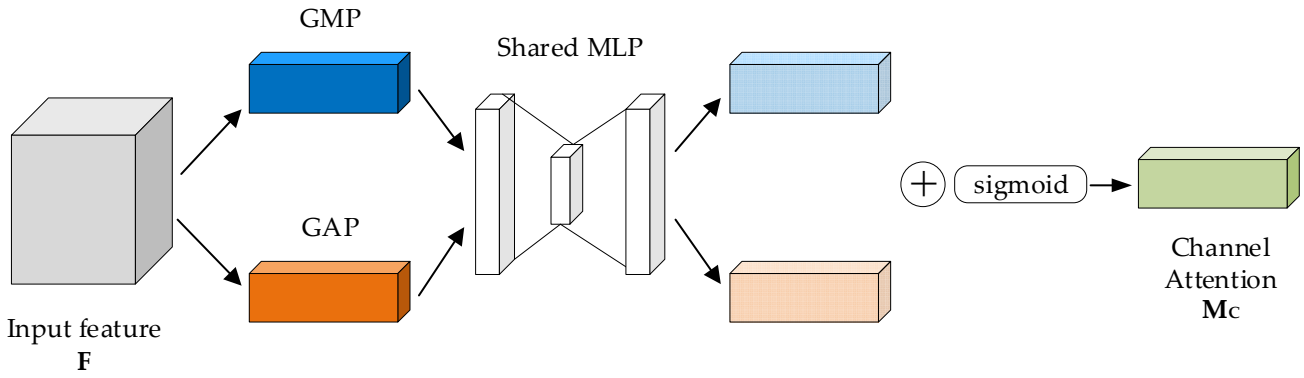


Figure 8. Structure diagram of CAM. GMP represents the global max pooling; GAP represents the global average pooling; The shared MLP is composed of multi-layer perceptron with one hidden layer.

The second half of CBAM is the SAM module. It takes F'(C × H × W) through a global maximum pooling and global average pooling based on the number of channels to obtain two H × W × 1 feature maps. The two feature maps are channel-spliced and then subjected to a 7 × 7 convolution operation to reduce the dimension to H × W × 1. Finally, the weight (M_S) of each channel of the input feature is obtained by sigmoid. Then, it multiplies M_S with the corresponding elements of the input features F' to obtain the

final feature map F'' (C × H × W) and complete the spatial attention operation. The SAM structure is shown in Figure 9.

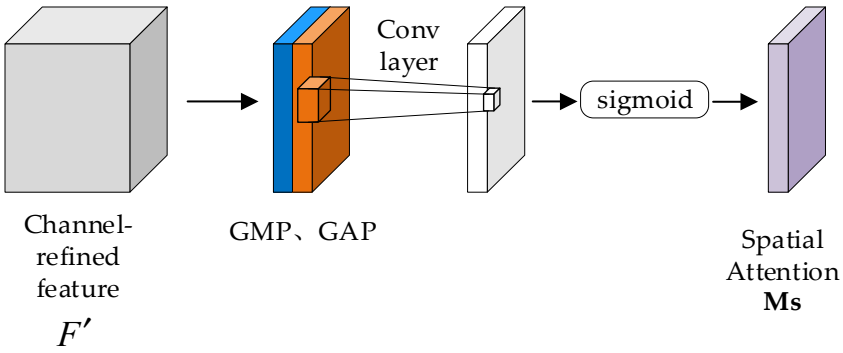


Figure 9. Structure diagram of SAM. GMP represents the global max pooling; GAP represents the global average pooling.

2.2.7. MC-YOLOv4 Network Model

The network structure model of the improved YOLOv4 (MC-YOLOv4) is shown in Figure 10.

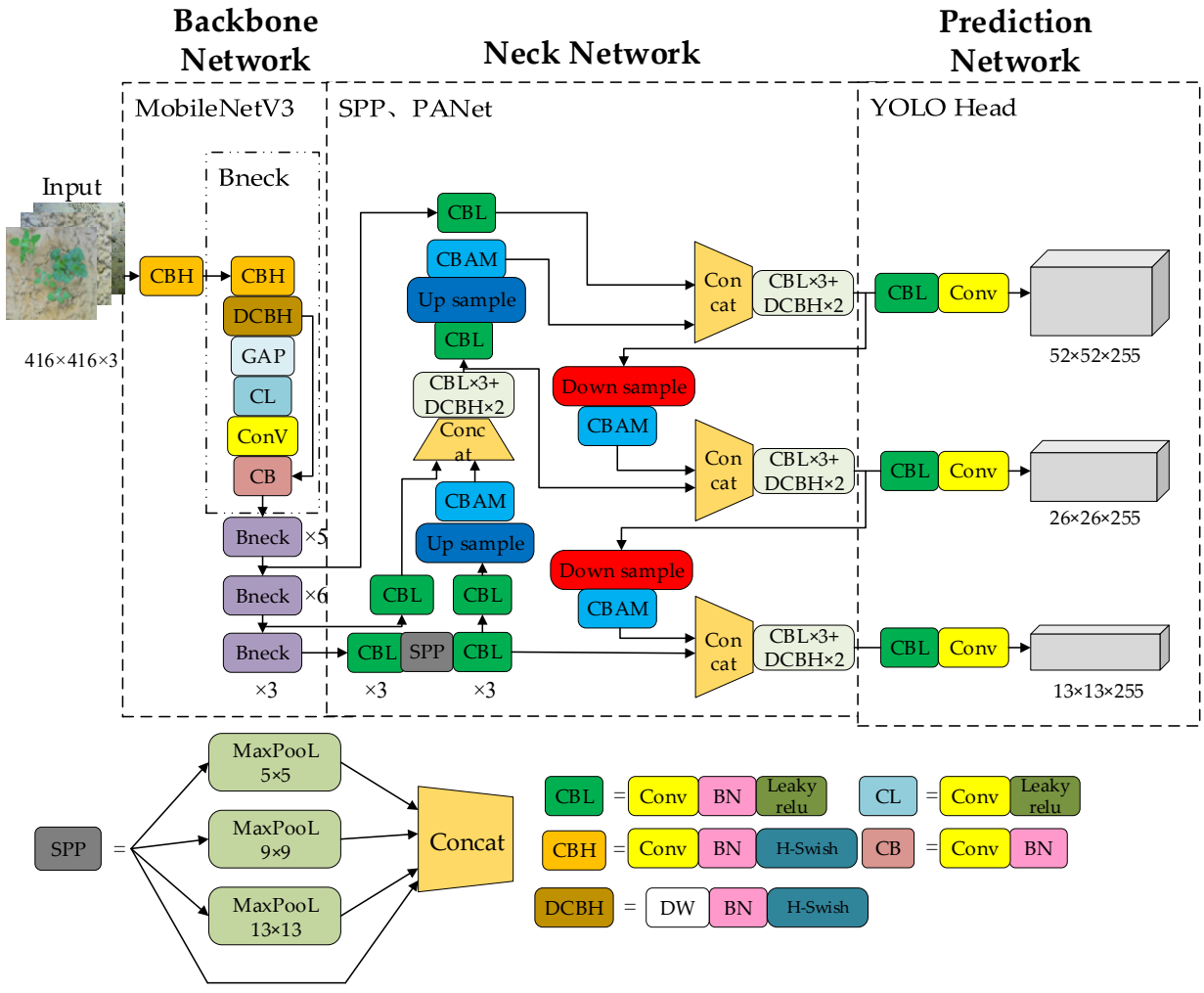


Figure 10. MC-YOLOv4 network structure model. Conv represents the convolution; BN represents the batch normalization; DW represents the depthwise separable convolution; GAP represents the global average pooling; Up sample represents the upsampling; Down sample represents the downsampling;

MaxPooL represents the maximum pooling; Concat represents a feature fusion method of adding the number of channels; CBAM represents the convolutional block attention module; Leaky relu and H-Swish represents two kinds of activation function.

2.2.8. K-Means++ Clustering Algorithm

The choice of anchor box is one of the important prerequisites for accurate target prediction. The anchor box used in the original YOLOv4 model was obtained based on the COCO dataset using a typical K-means clustering algorithm. However, the COCO dataset contains 80 classes, and the anchor box sizes of different classes vary. So, the anchor box sizes of the YOLOv4 model also vary considerably, and the K-means clustering algorithm tends to fall into local optima. In this paper, for the potato and weed datasets containing only two classes, most anchor box sizes should be on the small side. In order to make the MC-YOLOv4 target detection network more accurate in predicting the location of the target to be detected, this study used the K-means++ clustering algorithm instead of the K-means clustering algorithm to re-cluster the already labeled potato and weed dataset. The K-means++ clustering algorithm flows as follows:

(1) Enter the set N of the width and height of the potatoes and weeds in the dataset and the number of clustering centers k.
(2) A point is randomly selected from the input set of data points N as the first cluster center O_1.
(3) Calculate the distance $d(x)$ from each of the remaining points x in the set N to its nearest cluster center O_X; the more distant the point x becomes, the next cluster center has a greater probability $P(x)$. Repeat this step until the k clustering centers are found.
(4) Calculate the distance $d(x)$ from all the points in the set N to each of the k clustering centers and classify the point into the category of the clustering center with the smallest distance. For the clustering results, the center Ci of each cluster category is recalculated.
(5) Repeat step (4) until each clustering category center Ci no longer changes, and the output k clustering center results are O_1, O_2, $O_3 \cdots O_{k-1}$, O_k.

The final nine anchor boxes were obtained as (8,17), (10,25), (15,30), (19,40), (22,55), (28,48), (33,65), (42,89), and (68,139), respectively, and the updated anchor boxes were beneficial to improve the accuracy of weed identification in the potato field. Figure 11 shows the visualization of the K-means clustering algorithm.

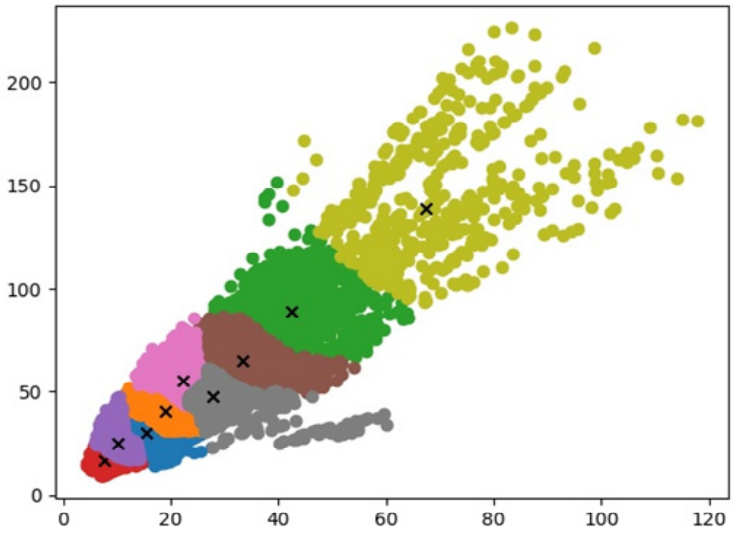


Figure 11. The visualization of K-means clustering algorithm.

3. Results and Analysis

This section mainly introduces two aspects. First, we present the training results and detection effects of each model on the dataset. Secondly, we perform a detailed analysis of the different models with suitable evaluation metrics, 3.1. and test the environment and parameter setting. Table 2 shows the training and test environment configuration.

Table 2. Training and test environment configuration table.

Configuration	Parameter
Operating System	Windows10 (Microsoft, Redmond, WA, USA)
CPU	Intel Core i5-12400F CPU@4.4 GHz (Intel, Santa Clara, CA, USA)
GPU	GeForce RTX 1080Ti 11G (GIGABYTE, New Taipei City, Taiwan)
Running Memory	16G
Accelerate Environment	CUDA11.0 CuDNN7.6.5
Pytorch	1.7.1

For the MC-YOLOv4 model, the Batchsize during training was set to 32, the maximum learning rate was set to 0.001, the momentum was set to 0.937, the Adam optimizer was used for optimization, the Focal loss was used to balance the positive and negative samples, and the positive and negative sample balance parameter was set to 0.25.

3.1. Model Performance Evaluation Indices

In this study, the average precision AP, mean average precision mAP, single image detection speed, and a number of model parameters were used as evaluation indexes. AP is related to the P (precision) and R (recall) of the model, and the formulas for calculating P, R, AP, and mAP are shown in Equations (11)–(14).

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\% \tag{11}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\% \tag{12}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{13}$$

$$mAP = \frac{\int_0^1 P(R)\mathrm{d}R}{N} \tag{14}$$

where TP is the number of correct model detections, FP is the number of model detection errors and target classification errors, FN is the number of model misses, and N is the number of categories. In this paper, the threshold was set to 0.7 and only when the IoU between the prediction box and the ground truth box exceeded 0.7, was the prediction box considered a positive sample, otherwise, it was a negative sample. In general, researchers usually set the threshold of the model to 0.5, which will result in higher values of the evaluation metrics of the model, but this also made the model's effectiveness in training and practice much different. Given that the methods presented in this study may be subsequently applied to weed control systems in crop fields, it is desirable to evaluate the models of this paper with a higher standard.

3.2. Training Results of MC-YOLOv4 Model

The MC-YOLOv4 model was used for the training and validation of the potato and weed datasets. The results of the loss functions for the training and validation sets during the training process are shown in Figure 12.

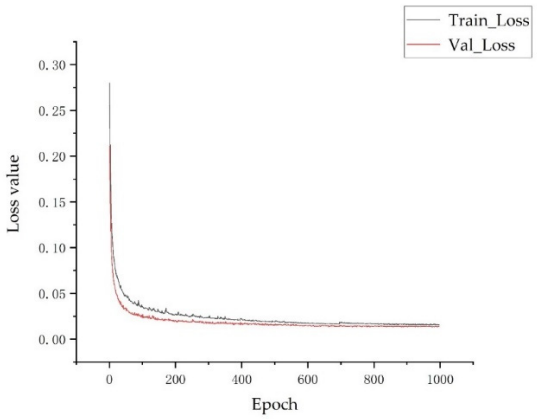


Figure 12. The loss value of MC-YOLOv4.

As can be seen from Figure 12, the MC-YOLOv4 network model decreased the val_loss and train_loss as the number of epochs increased. After 800 epochs, the train_loss and val_loss leveled off, and the model converged due to the comprehensive preprocessing of the dataset, and the model did not overfit, which is very rare.

3.3. Performance Comparison of Different Attention Mechanisms

In order to verify the advantage of the CBAM attention mechanism module used in this study, the CBAM attention mechanism module was replaced by the SE attention mechanism module and the ECA attention mechanism module in the same location of the network for separate experiments. The experimental results are shown in Table 3.

Table 3. Performance comparison of the three attention mechanism modules.

Models	Average Precision/%		mAP/%	Time/ms	Parameters/$\times 10^6$	Volume/MB
	Weed	Potato				
Base	95.47	98.41	96.94	11.7	11.32	53.7
Base-SE	96.85	98.36	97.60	11.9	11.37	54.6
Base-ECA	97.22	98.46	97.84	11.8	11.33	53.7
Base-CBAM(MC-YOLOv4)	98.63	98.42	98.52	12.4	11.53	55.4

Note: Base represents the MC-YOLOv4 model without the attention mechanism.

As shown in Table 3, all the metrics of the model changed after adding different attention mechanism modules, but all of them were improved compared with the original model. The model with the CBAM attention mechanism module that performed the best increased the mAP value by 1.58% compared to the original model and by 0.92% and 0.68% compared to the models with SE and ECA, respectively. This indicates that the CBAM attention mechanism module is most suitable for this study. It can be seen that the main improvement with the inclusion of the attention mechanism module is the accuracy of weed recognition. SENet adds an attention mechanism in the channel dimension and the key operations are squeeze and excitation. After processing by SENet, the importance of each feature channel becomes different, so the neural network focuses on certain channels with large weight values. ECANet can be seen as an improvement of SENet, which also increases the attention mechanism for channel dimensionality. It removes the fully connected layer, avoids dimensionality reduction, and accomplishes cross-channel information interaction through one-dimensional convolution. The size of the convolution kernel adaptively changes through the function, allowing more cross-channel interaction for the layers with a larger number of channels. The CBAM attention mechanism module starts from two scopes: channel and spatial. It introduces two analysis dimensions spatial attention, and channel attention, to realize a sequential attention structure from the channel to space. Spatial attention allows the neural network to pay more attention to the pixel regions in the image

that play a decisive role in classification and ignore irrelevant regions. Channel attention is used to deal with the assignment relationship of the feature map channels. Simultaneous attention allocation to both dimensions enhances the effect of the attention mechanism on the model performance. Of course, relative to the original model, the addition of the attention mechanism module increases the number of total parameters and the weight size of the model. However, the detection speed is not affected too much, and it can fully meet the requirements of the field weeding system.

In order to visually compare the effect of adding the attention mechanism to the model, the effect of adding the three attention mechanisms is visually analyzed in this study. Figure 13 shows three sets of visual heatmaps in the attention mechanism. Figure 13 shows that by adding the attention mechanism module, the focus of the model on the classification target is enhanced, and in particular, the CBAM attention mechanism module achieves the best results.

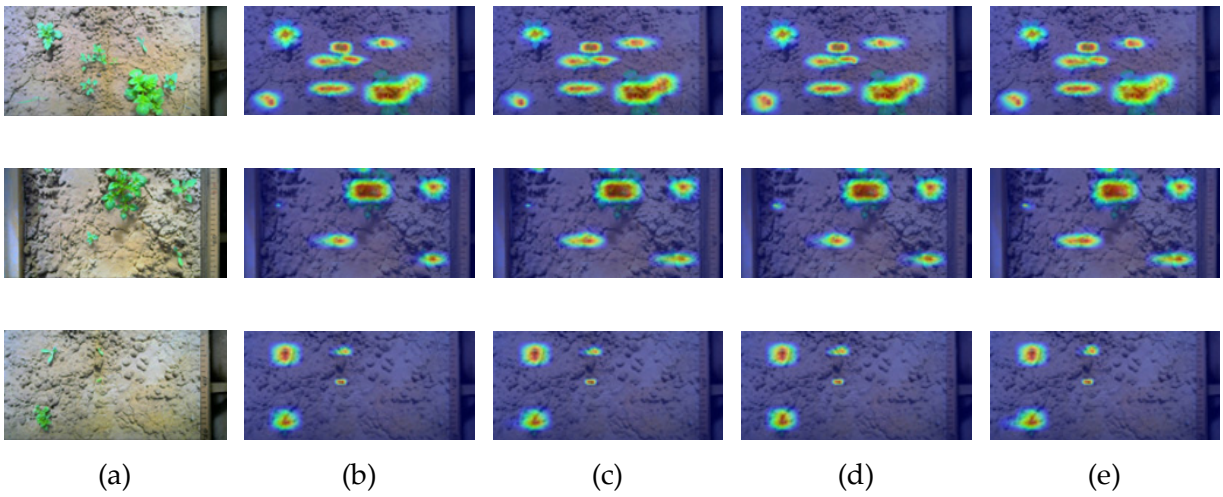


Figure 13. Three sets of visual heatmaps of the attention mechanism. (**a**) Original images representing three groups of potatoes and weeds, (**b**) represents the results of the heat map using the base model, (**c**) represents the heat map results using the base model and the SENet attention mechanism module, (**d**) represents the heat map results using the base model and the ECANet attention mechanism module, (**e**) represents the heat map results using the base model and the CBAM attention mechanism module.

3.4. Comparison of Performance with Different Network Models

The MC-YOLOv4 model, the original YOLOv4 model, the YOLOv4-tiny model, the Faster R-CNN model, the YOLOv5 l model, the SSD (MobilenetV2) model, the SE-YOLOv5x model, the EM-YOLOv4-Tiny model, and the AD-YOLOv3 model were iterated 1000 times in this paper, and Table 4 shows the comparison of the training results of the above nine models. Compared with the original YOLOv4 model, the MC-YOLOv4 model in this paper improves the mAP by 3.2 percentage points, reduces the single image detection speed by 18.7 ms, reduces the weight size by 188.6 MB, and greatly reduces the number of network parameters. Compared to the Faster R-CNN model, the MC-YOLOv4 model has 14.36 percentage points higher mAP, 59.7ms less detection time for a single image, 52.6MB less weight size, and a much smaller number of parameters than Faster R-CNN. Compared to the SSD model, which replaces the backbone network with MobilenetV2, the MC-YOLOv4 model in this paper has 19.86 percentage points higher mAP, 5.6 ms higher single image detection time, 40.6MB larger weight file size, and a larger number of network parameters than the SSD model. Compared to the YOLOv4-tiny model, the MC-YOLOv4 model in this paper has 4.48 percentage points higher mAP, 6.14 ms higher single image detection speed, 30MB higher weight file size, and a larger number of network parameters than the YOLOv4-tiny model. Compared to the YOLOv5 l model, the MC-YOLOv4 model in this paper has 0.06 percentage points higher mAP, 16.8 ms lower single image detection

speed, 120.6 MB lower weight file size, and a much smaller number of network parameters than the YOLOv5 l model.

Table 4. Training results of different models.

Models	Average Precision/%		mAP/%	Time/ms	Parameters/$\times 10^6$	Volume/MB
	Weed	Potato				
YOLOv4	93.29	97.35	95.32	31.1	63.9	244
YOLOv4-tiny	90.37	97.71	94.04	6.26	6	22.4
Faster R-CNN	72.12	96.20	84.16	72.1	28.3	108
YOLOv5 l	98.64	98.28	98.46	29.2	46.5	176
SSD(MobilenetV2)	77.36	79.96	78.66	6.8	6.2	14.8
SE-YOLOv5x	98.62	98.90	98.76	50.5	86.8	330
EM-YOLOv4-Tiny	91.24	97.44	94.34	9.8	6.8	28.8
AD-YOLOv3	90.46	96.40	93.43	30.8	62.3	236.6
MC-YOLOv4	98.63	98.42	98.52	12.4	11.53	55.4

After comparison, we found that the MC-YOLOv4 model performed very well in terms of detection accuracy and detection speed; especially in terms of detection accuracy, the mAP value of MC-YOLOv4 was higher than all the other models. Although inferior to the lightweight SSD and YOLOv4-tiny in terms of the number of network parameters and detection speed, both this SSD and YOLOv4-tiny performed poorly in terms of detection accuracy and may not be able to meet the requirements of field weeding systems. Compared to YOLOv5 and the original YOLOv4, which have a large number of network parameters, MC-YOLOv4 not only has the advantage of accuracy but also has the advantage of being lightweight, which is suitable for the requirements of field weeding systems.

3.5. Ablation Experiment

The MC-YOLOv4 algorithm replaces the backbone network with MobilenetV3 based on the original YOLOv4 algorithm and introduces depthwise separable convolutions instead of a part of the traditional convolution in the PANet structure while inserting a CBAM convolutional attention mechanism module after upsampling and downsampling and using the K-means++ algorithm instead of the K-means algorithm for anchor frame clustering. In order to be able to analyze more obviously the impact of the improved YOLOv4 algorithm (MC-YOLOv4) proposed on the original YOLOv4 algorithm, eight groups of experiments were conducted, respectively, and their experimental results are shown in Table 5.

As can be seen from Table 5, after adding the CBAM attention mechanism module, the network model paid more attention to the target to be detected, and the accuracy of the model improved for both potatoes and weeds, especially for the weeds, but it also increased the computational load of the model and increased the detection time. After replacing the backbone network and a part of the traditional convolution, the size and computation of the model were greatly reduced, which led to a significant improvement in the detection speed of the model but also a slight decrease in the detection accuracy of the weeds. After re-clustering the anchor boxes with the K-means++ algorithm, the anchor boxes were made to fit the data set more closely, and therefore, the accuracy was greatly improved for both the potato and weed targets.

3.6. Experiments with Different Models for Weed Detection in Potato Fields

In order to test the performance of the trained models for weed detection in the potato fields under different conditions, several images of different types of weeds in potato fields were taken again, and Figure 14 shows the detection results of the trained models. The YOLOv4 model had one false detection and four missed detections in four sets of tests. The YOLOv4-tiny model had two false detections and five missed detections in four sets of tests. The YOLOv5 l model had no false detection and four missed detections in four sets of

tests. The Faster R-CNN model had three false detections and seven missed detections in four sets of tests. The SSD model had three false detections and twelve missed detections in four sets of tests. The MC-YOLOv4 model had no cases of false detections or missed detections. The experimental results show that the improved model has good robustness and can solve the detection problem in the complex environment of the field.

Table 5. Results of ablation experiment.

Experiment	CBAM	MobileNetV3	K-Means++	Average Precision/%		mAP/%	Time/ms
				Weed	Potato		
1				93.29	97.35	95.32	31.1
2	√			96.23	97.51	96.87	31.7
3		√		93.26	97.78	95.52	11.7
4			√	95.72	97.36	96.54	31.1
5	√	√		96.18	97.63	96.90	12.4
6	√		√	98.66	97.64	98.15	31.8
7		√	√	95.47	98.41	96.94	11.7
8	√	√	√	98.63	98,42	98.52	12.4

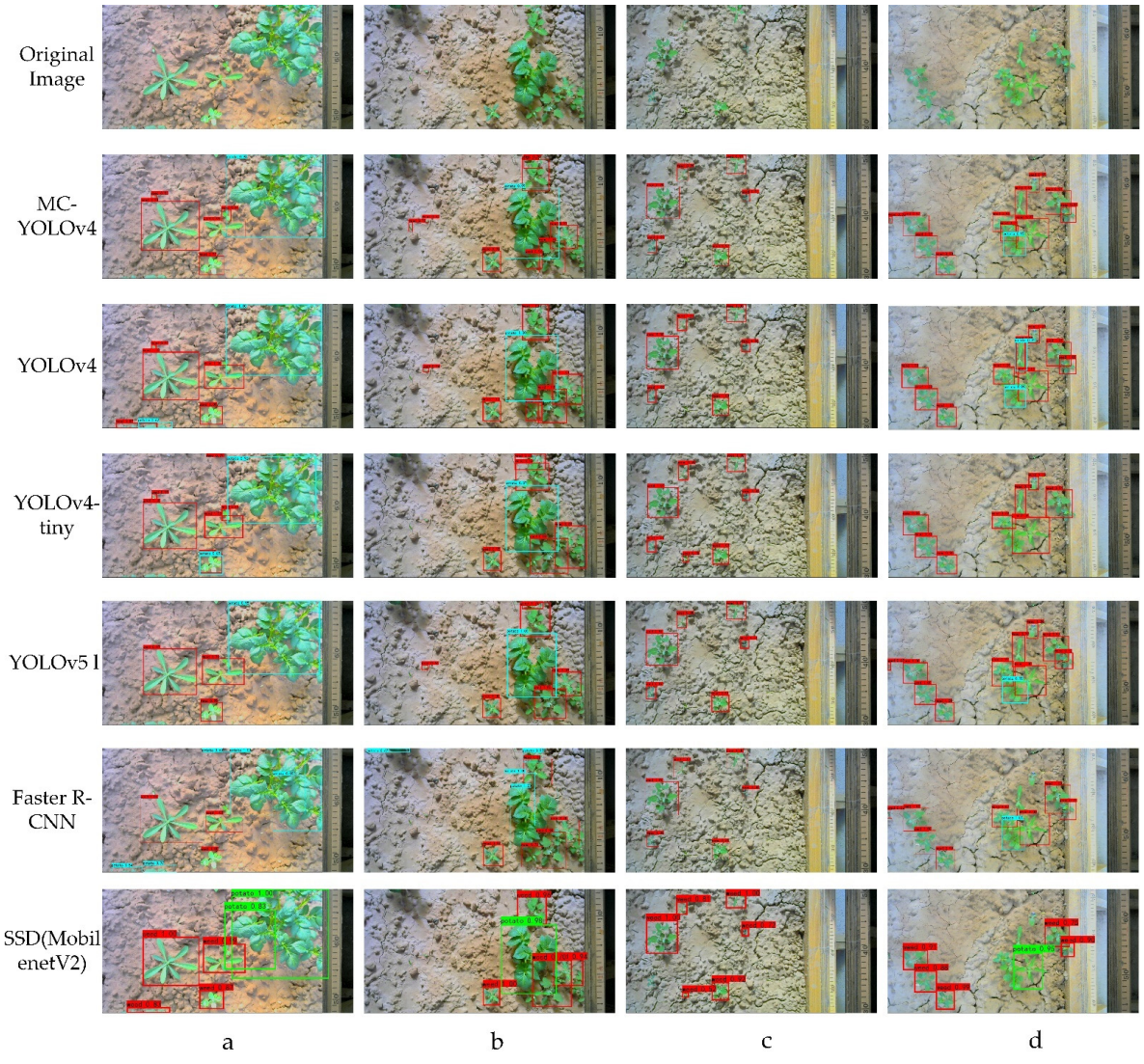


Figure 14. Results of weed detection in potato fields under different conditions with different models. (**a**) Represents images of potatoes and weeds without overlap, (**b**) Represents the image of the weeds

shaded by the potatoes, (**c**) Represents an image of multiple weeds, (**d**) Represents an image of potatoes obscured by weeds.

4. Discussion

4.1. Development of DLin Agriculture

With the demand for precision agriculture, how to detect plant diseases, orchard pickings, and field weed detection faster and more accurately has become a research direction for many scholars. DL-based object detection techniques were applied to weed detection in potato fields, which achieved very good results. DL is an area that has developed in recent years [35,36] and has been widely used for plant disease detection, orchard picking, and weed identification in the field. Peng et al. [37] selected ResNet50 to extract image features, introduced a feature pyramid network in the RPN network to generate target candidate boxes, optimized the structure of the Faster RCNN network, and the improved network could reach 95.5% for mAP0.5 and 87.9% for mAP0.75 when applied to weed identification in cotton fields. Zhao et al. [38] proposed an improved YOLOv3-based fruit recognition method to enable a picking robot to perform fast and accurate apple ripeness recognition under complex environments, such as different light, overlapping shade, and a large field of view in orchards. It combined the residual module in the DarkNet53 network with CSPNet, added the SPP module to the original network, and used Soft NMS instead of NMS. The improved YOLOv3 model was applied to the detection set of the apples with the mAP value of 96.3%. Wen et al. [39], in order to solve the problem of inaccurate detection in dense diseases and small area diseases of Panax pseudoginseng leaves, used an attention feature pyramid instead of the original feature pyramid of YOLOv3 to solve the interference problem of the feature fusion process, and also used Dual Bottleneck to filter the attention feature It also used Dual Bottleneck to filter the features extracted by the attention feature pyramid to enhance the specificity of the features and improve the robustness of the algorithm, and finally applied this to the detection of six types of Panax notoginseng leaf diseases with the mAP value of 80.6%. The development of DL depends to some extent on the optimization of detection algorithms and the updates of computer hardware such as GPU.

4.2. Limitations

The MC-YOLOv4 algorithm proposed in this study performs particularly successfully in weed identification in potato fields. Not only are the recognition accuracy and recognition speed considerably improved, but also the number of model parameters is greatly reduced, but there are still some limitations. First, since the images of the potatoes and weeds used in this study were obtained at the experimental site, the environment of the experimental field lacks some disturbing factors compared to a real potato field. Second, the weed species in this study are not comprehensive enough, and weed traits can vary from region to region due to climate and other factors. Future research will collect potato and weed images from more regions to expand our dataset. Finally, the MC-YOLOv4 model was not applied to weed identification in other crop fields due to time and field constraints. For future research, we will apply the MC-YOLOv4 model to weed identification in different crop fields to further improve our model.

5. Conclusions

There are no studies on weed detection in potato fields. Other existing models have the following two problems: (1) although the use of complex network models for weed detection in crops improves recognition accuracy, the recognition speed cannot meet the real-time requirements due to the large size of its network models; (2) the detection speed is improved by simplifying the network models, but frequently, the recognition accuracy cannot meet the requirements. In view of the above issues, in order to obtain a weed detection method in potato fields with both recognition accuracy and recognition speed,

we choose to improve on the YOLOv4 algorithm in this paper. The proposed algorithm replaces the backbone network CSPDarknet53 in the YOLOv4 network structure with the lightweight MobileNetV3 network. We introduce Depthwise separable convolutions instead of partial traditional convolutions in the PANet, which reduces the computational cost of the model and speeds up its detection. The K-means++ clustering algorithm is used instead of the K-means clustering algorithm to update the anchor box information of the model so that the anchor boxes are more suitable for the datasets in this study. Various image processing methods, such as CLAHE, MSR, SSR, and gamma, are used to increase the robustness of the model, which eliminates the problem of overfitting. CIoU is used as the loss function, and the cosine annealing decay method is used to adjust the learning rate to make the model converge faster.

The results show that compared with the original YOLOv4 model, the MC-YOLOv4 model in this paper improves the mAP by 3.2 percentage points, reduces the single image detection speed by 18.7 ms, reduces the weight size by 188.6 MB, and greatly reduces the number of network parameters. The results indicate that the improvement of this study has been extremely successful, not only in considerably reducing the computational cost of the YOLOv4 model but also in greatly increasing its detection accuracy and speed. Compared with the YOLOv4-tiny model, Faster R-CNN model, YOLOv5 l model and SSD(MobilenetV2) model, the MC-YOLOv4 model all have a higher recognition accuracy. The MC-YOLOv4 model achieves the mAP of 98.52% for weed identification in potato fields, a detection speed of 12.49ms for a single image, and a weight file size of 55.4MB, and can be fully deployed in field-smart weed identification systems to meet accuracy and speed requirements.

This study presents an efficient weed identification model for potato fields, which also serves as a reference for efficient weed detection in other crop fields and provides theoretical and technical support for automated weed control. In future research, the MC-YOLOv4 model will be applied to weed control systems in the field of precision agriculture for trial improvement and optimization. In agriculture, DL algorithmic models, in combination with intelligent agricultural equipment, will definitely take agricultural development to a higher level.

Author Contributions: Conceptualization, J.Z. and G.T.; methodology, J.Z. and G.T.; data curation, J.Z., C.Q. and Q.L.; validation, J.Z., K.Z. and B.G.; writing—original draft preparation, J.Z. and G.T.; writing—review and editing, J.Z., C.Q. and Q.L.; funding acquisition, G.T.; visualization, J.Z. and G.T.; supervision, J.Z. and G.T.; project administration, J.Z. and G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Natural Science Foundation of China (31401291) and The 10th Batch of Changzhou Science and Technology Planning Projects (International Science and Technology cooperation/Hong Kong, Macao and Taiwan Science and Technology cooperation) (CZ20220010).

Acknowledgments: The author would like to thank all contributors to this study.

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

References

1. Zhang, K.; Wang, X.; Wang, Z. The danger of field weeds and control technology. *Mod. Agric.* **2011**, *9*, 40–41.
2. Llewellyn, R.; Ronning, D.; Clarke, M.; Mayfield, A.; Walker, S.; Ouzman, J. Impact of weeds in Australian grain production. *Grains Res. Dev.* **2016**. Available online: www.grdc.com.au/ImpactOfWeeds (accessed on 9 March 2016).
3. Li, Y.; Xu, X.; Guo, J. Weed identification in maize fields based on a full convolutional neural network approach. *Jiangsu Agric. Sci.* **2022**, *50*, 93–100.
4. Gardee, Y.; Singh, P.K.; Dubey, R.P.; Gupta, P.K. Assessment of yield and economic losses in agriculture due to weeds in India. *Crop Protection* **2018**, *107*, 12–18. [CrossRef]
5. Pérez-Ruiz, M.; Slaughter, D.; Gliever, C.; Upadhyaya, S. Automatic GPS-based intra-row weed knife control system for transplanted row crops. *Comput. Electron. Agric.* **2012**, *80*, 41–49. [CrossRef]
6. Tillett, N.; Hague, T.; Grundy, A.; Dedousis, A. Mechanical within-row weed control for transplanted crops using computer vision. *Biosyst. Eng.* **2008**, *99*, 171–178. [CrossRef]

7. Midtiby, H.S.; Mathiassen, S.K.; Andersson, K.J. Performance evaluation of a crop/weed discriminating microsprayer. *Comput. Electron. Agric.* **2011**, *77*, 35–40. [CrossRef]
8. Utstumo, T.; Urdal, F.; Brevik, A.; Dørum, J.; Netland, J.; Overskeid, Ø.; Berge, T.W.; Gravdahl, J.T. Robotic in-row weed control in vegetables. *Comput. Electron. Agric.* **2018**, *154*, 36–45. [CrossRef]
9. Søgaard, H.; Lund, I. Application Accuracy of a Machine Vision-controlled Robotic Micro-dosing System. *Biosyst. Eng.* **2007**, *96*, 315–322. [CrossRef]
10. Gu, B.; Liu, Q.; Tian, G.; Wang, H.; Li, H.; Xie, S. Recognizing and locating the trunk of a fruit tree using improved YOLOv3. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 122–129.
11. Zhang, M.; Ji, Y.; Li, S. Progress of Agricultural Machinery Navigation Technology. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 1–18.
12. Yuan, X.; Huang, E. Experience and Reference of Smart Agriculture Development in European Union. *World Agric.* **2022**, *5*, 27–36.
13. Xu, J.; Gu, B.; Tian, G. Review of agricultural IoT technology. *Artif. Intell. Agric.* **2022**, *6*, 10–22. [CrossRef]
14. Yu, K.; Jia, L.; Chen, Y. Deep Learning: Yesterday, Today, and Tomorrow. *J. Comput. Res. Dev.* **2013**, *50*, 1799–1804.
15. Zhang, H.; Zhou, H.; Zheng, J. Research Progress and Prospect in Plant Phenotyping Platform and Image Analysis Technology. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 1–17.
16. Elstone, L.; How, K.Y.; Brodie, S.; Ghazali, M.Z.; Heath, W.P.; Grieve, B. High Speed Crop and Weed Identification in Lettuce Fields for Precision Weeding. *Sensors* **2020**, *20*, 455. [CrossRef]
17. Pulido, C.; Solaque, L.; Velasco, N. Weed recognition by SVM texture feature classification in outdoor vegetable crop images. *Ing. E Investig.* **2017**, *37*, 68–74. [CrossRef]
18. Peteinatos, G.G.; Reichel, P.; Karouta, J.; Andújar, D.; Gerhards, R. Weed Identification in Maize, Sunflower, and Potatoes with the Aid of Convolutional Neural Networks. *Remote Sens.* **2020**, *12*, 4185. [CrossRef]
19. Ferreira, A.; Freitas, D.; Silva, G.; Pistori, H.; Folhes, M. Weed detection in soybean crops using ConvNets. *Comput. Electron. Agric.* **2017**, *143*, 314–324. [CrossRef]
20. Partel, V.; Kakarla, S.; Ampatzidis, Y. Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence. *Comput. Electron. Agric.* **2019**, *157*, 339–350. [CrossRef]
21. Bakhshipour, A.; Jafari, A. Evaluation of support vector machine and artificial neural networks in weed detection using shape features. *Comput. Electron. Agric.* **2018**, *145*, 153–160. [CrossRef]
22. Espejo-Garcia, B.; Malounas, I.; Vali, E.; Fountas, S. Testing the Suitability of Automated Machine Learning for Weeds Identification. *AI* **2021**, *2*, 34–47. [CrossRef]
23. Liu, M.; Gao, T.; Ma, Z.; Song, Z.; Li, F.; Yan, Y. Target Detection Model of Corn Weeds in Field Environment Based on MSRCR Algorithm and YOLOv4-tiny. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 246–255.
24. Zhang, J.; Su, W.; Zhang, H.; Peng, Y. SE-YOLOv5x: An Optimized Model Based on Transfer Learning and Visual Attention Mechanism for Identifying and Localizing Weeds and Vegetables. *Agronomy* **2022**, *12*, 2061. [CrossRef]
25. Wang, Q.; Cheng, M.; Huang, S.; Cai, Z.; Zhang, J.; Yuan, H. A deep learning approach incorporating YOLO v5 and attention mechanisms for field real-time detection of the invasive weed Solanum rostratum Dunal seedlings. *Comput. Electron. Agric.* **2022**, *199*, 107194. [CrossRef]
26. Dong, H.; Chen, X.; Sun, H.; Yao, L. Weed detection in vegetable field based on improved YOLOv4 and image processing. *J. Graph.* **2022**, *43*, 559–569.
27. Zhang, H.; Wang, Z.; Guo, Y.; Ma, Y.; Cao, W.; Chen, D.; Yang, S.; Gao, R. Weed Detection in Peanut Fields Based on Machine Vision. *Agriculture* **2022**, *12*, 1541. [CrossRef]
28. Tang, J. Weed identification based on K-means feature learning combined with convolutional neural network. *Comput. Electron. Agric.* **2017**, *135*, 63–70. [CrossRef]
29. Wang, Q.; Gu, W.; Cai, P. Detection Method of Double Side Breakage of Population Cotton Seed Based on Improved YOLOv4. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 389–397.
30. Wang, X.; Li, Y.; Yang, Z. Method of Clods and Stones from Impurified Potatoes Based on Improved YOLO v4 Algorithm. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 241–247.
31. Li, J.; Li, Y.; Niu, F. Diet Behavior Detection Method Based on YOLOv4. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 251–256.
32. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An advanced object detection network. *Assoc. Comput. Mach.* **2016**, *5*, 516–520.
33. 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), Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.
34. Woo, S.; Park, J.; Lee, J.Y. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
35. Hassan, S.A.; Rahim, T.; Shin, S.Y. An improved deep convolutional neural network-based autonomous road inspection scheme using unmanned aerial vehicles. *Electronics* **2021**, *10*, 2764. [CrossRef]
36. Roy, S.; Menapace, W.; Oei, S.; Luijten, B.; Fini, E.; Saltori, C.; Huijben, I.; Chennakeshava, N.; Mento, F.; Sentelli, A.; et al. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE Trans. Med. Imaging* **2020**, *39*, 2676–2687. [CrossRef]

37. Peng, M.; Xia, J.; Peng, H. Efficient recognition of cotton and weed in field based on Faster R-CNN by integrating FPN. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 202–209.
38. Zhao, H.; Qiao, Y.; Wang, H.; Yue, Y. Apple fruit recognition in complex orchard environment based on improved YOLOv3. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 127–135.
39. Wen, B.; Cao, R.; Yang, Q. Leaf disease for Panax notoginseng using an improved YOLOv3 algorithm. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 164–172.