agriculture

MDPI

Article

Maize Kernel Quality Detection Based on Improved Lightweight YOLOv7

Lili Yang [1,2], Chengman Liu [3], Changlong Wang [1] and Dongwei Wang [1,2,*]

[1] College of Mechanical and Electrical Engineering, Qingdao Agricultural University, Qingdao 266109, China; y_lili1980@qau.edu.cn (L.Y.); 20190200861@stu.qau.edu.cn (C.W.)
[2] Yellow River Delta Intelligent Agricultural Equipment Industry Academy, Dongying 257300, China
[3] College of Electrical and Control Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China; 20190200418@stu.qau.edu.cn
* Correspondence: 200701031@qau.edu.cn

Abstract: As an important cereal crop, maize is a versatile and multi-purpose crop, primarily used as a feed globally, but also is important as a food crop, and has other uses such as oil and industrial raw materials. Quality detection is an indispensable part of functional and usage classification, avoiding significant waste as well as increasing the added value of the product. The research on algorithms for real-time, accurate, and non-destructive identification and localization of corn kernels based on quality classification and equipped with non-destructive algorithms suitable for embedding in intelligent agricultural machinery systems is a key step in improving the effective utilization rate of maize kernels. The difference in maize kernel quality leads to significant differences in price and economic benefits. This algorithm reduced unnecessary waste caused by the low efficiency and accuracy of manual and mechanical detection. Image datasets of four kinds of maize kernel quality were established and each image contains a total of about 20 kernels of different quality randomly distributed. Based on the self-built dataset, the YOLOv7-tiny, as the backbone network, was used to design a maize kernel detection and recognition model named "YOLOv7-MEF". Firstly, the backbone feature layer of the algorithm was replaced by MobileNetV3 as the feature extraction backbone network. Secondly, ESE-Net was used to enhance feature extraction and obtain better generalization performance. Finally, the loss function was optimized and replaced with the Focal-EOIU loss function. The experiment showed that the improved algorithm achieved an accuracy of 98.94%, a recall of 96.42%, and a Frame Per Second (FPS) of 76.92 with a model size of 9.1 M. This algorithm greatly reduced the size of the model while ensuring high detection accuracy and has good real-time performance. It was suitable for deploying embedded track detection systems in agricultural machinery equipment, providing a powerful theoretical research method for efficient detection of corn kernel quality.

Keywords: quality detection; YOLOv7-tiny; MobileNetV3; ESE-Net; *Focal-EIoU* loss

Citation: Yang, L.; Liu, C.; Wang, C.; Wang, D. Maize Kernel Quality Detection Based on Improved Lightweight YOLOv7. *Agriculture* **2024**, *14*, 618. https://doi.org/10.3390/agriculture14040618

Academic Editor: Jiangbo Li

Received: 18 March 2024
Revised: 8 April 2024
Accepted: 10 April 2024
Published: 16 April 2024

1. Introduction

Maize is one of the most important cereal crops in the world [1] and is widely distributed in the United States, China, Brazil, and other countries. Global maize cultivation has reached 197 million hectares with an annual production of more than 1 billion tons [2]. Maize accounts for 40% of cereal production in Sub-Saharan Africa (SSA), where more than 80% is used as food [3]. Also, maize is an important energy source, an important feed source for livestock farming, and one of the indispensable raw materials in many industries [4]. The corn combine harvester is used to harvest and thresh maize when the water content is lower than 25% and then it is stored safely [5]. However, in the process of natural dehydration and mechanical harvesting in the field, maize will produce mildew, germination, and breakage, which affects the quality. In particular, aflatoxin contamination

of moldy maize can adversely affect global trade and health and cause economic losses to the country [6]. In order to improve the efficient utilization of maize kernels, it is necessary to identify and classify the maize kernels for different uses such as edible, oil extraction, animal feed, and industrial raw materials. The system performance of maize kernel batch sorting directly affects the efficiency and quality. Therefore, the design of an efficient, fast, non-destructive, and reliable quality sorting system is of great scientific and social importance to improve the automation and commercial value of maize harvesting.

In recent years, machine vision and deep learning technologies have been widely applied in different fields, and more and more scholars are applying them to various agricultural fields for crop identification and classification tasks [7,8]. Ju et al. [9] proposed a jujube defect classification model for small datasets based on a convolutional neural network and transfer learning. The original CNN model was improved by embedding the SE module and using a triple loss function and a central loss function instead of a softmax loss function. The designed SE-ResNet50-CL model optimized the fine-grained classification problem of jujube defect recognition, with a testing accuracy of 94.15%. Wang et al. [10] studied the design of an improved Faster R-CNN model for tomato ripeness detection. The model achieved better detection results in branch occlusion, fruit overlap, and lighting effects, with an Average Precision (mAP) of 96.14%. Ni et al. [11] improved AlexNet using classical networks such as VGGNet and GoogLeNet. The new structure was proposed to classify strawberry varieties of different qualities with an average accuracy of 95.75%. Huang et al. [12] designed a pipeline that completely followed the segmentation–classification procedure to classify soybean seeds. Image segmentation is performed by the deep learning method Mask R-CNN, and the classification stage is performed by a novel network of self-designed Soybean Network (SNet). The SNet model achieved a recognition accuracy of 96.2% with only 1.29 M parameters. Zhang et al. [13] proposed a high-throughput corn ear screening method based on a dual-pathway convolutional neural network. The dual-pathway convolutional neural network combined the advantages of VGG-16 and Resnet-50, which could greatly reduce the input parameters and improve the accuracy. The average classification accuracy of this algorithm reached 97.23%. Zhao et al. [14] proposed a wheat detection framework, WGNet. The accuracy of the proposed method for wheat kernels including germinated, scab, moldy, and normal kernels reached 97.0%, and it took 10 s to measure 2500 wheat kernels. Yang et al. [15] developed a method based on multispectral (MS) images combined with the improved YOLOv3 Tiny Detection module (MDDNet) to automatically detect and classify multiple types of defects in potatoes. The average accuracy of this model for potato defects was up to 90.26%, and the detection time for each MS image was about 75 ms. Kurtulmuş et al. [16] proposed, trained, and tested a computer vision system comparing three deep learning architectures, AlexNet, GoogleNet, and ResNet, used for individual recognition of about 4800 sunflower seeds. The GoogleNet algorithm achieved a classification accuracy of up to 95%. Jeyaraj PR et al. [17] developed a non-contact and cost-effective rice grading system based on accurate deep learning according to the appearance and characteristics of rice. Using AlexNet architecture, they obtained an average accuracy of 98.2% with 97.6% sensitivity and 96.4% specificity.

In conclusion, researchers at home and abroad have undertaken a series of studies on the detection of agricultural products and achieved good results [18–21], which provided a reference for the detection of maize kernel quality. For example, Bi et al. [22] combined deep learning with machine vision and used the basis of Swin Transformer to improve maize seed recognition. The Average Precision, recall, and F1 score of the model on the test set reached 96.53%, 96.46%, and 96.47%, respectively. Yang et al. [23] combined hyperspectral imaging (HSI) with sparse Autoencoder (SAE) and Convolutional neural Network (CNN) algorithms to perform grade detection of moldy maize kernels. The constructed SAE-CNN-SVM model had an accurate recognition rate of 99.47% and 98.94% in the training and test sets, respectively, with good recognition ability for the early detection of moldy maize kernels. Zhao et al. [24] explored the application of electromagnetic vibration and deep learning techniques in maize seed detection and sorting. The precision, recall, and F1

score of the improved Faster R-CNN model were increased by 3.73%, 3.55%, and 3.79%, respectively, and the false positive rate was reduced by 1.31% on average. Xu et al. [25] established a maize seed defect detection method based on a deep learning algorithm by Hyperspectral imaging (HSI) technology. This classification accuracy of the convolutional neural network architecture based on the attention classification mechanism (CNN-ATM) was more than 90%, and the sensitivity and specificity were 97.50% and 98.28%, respectively. Xu et al. [26] proposed an improved network P-ResNet for identifying maize seeds for transfer learning and obtained better detection results with loss maintained around 0.01. This model performed generalization experiments on the classification of Baoqiu, Shantong, Xinnuo, Liaoge, and Kuexian varieties, and the accuracy reached 99.74%, 99.68%, 99.68%, 99.61%, and 99.80%, respectively. Jiao et al. [27] proposed a method using μCT technology and R-YOLOv7-tiny for detecting corn endosperm cracks. An algorithm was developed to automatically extract crack information in the detection head of the R-YOLOv7 tiny model by improving the model. The detection speeds are 93.80%, 87.90%, 92.10%, 9.70 MB, and 67.11 fps, respectively. This method improves the overall crack detection and crack omission rates by 7.86% and 7.29%, respectively. Wei et al. [28] developed a novel modeling approach based on deep learning based on a backpropagation neural network–genetic algorithm model. The original spectrum was converted into a two-dimensional matrix to construct an improved convolutional neural network with dilated convolution to classify the maize kernels, and the accuracy reached 96.1%.

To sum up, most of the detection models for maize kernels were built by classical classification networks such as Faster R-CNN and ResNet at present. Relatively little research has been undertaken on the application of the newly introduced lightweight network. The sampling equipment for establishing the dataset was mostly expensive spectral imaging equipment. At the same time, most maize kernels used single-grain image acquisition or multi-grain regular image acquisition, which was not suitable for the field environment.

Thus, in response to the above issues in current research on maize kernel quality detection, this study proposed a detection algorithm based on the improved YOLOv7-tiny. The algorithm has the characteristics of low cost, high precision, real-time, and light weight, and it is suitable for embedded hardware equipment. The key findings and highlights of the specific goals of the research are summarized as follows:

- We established a low-cost data acquisition system. After passing through the corn combine harvester, the maize kernels are randomly distributed through electromagnetic vibration and sampled by ordinary RGB industrial cameras. Also, we established a standardized maize kernel quality dataset, including four categories: moldy, germinant, intact, and broken.
- A maize kernel quality detection model, YOLOv7-MEF, was developed. In this algorithm, MobileNetV3 was used to replace the original feature extraction backbone network, ESE-Net was integrated to enhance feature extraction, and *Focal-EIoU* was used to optimize the original loss function. The algorithm is made with high accuracy, fast detection speed, and small model size.
- The self-established maize kernel database was used to evaluate the model, and ablation experiments were carried out to verify the algorithm's recognition and location effect on low-cost sampling images, providing a theoretical basis for related research.

2. Materials and Methods

2.1. Materials

2.1.1. Dataset Acquisition

In this study, the research object was maize kernel sample Denghai W385. The original data images for the maize kernel quality detection experiment were taken by ourselves. The maize kernels were collected from the corn base in Caoxian County, Heze City, Shandong Province, China. After the maize kernels were harvested and threshed by a corn combine harvester, four kinds of quality seeds were collected: normal, broken, germinated, and

moldy. The RGB images of corn kernels were captured using a machine vision experimental rack combined with a MOKOSE c100 camera from Shenzhen Yunchuang Technology Co., Ltd. in Shenzhen, China, with an initial pixel resolution of 1920 × 1080. After comparison and selection, about 20,000 kernels of maize were selected for test sampling, with an average of about 5000 kernels per category. We used electromagnetic vibration to disperse maize kernels and used a MOKOSE c100 industrial camera with an RH-MVT3-900-1 fixed bracket on the transportation track for photography. The initial pixel setting of the camera was 1920 × 1080, and S-EYE 2.0 was used to realize the communication between the camera and the computer. Other adjustments of the camera, such as contrast, exposure, brightness, etc., should be set reasonably according to the test site. When shooting, we adjusted the camera angle to the horizontal position and adjusted the height to be 10 cm away from the track. In the same light source environment, the black track background was selected for image acquisition, and a total of 775 sample images were initially collected, and each image contained about 20 maize kernels.

2.1.2. Dataset Labeling

The 775 maize kernel images in the dataset were labeled by LabelImg v1.8.1, and each kernel in the images was boxed out by category using a horizontal rectangular box of a specific color to their own. LabelImg Software is an open-source tool compiled in the Python language to label the input image with a rectangular box to generate an XML file corresponding to the image name. The position of the rectangular box in the software is determined by the coordinates of the upper left corner and the lower right corner. Since the shape of each corn grain is different, it is easy for it to have adverse effects on the model training, so it is necessary to minimize the irrelevant factors entering the rectangular box. Each image is labeled with about 20 different locations of maize kernels, and all images are labeled to produce 16,226 labeled boxes. Intact quality maize kernels are labeled with blue rectangular boxes, indicating intact; moldy maize kernels are labeled with pink rectangular boxes, marking moldy; individual broken maize kernels should be marked with an orange rectangle box to mark broken; the ones that are sprouted are labeled with green rectangles, identifying germinant. The labels are shown in Figure 1.

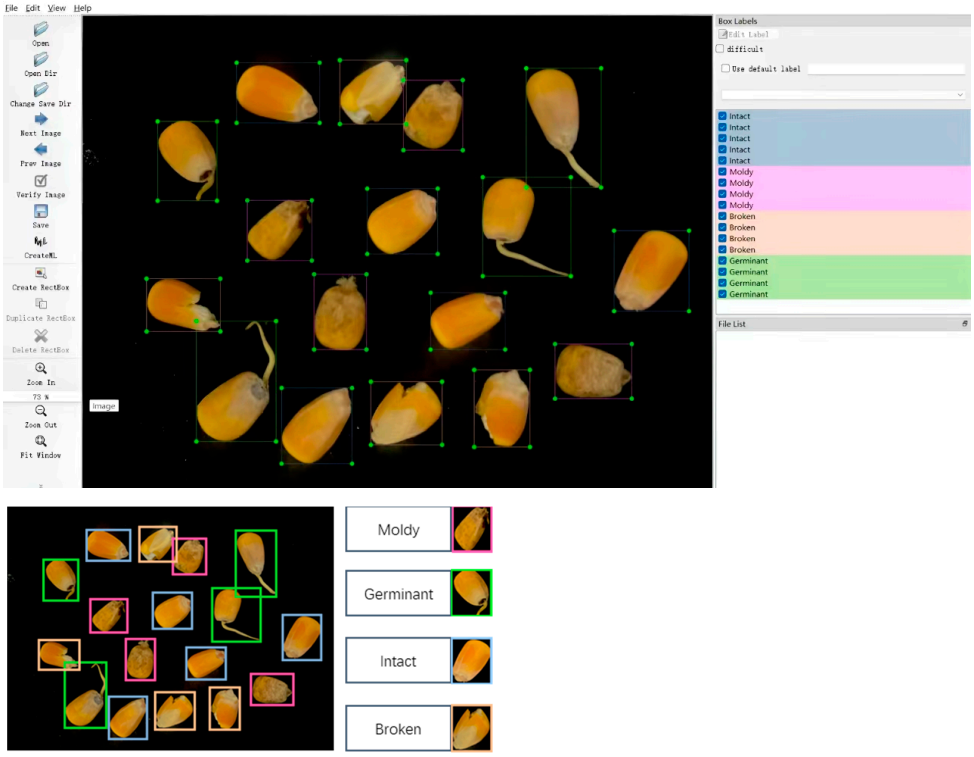
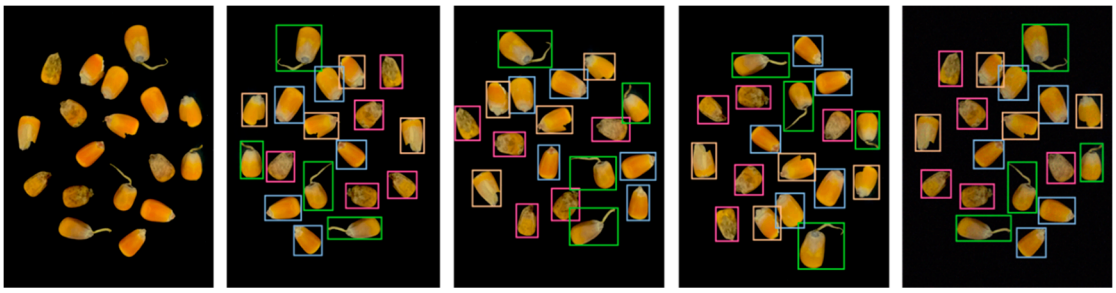


Figure 1. Example of maize kernel label image, including four categories.

2.1.3. Data Augmentation

In order to improve the training effect, reduce overfitting, better extract the features of maize kernel images, and improve the generalization ability of the model, this study used random data augmentation [29] to enhance the labeled images. Maize kernel images were augmented by Gaussian noise, translation, and angle inversion. The results of data augmentation are shown in Figure 2. In order to highlight the image features, further manual cropping and screening are carried out on the image after data augmentation. Finally, the number of images increased from 775 to 2873 after selection and the sample size increased from 16,186 to 64,744. The augmented dataset was divided into training set, test set, and validation set in the ratio of 8:1:1 for subsequent model training. The quantity distribution of label categories in the dataset is shown in Table 1.



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

Figure 2. Data augmentation example (**a**) original image; (**b**) horizontal flip image; (**c**) rotate 40° image; (**d**) vertical flip image; and (**e**) Gaussian noise image.

Table 1. Category and quantity of maize quality dataset.

Category	Number	Training Set	Test Set	Validation Set
Intact	15,684	12,548	1568	1568
Moldy	16,104	12,884	1612	1612
Broken	16,660	13,328	1664	1664
Germinant	16,296	13,036	1628	1628
total	64,744	51,796	6472	6472

2.2. Training Environment and Methods

The models in this study were trained under the Windows 10 Professional operating system with Intel(R) Core(TM) i9-10920X CPU, RTX 2080 Ti GPU, 11 GB of graphics memory, and 64 GB of host memory. The Cuda version is 10.0, the Cudnn version is 8.4.0, the Python version is 3.9.7, and we use the Python deep learning framework. The Python version is 1.12.1, the Torchvision version is 0.13.1, and the Torchaudio version is 0.12.1.

In the training process, the input size of the image was set to 640 × 640 pixels, the batch size was set to 16, Adaptive Moment Estimation (Adam) was used as the optimizer, the learning rate was set to 0.01, the momentum was set to 0.937, the weight recession coefficient was set to 0.01, and the number of training rounds was set to 100.

2.3. Performance Indexes

In this study, *Precision*, *Recall*, mean Average Precision (*mAP*), and *F1 score* were used as the model evaluation indexes, along with the model parametric size to comprehensively evaluate the model performance. The *mAP* is the mean of the Average Precision (*AP*) and the Average Precision (*AP*) is the area of the P-R curve, where map@0.5 is an indicator for evaluating the accuracy of the model, which represents the *mAP* value when the IoU threshold is set at 50%. The FPS reflects the refresh rate of the model inference speed. The calculation formula for each index is as follows:

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$mAP = \frac{\sum AP}{K_{Class}} \quad (3)$$

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$FPS = \frac{Framenum}{Elapsed\ Time} \quad (5)$$

where TP denotes positive samples with positive model predictions, FP denotes negative samples with positive model predictions, FN denotes positive samples with negative model predictions, and TN denotes negative samples with negative model predictions. P_x is the precision of a specific class and K_{class} is the number of quality categories.

3. Results

3.1. Comparison of Models

At present, the mainstream target detection models are divided into two-stage models and one-stage models. The representative network models for two-stage target detection include R-CNN, SPP-Net, Faster R-CNN, and R-FCN. The one-stage target detection model extracts features directly in the network to predict the class and location of objects, which features reduced training time and model complexity of the network, is faster, and is more suitable for mobile deployment. The representative network models in one-stage include SSD, RetinaNet and YOLO series. According to the related literature, we selected Faster-RCNN, SSD, YOLOv5, and YOLOv7 series for comparison and selection.

YOLOv7-tiny [30] is an efficient, lightweight target detection algorithm with a more compact network architecture and an optimized training strategy based on YOLOv7. By reducing model size and computation, YOLOv7-tiny is more suitable for real-time operation on embedded devices and mobile terminals. In this study, different models were trained using the same self-made maize kernel dataset, and the evaluation indexes of each model are shown in Table 2. Experimental data showed that the YOLOv7-tiny model was more suitable as the backbone network for this study. Compared with other models, the YOLOv7-tiny model had the smallest size parameters, only 11.72 MB, making it easier for embedded deployment. Meanwhile, it also had the highest prediction accuracy for the training set, reaching 97.21%. Considering the Recall, map@0.5/%, and the size of parameters, it had the best performance. Figure 3 shows the Precision and Recall curves of different models after training, and it was seen from the figure that YOLOv7-tiny was relatively better trained on the maize kernel dataset. Therefore, YOLOv7-tiny was chosen as the backbone network for further study.

Table 2. Comparison of detection performance of different models.

Model	Precision	Recall	map@0.5	Model Size/M
Faster-RCNN	88.51%	88.54%	86.56%	108.29
SSD	92.89%	92.66%	92.83%	92.13
YOLOv5	89.13%	91.3%	91.75%	27.14
YOLOv7	97.66%	91.93%	94.35%	73.38
YOLOv7x	98.83%	88.35%	96.62%	138.7
YOLOv7-tiny	97.21%	92.3%	94.95%	11.72

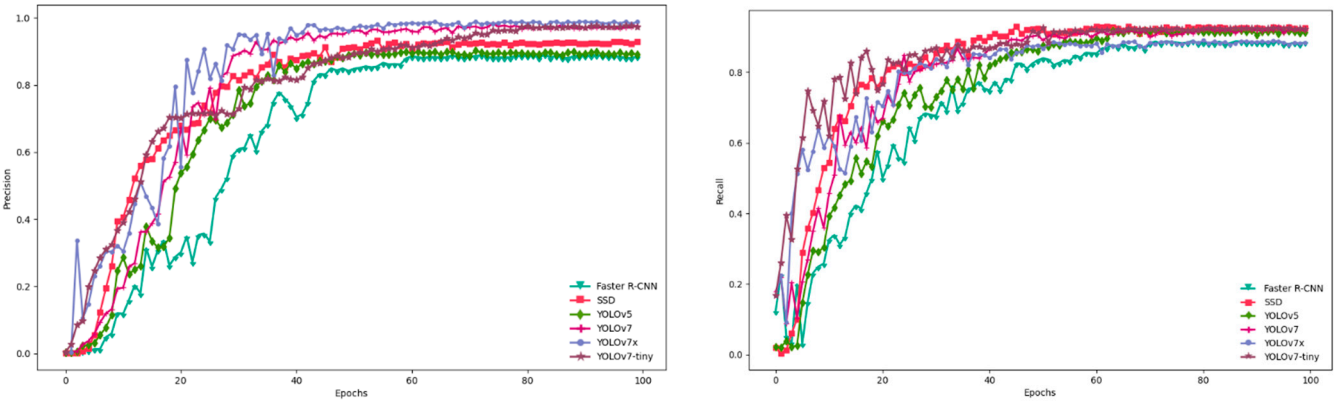


Figure 3. Precision and Recall curve of training set for different networks.

3.2. YOLOv7-Tiny Structure

YOLOv7-tiny is lightweight and adjusted on the basis of the YOLOv7 model, whose body consists of a backbone, neck, and head. In the backbone structure, YOLOv7-Tiny uses the Effective Long-range Aggregation Network (ELAN)structure to replace the E-ELAN structure in the original YOLOv7 model. ELAN is composed of multiple convolution layers, BN layers, and a Leaky ReLU Activation function. The feature maps output by the three CBLs of its main branch are fused with the feature maps output by the one CBL in the periphery through the Concat layer, and the final result is output through the final CBL. In the ELAN structure, in addition to the change in the number of channels of the two CBL modules at the front, the number of input channels in each of the remaining branches is consistent with the number of output channels, which can reduce memory access, reduce running time, and improve training efficiency. The backbone structure is shown in Figure 4.

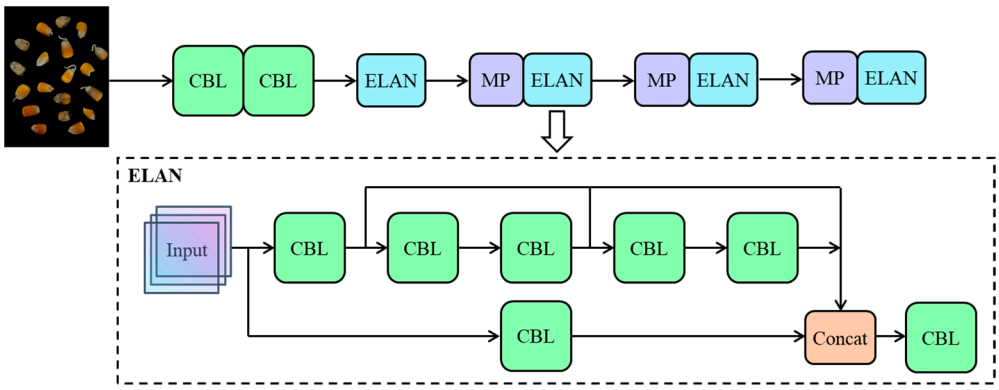


Figure 4. Structure of the backbone in YOLOv7-tiny.

The YOLOv7-tiny model connects its backbone part with the neck part through the SPCSP module, which is composed of numerous CBL structures together with three Max-Pool layers. Adding this structure can achieve a richer gradient combination while reducing the computational complexity of the model. In the neck part, the original YOLOv7 PANet structure is still used for feature aggregation, and the PANet supports top-down and bottom-up bi-directional information flow. With the CBL module and ELAN structure, the output feature information of different layers in the backbone part can be fused to enhance the ability of the model to extract deep and shallow feature information and ensure the integrity and diversity of the extracted features. In the head section, the YOLOv7-tiny model uses the CBL module to replace the REPConv module in the original YOLOv7 structure to achieve its function of adjusting the number of channels, thus further simplifying the model structure. The complete YOLOv7-tiny structure is shown in Figure 5.

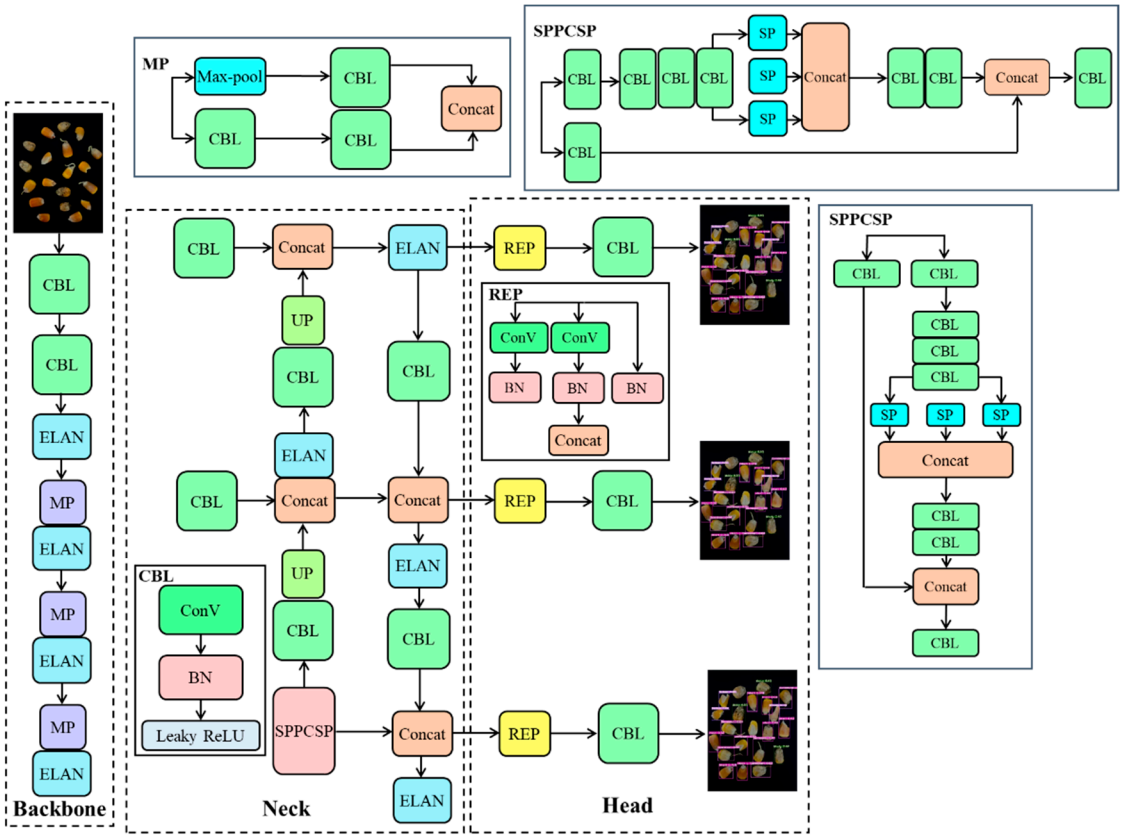


Figure 5. Structure of YOLOv7-tiny.

3.3. YOLOv7-MEF

Compared with the YOLOv7 model, YOLOv7-tiny has advantages in terms of speed and light weight, but its prediction accuracy is not as good as YOLOv7, and its own structure also has the following disadvantages. Firstly, a large number of ELANs in the backbone will increase the calculation amount and the number of parameters. Secondly, because the size of the network layers of YOLOv7-tiny is too small, it is not conducive to feature extraction, which will lead to the decline of model prediction accuracy. In view of the above shortcomings, we improved the YOLOv7-tiny model and proposed a YOLOv7-MEF model with better performance.

3.3.1. MobileNetV3

To make the model easier for embedded deployment, we added the MobileNetV3 [31] to the backbone of YOLOv7-tiny. Compared with other networks, MobileNetV3 has the advantages of short training time, small model weight files, and good convergence in crop recognition and classification, making it more suitable for mobile device deployment. It made the whole model more lightweight, reduced the number of parameters and calculations of the model, and improved the real-time detection.

As a lightweight network structure, MobileNetV3 uses depth-separable convolution to replace the original convolution layer. By separating the spatial filter from the feature generation mechanism, the traditional convolution is effectively decomposed, thus greatly reducing the computational effort. Figure 6 shows the convolution modes of the two convolution modes, the traditional convolution on the left and the depth-separable convolution on the right. Compared with the traditional convolution of YOLOv7-tiny, depth-separable convolution is divided into two stages: channel-to-channel convolution and point-to-point convolution. In the channel-to-channel convolution phase, each channel in the input feature map was convolved only with its corresponding single-channel convolution kernel. Each input channel was independent of the other, and the feature fusion between channels was

eliminated, so that the number of channels in the output feature map after convolution remained unchanged. Point-to-point convolution uses the traditional convolution method, uses a 1 × 1 convolution kernel to check the output feature maps of all channels for integration, and then changes the number of channels in the output feature maps. The ratio Q of the number of depth-separable convolution parameters to the number of conventional convolution parameters and the computational ratio R is

$$Q = R = \frac{1}{N} + \frac{1}{{D_K}^2} \tag{6}$$

where N is the number of output feature maps and D_K is the size of output feature maps.

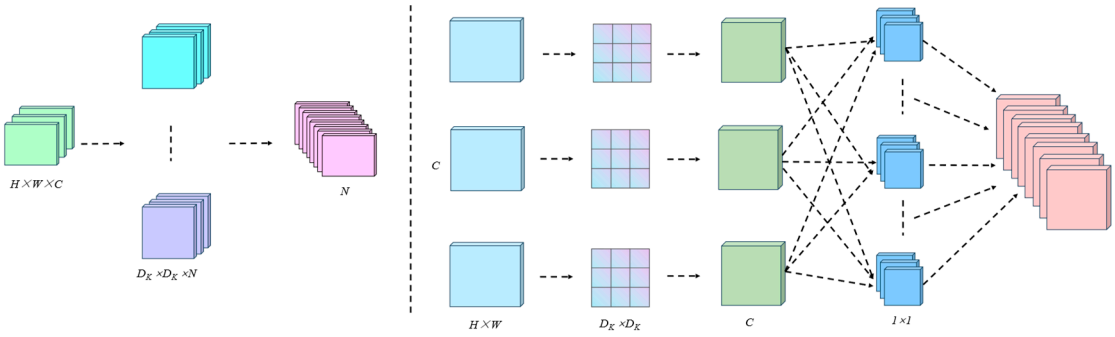


Figure 6. Standard convolution and depth-separable convolution.

The MobileNetV3 network was designed with an inverse residual structure with linear bottlenecks, and the ReLU function was replaced with a linear function at the end of the bottlenecks to prevent the features from being corrupted. Additionally, the bottlenecks are designed in the opposite way to ResNet, which first expands, extracts features, and then compresses, thereby improving the efficiency of feature extraction. In addition, the SE attention module in the MobileNetV3 network will enable the network to focus greater attention on more useful information. Using the h-swish activation function instead of the original swish function made the computation faster and more efficient. The overall network structure is shown in Figure 7.

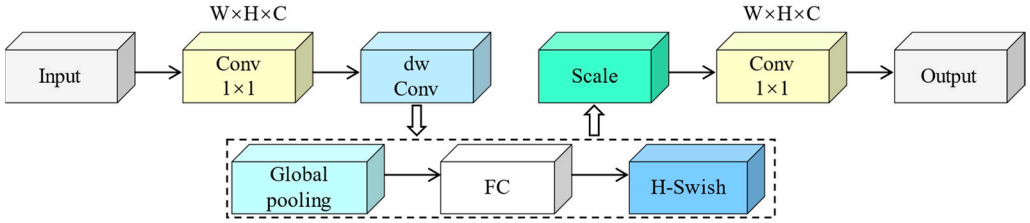


Figure 7. Structure of improved MobileNetV3 algorithm.

3.3.2. ESE-Net Efficient Attention Mechanism

To further improve the ability of the model to extract features, we used the Effective Squeeze-and-Excitation Network (ESE-Net) [32] efficient attention module instead of the SE-Net module in MobileNetV3. The structure of the original Squeeze-and-Excitation Network (SE-Net) module is shown in Figure 8a. A three-channel feature map with the size of H × W × C was input, and the image size was changed to 1 × 1 × C by the Squeeze module (Global pooling). Next, the Excitation operation was performed to convert the Squeeze module's output feature map to a 1 × 1 × C/r (r > 1, which was a positive integer) size through the first fully connected layer. After that, the ReLU activation function was connected, and then the second fully connected layer was connected to convert the number of channels into C. After that, the weight of each channel could be obtained by the Sigmoid activation function, and the feature map output with the SE-Net attention mechanism could be obtained by multiplying the weight with the input.

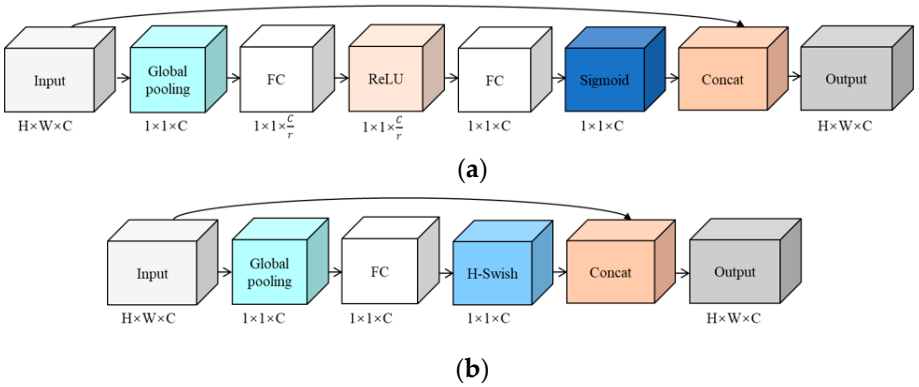


Figure 8. (**a**) Structure of the original SE-Net module; (**b**) structure of the ESE-Net module.

In order to reduce the computational complexity of the model, the scale factor r was introduced in SE Net, but this also reduced the feature extraction weight ability of the SE module, resulting in feature loss and weakening of the weight expression ability of the module. Therefore, we used the ESE-Net module instead of SE-Net. The structure of the ESE-Net module is shown in Figure 8b. Compared with SE-Net, the first fully connected layer with scale factor r and the ReLU activation function were removed, and the channel attention weight generated by the module was increased, while the calculation amount of the model was reduced.

To verify the influence of substituting ESE-Net for SE-Net on the detection performance of maize kernel characteristics, we compared the heat maps output by the above two different modules with the same dataset, as shown in Figure 9. The heat map visualization [33] is used to visually display the areas that the network pays more attention to. The darker the color, the more obvious the network extraction features. For the heat map of broken maize kernels, the ESE-Net module had a detection error rate of 0, and all features of broken maize kernels have been extracted. For the heat map of moldy maize kernels, SE-Net missed one kernel, and ESE-Net identified all moldy ones with much darker colors. For the heat map of germinated maize kernels, all four germinated ones were detected, but the ESENe color was darker, indicating a deeper feature extraction. For the heat map of intact maize kernels, ESE-Net had no false detection, the color became reddish brown, and the feature extraction was more obvious. In summary, it could be seen from the figure that after adding the ESE-Net module, the model paid more attention to the refined features of different qualities of maize kernels and improved the prediction accuracy and feature extraction ability of kernels.

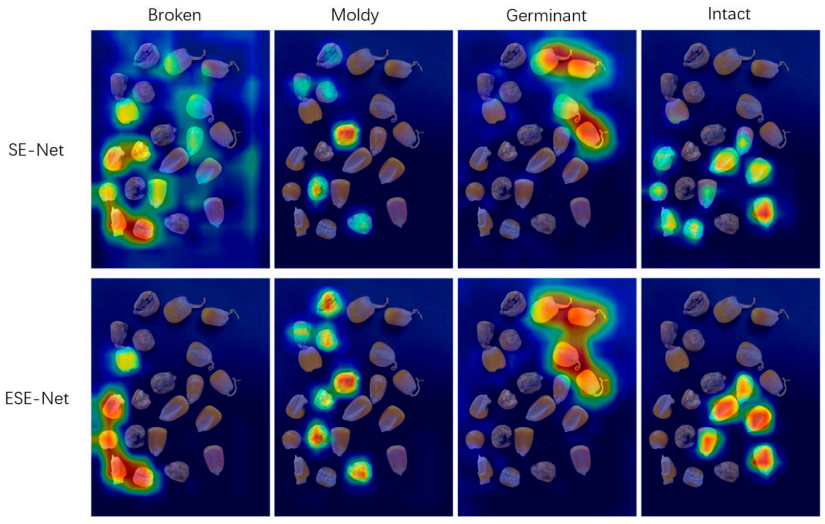


Figure 9. Comparison of heat maps using SE-Net and SES-Net network of different categories.

3.3.3. *Focal-EIoU* Loss

YOLOv7-tiny uses Complete Intersection over Union Loss (CIoU-Loss) [33] as the border loss function, and the penalty term of *CIoU* is the distance and relative proportion of the rectangular frame, which is calculated as follows:

$$CIoU = IoU - \frac{D_2^2}{D_C^2} - \alpha v \tag{7}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{8}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{9}$$

where IoU is the ratio of intersection and union between prediction box and real box and D_2 is the distance between prediction box and target box center point. D_C is the diagonal distance of the smallest external rectangle C, α is the balance parameter, and v is a parameter that measures the consistency of the relative proportions of two rectangular boxes.

The penalty term for $CIoU$ was the relative ratio of width to height, rather than the value of width to height. However, when the width and height were satisfied $\left\{(w = kw^{gt}, h = kh^{gt})\middle|k \in R^+\right\}$, using the relative scale as a penalty term no longer worked.

EIoU improved the shortcomings of *CIoU* by dividing the loss function into three parts: distance loss, direction loss, and *IoU* loss. The α and v of *CIoU* have been modified, where C_w and C_h represent the length and height of the minimum box that covers the real box and anchor box, thus solving the problem caused by *CIoU* using aspect ratio. When there is a lot of noise in the drawing and the contour is not very clear, generally deep learning models may experience false positives and false negatives. During training, there will be many easily distinguishable candidate boxes and negative examples, resulting in an imbalance of positive and negative examples and difficult to distinguish samples. The trained model can recognize easily distinguishable samples, but the recognition effect for difficult to distinguish samples is still very poor. Therefore, in order to better improve the performance of the model, FocalL1 loss is used to set different gradients, as shown in Equation (11). Higher gradients are set in areas with higher error rates, which focus more on identifying difficult samples and can reduce the impact of low-quality samples on model performance. By integrating Enhanced Intersection over Union (*EIoU*) loss and FocalL1 loss, the final Focal *EIoU* loss is obtained.

Loss function *EIoU* directly used the prediction results of width and height as penalty terms and the formula is as follows:

$$\begin{aligned} L_{EIoU} &= L_{IoU} + L_{dis} + L_{asp} \\ &= 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \frac{\rho^2\left(w, w^{gt}\right)}{C_w^2} + \frac{\rho^2\left(h, h^{gt}\right)}{C_h^2} \end{aligned} \tag{10}$$

where C_w and C_h are, respectively, the width and height of the two rectangles. It can be seen from the formula that *EIoU* directly took the side length as the penalty term and divided the loss function into three parts: Iou loss, distance loss, and side length loss. *EIoU* theoretically solved the problem that the width and height of *CIoU* could not be enlarged or reduced at the same time and optimized and improved *CIoU*.

For object detection, most of the prediction boxes obtained from anchor points have little difference in *IoU* compared to the ground truth box, and training on such samples can easily cause significant fluctuations in loss values. By using FocalL1 Loss [34], we made such samples have a smaller gradient, so as to reduce the negative impact of such samples on the gradient. The expression is as follows:

$$FocalLoss = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{11}$$

where α_t is used to solve the problem of imbalance between positive and negative samples and γ is used to solve the problem of imbalance between hard and easy samples. By integrating *EIoU* Loss and FocalL1 Loss, *Focal-EIoU* Loss used in the YOLOv7-MEF was obtained [34], and its formula is as follows:

$$L_{FocalE-IoU} = IoU^{\gamma} L_{EIoU} \tag{12}$$

3.3.4. YOLOv7-MEF

This study took Yolov7-tiny as the original model and introduced lightweight MobileNetV3 network in the backbone of the model to reduce model parameters and reduce network complexity. Then, the SE-Net attention mechanism in the MobileNetv3 model is replaced with the ESE-Net attention mechanism to reduce the channel information loss and improve the feature extraction ability of the model. Finally, by improving the loss function and replacing the original *CIoU* Loss with *Focal-EIoU* Loss, the convergence speed of the model is improved. The structure diagram of the YOLOv7-MEF model is shown in Figure 10.

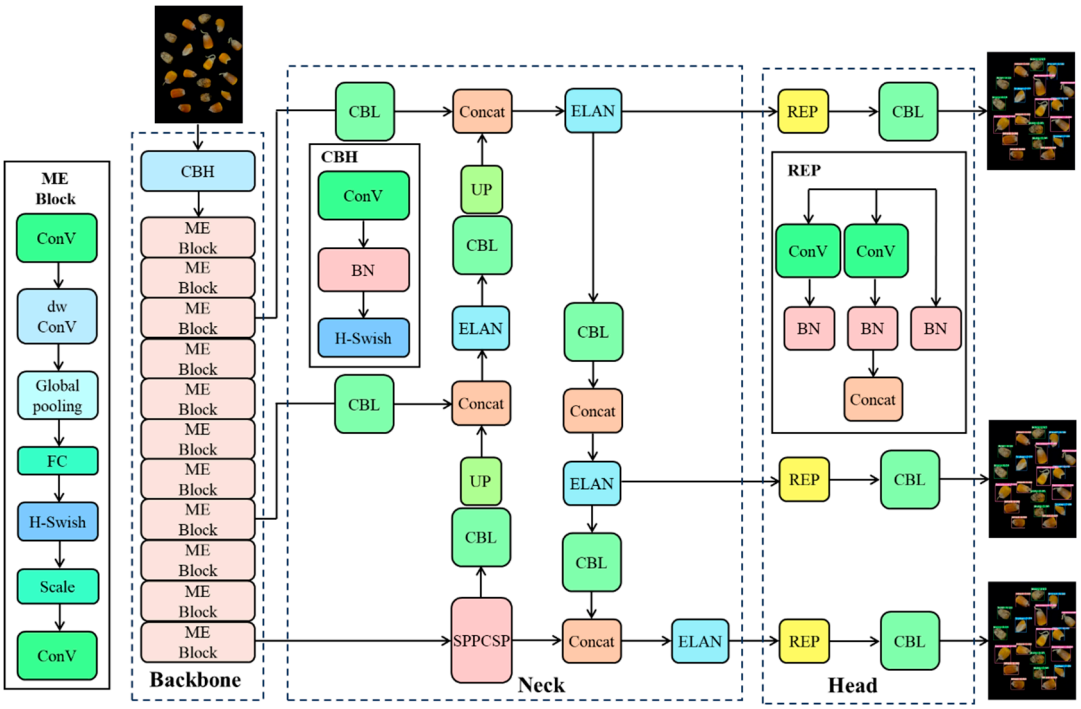


Figure 10. Structure of YOLOv7-MEF.

4. Model and Algorithm Test

4.1. Ablation Experiment

We use the same image dataset to verify the performance of the improved model. During the research process, the modules were embedded into the feature networks generated by YOLOv7-tiny, and the results of the ablation experiment are shown in Table 3. Under the same experimental parameters, the accuracy and recall of YOLOv7-MobileNetV3 added to MobileNetV3 were decreased, the model size was reduced, and the real-time detection speed was enhanced. Based on this model, the SE-Net in the model was replaced by ESE-Net to obtain the model YOLOv7-ME. The accuracy of the model was improved from 93.13% to 95.32%, the recall increased by 6.17%, the model size was further reduced, the detection and training speed was again accelerated, and the FPS value reached 71.43. Then, the function of *Focal-EIoU* Loss was verified. Based on the YOLOv7-MobileNetV3 model, the YOLOv7-MF model with *CIoU* Loss replaced by *Focal-EIoU* Loss directly adjusted the edge length of the prediction frame. The dramatic fluctuations in loss values due to a poor-

quality anchor were reduced. As shown in Figure 11, compared with the above models, the loss curve of the YOLOv7-MF model was smoother and the convergence speed was faster. Finally, the YOLOv7-MEF model was trained, and, compared with each evaluation index, the accuracy of the YOLOv7-MEF model was the highest, reaching 98.94%. Its recall was optimal at 96.42% and the model size was further reduced to 9.1 MB, while its real-time detection speed reached 76.92, which was more favorable for embedded deployment. The experimental data showed that the improved YOLOv7-MEF model not only reduced the model parameters, but also improved the efficiency of real-time detection of maize seed quality, and the detection quality of the model was also improved.

Table 3. Comparison of ablation experiment performance evaluation index.

Model	Precision	Recall	Model Size/M	FPS
YOLOv7-tiny	97.21%	93.14%	11.72	47.62
YOLOv7-MobileNetV3	93.13%	81.3%	8.25	64.52
YOLOv7-ME	95.32%	87.47%	8.17	71.43
YOLOv7-MF	94.43%	91.3%	8.23	67.11
YOLOv7-MEF	98.94%	96.42%	9.1	76.92

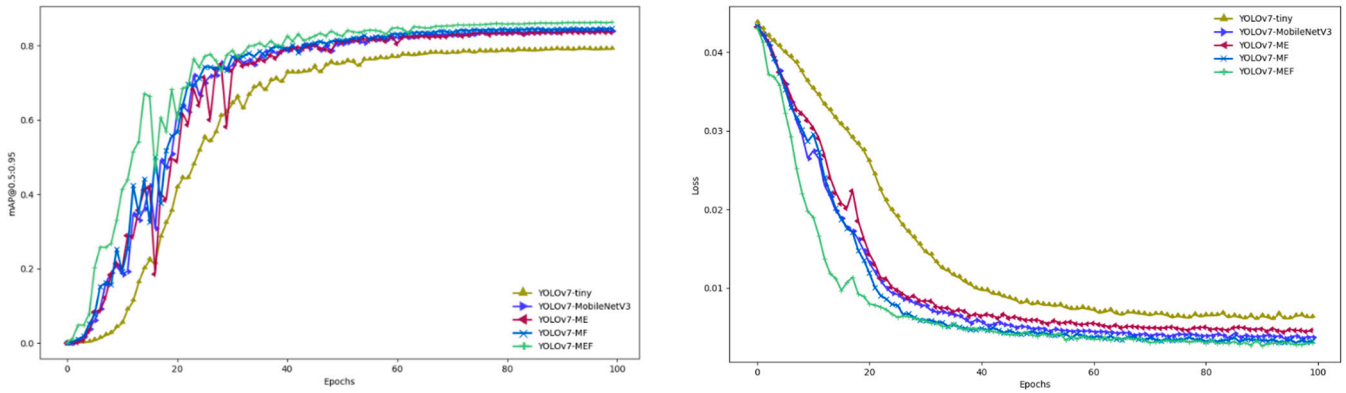


Figure 11. mAP@0.5:0.95 and loss curves of the training set for different improved networks.

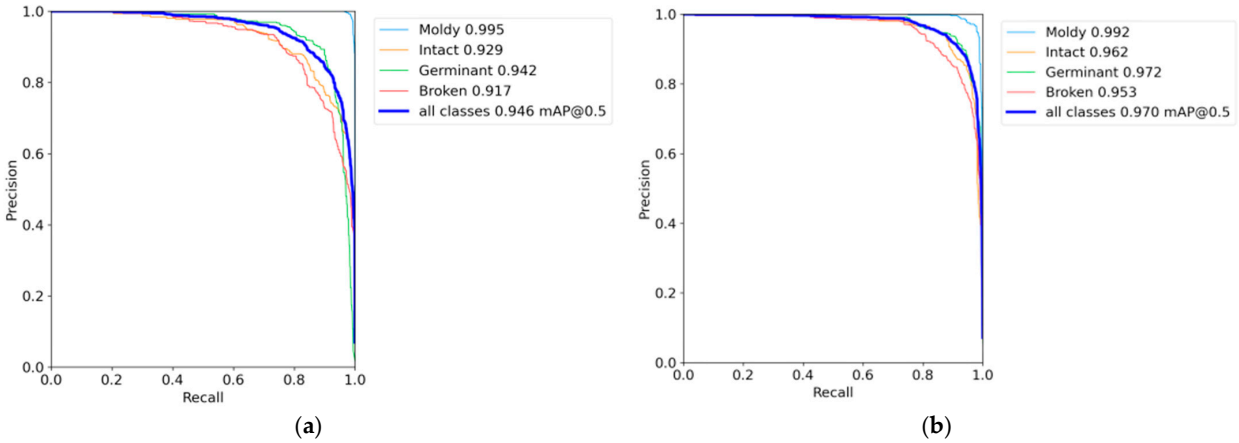
To further verify the detection performance of YOLOv7-MEF, the experiment also output the map@0.5:0.95 curve and loss curve of five models, as shown in Figure 11. Compared with the other four models, YOLOv7-MEF had the highest value of map@0.5:0.95. Meanwhile, compared with the loss curve, it was seen that the loss curve of the YOLOv7-MEF model was the smoothest, had the fastest convergence, and had the best comprehensive performance compared with the others.

4.2. Comparative Analysis of YOLOv7-MEF and YOLOv7-Tiny Model

In order to further study the prediction of each quality maize kernel by the model, the P-R curves of YOLOv7-tiny model and YOLOv7-MEF model were output based on the same maize kernel dataset. As shown in Figure 12, for the YOLOv7-tiny model, the model had the best prediction effect on moldy kernels, and its map@0.5/% value reached 99.5%. However, the prediction effect of the model for intact and broken maize kernels needed to be improved. The map@0.5/% value of intact ones was 92.9%, and the map@0.5/% value of broken ones was only 91.7%. The map@0.5/% value of the whole model was 94.6%, indicating that the model still had great room for improvement.

For the improved YOLOv7-MEF, the prediction map@0.5/% value of the model for all quality maize kernels reached over 95%. Among them, the prediction effect for moldy maize kernels was the best; the map@0.5/% was 99.2%. The model also improved the detection of broken maize kernels, from 91.7% to 95.3%, an improvement of 3.6%. The detection of intact maize kernels was also improved by 3.3%. In general, the prediction

effect of the YOLOv7-MEF model was further improved on the basis of the YOLOv7-tiny model, and the predicted map@0.5% value increased by 2.4%.



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

Figure 12. Comparison of PR curves: (**a**) PR curve of YOLOv7-tiny; (**b**) PR curves of YOLOv7-MEF.

We randomly selected three captured images from the orbit and output the actual prediction results of YOLOv7-tiny and YOLOv7-MEF to compare the effectiveness of algorithms before and after improvement. The improved YOLOv7-MEF model reduced the loss of channel information by using the ESE-Net efficient attention mechanism, which could better determine the quality and position information of maize kernels, thus enabling higher quality anchor boxes and improving the accuracy of the model in detecting maize kernels in different regions. From Figure 13, it was seen that there were five missed cases in YOLOv7-tiny, including two germinated samples, two broken samples, and one intact sample, and no missed cases in the improved model. YOLOv7-tiny had three prediction errors, in that two germinated samples were predicted as intact samples and one intact sample was predicted as a broken sample, and YOLOv7-MEF predicted all correctly. At the same time, the confidence of the original model for the detection of maize kernels of different quality was relatively low, and the confidence of the improved YOLOv7-MEF for the detection of various qualities was more than 85%, and the majority of the confidence was about 95%. This showed that the detection accuracy of the improved YOLOv7-MEF model met the practical requirements and was suitable for deployment into a practical production environment.

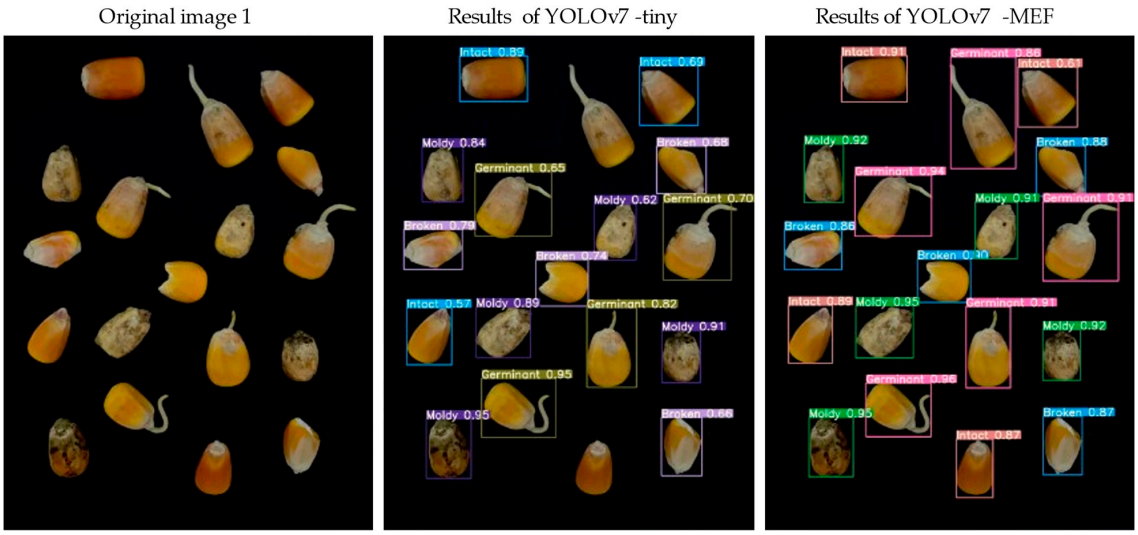


Figure 13. *Cont.*

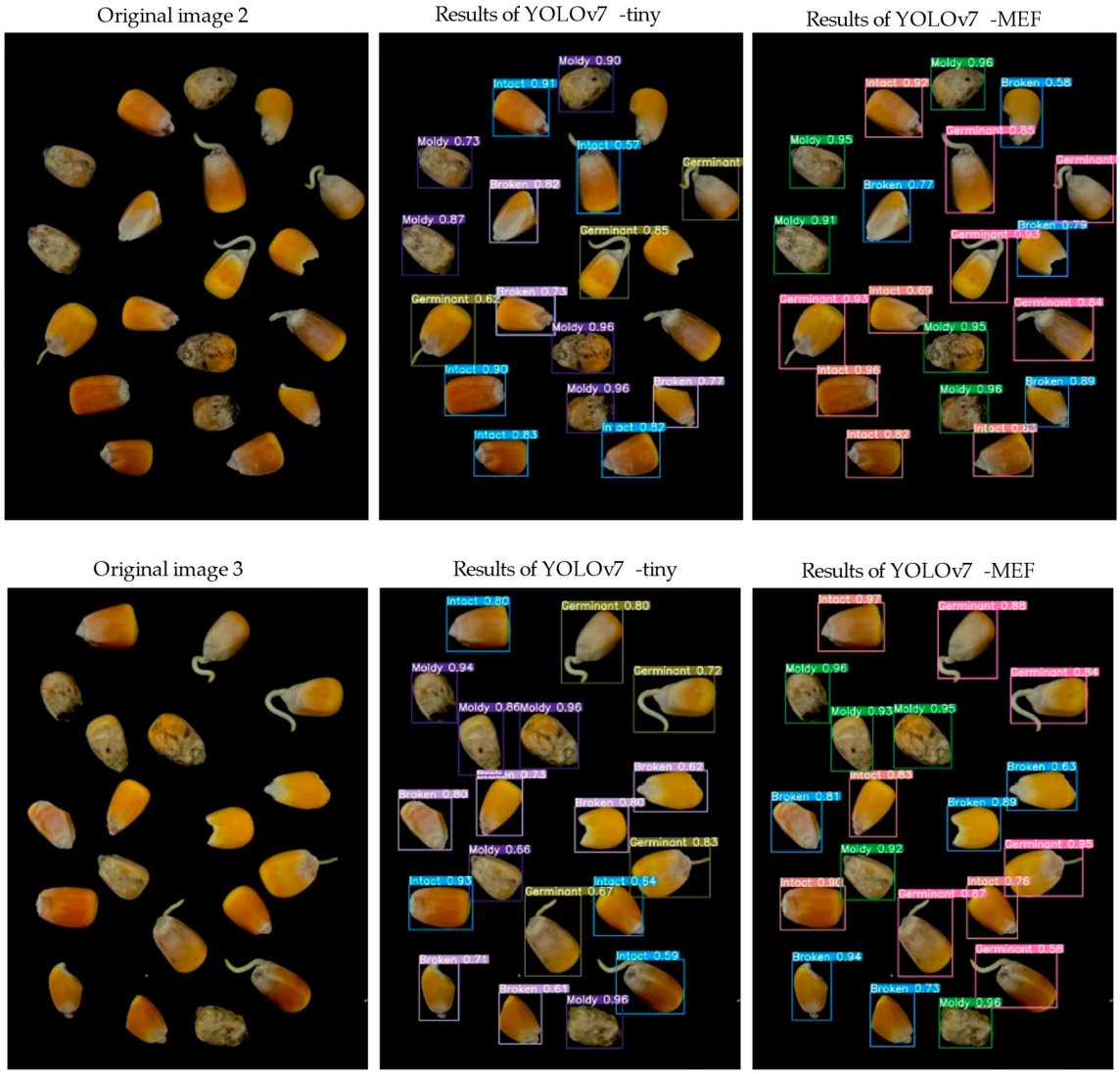


Figure 13. Comparison of visual test results of maize kernels before and after improved network.

5. Discussion

This study researched that after maize kernels were harvested and threshed by corn combine, an electromagnetic vibration track was used to eliminate the stacking and adhesion of them, and batch sampling was performed by vibrational field. Each batch of about 20 kernels of different quality was randomly arranged on the track and sampled by an industrial camera to analyze the effectiveness of the model in maize quality detection. We proposed a recognition and localization method based on the improved YOLOv7-tiny algorithm.

A self-constructed maize kernel dataset was created, with each image consisting of a randomized arrangement of about 20 kernels containing germinated, broken, moldy, and intact ones. In order to enhance the generalization ability of the algorithm, the data were augmented to cover all aspects of different kernel quality.

The lightweight target detection algorithm proposed in this paper identified and localized maize kernels on the electromagnetic vibration track with a detection accuracy of 98.94%, a model size of 9.1 M, and an FPS/(frame/s) of 76.92. It was seen from the results that the improved maize kernel quality detection model could meet the detection requirements of real-time, accurate, and non-destructive identification and positioning according to kernel quality, which was conducive to model deployment, providing technical support for this kind of research. Compared with the original model, the size of the model

was reduced by 22.27%, the accuracy and recall were increased by 1.73 and 3.28 percentage points, respectively, and the FPS was increased by 61.5%.

Compared with the literature [22–27] on maize kernel research, the accuracy of the YOLOv7-MEF (each indicator in Table 3) proposed in this study was significantly improved by about 2%, and the model size is also significantly smaller. Moreover, data collection can be achieved using industrial cameras with brackets, and corn kernels do not require manual placement in fixed positions. Compared to classification algorithms that use hyperspectral imaging for research, it reduces operating costs and improves the applicability of the algorithm.

The method developed in this study is aimed at embedding agricultural machinery systems and becoming a part of smart agricultural machinery. The entire system required an image acquisition system, including a high-speed visual inspection camera for about 1000 RMB, a visual bracket for about 600 RMB, and a CPU with Raspberry Pi with an 8 G storage card for about 800 RMB. The core components for a total of 2400 RMB are completed. Meanwhile, GUI interfaces can run the trained model directly on the Raspberry PI.

This study only conducted image acquisition and algorithm experiments on one variety of corn. Although the results were good, further work is still needed. For example, collecting more varieties of corn for experimentation, and switching to other grains similar to corn such as peanuts and beans for generalization experiments. This will further enhance the promotion of algorithms and the applicability of agricultural machinery products.

6. Conclusions

This paper developed an algorithm based on corn kernel classification recognition, which reduced the model size by 22.27%, improved Recall and Precision by 1.73 and 3.28 percentage points, respectively, and improved FPS by 61.5%. This algorithm greatly reduced the number of model parameters while ensuring high detection accuracy, and has good real-time performance. It is suitable for deploying embedded track detection systems in agricultural machinery equipment, providing a powerful theoretical research method for efficient detection of corn kernel quality. Meanwhile, this algorithm is currently only applicable to the classification and recognition of different varieties of corn kernels. Further research is needed for the classification and recognition of corn seeds and other grains, which is also our next step.

The lightweight backbone network MobileNetV3 was used instead of the feature backbone network of the original model. This improvement significantly reduced the size of the network, resulting in a nearly 30% reduction in model size and a 16.9% increase in FPS.

The ESE-Net efficient attention mechanism was used to enhance feature extraction to obtain better generalization and more accurate recognition effects for small targets. This improvement led to a 5.67% increase in model recall and a 23.81 increase in FPS.

The focal-EIoU regression Loss function was used to replace CIoU, which improved the convergence and regression accuracy of the model. After adding this, the detection accuracy was 2.78 percentage points higher, and FPS was 19.49 higher.

Regarding the current situation of corn kernels, intact kernels should account for over 95%, while the remaining three types should not exceed 5% in total. Among them, moldy ones had the highest proportion and greater harm, followed by broken ones, and the least prevalent were those showing germination. Through multiple experiments, the confidence level of moldy detection was over 90%, which fully conforms to the current recognition and classification of corn kernels.

Author Contributions: Conceptualization, L.Y.; formal analysis, D.W.; investigation, C.W. and L.Y.; data curation, C.L.; funding acquisition, L.Y. and C.L.; methodology, C.L.; writing—original draft, L.Y. and D.W.; writing—review, C.L. and D.W.; visualization, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program "Research and Integration Demonstration of Light Simplified High Yield Technology for Soybean and Other Oil Crops" Project "Research and Development of Common Light Simplified High Yield Technology for Main Oil Crops" (2022YFD2300101), and the Key Research and Development Plan of Shandong Province (Major Science and Technology Innovation Project) "High-Performance Seeding and Harvesting Key Components and Intelligent Work Machines Creation" (2021CXGC010813).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding authors upon reasonable request.

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

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

References

1. Dai, D.; Ma, Z.; Song, R. Maize kernel development. *Mol. Breed.* **2021**, *41*, 2. [CrossRef] [PubMed]
2. Erenstein, O.; Jaleta, M.; Sonder, K.; Mottaleb, K.; Prasanna, B. Global maize production, consumption and trade: Trends and R&D implications. *Food Secur.* **2022**, *14*, 1295–1319. [CrossRef]
3. Ekpa, O.; Palacios-Rojas, N.; Kruseman, G.; Fogliano, V.; Linnemann, A.R. Sub-Saharan African Maize-Based Foods—Processing Practices, Challenges and Opportunities. *Food Rev. Int.* **2019**, *35*, 609–639. [CrossRef]
4. Klopfenstein, T.; Erickson, G.; Berger, L. Maize is a critically important source of food, feed, energy and forage in the USA. *Field Crop. Res.* **2013**, *153*, 5–11. [CrossRef]
5. Wang, K.; Xie, R.; Ming, B.; Hou, P.; Xue, J.; Li, S. Review of combine harvester losses for maize and influencing factors. *Int. J. Agric. Biol. Eng.* **2021**, *14*, 1–10. [CrossRef]
6. Wu, F. Global impacts of aflatoxin in maize: Trade and human health. *World Mycotoxin J.* **2015**, *8*, 137–142. [CrossRef]
7. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
8. Adige, S.; Kurban, R.; Durmuş, A.; Karaköse, E. Classification of apple images using support vector machines and deep residual networks. *Neural Comput. Appl.* **2023**, *35*, 12073–12087. [CrossRef]
9. Ju, J.; Zheng, H.; Xu, X.; Guo, Z.; Zheng, Z.; Lin, M. Classification of jujube defects in small data sets based on transfer learning. *Neural Comput. Appl.* **2021**, *34*, 3385–3398. [CrossRef]
10. Wang, Z.; Ling, Y.; Wang, X.; Meng, D.; Nie, L.; An, G.; Wang, X. An improved Faster R-CNN model for multi-object tomato maturity detection in complex scenarios. *Ecol. Inform.* **2022**, *72*, 101886. [CrossRef]
11. Ni, J.; Gao, J.; Li, J.; Yang, H.; Hao, Z.; Han, Z. E-AlexNet: Quality evaluation of strawberry based on machine learning. *J. Food Meas. Charact.* **2021**, *15*, 4530–4541. [CrossRef]
12. Huang, Z.; Wang, R.; Cao, Y.; Zheng, S.; Teng, Y.; Wang, F.; Wang, L.; Du, J. Deep learning based soybean seed classification. *Comput. Electron. Agric.* **2022**, *202*, 107393. [CrossRef]
13. Zhang, J.; Ma, Q.; Cui, X.; Guo, H.; Wang, K.; Zhu, D. High-throughput corn ear screening method based on two-pathway convolutional neural network. *Comput. Electron. Agric.* **2020**, *175*, 105525. [CrossRef]
14. Zhao, W.; Liu, S.; Li, X.; Han, X.; Yang, H. Fast and accurate wheat grain quality detection based on improved YOLOv5. *Comput. Electron. Agric.* **2022**, *202*, 107426. [CrossRef]
15. Yang, Y.; Liu, Z.; Huang, M.; Zhu, Q.; Zhao, X. Automatic detection of multi-type defects on potatoes using multispectral imaging combined with a deep learning model. *J. Food Eng.* **2023**, *336*, 111213. [CrossRef]
16. Kurtulmuş, F. Identification of sunflower seeds with deep convolutional neural networks. *J. Food Meas. Charact.* **2020**, *15*, 1024–1033. [CrossRef]
17. Jeyaraj, P.R.; Asokan, S.P.; Nadar, E.R.S. Computer-Assisted Real-Time Rice Variety Learning Using Deep Learning Network. *Rice Sci.* **2022**, *29*, 489–498. [CrossRef]
18. Amatya, S.; Karkee, M.; Gongal, A.; Zhang, Q.; Whiting, M.D. Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting. *Biosyst. Eng.* **2016**, *146*, 3–15. [CrossRef]
19. Ye, W.; Yan, T.; Zhang, C.; Duan, L.; Chen, W.; Song, H.; Zhang, Y.; Xu, W.; Gao, P. Detection of Pesticide Residue Level in Grape Using Hyperspectral Imaging with Machine Learning. *Foods* **2022**, *11*, 1609. [CrossRef] [PubMed]
20. Gao, H.; Zhen, T.; Li, Z. Detection of Wheat Unsound Kernels Based on Improved ResNet. *IEEE Access* **2022**, *10*, 20092–20101. [CrossRef]
21. Jin, B.; Zhang, C.; Jia, L.; Tang, Q.; Gao, L.; Zhao, G.; Qi, H. Identification of Rice Seed Varieties Based on Near-Infrared Hyperspectral Imaging Technology Combined with Deep Learning. *ACS Omega* **2022**, *7*, 4735–4749. [CrossRef]
22. Bi, C.; Hu, N.; Zou, Y.; Zhang, S.; Xu, S.; Yu, H. Development of Deep Learning Methodology for Maize Seed Variety Recognition Based on Improved Swin Transformer. *Agronomy* **2022**, *12*, 1843. [CrossRef]

23. Yang, D.; Jiang, J.; Jie, Y.; Li, Q.; Shi, T. Detection of the moldy status of the stored maize kernels using hyperspectral imaging and deep learning algorithms. *Int. J. Food Prop.* **2022**, *25*, 170–186. [CrossRef]
24. Zhao, C.; Quan, L.; Li, H.; Liu, R.; Wang, J.; Feng, H.; Wang, Q.; Sin, K. Precise Selection and Visualization of Maize Kernels Based on Electromagnetic Vibration and Deep Learning. *Trans. ASABE* **2020**, *63*, 629–643. [CrossRef]
25. Xu, P.; Sun, W.; Xu, K.; Zhang, Y.; Tan, Q.; Qing, Y.; Yang, R. Identification of Defective Maize Seeds Using Hyperspectral Imaging Combined with Deep Learning. *Foods* **2022**, *12*, 144. [CrossRef] [PubMed]
26. Xu, P.; Tan, Q.; Zhang, Y.; Zha, X.; Yang, S.; Yang, R. Research on Maize Seed Classification and Recognition Based on Machine Vision and Deep Learning. *Agriculture* **2022**, *12*, 232. [CrossRef]
27. Jiao, Y.; Wang, Z.; Shang, Y.; Li, R.; Hua, Z.; Song, H. Detecting endosperm cracks in soaked maize using μCT technology and R-YOLOv7-tiny. *Comput. Electron. Agric.* **2023**, *213*, 108232. [CrossRef]
28. Wei, Y.; Yang, C.; He, L.; Wu, F.; Yu, Q.; Hu, W. Classification for GM and Non-GM Maize Kernels Based on NIR Spectra and Deep Learning. *Processes* **2023**, *11*, 486. [CrossRef]
29. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
30. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
31. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.-C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. *arXiv* **2019**, arXiv:1905.02244.
32. Lee, Y.; Park, J. CenterMask: Real-Time Anchor-Free Instance Segmentation. *arXiv* **2020**, arXiv:1911.06667.
33. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *arXiv* **2016**, arXiv:1610.02391.
34. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]